# COSMI: COMPOSITIONAL SYNTHESIS OF MULTI-OBJECT INTERACTIONS

**Anonymous authors**
Paper under double-blind review

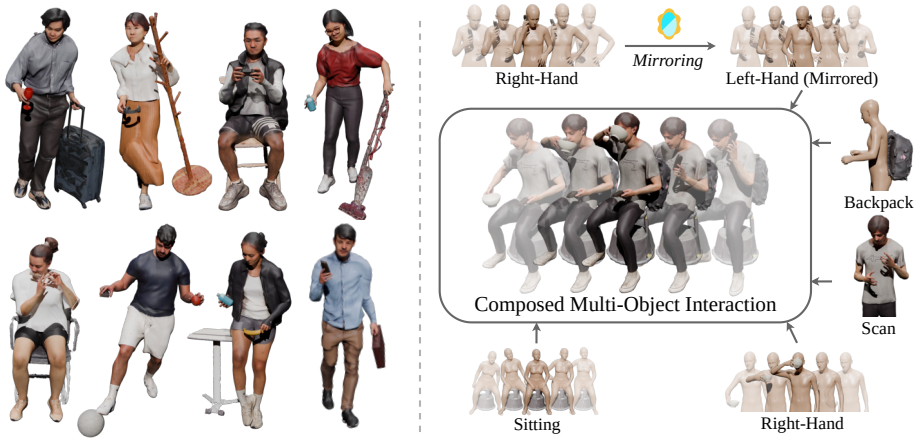



Figure 1: **COSMI.** Left: sequences of the COSMI dataset, composed from single-object captures and rendered with clothed scans. Right: composition of one sequence from a sitting clip, a right-hand action, a left-hand action obtained by mirroring, and a backpack, rendered with a clothed scan.

## ABSTRACT

Generative models of human-object interaction are bounded by the data that exists: everyday activities involve several objects, but most captured datasets record one at a time, as multi-object capture is combinatorially expensive. Our observation is that interactions are local, so single-object captures already contain the parts of multi-object activities. We compose them: contact-consistent clips of single interactions, mirrored to balance the hands, transfer between bodies, and a language model and geometric checks admit only the pairings that are plausible, semantically and physically. Therefore, the dataset grows combinatorially with the clips rather than recording time. The COSMI dataset holds 222k sequences and 275 hours with up to five objects, nearly thirty times the largest multi-object capture, and can be extended by adding datasets or even hand-object recordings. On this data we train the COSMI method, a text-to-interaction diffusion transformer that follows how the data is built: weight-shared object slots generate a variable number of objects, predicted relative to the body parts that move them. On a benchmark with an unseen object and unseen interaction combinations, models trained on the dataset generalize to the unseen combinations. COSMI outperforms baselines in text alignment and contact accuracy, where its margin is largest on the unseen object. Code, models, and the dataset pipeline will be released.

## 1 INTRODUCTION

Generating 3D human-object interaction (HOI) serves animation, embodied AI, and augmented reality, from humanoid controllers and robot policies that learn from HOI motion (Xu et al., 2025b; Wang et al., 2025b) to egocentric assistants (Banerjee et al., 2025; Ma et al., 2024) and composite animation such as sitting down while carrying an object (Pan et al., 2025). The activities these applications target are frequently multi-object: in the KIT whole-body database, 99 of the 262 object-related motions involve two or more objects (Mandery et al., 2016). Like other generative models,



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.

motion generators improve with data (Wang et al., 2025a; Xu et al., 2025a), so what HOI generators can produce is bounded by the interaction data that exists.

On the data side, HOI datasets mostly record one person with one object (Taheri et al., 2020), and the generators trained on them inherit this scope (Xu et al., 2023; Diller & Dai, 2024; Peng et al., 2025). Capturing several objects at once is expensive, and the cost grows combinatorially with the objects, body parts, and pairings (Xie et al., 2024), so captured multi-object datasets stay below ten hours (Lv et al., 2024; Kim et al., 2025; Lu et al., 2025), against 21.8 in the single-object collection InterAct (Xu et al., 2025a). On the model side, generators are built for one object (Diller & Dai, 2024; Pi et al., 2025; Xu et al., 2023), and the few multi-object ones train one model per object count (Lv et al., 2024; He et al., 2025) or pad a fixed state (Zhang et al., 2024a) and leave implicit which body part moves which object. The concurrent SK-HOI (Peng et al., 2026) handles a variable number of objects in three separate stages. There is no multi-object data at single-object scale, and no end-to-end generator for a variable number of objects grounded to the body parts that move them.

Synthetic and composed data has closed comparable gaps before (Black et al., 2023; Xie et al., 2024; Mandlekar et al., 2023), and closest to our setting, SINC composes simultaneous actions by taking each body part from the clip whose action uses it (Athanasiou et al., 2023). Our observation is that interactions with objects are local in the same way: a cup is held by a hand, a chair supports the pelvis, a ball is kicked by a foot, and the rest of the body is free. The limb that performs one interaction can therefore be transferred, together with its object, to a body that performs another, and multi-object interactions can be composed from single-object captures (Fig. 1), which exist at scale. The composition is only as good as its gating: from contact-consistent clips of single interactions, a language model decides which can plausibly happen at once, and geometric checks keep only the compositions that stay in contact and free of penetration (Fig. 2). Mirroring the one-handed and one-footed clips balances the left and right hands, which captured interactions skew to the right. The resulting COSMI dataset contains 222k sequences and 275 hours of motion, nearly thirty times the largest captured multi-object dataset (Lv et al., 2024). Its physical plausibility stays close to the parent captures (Section A.1), and since nothing in the composition is specific to a dataset, it grows by adding a rule rather than a recording session, even from hand-object recordings (Section 5).

Such data calls for a generator for a variable number of objects. COSMI is a single diffusion transformer that denoises the person and a set of objects from a prompt and initial poses. Its object slots share their weights, so no parameter depends on the object count, and one model generates one to five objects. Each object is predicted relative to the joint that drives it, which a part-assignment head reads from these conditions, and text tokens link the words of the prompt to the objects. On a benchmark that holds out an unseen object and five unseen combinations, such as walking-with-an-apple (while walking-only and apple-eating stay in training), COSMI and the baselines trained on the COSMI dataset alike generalize to the unseen combinations. COSMI outperforms HIMO (Lv et al., 2024) and multi-object adaptations of MDM and PriorMDM (Tevet et al., 2023; Shafir et al., 2024) in text alignment and contact accuracy, where its margin is largest on the unseen object.

In summary, our contributions are:

- Composition of multi-object interactions from single-object captures, and the COSMI dataset: 222k sequences with up to five objects, left and right balanced by mirroring, extensible by design.
- The COSMI method: a text-conditioned diffusion transformer for a variable number of objects, with joint-relative prediction, part assignment, auxiliary losses, and contact-consistent guidance.
- A benchmark on the COSMI dataset that tests generalization to an unseen object and to unseen interaction combinations. Code and models will be released.

## 2 Related Work

**Human-object interaction datasets.** Most full-body interaction datasets, captured with marker-based motion capture or multi-view video, record one person with one object (Taheri et al., 2020; Bhatnagar et al., 2022; Huang et al., 2024; Li et al., 2023; Zhang et al., 2023b; Zhao et al., 2024), later also with articulated objects (Jiang et al., 2023; Fan et al., 2023). Full-body datasets with several objects per sequence are recent and small (Table 1): HIMO records two or three small rigid objects (Lv et al., 2024), ParaHome and HUMOTO articulated household objects (Kim et al., 2025; Lu et al., 2025), and HOI-M3, CORE4D, and MMHOI several people (Zhang et al., 2024a; Liu

et al., 2025; Kogashi et al., 2026). Hand-object datasets omit the body but record more hours and objects, often several at once (Hampali et al., 2020; Chao et al., 2021; Zhan et al., 2024; Liu et al., 2024; Banerjee et al., 2025; Fu et al., 2025), and a body fitted to such a recording turns it into a source clip for our pipeline (Section 5). Interaction data has also been generated procedurally, placed into scenes, augmented, or recovered from video (Xie et al., 2024; Wang et al., 2022; Xu et al., 2025a; Zhang et al., 2026; Xie et al., 2026), always with one object. The COSMI dataset instead composes these single-object captures into multi-object sequences: every source clip pairs with every compatible target and composed sequences serve as targets again, so the number of sequences grows combinatorially with the clips rather than linearly with recording time (Table 1).

**Compositionality in motion and interaction modeling.** Animation has long composed captured clips without objects, in time with motion graphs (Kovar et al., 2002) and across body parts by splicing (Ikemoto & Forsyth, 2004; Heck et al., 2006). SINC revisits spatial composition with language, composing the body parts of two clips into training data from GPT-3 part labels, and leaves the semantic compatibility to future work (Athanasiou et al., 2023). Later models annotate, coordinate, or blend parts (Li et al., 2026; Zou et al., 2024; Jiang et al., 2025) or compose actions in time at inference (Athanasiou et al., 2022; Shafir et al., 2024; Petrovich et al., 2024; Barquero et al., 2024). Language models also plan compositions at inference, as contact chains, joint-contact pairs, affordance graphs, or plans (Xiao et al., 2024; Wang et al., 2024b; Li & Dai, 2026; Xu et al., 2024; Sun et al., 2026), while we use one offline to infer the semantic plausibility of clip combinations (Section 3). Composed training data exists for multi-person motion, by random placement (Shan et al., 2024) or with LLM guidance and a learned filter (Wu et al., 2026), and for interactions only as static poses (Zhao et al., 2022) or as segments and models composed at inference (Lv et al., 2024; Na et al., 2025). We extend body-part composition to interactions with objects, at the data level and across datasets, with object transfer across bodies, semantic gating, and physical checks.

**Text-driven interaction synthesis.** Text-to-motion generation rests on diffusion and token models trained on HumanML3D (Guo et al., 2022; Tevet et al., 2023; Zhang et al., 2024b; Chen et al., 2023; Zhang et al., 2023a; Guo et al., 2024) and, recently, on million-clip corpora (Wang et al., 2025a), two orders of magnitude larger than any captured interaction dataset. Interaction generators extend them to a body and one object (Ghosh et al., 2023; Xu et al., 2023; Li et al., 2023; 2024; Peng et al., 2025; Diller & Dai, 2024; Xu et al., 2024; Petrov et al., 2025; Ron et al., 2026; Ji et al., 2026; Pan et al., 2026; Jung et al., 2026; Cai et al., 2026), including articulated ones (Pi et al., 2025), and grasp synthesis models the fingers in contact (Taheri et al., 2022; Wu et al., 2022; Zhang et al., 2021; Cha et al., 2024; Christen et al., 2024; Zhang et al., 2025). Generators for several objects are few: HIMO trains separate models for two and three objects (Lv et al., 2024), SyncDiff one model per object configuration, conditioned on label phrases (He et al., 2025), and Dex2HOI one stream per object for at most two, both in the hands (Pratikaki et al., 2026). HOI-M3 pads a preset state and takes no text (Zhang et al., 2024a), and HOI-PAGE optimizes each sequence zero-shot for minutes (Li & Dai, 2026). The concurrent SK-HOI covers a variable number of objects with zero-padded slots, but generates in three separately trained stages followed by minutes of test-time optimization per sequence (Peng et al., 2026). COSMI instead is one text-conditioned diffusion transformer with no parameter tied to the object count, which generates one to five objects, articulated and worn, in a single stage and grounds the objects to the body parts that move them.

## 3 The COSMI Dataset

We build every sequence of the COSMI dataset from captured single-object interactions and motions, combined on one body. Our goal is to obtain semantically plausible multi-object interactions from these atomic clips. We therefore extract contact-consistent clips from the source sequences, annotate them with semantic information from a language model, and mirror them to balance the left and right hands. To combine the clips, we rely on this semantic information to pair plausible matches, followed by geometric checks of the physical plausibility of the result (Fig. 2). We build on GRAB (Taheri et al., 2020), BEHAVE (Bhatnagar et al., 2022), InterCap (Huang et al., 2024), OMOMO (Li et al., 2023), and ARCTIC (Fan et al., 2023), with BEHAVE and InterCap in their InterAct (Xu et al., 2025a) versions, and on CMU walking sequences from AMASS (Mahmood et al., 2019) selected with BABEL labels (Punnakkal et al., 2021). All sources are unified to a gender-

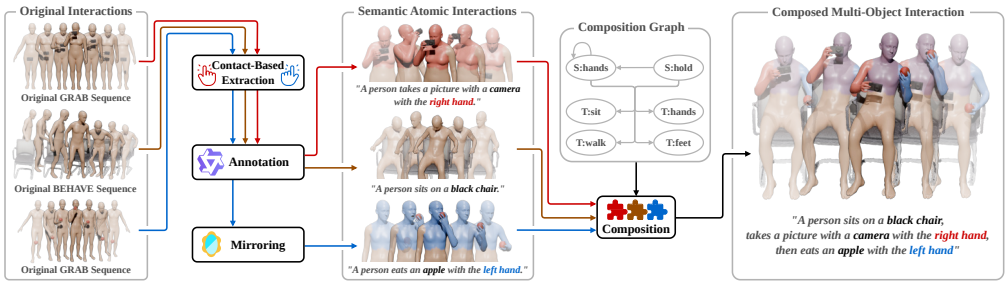



Figure 2: **Composition of one sequence.** Atomic clips are extracted by interaction rules, annotated by a language model, and mirrored. A sitting target receives two compatible hand actions, their limbs and objects are transferred onto its body, and the caption is assembled accordingly.

neutral SMPL-X (Pavlakos et al., 2019) body at 30 fps in a common frame, and every object is a template mesh with a rigid transform per frame, one per part for articulated objects (Section C.1).

**Contact-based extraction.** We build the dataset from *atomic clips*, segments in which a person performs a single interaction and stays in contact with the object from the first frame to the last, so that the acting limb and the object can be lifted out with their contact intact. Thirteen *interaction rules* specify the interaction types by the body parts that must and must not touch the object, by pose constraints such as an upright torso, and by the *local reference* joints relative to which the object is expressed (Section C.2). We evaluate every rule per frame, with contact defined as proximity below 5 cm to the object surface, and take contiguous windows of at least 3 s as atomic clips, 1,491 in total. Holds and worn objects reduce to a single representative frame, and the walking rule admits whole sequences. In composition, *source* rules cover the manipulation of small objects with one or both hands and passive holds (GRAB, ARCTIC). *Target* rules cover the interactions that engage the whole body, sitting, moving large objects, foot interactions, and walking (BEHAVE, InterCap, OMOMO, AMASS), and *augmentation* rules the worn backpack and eyeglasses (BEHAVE, GRAB).

Global trajectories do not transfer between bodies, so we express each object relative to the body part performing its interaction, the local reference joints $\mathcal{J}$ its rule specifies, such as the wrist or the foot. For a joint $j \in \mathcal{J}$ with world rotation $\boldsymbol{R}_j^t$ and position $\boldsymbol{t}_j^t$, the relative transform is

$$\boldsymbol{R}_{\mathcal{O}|j}^t = (\boldsymbol{R}_j^t)^\top \boldsymbol{R}_{\mathcal{O}}^t, \qquad \boldsymbol{t}_{\mathcal{O}|j}^t = (\boldsymbol{R}_j^t)^\top (\boldsymbol{t}_{\mathcal{O}}^t - \boldsymbol{t}_j^t), \tag{1}$$

and inverting it with the joints of *any* body carries the object over to a mirrored or a composed body.

**Annotation.** Whether two interactions can plausibly be performed simultaneously is a semantic question: a person can browse a phone while sitting, but not browse a phone while taking a picture with a camera, although one hand is free for each. Composition therefore needs to know what every clip captures, which the labels of the source datasets, written for whole recordings in differing forms, do not tell. We extract this information with a language model, Qwen3 (Yang et al., 2025), which names the sustained interaction with a verb, correcting labels that are wrong or too general, and writes one sentence, such as *"A person drinks from a cup with the right hand."* (prompts in Section C.3). Per action and object, it further marks whether the action *requires gaze* at the object and *involves the head*, as drinking, eating, and calling do, which decides what can happen at once.

**Mirroring.** Captured interactions favor the right hand (Table 1), so we mirror every one-handed and one-footed clip and swap its rule, which equalizes the two sides and brings the number of clips to 2,663. Reflecting the body is straightforward, $\tilde{\boldsymbol{R}}_j^t = \boldsymbol{M}\boldsymbol{R}_j^t\boldsymbol{M}$ with $\boldsymbol{M} = \mathrm{diag}(-1, 1, 1)$, followed by swapping the left and right joints. The object cannot be reflected in the same way: a reflected mesh is a new object, and reflecting only its world pose breaks the contact with the mirrored hand, because the SMPL-X body is not symmetric. We keep the template $\boldsymbol{V}_{\mathcal{O}}$ and fit a proper rotation $\boldsymbol{Q}_{\mathcal{O}}$ and translation $\boldsymbol{p}_{\mathcal{O}}$ that pose it like its reflection,

$$\boldsymbol{Q}_{\mathcal{O}}(\boldsymbol{V}_{\mathcal{O}} - \bar{\boldsymbol{v}}_{\mathcal{O}}) + \boldsymbol{p}_{\mathcal{O}} \;\approx\; \boldsymbol{M}(\boldsymbol{V}_{\mathcal{O}} - \bar{\boldsymbol{v}}_{\mathcal{O}}), \tag{2}$$

with $\bar{\boldsymbol{v}}_{\mathcal{O}}$ the centroid of the template, which is exact for symmetric objects. Composed with the reflected joint-relative pose, this alignment yields a proper transform of the original mesh that preserves the contact, and Eq. (1) places the object on the opposite limb (details in Section C.5).

Table 1: **Dataset comparison.** L / R: share of left- and right-hand interactions. The COSMI dataset combines the scale of single-object captures with several objects per sequence and balanced hands.

| Dataset | Multi-Object | Interaction Coverage | | | | Hand Bias | Scale | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Small-Object Manipulation | Large-Object Manipulation | Sitting / Body Support | Articulated Objects | L / R (%) | # Subjects | # Objects | # Sequences | # Frames (M) | Hours |
| GRAB (Taheri et al., 2020) | ✗ | ✓ | ✗ | ✗ | ✗ | 15 / 85 | 10 | 51 | 1,335 | 1.62 | 3.76 |
| BEHAVE (Bhatnagar et al., 2022) | ✗ | ✗ | ✓ | ✓ | ✗ | 45 / 55 | 8 | 18 | 299 | 0.45 | 4.13 |
| InterCap (Huang et al., 2024) | ✗ | ✓ | ✓ | ✓ | ✗ | 7 / 93 | 10 | 10 | 223 | 0.07 | 0.62 |
| OMOMO (Li et al., 2023) | ✗ | ✗ | ✓ | ✗ | ✓ | 47 / 53 | 16 | 15 | 5,280 | 0.96 | 8.91 |
| ARCTIC (Fan et al., 2023) | ✗ | ✓ | ✗ | ✗ | ✓ | 36 / 64 | 9 | 11 | 301 | 0.22 | 2.02 |
| HIMO (Lv et al., 2024) | ✓ | ✓ | ✗ | ✗ | ✗ | 47 / 53 | 34 | 53 | 3,376 | 4.08 | 9.44 |
| HOI-M$^3$ (Zhang et al., 2024a) | ✓ | ✓ | ✓ | ✓ | ✗ | 45 / 55 | 31 | 67 | 86 | 1.87 | 8.64 |
| ParaHome (Kim et al., 2025) | ✓ | ✓ | ✓ | ✓ | ✓ | 42 / 58 | 38 | 22 | 207 | 0.88 | 8.11 |
| HUMOTO (Lu et al., 2025) | ✓ | ✓ | ✓ | ✓ | ✓ | 44 / 56 | 1 | 63 | 735 | 0.24 | 2.18 |
| **COSMI (Ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | 49 / 51 | 59 | 36 | 221,677 | 29.74 | 275.4 |

**Composition.** We pair a *target* clip, which provides the body motion, with a *source* clip, whose action is added on a free limb, if the two are compatible at two levels. The composition graph settles which interactions can share a body (Fig. 2): whole-body targets accept every source, and a one-hand target or a hand action accepts the other hand. Within a permitted pair, the semantic annotations decide whether the two actions can plausibly happen at once: a person cannot attend to two objects simultaneously, so two actions that both require gaze or both involve the head follow each other in sequence, a right- and a left-hand action on a sitting, walking, or standing body with a blended transition. Other pairs are composed simultaneously: the source limb, with the head and spine when the action needs them, replaces its counterpart on the target body, and its object follows through Eq. (1). We keep a composition only if it is physically plausible: every object keeps the contacts of its rule, the body does not penetrate itself, and no object penetrates another (Section C.7). Enumerating the graph exhaustively, with composed sequences serving as targets again, yields one- to three-object levels and a sequential level, and worn objects are added at every level, up to five objects per sequence. Captions are assembled from the annotations by a template and paraphrased.

**Statistics and quality.** The COSMI dataset contains 221,677 sequences and 275 hours of motion from 59 subjects and 36 objects, seven of them articulated, with up to five objects per sequence and left and right hands balanced at 49/51 (Table 1). Every composed sequence inherits the capture quality of its parents, which we verify with the metrics of HUMOTO (Lu et al., 2025) on the composed sequences and on their parent recordings (Section A.1). Foot sliding is unchanged, since composition alters neither the lower body nor the global trajectory, and jerk and penetration change by less than the captured datasets differ among themselves, human jerk the most, by up to half.

**Extensibility and benchmark.** Nothing in the composition is specific to a source dataset: the rules are stated by contacts, pose constraints, and reference joints on the unified body, so any clip on this body enters through a rule, as the AMASS walking sequences do. Partial observations enter as well, since every interaction is stored relative to the body part that performs it. A hand-object recording becomes a source clip once a body is fitted to it, which is less obvious, so we demonstrate it with HOT3D (Section 5). The exhaustive enumeration also makes generalization testable: a pairing such as walking with an apple can be held out as a whole while walking and eating an apple remain in training. Our benchmark builds on this and tests how well a model generalizes to five such unseen combinations and to an unseen object, held out with all its sequences (Section 5).

## 4 The COSMI Method

COSMI is a single text-conditioned model for a variable number of objects (Fig. 3): no parameter depends on the object count, so one model generates interactions with one to five objects, articulated and worn ones included, in a single denoising pass. Given a prompt $\mathcal{W}$, the initial pose of the person, and a set of $N$ objects, each defined by mesh, name, and initial pose, it generates the human motion $\mathcal{H}$ and the object motions $\mathcal{O}_1, \ldots, \mathcal{O}_N$ over $W{=}120$ frames, 4 s at 30 fps.

**Human representation.** We represent the person with the gender-neutral SMPL-X body model (Pavlakos et al., 2019) with a given shape, as in the COSMI dataset, in a canonical frame defined by the initial pose of the window. The human modality is

$$\mathcal{H} = \left\{ \boldsymbol{J}^t_{\mathcal{H}},\, \boldsymbol{\theta}^t_{\mathcal{H}},\, \boldsymbol{\theta}^t_{\text{hands}},\, \boldsymbol{c}^t_{\mathcal{H}} \right\}_{t=1}^{W}, \tag{3}$$

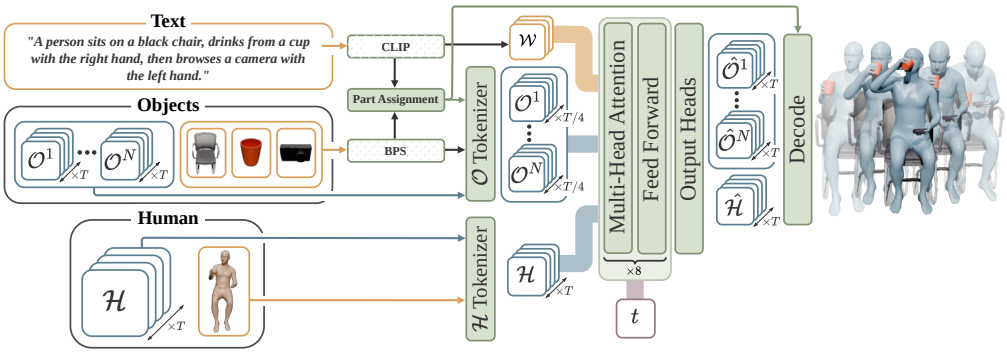



Figure 3: **COSMI method.** One transformer denoises the **human tokens** $\mathcal{H}$, one per frame, and the **object tokens** $\mathcal{O}_n$, one per four frames, at **diffusion step** $t$. Conditioned on the CLIP **text tokens** $\mathcal{W}$, the **initial poses**, and the **name** and **BPS** geometry of every object, it generates the human motion and each object motion relative to the joint that drives it.

where, for every frame $t$, $\boldsymbol{J}^t_{\mathcal{H}}$ holds global joint positions, $\boldsymbol{\theta}^t_{\mathcal{H}}$ the global rotations of the body joints in the 6D representation (Zhou et al., 2019), $\boldsymbol{\theta}^t_{\text{hands}}$ the PCA coefficients of the finger pose of both hands, and $\boldsymbol{c}^t_{\mathcal{H}}$ floor-contact scores of the feet, 256 values per frame in total (Section D.1).

**Object representation.** Each object $\mathcal{O}_n$ is defined analogously, with the index $n$ omitted:

$$\mathcal{O} = \left\{\boldsymbol{T}^t_{\mathcal{O}},\, \boldsymbol{a}^t_{\mathcal{O}},\, \{\boldsymbol{T}^t_{\mathcal{O}|j}\}_{j\in\mathcal{J}_{\text{rel}}},\, \boldsymbol{c}^t_{\mathcal{O}}\right\}_{t=1}^{W}. \tag{4}$$

The *absolute pose* $\boldsymbol{T}^t_{\mathcal{O}}$ is the rotation and the center of mass of the object in the canonical frame of the window, and the *hinge channel* $\boldsymbol{a}^t_{\mathcal{O}}$ encodes the articulation angle of objects with a moving part. The *relative poses* $\boldsymbol{T}^t_{\mathcal{O}|j}$ express the same pose in the frames of six candidate driving joints $\mathcal{J}_{\text{rel}} = \{\text{spine}, \text{head}, \text{both wrists}, \text{both feet}\}$, through Eq. (1), so that an object held in a hand has a nearly constant pose in the wrist frame whatever the rest of the body does. Finally, the *contact scores* $\boldsymbol{c}^t_{\mathcal{O}}$ are soft indicators of contact between the object and 15 body regions (Section D.2).

**Conditioning.** The text prompt is encoded with CLIP (Radford et al., 2021) and enters the model both as a pooled sentence embedding and as the sequence of word-piece embeddings, the text tokens $\mathcal{W}$. Following HIMO (Lv et al., 2024), we also condition on the first frame of the window, the initial pose of the person and of every object. Each object is further described by its name and by an embedding of its geometry, computed from basis point set features (Prokudin et al., 2019) with a hierarchical encoder (Section D.3). Since the object names appear in the captions and are encoded with the same CLIP model, the text tokens, as in MotionDiffuse (Zhang et al., 2024b), let every object be related to the words that refer to it, which the pooled embedding cannot.

**Denoiser.** The denoiser $\text{COSMI}_\psi$ is a transformer (Vaswani et al., 2017) with one token per frame of the human motion $\mathcal{H}$ and of every object motion $\mathcal{O}_n$, all attending to each other and to the conditioning tokens. Since attention cost grows quadratically with the tokens, we chunk every four frames of an object into one token. A smoothing convolution decodes the output tokens to per-frame parameters, without which the object motion has a seam every four frames (Section D.4).

**Object slots.** The token streams of the objects, which we call slots, receive no slot-specific embedding: all slots share the input and output projections and differ only in their conditioning, and masks exclude empty slots from attention and from the per-object losses. The network is therefore permutation-equivariant with respect to the objects, and none of its parameters depends on their number, which is set at run time rather than in the architecture, with up to five slots in training.

**Part assignment.** We say that a joint $j \in \mathcal{J}_{\text{rel}}$ *drives* an object if the object moves with this joint: an object held in a hand, worn on the head or back, or kicked with a foot is driven by the wrist, head, spine, or foot. A chair, in contrast, is not driven but only supports the person who sits on it. A lightweight transformer head predicts the driving joints of every object from the conditions alone, supervised with labels derived from contacts and joint motion (Section D.5). Its prediction enters the denoiser as an additional input and selects the decoding: a driven object is placed through the relative pose $\boldsymbol{T}^t_{\mathcal{O}|j}$ of its driving joint, and an undriven object keeps its absolute pose $\boldsymbol{T}^t_{\mathcal{O}}$.

**Objective.** We use the DDPM formulation (Ho et al., 2020) with a cosine noise schedule of 1000 steps (Nichol & Dhariwal, 2021), and the network predicts the clean sample (Ramesh et al., 2022). Let $\boldsymbol{x} = (\mathcal{H}, \mathcal{O}_1, \ldots, \mathcal{O}_N)$ denote a window, $\boldsymbol{x}^\tau$ its noisy version at diffusion step $\tau$, and $\mathcal{C}$ the conditions. The denoiser $\mathrm{COSMI}_\psi$ is trained with

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x},\tau}\; w(\tau) \sum_k \lambda_k \, \mathcal{L}_{\mathrm{k}}\big(\mathrm{COSMI}_\psi(\boldsymbol{x}^\tau; \tau, \mathcal{C}), \boldsymbol{x}\big), \tag{5}$$

where $w(\tau)$ is the min-SNR weight (Hang et al., 2023). The terms $\mathcal{L}_{\mathrm{k}}$ cover the reconstruction of the parameters and their frame-to-frame differences, the contact scores, and the plausibility of the body through foot-skating and forward-kinematics penalties. A consistency term further poses points of the object surface with the absolute and with each relative pose, which describe the same motion, and penalizes their distance to the ground truth (all weights $\lambda_k$ in Section D.6).

**Training and inference.** Windows are cut at random offsets and paired with one of their captions. We drop the text prompt for classifier-free guidance (Ho & Salimans, 2022) and, for robustness to missing inputs, also worn objects, names, and the part-assignment input (Section D.7). We sample with 100 DDPM steps respaced from the 1000 training steps and a classifier-free guidance scale of 2.5, and generate longer sequences window by window with half-window blended overlaps.

**Contact-consistent guidance.** The model predicts finger-object contacts in $\boldsymbol{c}^t_{\mathcal{O}}$ and foot-floor contacts in $\boldsymbol{c}^t_{\mathcal{H}}$, but its geometry does not always agree with them. Since the contact labels decrease with the distance between a finger and the object surface and with the speed of a foot, a predicted contact implies an upper bound on this distance or speed, which holds exactly on the training data. In the last sampling steps, gradient steps on the predicted clean sample penalize violations of these bounds (Eq. (10)), which closes the gaps between fingers and objects and stops planted feet.

## 5 Experiments

We evaluate on our benchmark, compare with HIMO on its dataset, and ablate COSMI. Figure 4 shows HOT3D (Banerjee et al., 2025) recordings turned into source clips by fitting a body to the headset and the MANO (Romero et al., 2017) hands (Section A.7), and Fig. 5 compares COSMI with the baselines. Composed sequences also animate clothed scans (Wang et al., 2024a) registered as SMPL-X+D (Kostyrko et al., 2026), held objects following the hands (Figs. 1 and 9).

**Data split.** A split by subject, the common choice for captured datasets, says little about generalization here, because the clips of a held-out subject would still occur in thousands of training sequences. Instead, we split by composition and hold out what a multi-object model is expected to generalize to. The *unseen object* is the black chair, held out with all its sequences, a new instance of a seat category whose other members remain in training, so the model has to keep the person seated on it from its geometry, name, and initial pose. The *unseen combinations* are five pairings whose parts are common in training but never occur together: walking with an apple, sitting with a flip phone, a camera together with a mug, a foot interaction while drinking, and standing with a water bottle. The held-out sequences form the test set, 35k sequences against 187k for training.

**Protocol.** We train a single model on all training sequences and evaluate the same weights throughout. The *two-object* and *three-object* tiers follow the setting of HIMO (Lv et al., 2024), test sequences with exactly two or three rigid interacted objects and no worn objects, and the *all-objects* tier contains every test sequence, with one to five objects, articulated and worn included. Every test sequence is generated three times over its full length from its caption and initial poses.

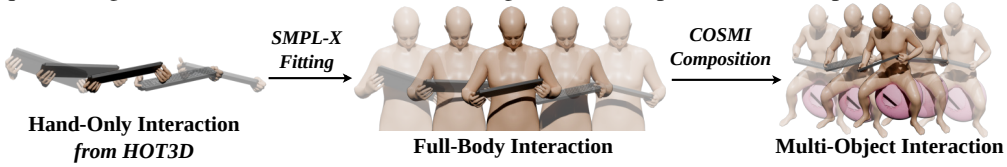



Figure 4: **Hand-object recordings as source clips.** A HOT3D recording of the hands and the object (left) becomes a full-body clip once SMPL-X is fitted to the headset and the MANO hands (middle), and the composition pairs it with a sitting clip like any captured source (right).

Table 2: **All-objects tier.** One to five objects per sequence. COSMI is closest to the text and to the ground-truth contacts and has the lowest hinge error. Data denotes the ground-truth test sequences, $\mathcal{H}$ and $\mathcal{O}$ the human and the object metrics. Values are means over three generations, with 95 % confidence intervals as superscripts for the learned evaluator. Best in bold, second best underlined, diversity best when closest to the data and the hinge sweep when closest to 1.

| Method | Text-to-motion | | | | | $\mathcal{H}$ | $\mathcal{O}$ | Contact | | | Articulated | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FID↓ | Top-1↑ | Top-3↑ | MM-Dist↓ | Div.→ | Foot sl.↓ | Jerk↓ | T-Acc↑ | B-Acc↑ | Pen. (%)↓ | Hinge err.↓ | Sweep→ |
| Data | – | 0.740 | 0.919 | 0.740 | 4.31 | 1.879 | 0.366 | – | – | 34.60 | – | – |
| MDM | $0.148^{\pm.002}$ | $0.619^{\pm.001}$ | $0.861^{\pm.002}$ | $1.089^{\pm.003}$ | $4.11^{\pm.28}$ | 3.290 | 0.495 | 0.901 | 0.934 | 37.49 | 0.547 | **0.846** |
| PriorMDM | $0.077^{\pm.000}$ | $0.664^{\pm.001}$ | $0.876^{\pm.001}$ | $0.933^{\pm.002}$ | **4.18**$^{\pm.26}$ | 1.874 | **0.390** | 0.938 | 0.936 | 35.08 | 0.557 | 0.707 |
| COSMI (ours) | **0.049**$^{\pm.000}$ | **0.696**$^{\pm.003}$ | **0.900**$^{\pm.001}$ | **0.827**$^{\pm.002}$ | $4.15^{\pm.32}$ | **1.395** | 0.409 | **0.980** | **0.968** | **33.96** | **0.437** | 0.665 |

Table 3: **Two- and three-object tiers.** Rigid objects only and no worn objects, the setting of HIMO. Format as in Table 2.

| | Method | Text-to-motion | | | | | $\mathcal{H}$ | $\mathcal{O}$ | Contact | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FID↓ | Top-1↑ | Top-3↑ | MM-Dist↓ | Div.→ | Foot sl.↓ | Jerk↓ | T-Acc↑ | B-Acc↑ | Pen. (%)↓ |
| 2 obj. | Data | – | 0.510 | 0.799 | 0.539 | 2.91 | 2.478 | 0.635 | – | – | 11.92 |
| | MDM | $0.260^{\pm.005}$ | $0.294^{\pm.003}$ | $0.544^{\pm.005}$ | $1.132^{\pm.003}$ | $2.99^{\pm.18}$ | 6.598 | 0.863 | 0.919 | 0.915 | 21.49 |
| | PriorMDM | $0.090^{\pm.002}$ | $0.398^{\pm.005}$ | $0.692^{\pm.002}$ | $0.768^{\pm.001}$ | $3.03^{\pm.23}$ | 2.641 | 0.570 | 0.969 | 0.930 | 16.70 |
| | HIMO | $0.095^{\pm.003}$ | $0.391^{\pm.003}$ | $0.670^{\pm.002}$ | $0.964^{\pm.002}$ | **2.96**$^{\pm.17}$ | 5.406 | 1.450 | 0.818 | 0.885 | 28.68 |
| | COSMI (ours) | **0.066**$^{\pm.002}$ | **0.405**$^{\pm.007}$ | **0.699**$^{\pm.003}$ | **0.737**$^{\pm.007}$ | $2.78^{\pm.21}$ | **1.974** | **0.466** | **0.993** | **0.962** | **11.94** |
| 3 obj. | Data | – | 0.298 | 0.580 | 1.032 | 2.77 | 1.407 | 0.271 | – | – | 37.15 |
| | MDM | $0.424^{\pm.010}$ | $0.192^{\pm.002}$ | $0.424^{\pm.001}$ | $1.263^{\pm.003}$ | $2.82^{\pm.18}$ | 1.191 | 0.400 | 0.836 | 0.923 | 38.62 |
| | PriorMDM | $0.313^{\pm.003}$ | $0.218^{\pm.001}$ | $0.454^{\pm.002}$ | $1.204^{\pm.003}$ | $2.74^{\pm.09}$ | 1.247 | **0.308** | 0.901 | 0.921 | 39.97 |
| | HIMO | $0.284^{\pm.003}$ | $0.243^{\pm.002}$ | **0.516**$^{\pm.000}$ | $1.233^{\pm.001}$ | $2.84^{\pm.14}$ | 2.311 | 0.723 | 0.808 | 0.917 | 42.85 |
| | COSMI (ours) | **0.157**$^{\pm.002}$ | **0.244**$^{\pm.001}$ | $0.507^{\pm.002}$ | **1.056**$^{\pm.003}$ | **2.76**$^{\pm.14}$ | **0.932** | 0.399 | **0.968** | **0.966** | **37.90** |

**Metrics.** A learned evaluator, a text-motion co-embedding trained on our training split, provides FID (Heusel et al., 2017), R-precision (top-1 and top-3), multimodal distance (MM-Dist), and diversity, the metrics of text-to-motion generation (Guo et al., 2022; Lv et al., 2024). Foot sliding and object jerk follow HUMOTO (Lu et al., 2025), penetration counts the frames in which the body enters an object, and all three are measured on the decoded meshes. Temporal and body contact accuracy (T-Acc, B-Acc) (Peng et al., 2026) compare the generated body-object contacts with those of the ground truth per frame and per body marker. For articulated objects, we report the hinge-angle error and the sweep ratio of the generated to the ground-truth hinge motion, and all values are means over three generations (definitions in Section B).

**Baselines.** The closest prior work is HIMO (Lv et al., 2024), whose benchmark also adapts MDM (Tevet et al., 2023) and PriorMDM (Shafir et al., 2024) to several objects. We re-implement MDM and PriorMDM on our representation with our initial-pose conditioning, schedule, and losses, but with a pooled text embedding and a small geometry encoder, so that each is a single model evaluated in all tiers. MDM has one token per frame, whereas PriorMDM has separate human and object streams with one communication block. HIMO supports rigid objects only, so its two models are trained on the two- and three-object tiers in its format and scored with our evaluator. The concurrent SK-HOI (Peng et al., 2026) is not compared, since its full code is unavailable and its test-time optimization of about four minutes per sequence is out of reach for 35k sequences.

**Implementation details.** COSMI has 30.4M parameters, between MDM (20.7M) and PriorMDM (39.6M), and trains in eight hours on one RTX 5090. A window is generated in 0.16 s with classifier-free guidance, and in 1.1 s with the contact-consistent guidance (Section D.8).

**Comparison with baselines.** Table 2 and Fig. 5 compare the methods in the all-objects tier, COSMI with the contact-consistent guidance of Section 4. COSMI obtains the best FID, R-precision, MM-Dist, and foot sliding, and its penetration stays at the ground-truth level, the residual penetration of the captures, which the baselines exceed. COSMI has the lowest hinge-angle error but the smallest hinge sweep, two thirds of the ground truth (Section A.6). The two- and three-object tiers agree (Table 3), the baselines up to seventeen points above the ground-truth penetration with two objects, except that HIMO, trained per tier, has the best top-3 R-precision with three objects.

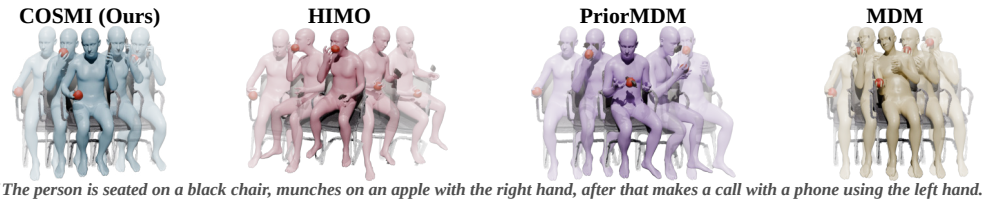



Figure 5: **Qualitative comparison.** Generations for a test prompt with the unseen black chair and two hand actions in sequence. COSMI keeps the person seated, and generates plausible contacts.

Table 4: **Unseen object and combinations.** Contact accuracy on the test sequences with the unseen object, the black chair, and with the five unseen combinations.

| Method | black chair | | walk + apple | | sit + flip phone | | camera + mug | | foot + drink | | stand + bottle | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T-Acc↑ | B-Acc↑ | T-Acc↑ | B-Acc↑ | T-Acc↑ | B-Acc↑ | T-Acc↑ | B-Acc↑ | T-Acc↑ | B-Acc↑ | T-Acc↑ | B-Acc↑ |
| MDM | 0.823 | 0.917 | 0.971 | 0.946 | 0.903 | 0.930 | 0.955 | 0.953 | 0.926 | 0.958 | 0.977 | 0.929 |
| PriorMDM | 0.895 | 0.916 | 0.983 | 0.950 | 0.941 | 0.937 | 0.965 | 0.956 | 0.942 | 0.954 | 0.969 | 0.939 |
| COSMI (ours) | **0.975** | **0.960** | **0.995** | **0.974** | **0.981** | **0.959** | **0.984** | **0.984** | **0.953** | **0.981** | **0.993** | **0.967** |

**Comparison on the HIMO dataset.** To compare with HIMO on captured data, we also re-train COSMI on the HIMO dataset and evaluate it with the evaluator of HIMO (Section A.3). COSMI has the highest top-3 R-precision in both tiers and, with three objects, also the best FID.

**Ablations.** We ablate the guidance, the auxiliary losses, the part assignment, the text tokens and object names, the geometry encoder, and the chunk size, with results in Section A.2. The auxiliary losses matter most, and the part assignment improves both text alignment and contact accuracy. Text tokens and object names improve text alignment, and the guidance significantly decreases foot sliding and hand-object gaps without damaging the other metrics.

**Generalization.** Table 4 reports the contact accuracy on the unseen object and on each unseen combination. On the combinations, all methods match or exceed their accuracy on the whole test set, with one drop of three points, so they appear to learn the base motion and the hand interaction independently of each other. COSMI is the most accurate on every subset, and the unseen object is harder: the baselines lose four to eight points of temporal contact accuracy on the black chair, whereas COSMI stays within one point of its value on the whole test set.

## 6 Conclusions

Everyday activities combine objects in countless ways, from drinking while sitting to eating an apple while walking, and the combinations to capture grow faster than any recording effort. The COSMI dataset shows that they need not all be recorded: interactions are local, so a contact-consistent clip of one interaction carries its limb and object onto a body that performs another. A language model and geometric checks admit only the combinations that a person would plausibly perform and that remain physically valid, so what emerges is not stitched clips but semantically meaningful, realistic multi-object interactions. The result is 222k sequences and 275 hours with up to five objects, balanced between the hands by mirroring and nearly as plausible as their captures. Nothing in the composition is tied to a source dataset, so every further capture multiplies the combinations, full-body datasets through a rule and hand-object recordings once a body is fitted to them, as HOT3D shows. The COSMI method, a text-driven diffusion transformer grounding the objects to the body parts that move them, generates a variable number of objects end to end. The baselines and COSMI alike generalize to combinations never seen in training, and COSMI also to an unseen object.

**Limitations and future work.** The composition covers sustained single interactions at once or in sequence. Extending it to multi-stage activities such as cooking or assembling, whose steps and objects depend on each other, and bringing the data together with interactable scenes, where a cup is taken from a shelf on the way, are the directions we consider most promising. On the method side, the guidance closes the gaps between fingers and objects, but the finger motion has yet to reach the precision of dedicated grasp synthesis. Finally, text is a coarse interface: the prompt fixes the objects and actions, but not the path of the person or the timing of sequential actions.

### AI USE STATEMENT

In this work, we used generative AI tools for dataset annotation, method implementation, background research and writing polishing. A language model, Qwen3-14B (Yang et al., 2025), is a component of the dataset pipeline (Section 3): it corrects the action label of every clip, writes the sentence describing it, and states per action-object pair whether the action requires gaze at the object and involves the head, which decides which clips are composed and seeds the captions. It neither generates nor selects motion, its prompts and an example answer are shown in Fig. 11, and every composition it admits passes the geometric checks of Section C.7. LLM-based coding assistants helped implement the method and the evaluation code, which the authors reviewed and tested. We have not used generative AI tools to develop the conceptual framework, to propose hypotheses, to design the experiments, or to interpret the results. Additionally, we used generative AI tools for writing: drafting and editing passages for readability, searching and summarizing related work, and formatting references. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

No new data of human subjects were recorded: the COSMI dataset is composed from the public motion-capture datasets named in Section 3, used under their research licenses, as are HOT3D and 4D-DRESS in the qualitative demonstrations and HIMO in the comparison. The subjects are represented by a gender-neutral SMPL-X body, which carries neither face nor appearance, so that no person is identifiable, except in the scans of 4D-DRESS, which are shown as their authors released them. Since the composed sequences derive from these sources, the dataset pipeline will be released, so that the COSMI dataset can be rebuilt from the sources under their own terms. The models generate the motion of a body model for animation, and we see no avenue for misuse beyond that of motion synthesis in general.

### REPRODUCIBILITY STATEMENT

Code, models, and the dataset pipeline will be released. The construction of the COSMI dataset is specified in Sections 3 and C: the sources and their preprocessing, the thirteen interaction rules (Table 8), the annotation prompts with the language model used (Fig. 11), the mirroring (Eq. (6)), the composition graph (Fig. 13), and the feasibility checks with their thresholds. The benchmark split, with the unseen object and the five unseen combinations, is defined in Section 5 and Table 4. The method is specified in Sections 4 and D: the representation of the person and the objects, the architecture, the loss terms and their weights, all training hyperparameters with the hardware and the training time, the sampler, and the guidance, with generation times in Table 9. The evaluator and its training protocol, and the physical metrics with their thresholds and units, are defined in Section B, and the baselines are re-implementations described in Section 5, with their sizes in Section D.4.

## REFERENCES

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal action composition for 3D humans. In *International Conference on 3D Vision*, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9950–9961, 2023.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7071, 2025.

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 457–469, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15914–15925, 2022.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8726–8737, 2023.

Songjin Cai, Linjie Zhong, Ling Guo, and Changxing Ding. ViHOI: Human-object interaction synthesis with visual priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9044–9053, 2021.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023.

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.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954, 2023.

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. GigaHands: A massive annotated dataset of bimanual hand activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17461–17474, 2025.

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*, 42(2):1–12, 2023.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5142–5151, 2022.

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

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3193–3203, 2020.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7407–7417, 2023.

Wenkun He, Yun Liu, Ruitao Liu, and Li Yi. SyncDiff: Synchronized motion diffusion for multi-body human-object interaction synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11731–11743, 2025.

Rachel Heck, Lucas Kovar, and Michael Gleicher. Splicing upper-body actions with locomotion. *Computer Graphics Forum*, 25(3):459–466, 2006.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30:6626–6637, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. *International Journal of Computer Vision*, 132(7):2551–2566, 2024.

Leslie Ikemoto and David A. Forsyth. Enriching a motion collection by transplanting limbs. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 99–108, 2004.

Yihong Ji, Jinsong Zhang, He Hu, and Hongbo Xu. HOIMask: Towards generative masked modeling for human object interaction generation. In *European Conference on Computer Vision*, pp. 450–468, 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9331–9342, 2023.

Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22735–22745, 2025.

Hwanhee Jung, Seunggwan Lee, Jeongyoon Yoon, SeungHyeon Kim, Giljoo Nam, Qixing Huang, and Sangpil Kim. Decoupled generative modeling for human-object interaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing everyday home activities towards 3D generative modeling of human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1816–1828, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kaen Kogashi, Anoop Cherian, and Meng-Yu Jennifer Kuo. MMHOI: Modeling complex 3D multi-human multi-object interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1512–1521, 2026.

Margaret Kostyrko, Yuxuan Xue, Garvita Tiwari, and Gerard Pons-Moll. Revisiting avatar-as-image: High-fidelity registration is all you need. In *European Conference on Computer Vision*, pp. 97–115, 2026.

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. *ACM Transactions on Graphics*, 21(3):473–482, 2002.

Chuqiao Li, Xianghui Xie, Yong Cao, Andreas Geiger, and Gerard Pons-Moll. FrankenMotion: Part-level human motion generation and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 42(6):1–11, 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*, pp. 54–72, 2024.

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

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual Tool-ACtion-Object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024.

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. CORE4D: A 4D human-object-human interaction dataset for collaborative object REarrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1769–1782, 2025.

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

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*, pp. 300–318, 2024.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465, 2024.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451, 2019.

Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809, 2016.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning*, pp. 1820–1864, 2023.

Jeonghyeon Na, Sangwon Baik, Inhee Lee, Junyoung Lee, and Hanbyul Joo. Learning to generate human-human-object interactions from textual descriptions. *Advances in Neural Information Processing Systems*, 38:142894–142912, 2025.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021.

Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. TokenHSI: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5379–5391, 2025.

Xiaokang Pan, Zhizhong Zhang, Yangyuan Liu, Zhuoran Chen, Zhiwei Zhang, Bin Ji, Mingang Chen, Yong Xie, Jingyu Gong, Xuhong Wang, Xin Tan, and Yuan Xie. Fast-HOI: Fast human-object interaction synthesis via distilled interaction prior and physical constrains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 3719–3728, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2869–2879, 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. *arXiv preprint arXiv:2608.03158*, 2026.

Ilya A. Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. TriDi: Trilateral diffusion of 3D humans, objects, and interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5523–5535, 2025.

Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3D human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1911–1921, 2024.

Huaijin Pi, Zhi Cen, Zhiyang Dou, and Taku Komura. CoDA: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects. *Advances in Neural Information Processing Systems*, 38:35849–35871, 2025.

Chrysa Pratikaki, Pablo Ruiz-Ponce, Jiankang Deng, Stefanos Zafeiriou, and Rolandos Alexandros Potamias. Dex2HOI: Dexterous bimanual two-object interaction generation. *arXiv preprint arXiv:2605.30444*, 2026.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4331–4340, 2019.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with English labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 722–731, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):1–17, 2017.

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit H. Bermano. HOIDiNi: Human-object interaction through diffusion noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. *Advances in Neural Information Processing Systems*, 37:140032–140065, 2024.

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

Mengyi Shan, Lu Dong, Yutao Han, Yuan Yao, Tao Liu, Ifeoma Nwogu, Guo-Jun Qi, and Mitch Hill. Towards open domain text-driven synthesis of multi-person motions. In *European Conference on Computer Vision*, pp. 67–86, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shanlin Sun, Jiaqi Xu, Gabriel De Araujo, Shenghan Zhou, Hanwen Zhang, Ziheng Huang, Chenyu You, and Xiaohui Xie. CoMA: Compositional human motion generation with multi-modal agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9206–9214, 2026.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, pp. 581–600, 2020.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13253–13263, 2022.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4D-DRESS: A 4D dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 550–560, 2024a.

Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Weishuai Zeng, Qin Jin, and Zongqing Lu. Scaling large motion models with million-level human motions. In *International Conference on Machine Learning*, pp. 65802–65827, 2025a.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. SkillMimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17540–17549, 2025b.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3D scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. InterControl: Zero-shot human interaction generation by controlling every joint. *Advances in Neural Information Processing Systems*, 37:105397–105424, 2024b.

Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, and Lingjie Liu. Text2Interact: High-fidelity and diverse text-to-two-person interaction generation. In *International Conference on Learning Representations*, 2026.

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

Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *International Conference on Learning Representations*, 2024.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. GRAIL: Generating humanoid loco-manipulation from 3D assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026.

Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10003–10015, 2024.

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. 14882–14894, 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. InterDreamer: Zero-shot text to 3D dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37: 52858–52890, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7048–7060, 2025a.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12266–12277, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456, 2024.

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

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14730–14740, 2023a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8834–8845, 2023b.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-$M^3$: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 516–526, 2024a.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024b.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27694–27705, 2025.

Yangsong Zhang, Abdul Ahad Butt, Gül Varol, and Ivan Laptev. InterPose: Learning to generate human-object interactions from large-scale web videos. In *International Conference on 3D Vision*, pp. 750–759, 2026.

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'M HOI: Inertia-aware monocular capture of 3D human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 729–741, 2024.

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

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.

Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. ParCo: Part-coordinating text-to-motion synthesis. In *European Conference on Computer Vision*, pp. 126–143, 2024.

# APPENDIX

**Abstract.** This appendix extends the experiments of the main paper and completes the description of the dataset and the method. In Section A, we measure the quality of the composed sequences against their parent recordings, ablate the components of the COSMI method, compare with HIMO on its own dataset, and present qualitative results and failure cases. We also show that hand-object recordings enter the composition once a body is fitted to them, and we render composed sequences on clothed scans. In Section B, we define the evaluator behind the learned metrics and the physical metrics with their thresholds and units. In Section C, we detail the construction of the dataset, from the preprocessing of the sources and the interaction rules to the mirroring that balances the hands and the feasibility checks of the composition. Finally, in Section D, we describe the COSMI method in full, including the representations of the person and the objects, the architecture, the training objective, and the contact-consistent guidance.

**A Additional Results** **19**
A.1 Quality of composed sequences . . . 19
A.2 Ablations . . . 19
A.3 Comparison on the HIMO dataset . . . 19
A.4 Held-out subsets . . . 19
A.5 Qualitative comparison . . . 20
A.6 Failure cases . . . 20
A.7 Hand-object recordings . . . 22
A.8 Clothed scans . . . 22

**B Metric Definitions** **22**
B.1 Evaluator . . . 22
B.2 Physical metrics . . . 23

**C Dataset Details** **26**
C.1 Preprocessing . . . 26
C.2 Interaction rules . . . 26
C.3 Annotation prompts . . . 26
C.4 Joint-relative objects . . . 26
C.5 Mirroring . . . 29
C.6 Composition graph . . . 29
C.7 Merge and feasibility . . . 29
C.8 Sequential composition and levels . . . 29

**D Method Details** **33**
D.1 Human representation . . . 33
D.2 Object representation . . . 33
D.3 Geometry encoder . . . 33
D.4 Architecture . . . 33
D.5 Part-assignment head . . . 33
D.6 Loss terms . . . 34
D.7 Training details . . . 34
D.8 Runtime . . . 34
D.9 Long sequences . . . 34
D.10 Contact-consistent guidance . . . 35

Table 5: **Quality of composed sequences.** Physical plausibility of the captured parent sequences of every source (Original) and of the composed sequences built from them (COSMI), measured with the metrics of HUMOTO (Lu et al., 2025). The difference to the parents is given in parentheses and colored by its relative size: green for an improvement, amber for an increase below 50 %, red for an increase of 50 % or more, and gray for a negligible change. Metrics that do not apply to a source are marked with –.

| Dataset *(parent sequences only)* | Human Motion | | | | Object Motion | | Human-Object Penetration | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Foot sliding (cm) ↓ | | Jerk (m/s$^3$) ↓ | | Jerk (m/s$^3$) ↓ | | Body (cm) ↓ | | Hands (cm) ↓ | |
| | Original | COSMI | Original | COSMI | Original | COSMI | Original | COSMI | Original | COSMI |
| GRAB (Taheri et al., 2020) | 0.03 | 0.03 (+0.00) | 0.33 | 0.44 (+0.11) | 1.67 | 1.37 (−0.30) | 0.08 | 0.05 (−0.03) | 0.54 | 0.53 (−0.01) |
| BEHAVE (Bhatnagar et al., 2022) | 0.13 | 0.11 (−0.02) | 1.15 | 1.11 (−0.04) | 1.71 | 2.03 (+0.32) | 1.64 | 1.93 (+0.29) | – | – |
| ARCTIC (Fan et al., 2023) | 0.01 | 0.01 (+0.00) | 0.31 | 0.45 (+0.14) | 1.16 | 1.23 (+0.07) | 0.00 | 0.01 (+0.01) | 0.79 | 1.03 (+0.24) |
| InterCap (Huang et al., 2024) | 0.08 | 0.09 (+0.01) | 1.96 | 2.34 (+0.38) | 2.35 | 2.58 (+0.23) | 0.17 | 0.40 (+0.23) | 0.52 | 0.23 (−0.29) |
| OMOMO (Li et al., 2023) | 0.29 | 0.29 (+0.00) | 1.25 | 1.32 (+0.07) | 1.44 | 1.35 (−0.09) | 0.01 | 0.01 (+0.00) | – | – |
| AMASS (Mahmood et al., 2019) | 0.17 | 0.18 (+0.01) | 2.56 | 3.42 (+0.86) | – | 5.30 | – | – | – | – |

# A Additional Results

## A.1 Quality of composed sequences

Table 5 compares, for each of the six sources, the parent recordings with the composed sequences built from them, with the metrics of HUMOTO (Lu et al., 2025): foot sliding and jerk of the human, jerk of the objects, and the depth to which objects penetrate the body and the hands. Foot sliding is unchanged, since composition alters neither the lower body nor the global trajectory. Human jerk rises by a fifth to nearly a half on GRAB, ARCTIC, InterCap, and AMASS, most in absolute terms for the walking sequences of AMASS, which carry the arm motion of a manipulation after composition, and object jerk and penetration change in both directions and worsen by less than the captured datasets differ among themselves, the largest relative rise being the body penetration of InterCap, from 0.17 to 0.40 cm. Objects on walking bodies have no captured counterpart and show the highest object jerk, since they move with the gait.

## A.2 Ablations

Table 6 changes one component of the model at a time. Guidance reduces foot sliding by a fifth and closes the finger-object gaps: on 512 test windows, the share of predicted contacts that violate their bound (Eq. (9)) by more than 5 mm drops from 18.9 % to 0.9 %. Text alignment and contact accuracy stay unchanged, since the guidance enforces the predicted contacts rather than correcting them, at the cost of slightly more object jerk and penetration. The other variants are sampled without guidance and compare with the unguided model. The auxiliary losses matter most: without them, foot sliding doubles and temporal contact accuracy drops by 17 points. The part assignment improves both text alignment and contacts, whereas the text tokens and the object names improve the text alignment only, by two and one points of top-1 R-precision. Replacing the hierarchical geometry encoder by an MLP changes little, slightly worse top-3 R-precision and multimodal distance against a better FID and object jerk, and the chunk size of four balances single-frame object tokens, which lower the temporal contact accuracy by two points, against chunks of eight, which lower the top-1 R-precision by two points.

## A.3 Comparison on the HIMO dataset

Table 7 compares COSMI, trained on the HIMO dataset and sampled without guidance, the setting closest to that of HIMO, with the values that HIMO reports for MDM, PriorMDM, and HIMO under its evaluator. COSMI has the highest top-3 R-precision in both tiers, 0.730 and 0.582 against 0.637 and 0.535 for HIMO, and with three objects also the lowest FID and MM-Dist and the diversity closest to the data. With two objects, HIMO, trained for this number of objects alone, has a lower FID, 1.481 against 4.896, and a lower MM-Dist, and our multimodality is the lowest in both tiers.

## A.4 Held-out subsets

The split assigns each test sequence to the first matching subset: black chair 12,883 sequences; walking with an apple 10,737; sitting with a flip phone 4,630; a foot interaction while drinking

Table 6: **Ablations.** Each row changes one component of the full model. The row w/o guidance samples without the contact-consistent guidance of Section 4, and all rows below it are sampled without guidance as well. Of these, w/o hier. enc. replaces the hierarchical geometry encoder by an MLP, $\mathcal{O}$ chunk $n$ uses $n$ frames per object token instead of 4, and aux. loss denotes the interaction and plausibility terms. The auxiliary losses matter most, part assignment improves the alignment with the text and the contacts, and text tokens and object names improve the alignment with the text. Format as in Table 2.

| Method | Text-to-motion | | | | $\mathcal{H}$ | $\mathcal{O}$ | Contact | | |
|---|---|---|---|---|---|---|---|---|---|
| | FID↓ | Top-1↑ | Top-3↑ | MM-Dist↓ | Foot sl.↓ | Jerk↓ | T-Acc↑ | B-Acc↑ | Pen. (%)↓ |
| Data | – | 0.740 | 0.919 | 0.740 | 1.879 | 0.366 | – | – | 34.60 |
| COSMI (ours) | $0.049^{\pm.000}$ | $0.696^{\pm.003}$ | $\mathbf{0.900}^{\pm.001}$ | $0.827^{\pm.002}$ | **1.395** | 0.409 | **0.980** | 0.968 | 33.96 |
| w/o guidance | $0.050^{\pm.000}$ | $\mathbf{0.697}^{\pm.003}$ | $\mathbf{0.900}^{\pm.001}$ | $\mathbf{0.825}^{\pm.002}$ | 1.744 | 0.397 | 0.979 | 0.970 | 33.55 |
| w/o hier. enc. | $0.041^{\pm.000}$ | $0.695^{\pm.001}$ | $0.893^{\pm.001}$ | $0.845^{\pm.000}$ | 1.757 | 0.364 | 0.975 | 0.971 | 33.62 |
| w/o names | $0.057^{\pm.000}$ | $0.687^{\pm.001}$ | $0.889^{\pm.001}$ | $0.877^{\pm.001}$ | 1.768 | 0.383 | 0.974 | 0.969 | 34.84 |
| $\mathcal{O}$ chunk 1 | $0.052^{\pm.000}$ | $0.686^{\pm.000}$ | $0.883^{\pm.001}$ | $0.886^{\pm.001}$ | 1.658 | **0.303** | 0.957 | **0.974** | **33.12** |
| $\mathcal{O}$ chunk 8 | $\mathbf{0.039}^{\pm.001}$ | $0.677^{\pm.001}$ | $0.881^{\pm.000}$ | $0.903^{\pm.002}$ | 1.714 | 0.393 | 0.979 | 0.970 | 33.66 |
| w/o $\mathcal{W}$ | $0.043^{\pm.000}$ | $0.675^{\pm.001}$ | $0.888^{\pm.000}$ | $0.907^{\pm.003}$ | 1.637 | 0.334 | **0.980** | 0.968 | 33.91 |
| w/o part assign. | $0.069^{\pm.000}$ | $0.644^{\pm.001}$ | $0.859^{\pm.002}$ | $0.982^{\pm.003}$ | 1.727 | 0.405 | 0.945 | 0.944 | 34.47 |
| w/o aux. loss | $0.256^{\pm.002}$ | $0.550^{\pm.002}$ | $0.783^{\pm.001}$ | $1.435^{\pm.003}$ | 3.633 | 0.449 | 0.810 | 0.910 | 36.25 |

Table 7: **Comparison on the HIMO dataset.** Values of the baselines as reported by HIMO (Lv et al., 2024), with the 95 % confidence intervals that HIMO reports, and the Data row is HIMO's Real row. COSMI is trained on the HIMO dataset and scored with the evaluator of HIMO over 20 replications, sampled without guidance, as the closest setting to that of HIMO. Format as in Table 2.

| | Method | Text-to-motion | | | | |
|---|---|---|---|---|---|---|
| | | FID↓ | Top-3↑ | MM-Dist↓ | Div.→ | MModal.↑ |
| 2 obj. | Data | $0.018^{\pm.006}$ | $0.799^{\pm.008}$ | $3.566^{\pm.011}$ | $11.40^{\pm.26}$ | – |
| | MDM | $6.846^{\pm.332}$ | $0.605^{\pm.010}$ | $8.019^{\pm.050}$ | $\mathbf{11.39}^{\pm.23}$ | $1.288^{\pm.211}$ |
| | PriorMDM | $7.852^{\pm.252}$ | $0.589^{\pm.003}$ | $7.251^{\pm.006}$ | $12.58^{\pm.15}$ | $1.591^{\pm.145}$ |
| | HIMO | $\mathbf{1.481}^{\pm.043}$ | $0.637^{\pm.003}$ | $\mathbf{3.649}^{\pm.010}$ | $11.66^{\pm.20}$ | $\mathbf{1.786}^{\pm.057}$ |
| | COSMI (ours) | $4.896^{\pm.038}$ | $\mathbf{0.730}^{\pm.005}$ | $4.436^{\pm.010}$ | $10.76^{\pm.12}$ | $0.691^{\pm.015}$ |
| 3 obj. | Data | $0.181^{\pm.044}$ | $0.699^{\pm.005}$ | $3.770^{\pm.028}$ | $9.77^{\pm.22}$ | – |
| | MDM | $4.571^{\pm.110}$ | $0.502^{\pm.014}$ | $6.314^{\pm.027}$ | $8.90^{\pm.29}$ | $0.789^{\pm.572}$ |
| | PriorMDM | $4.821^{\pm.203}$ | $0.514^{\pm.026}$ | $5.890^{\pm.023}$ | $9.34^{\pm.02}$ | $\mathbf{0.812}^{\pm.328}$ |
| | HIMO | $4.771^{\pm.164}$ | $0.535^{\pm.018}$ | $5.087^{\pm.041}$ | $8.95^{\pm.14}$ | $0.756^{\pm.055}$ |
| | COSMI (ours) | $\mathbf{2.815}^{\pm.043}$ | $\mathbf{0.582}^{\pm.011}$ | $\mathbf{4.366}^{\pm.022}$ | $\mathbf{9.63}^{\pm.14}$ | $0.741^{\pm.028}$ |

3,051; standing with a water bottle 2,018; a camera together with a mug 1,799. The remaining 186,559 sequences form the training set.

### A.5 Qualitative comparison

Figure 6 compares COSMI with the baselines on test prompts of the three-object and the all-objects tier, all from the unseen object or combinations. COSMI keeps every object in the hand that drives it over the whole sequence, an apple at the mouth or a camera in front of the face, whereas the baselines, which have no part assignment and attach an object to a joint only while its predicted contact score stays high, let them drift away from the hand or float beside the body once the arm moves. HIMO appears in the three-object tier only, as it only supports a fixed number of objects.

### A.6 Failure cases

Figure 7 shows the two failure modes of COSMI. Articulated objects are the first: the hinge moves less than in the ground truth, a sweep ratio of 0.665 in Table 2, and the moving part is sometimes off its true angle. The second is the unseen black chair, which is occasionally placed incorrectly relative

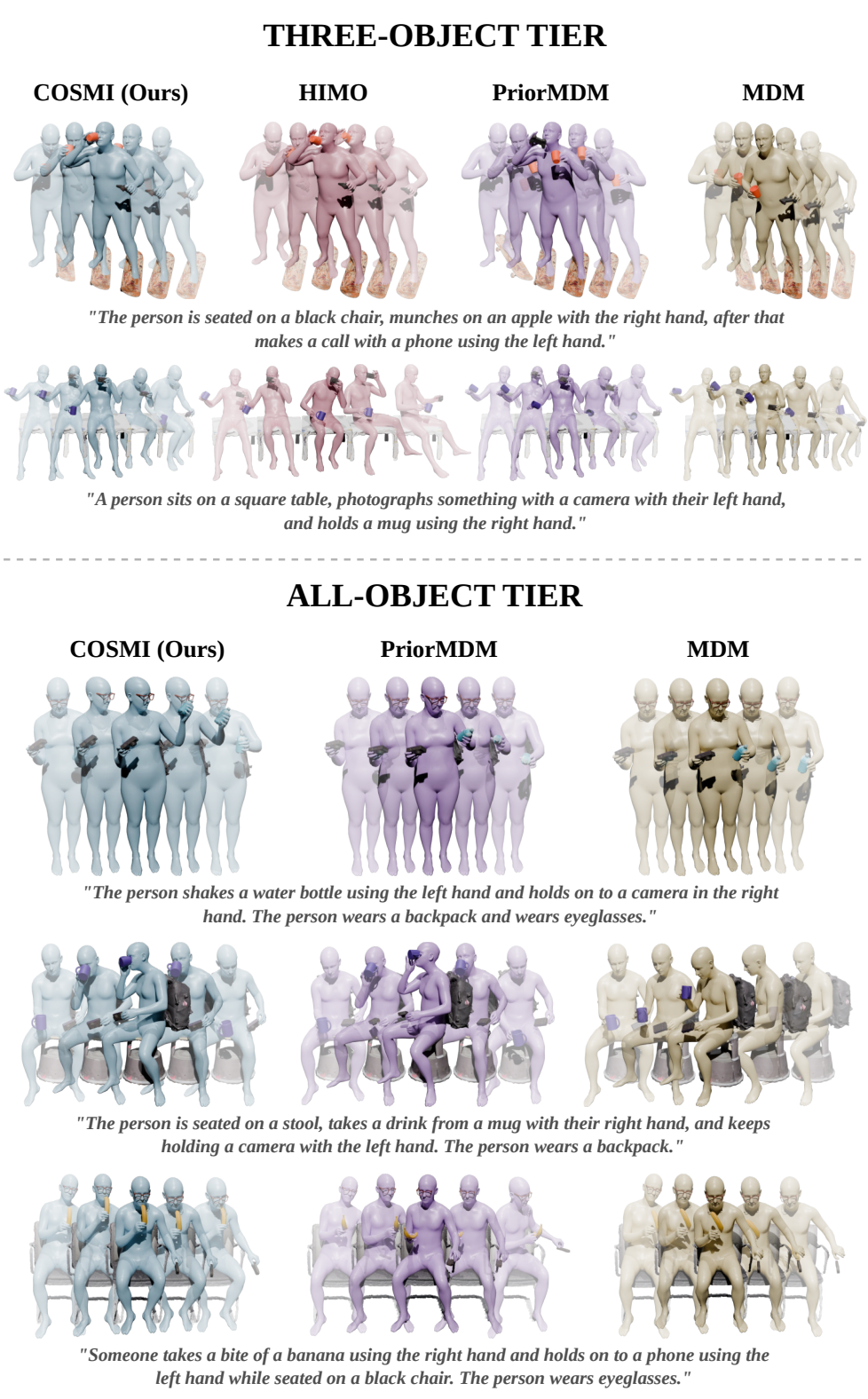
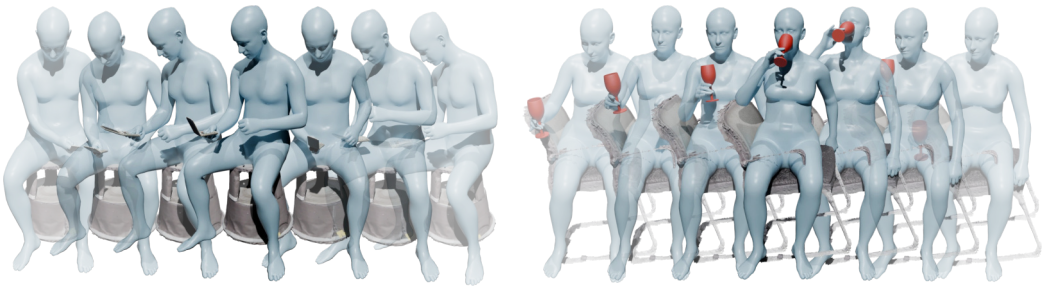



Figure 6: **Qualitative comparison with the baselines.** Generations for test prompts of the three-object tier (top) and of the all-objects tier (bottom), which includes more than 3 objects and therefore no HIMO. COSMI keeps every object in the hand that drives it, whereas the baselines let objects drift away from the hands or float beside the body.



*"The person manipulates a flip phone using the right hand, all while seated on a stool."*

*"Someone sips from a wine glass with the right hand while sitting down on a black chair."*

Figure 7: **Failure cases.** Left: the hinge of the flip phone has limited movement range. Right: the unseen black chair is placed off the body, which intersects its armrest.

to the body. The seats seen in training, such as the wooden chair, the yoga ball and the stool, have no armrests, so the model has no example of keeping the arms clear of them.

### A.7 Hand-object recordings

HOT3D (Banerjee et al., 2025) provides the trajectory of the head-mounted device, MANO (Romero et al., 2017) hand poses, and object poses. For every object, we keep the segments in which the head is stable and exactly one object is lifted and in contact with a hand, and place a neutral SMPL-X template with its head at the tracked head position, facing the object. The rotations of the upper body are then optimized so that the SMPL-X hand vertices match the MANO hands and the head faces the object, with a pose prior and a temporal smoothness term, and the finger poses are copied from MANO. The resulting clips contain a full body with the recorded hands and object, and the interaction rules process them like any other source clip, so HOT3D objects can be composed onto the walking, sitting, and large-object targets of the dataset (Fig. 8).

### A.8 Clothed scans

We use the SMPL-X+D registrations from (Kostyrko et al., 2026) of the 4D-DRESS scans (Wang et al., 2024a), per-vertex displacements on top of the body model, which we animate through the SMPL-X skinning weights with the joint transforms of the composed sequence. The displacements are attenuated along the forearms and removed on the hands, so that the recorded finger motion is preserved, and the objects are re-attached through their joint-relative transforms, so that a held object follows the hand of the scan (Figs. 1 and 9).

## B Metric Definitions

### B.1 Evaluator

The learned metrics follow the text-to-motion protocol (Guo et al., 2022) with an evaluator trained on our training split. A sequence is described per frame by the 24 joint positions and, per object slot, the 6D rotation, the center of mass, the hinge channel, and a presence flag, z-scored with the training statistics. A convolutional movement encoder, trained first as an autoencoder on 64-frame snippets and downsampling by four, feeds a bidirectional GRU motion encoder. The motion encoder is trained together with a projection of the frozen CLIP text embedding into a common 512-dimensional space with a contrastive margin loss, margin 3, where pairs with identical canonical captions are never negatives. R-precision and MM-Dist are computed on pools of 32 sequences, FID between the embeddings of the generations and of the ground-truth sequences, and diversity over 100 random pairs, each averaged over the three generations and reported with its 95 % confidence interval. As these metrics depend on the evaluator, we compare them only between methods scored with the same evaluator. Multimodality (MModal.), reported on the HIMO dataset only, is the mean distance between the embeddings of generations for the same caption, as in HIMO.

### B.2 PHYSICAL METRICS

Foot sliding (Li et al., 2024; Lu et al., 2025) is the horizontal displacement of the ankle and toe joints in frames where they are below a height threshold $H$, 8 cm for the ankles and 4 cm for the toes, which is one on the floor and zero at the threshold, summed, divided by the number of frames, and averaged over the four joints, in mm per frame. Object jerk is the mean norm of the third temporal difference of the object position, as in HUMOTO (Lu et al., 2025), in cm per frame$^3$. Penetration is the fraction of frames in which a body vertex lies deeper than 2 cm inside an object, the collision metric of SK-HOI (Peng et al., 2026). Table 5 instead uses the definitions of HUMOTO, whose penetration is the deepest intrusion of the body or hands into the object, averaged over frames. For the contact accuracy (Peng et al., 2026), a body marker is in contact when it lies within 2 cm of the object surface: T-Acc is the accuracy of the per-frame contact state, where a frame is in contact when any of the 138 markers is, and B-Acc the accuracy of the per-marker states. The hinge-angle error is the mean absolute difference between the generated and the ground-truth hinge angle, wrapped to $[-\pi, \pi]$, in radians, and the sweep ratio divides the total hinge motion of a generated sequence by that of the ground truth. The 95 % confidence intervals of these metrics are below 0.02, and below 0.2 percentage points for penetration.

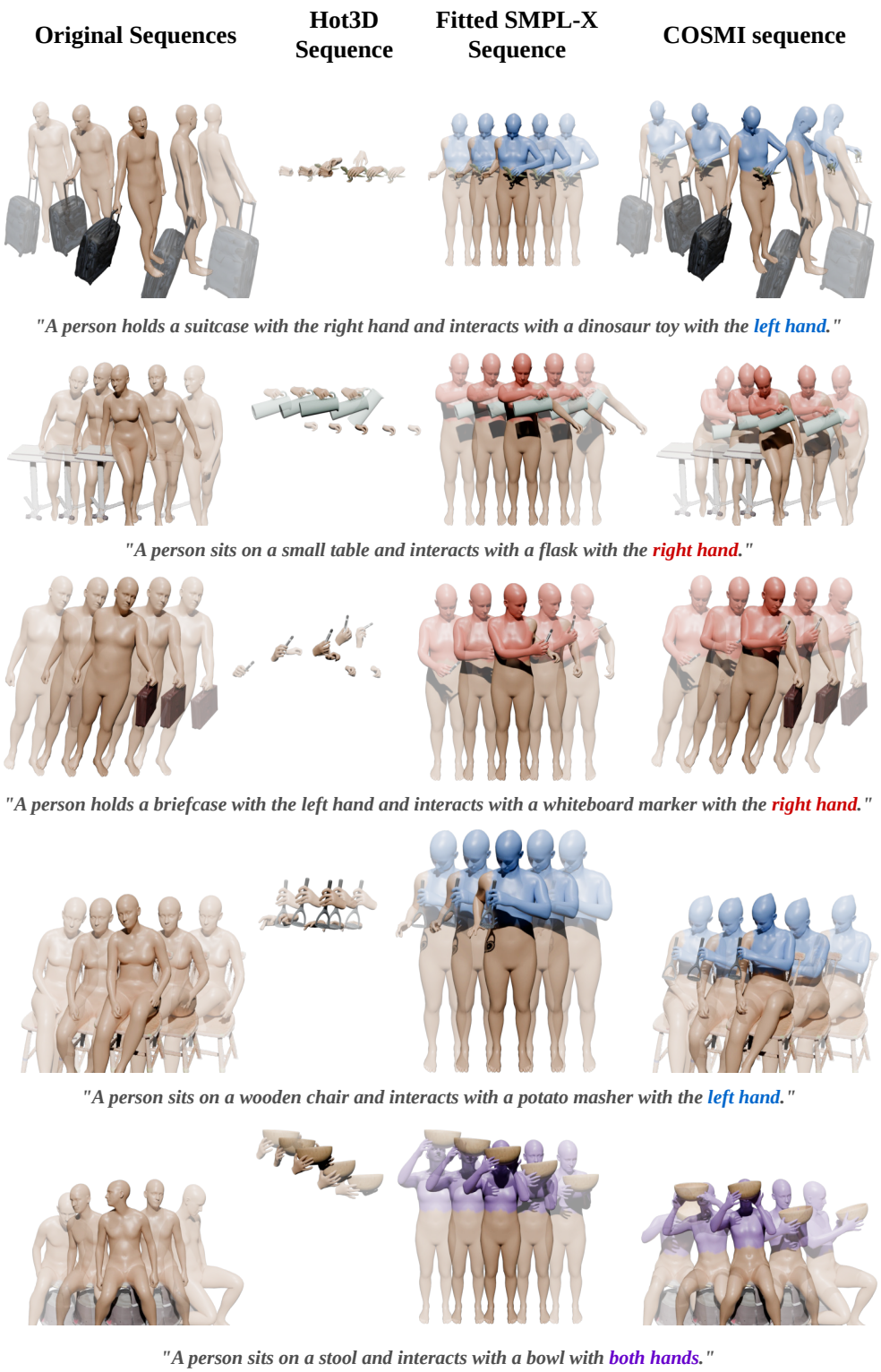



Figure 8: **Hand-object recordings as source clips.** A HOT3D recording (second column) provides the hands and the object. A body fitted to the head-mounted device and the MANO hands (third column) turns it into a source clip, which the rules and the composition process like any other, here onto sitting and large-object targets (first column), giving the composed sequences (fourth column).




Figure 9: **Clothed scans animated with composed sequences.** 4D-DRESS scans (left), registered as SMPL-X+D, are driven by the joint transforms of composed sequences, shown on the body model in the top row. Objects follow through their joint-relative transforms, and the finger motion is preserved on the hands.

Table 8: **Interaction rules.** Contact groups, pose constraints, local reference joints, and source datasets of the thirteen rules, with the number of atomic clips each yields before and after mirroring. The walking rule covers standing as well and selects whole AMASS sequences by their BABEL labels. T: target, S: source, A: augmentation.

| Rule | Interaction | Contact | No contact | Pose constraint | Reference $\mathcal{J}$ | Datasets | Clips |
|---|---|---|---|---|---|---|---|
| T:walk | Walking or standing | – | – | – | – | AMASS | 41 / 41 |
| T:sit | Sitting on a support object | lower back | lower front | upright torso | pelvis | BEHAVE, InterCap | 78 / 78 |
| T:rfoot | Right-foot manipulation | right foot | left foot | upright torso | right foot | InterCap | 19 / 24 |
| T:lfoot | Left-foot manipulation | left foot | right foot | upright torso | left foot | InterCap | 5 / 24 |
| T:rhand | Right-hand manipulation, large object | right hand | left hand | upright torso | right wrist | InterCap, OMOMO | 492 / 866 |
| T:lhand | Left-hand manipulation, large object | left hand | right hand | upright torso | left wrist | InterCap, OMOMO | 374 / 866 |
| S:rhand | Right-hand manipulation, small object | right hand | left hand | upright torso | right wrist | GRAB, ARCTIC | 68 / 71 |
| S:lhand | Left-hand manipulation, small object | left hand | right hand | upright torso | left wrist | GRAB, ARCTIC | 3 / 71 |
| S:hands | Two-handed manipulation | both hands | – | upright torso | both wrists | GRAB, ARCTIC | 187 / 187 |
| S:rhold | Passive right-hand hold | right hand | left hand | upright torso, elbow at 90° | right wrist | GRAB | 153 / 211 |
| S:lhold | Passive left-hand hold | left hand | right hand | upright torso, elbow at 90° | left wrist | GRAB | 58 / 211 |
| A:backpack | Backpack worn on the back | upper back | hands | upright torso, arms down | spine | BEHAVE | 11 / 11 |
| A:eyeglasses | Eyeglasses worn on the face | face | hands | – | both eyes | GRAB | 2 / 2 |
| Total | | | | | | | 1,491 / 2,663 |

# C Dataset Details

## C.1 Preprocessing

BEHAVE and InterCap enter in their InterAct (Xu et al., 2025a) versions, which correct jitter, contacts, and hand poses, and BEHAVE is converted from SMPL+H to SMPL-X. The subject shapes of GRAB and ARCTIC are fitted to the gender-neutral model with SMPLFitter (Sárándi & Pons-Moll, 2024), so that every source uses neutral shape parameters. All sequences are resampled to 30 fps and expressed in a Y-up frame with the floor at zero and the body centered on the ground plane in the first frame. Every object is a template mesh with a rotation and a translation per frame. The labels that the annotation corrects are the InterAct descriptions of GRAB, BEHAVE, and InterCap, the descriptions of OMOMO, a template sentence for ARCTIC, and the BABEL labels of the walking sequences.

## C.2 Interaction rules

Table 8 lists the thirteen rules with their contact groups, pose constraints, local references, and datasets, together with the number of clips each yields before and after mirroring, and Fig. 10 shows examples for every rule. A rule holds in a frame when every part of its contact group is within 5 cm of the object surface (Section 3), no part of its no-contact group is, and its pose constraint is met. An upright torso, the most common constraint, is a pelvis-to-head direction within 20 to 30 degrees of the vertical, depending on the rule.

## C.3 Annotation prompts

Qwen3 receives the rule description, the object name, the given label, and the narrative of the source dataset, and answers in two steps (Fig. 11). It first picks the verb that names the sustained interaction from a fixed inventory of 28 verbs, *hold, drink, move, lift, carry, push, pull, drag, switch on, take a picture, play, rotate, kick, dribble, pass, call, eat, see, browse, swing, manipulate, pour, toast, use, shake, clean, set down, interact*. It then writes one sentence with that verb, the object, and the acting hand in a fixed form, which the composition later parses. Whether an action requires gaze at the object and involves the head is asked once per verb and object pair in a third query of the same form, and the answer is shared by all clips with that pair.

## C.4 Joint-relative objects

The global transform of an object is recovered from its relative transform (Eq. (1)) and the world transform of the reference joint of any body, $\boldsymbol{R}^t_{\mathcal{O}} = \boldsymbol{R}^t_j \boldsymbol{R}^t_{\mathcal{O}|j}$ and $\boldsymbol{t}^t_{\mathcal{O}} = \boldsymbol{R}^t_j \boldsymbol{t}^t_{\mathcal{O}|j} + \boldsymbol{t}^t_j$. When the local reference holds several joints, as for two-handed manipulation, one relative transform per joint is stored, the global transform is recovered from every joint, and the results are averaged, with rotations averaged as unit quaternions and renormalized.

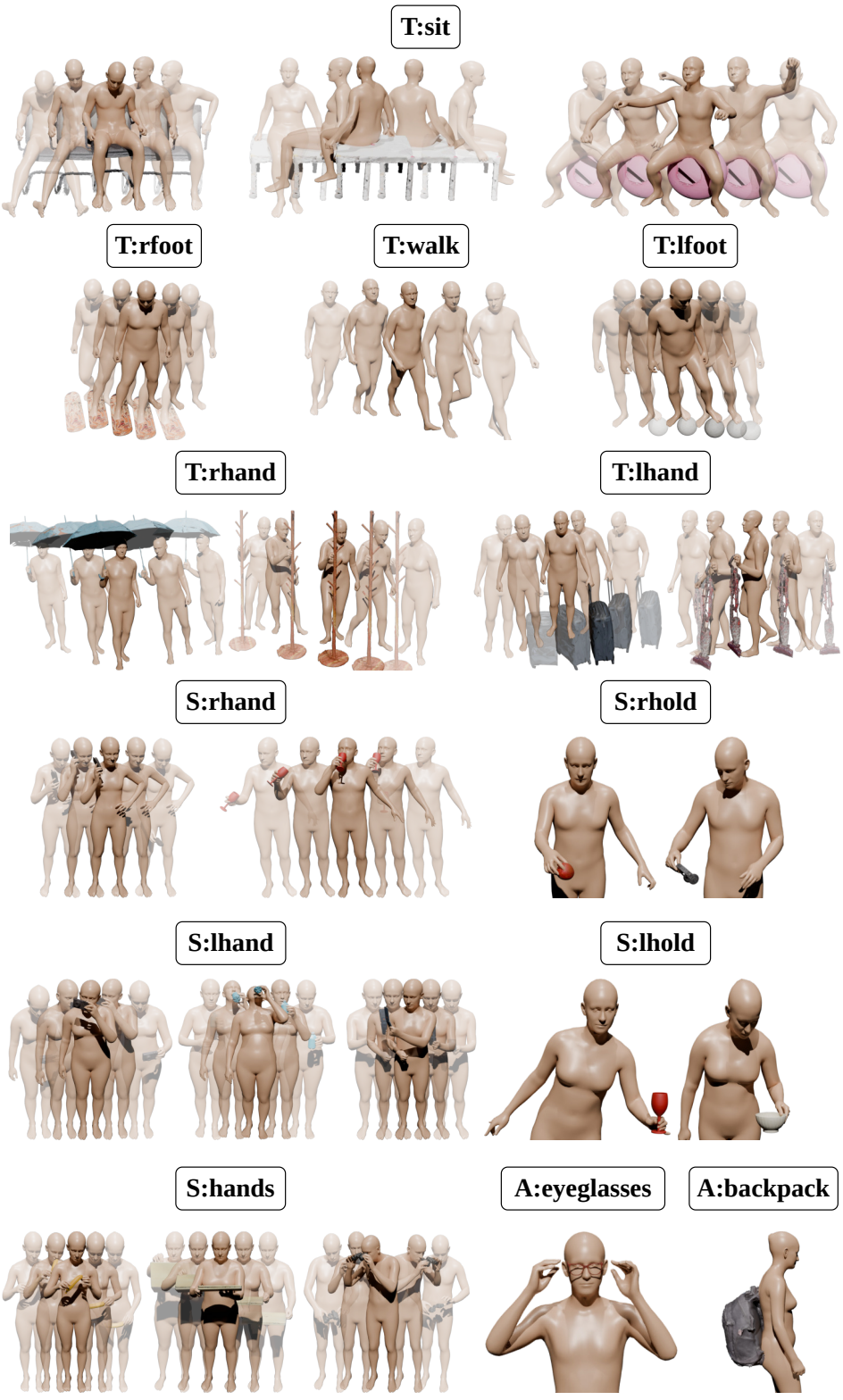



Figure 10: **Atomic clips per interaction rule.** Clips extracted by each of the thirteen rules of Table 8, the walking sequences of AMASS included. Target rules (T) engage the whole body, source rules (S) one or both hands, and augmentation rules (A) worn objects. Holds and worn objects are single frames.

**Action correction**

*Pick the single verb (from the list below) that best names the main, sustained interaction with "briefcase" in this narrative.*
**Rule:** The right hand is in contact with the object while the left hand is not.
**Object:** briefcase
**Given label (may be wrong):** `hold`
**Narrative:** A person holds a briefcase with their right hand, extends their right arm to move the briefcase from their right side to their left front, while taking a step to the left with their left foot.
**Allowed verbs** (choose exactly one, copied as written):
`hold, drink, move, lift, carry, push, pull, drag, switch on, take a picture, play, rotate, kick, dribble, pass, call, eat, see, browse, swing, manipulate, pour, toast, use, shake, clean, set down, interact`
**Guidelines:**
- The verb must describe something happening TO the object, not the person's general body movement while the grip stays fixed.
- `carry` = walking while holding; `push`/`pull`/`drag` = sliding without lifting.
- `move` = fallback when displaced but none of the above fit.
- Ignore setup/teardown (initial grab, final put-down).

*Reply with only the chosen verb.*

**Qwen3:** `move`

**Sentence generation**

*Write one short sentence describing a person's interaction with an object.*
**Rule:** The right hand is in contact with the object while the left hand is not.
**Object:** briefcase
**Action** (use exactly, do not swap for a synonym): `move`
**Requirements:**
- Conjugate `move` for subject "A person"; pick whatever preposition sounds natural.
- Present tense, one action, no direction or speed, no sub-parts.
- Refer to the object as "a briefcase".
- Start with "A person", end with "with the right hand", 4–10 words.

**Examples:**
- A person pours from a water bottle with the right hand.
- A person drinks from a cup with the right hand.
- A person eats an apple with the right hand.
- A person calls with a phone with the right hand.
- A person passes a game controller with the right hand.

*Reply with only the sentence.*

**Qwen3:** `A person moves a briefcase with the right hand.`

Figure 11: **Annotation prompts.** The two per-clip prompts of the annotation with the answer of Qwen3, for a clip of the InterCap briefcase whose given label *hold* is corrected to *move*.

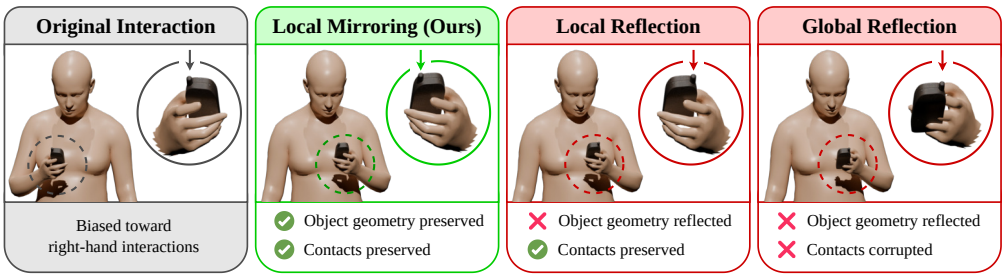



Figure 12: **Mirroring.** Global reflection of the object breaks the contact with the mirrored hand, and local reflection keeps the contact but mirrors the mesh. Our local mirroring poses the original mesh like its reflection with a proper rotation and preserves both.

### C.5 MIRRORING

Figure 12 compares three ways of mirroring the object of a clip whose body has been reflected with $\boldsymbol{M}$ and had its left and right joints swapped. Reflecting the world pose of the object breaks the contact, because SMPL-X is not symmetric and the mirrored hand is not exactly where the reflection puts the object. Reflecting the joint-relative pose keeps the contact but poses the mirror image of the mesh: a reflected posed object, $\boldsymbol{M}(\boldsymbol{R}^t_{\mathcal{O}|j}\boldsymbol{V}_{\mathcal{O}} + \boldsymbol{t}^t_{\mathcal{O}|j}) = (\boldsymbol{M}\boldsymbol{R}^t_{\mathcal{O}|j}\boldsymbol{M})(\boldsymbol{M}\boldsymbol{V}_{\mathcal{O}}) + \boldsymbol{M}\boldsymbol{t}^t_{\mathcal{O}|j}$, is the reflected template $\boldsymbol{M}\boldsymbol{V}_{\mathcal{O}}$ under a proper rotation and a translation. Our mirroring keeps the original template and uses the alignment of Eq. (2), which says that $\boldsymbol{M}\boldsymbol{V}_{\mathcal{O}}$ is, up to the residual, the original template rotated by $\boldsymbol{Q}_{\mathcal{O}}$ about its centroid and shifted by $\boldsymbol{p}_{\mathcal{O}}$. Substituting this for $\boldsymbol{M}\boldsymbol{V}_{\mathcal{O}}$ gives the mirrored relative transform

$$\tilde{\boldsymbol{R}}^t_{\mathcal{O}|j} = \boldsymbol{M}\boldsymbol{R}^t_{\mathcal{O}|j}\boldsymbol{M}\boldsymbol{Q}_{\mathcal{O}}, \qquad \tilde{\boldsymbol{t}}^t_{\mathcal{O}|j} = \boldsymbol{M}\boldsymbol{t}^t_{\mathcal{O}|j} + \boldsymbol{M}\boldsymbol{R}^t_{\mathcal{O}|j}\left[(\boldsymbol{I} - \boldsymbol{M}\boldsymbol{Q}_{\mathcal{O}})\,\bar{\boldsymbol{v}}_{\mathcal{O}} + \boldsymbol{M}\boldsymbol{p}_{\mathcal{O}}\right], \tag{6}$$

a proper rotation and a translation of the original template, so the mesh stays the same and the contact with the mirrored hand is preserved. The local reference is swapped to the opposite side, the global transform is recovered from the mirrored body, and the rule is swapped accordingly. The alignment of Eq. (2) is exact for symmetric objects. Over the 26 mirrored templates, articulated parts counted separately, the distance between the reflected and the aligned template averages 0.9 mm. The largest residuals belong to the largest objects, the upper part of the vacuum cleaner (4.0 mm), the umbrella (3.9 mm), and the skateboard (2.1 mm), which have no mirror symmetry. The fit then favors the regions that matter for the contact, the handle of the umbrella and the deck of the skateboard, at the cost of the rest. Relative to the bounding-box diagonal, every residual stays below 0.7 %.

### C.6 COMPOSITION GRAPH

Figure 13 shows which source rules each rule accepts when it serves as the target: whole-body targets accept every source, a one-hand target accepts the action and the hold of the other hand, and among sources a hand action or hold accepts the other hand, each unordered pair once. Pairs of the same role or with overlapping limbs are excluded by construction, and the remaining pairs pass through the semantic and physical checks of Section 3.

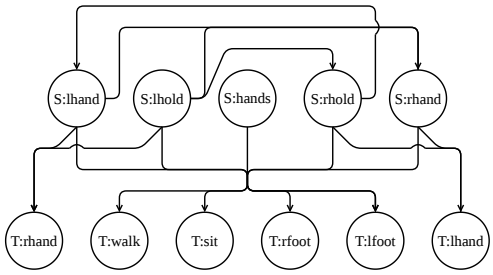



Figure 13: **Composition graph.** An arrow leads from every source rule to the rules it can be composed onto.

### C.7 MERGE AND FEASIBILITY

The composed pose is the target pose with the joints of the source action replaced: the wrist, elbow, shoulder, and collar of the acting arm, or both arms for a two-handed action. Every manipulation also replaces the two upper spine joints, because the torso bends with a reaching arm, whereas a passive hold replaces the arm alone. An action that requires gaze adds the neck and head, one that involves the head adds the spine as well, and an object the target already holds moves with its wrist when the spine changes. The source objects and worn objects then follow their reference joints through Eq. (1) (Fig. 14). The composition is kept if every object retains the contacts of its rule in at least half of the frames. It is rejected if, in any frame, an object comes within 1 cm of another object or enters the lower body, the hands enter the lower body, or the arms cross. Walking targets, whose arms swing, tolerate such self-intersections in up to a third of the frames.

### C.8 SEQUENTIAL COMPOSITION AND LEVELS

A sequential composition places a right-hand and a left-hand action one after the other on a target long enough for both, a walking clip, a sitting clip, or one of the two actions itself as a standing body. The second action starts 15 frames before the first ends, the spine and head are interpolated between the two over this overlap, and outside its period each source is frozen at its first or last frame together with its object. Since the two objects are attended to in turn, the gaze check is skipped, and the feasibility checks apply as above (rows two and four of Fig. 14).

Five steps enumerate the graph, the first feeding the one-object level: motion, walking clips composed with every compatible source, the one-hand ones serving as further targets; one object, every atomic clip and motion-step composition; two objects, every target with each of its valid sources; three objects, every two-object sequence with the sources valid for both of its rules; and sequential, the ordered pairs above. Worn objects are added at every level. Captions join the clause of each clip with "and" or, for sequential actions, "then", as in *"A person, wearing a backpack, sits on a wooden chair and drinks from a cup with the right hand."* A hand-written bank paraphrases the subject, the sentence frame, the verb, and the hand phrasing, but keeps the object noun phrases verbatim. Training draws one paraphrase per sample, and evaluation uses one fixed paraphrase per sequence (Fig. 15).

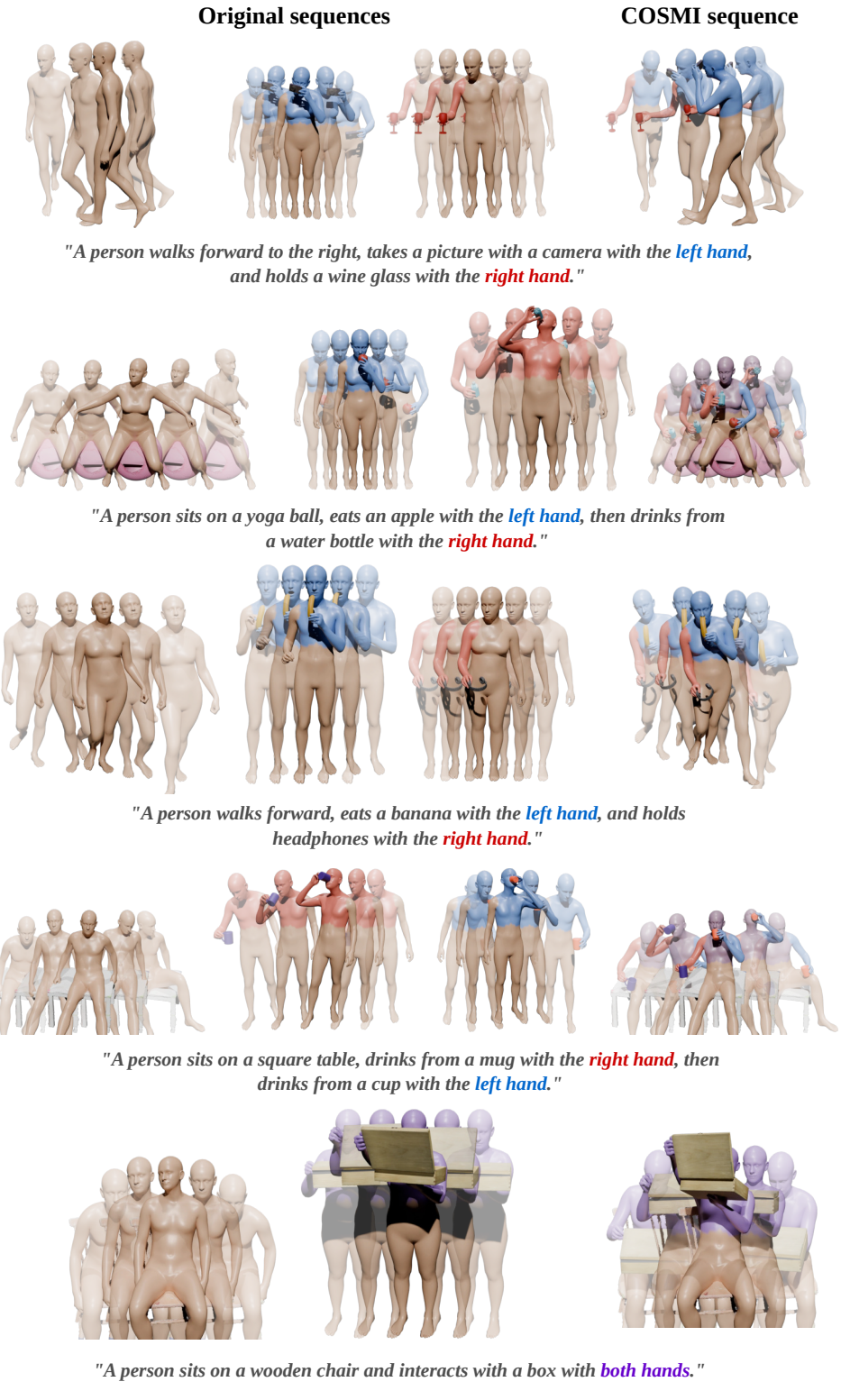


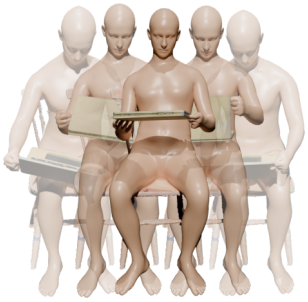

Figure 14: **Composition of sequences.** Left: the target clip, which provides the body motion, and the source clips, whose limbs are colored, blue for the left hand, red for the right hand, and purple for both hands. Right: the composed sequence, in which the source limbs replace their counterparts and every object follows its limb through Eq. (1). Rows two and four are sequential compositions.



*"A person, wearing a backpack and eyeglasses, walks forward and plays a flute with both hands."*



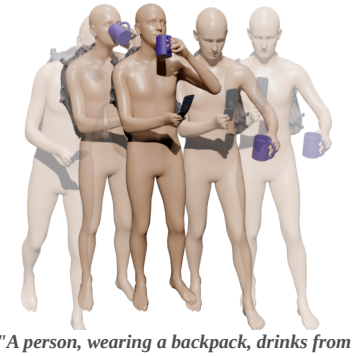
*"A person sits on a square table, drinks from a cup with the left hand, and holds a game controller with the right hand."*



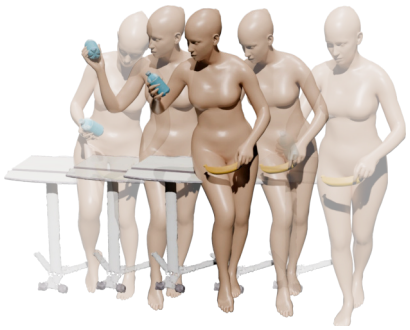
*"A person sits on a wooden chair and interacts with a laptop with both hands."*



*"A person, wearing a backpack and eyeglasses, sits on a yoga ball and plays with a game controller with both hands."*



*"A person, wearing a backpack, drinks from a mug with the left hand, then interacts with a flip phone with the right hand."*



*"A person sits on a small table, shakes a water bottle with the right hand, and holds a banana with the left hand."*



*"A person, wearing a backpack, holds an umbrella with the left hand and browses a camera with the right hand."*



*"A person, wearing a backpack, holds a skateboard with the left foot, eats an apple with the left hand, and holds headphones with the right hand."*

Figure 15: **Sequences of the COSMI dataset.** Composed sequences with their captions: simultaneous and sequential actions on walking, sitting, and standing bodies, with worn and articulated objects.

# D METHOD DETAILS

## D.1 HUMAN REPRESENTATION

Each frame of $\mathcal{H}$ (Eq. (3)) holds 256 values: the global positions $\boldsymbol{J}^t_{\mathcal{H}} \in \mathbb{R}^{24\times3}$ of the 22 body joints from the pelvis to the wrists and of the first joint of each middle finger, which fixes the wrist orientation, and the global rotations $\boldsymbol{\theta}^t_{\mathcal{H}} \in \mathbb{R}^{22\times6}$ of the body joints in the 6D representation (Zhou et al., 2019). The PCA coefficients $\boldsymbol{\theta}^t_{\text{hands}} \in \mathbb{R}^{2\times24}$ of both hands and the floor-contact scores $\boldsymbol{c}^t_{\mathcal{H}} \in \mathbb{R}^4$ of the ankles and toes complete the frame. At decoding, the mesh is posed from the rotations and the hand coefficients with the predicted pelvis as its root, and the other positions serve as a training signal tied to the rotations by the forward-kinematics loss.

## D.2 OBJECT REPRESENTATION

Each frame of $\mathcal{O}$ (Eq. (4)) holds 80 values. The absolute pose $\boldsymbol{T}^t_{\mathcal{O}} = (\boldsymbol{R}^t_{\mathcal{O}}, \boldsymbol{p}^t_{\mathcal{O}}) \in \mathbb{R}^9$ is the rotation relative to the orientation of the object in the first frame of the window, in the 6D representation, and the center of mass in the canonical frame. The hinge channel $\boldsymbol{a}^t_{\mathcal{O}} = (\cos\phi^t_{\mathcal{O}}, \sin\phi^t_{\mathcal{O}})$ encodes the articulation angle relative to the first frame, constant $(1, 0)$ for rigid objects. The relative pose with respect to a candidate joint $j \in \mathcal{J}_{\text{rel}}$ is

$$\boldsymbol{T}^t_{\mathcal{O}|j} = \Big((\boldsymbol{R}^t_j)^\top \boldsymbol{R}^t_{\mathcal{O}},\ (\boldsymbol{R}^t_j)^\top(\boldsymbol{p}^t_{\mathcal{O}} - \boldsymbol{p}^t_j)/r_{\mathcal{O}}\Big) \in \mathbb{R}^9, \qquad j \in \mathcal{J}_{\text{rel}}, \tag{7}$$

where $\boldsymbol{R}^t_j$ and $\boldsymbol{p}^t_j$ come from $\boldsymbol{\theta}^t_{\mathcal{H}}$ and $\boldsymbol{J}^t_{\mathcal{H}}$ and $r_{\mathcal{O}}$ is the bounding radius of the object plus 0.5 m. The normalized translation then lies in $[-1, 1]$ whenever the joint is near the bounding sphere, the only frames in which the stream is supervised. The contact scores $\boldsymbol{c}^t_{\mathcal{O}} \in \mathbb{R}^{15}$ of the ten fingers, the feet, the pelvis, the spine, and the head are soft indicators $\sigma\big(-100\,(d - \delta)\big)$ of the distance $d$ in meters to the object surface. Positions and hand coefficients are min-max normalized, and contact scores are scaled to $[-1, 1]$. An articulated object occupies one slot with the pose of its base part and the hinge angle, and its geometry is the assembled mesh at the hinge angle of the first frame.

## D.3 GEOMETRY ENCODER

The BPS representation (Prokudin et al., 2019) describes an object by the offsets from 1024 fixed basis points, sampled in a ball around its center of mass, to the nearest points of its surface, for the mesh oriented as in the first frame of the window. Our encoder aggregates these offsets over a fixed partition of the ball into 8 horizontal slabs of 16 cells each, split by azimuth and by distance to the vertical axis. Per-point features from a shared MLP are pooled within every cell and then within every slab, and the slab features are concatenated from bottom to top and projected to the geometry embedding, which preserves the vertical layout of the object.

## D.4 ARCHITECTURE

The denoiser consists of eight pre-LN transformer blocks with a model width of 512, eight attention heads, a feed-forward width of 2048, and a dropout of 0.1. Every human frame and every chunk of four object frames is a token, so a window with five objects has 270 motion tokens instead of 720. A one-dimensional convolution of kernel size three decodes the object tokens to per-frame parameters. The temporal order is encoded with rotary position embeddings (Su et al., 2024): a human frame and an object chunk at the same time receive the same position, whereas the step, sentence, and text tokens receive position zero. A learned type embedding distinguishes human, object, sentence, and text tokens. The denoiser has 27.3M parameters, 25.2M of them in the transformer blocks, the part-assignment head 2.6M, the geometry encoder 0.15M, and the CLIP projection 0.26M, 30.4M in total. The frozen CLIP ViT-B/32 text encoder adds 63.4M parameters, and our re-implementations of MDM and PriorMDM have 20.7M and 39.6M.

## D.5 PART-ASSIGNMENT HEAD

Per slot, the head receives the object conditioning, the initial pose of the person, and the sentence embedding, then two blocks of self-attention across slots and cross-attention to the text tokens produce one binary output per candidate joint. A joint is labeled as driving an object when it touches the

object for at least ten frames and the object travels at least half of the distance that the joint travels meanwhile, and the denoiser receives these labels during training. At decoding, the pose source is selected once per window: the driving joint with the highest average predicted score, a wrist scored by the maximum over its fingers. If no joint is predicted as driving or the highest score is below 0.5, the absolute pose is used. The baselines, which have no part-assignment head, select among all six candidate joints by the same contact scores, so an object follows a joint only while its predicted contact stays high.

### D.6 Loss terms

The terms $\mathcal{L}_{\mathrm{k}}$ of Eq. (5) are $\ell_1$ errors of the human parameters and of their frame-to-frame differences, of the absolute object pose, of the relative poses in their supervised frames, and of the hinge channel, and a cross-stream consistency term. Further terms are squared errors of the object and foot-floor contact scores and a foot-skating penalty on the predicted ankles and toes in frames where the ground-truth foot is in contact with the floor. A forward-kinematics term compares the joint positions of the predicted rotations with the ground-truth positions, and the binary cross-entropy of the part-assignment head completes the objective. The consistency term poses 100 surface points of the object with each of the seven predicted poses, points on a moving part with the predicted hinge angle, and penalizes the $\ell_1$ distance of the points and of their frame-to-frame differences to the ground truth. All weights $\lambda_k$ are 1, except 2 for the human-parameter term.

### D.7 Training details

We train for 80k steps with batches of 256 windows using Adam (Kingma & Ba, 2015) with a learning rate of $5{\cdot}10^{-4}$ and a cosine schedule with warm-up, and sample from an exponential moving average of the weights. Training takes eight hours on one RTX 5090. The text prompt is dropped with probability 0.1, replacing the sentence token by a learned null embedding and masking the text tokens. Each worn object is included with probability 0.5, an object name by a null embedding with probability 0.3, and the part-assignment input is removed with probability 0.1.

### D.8 Runtime

Table 9 reports generation times on one NVIDIA RTX 5090 with the sampler of Section 4, in fp16 with a compiled denoiser, where the conditional and unconditional passes of classifier-free guidance run as one doubled batch. Single sequences are timed at batch size one as the median of 20 runs after three warm-up runs. The median test sequence has 119 frames and fits one window. The denoising steps dominate without the guidance and its Adam iterations with it. The guidance therefore costs 7.2 times the generation time of a single sequence, but only 1.55 times in the batched evaluation of the test set, since its small kernels underuse the GPU on one sequence.

Table 9: **Runtime.** Generation time on one RTX 5090 without and with the contact-consistent guidance (G), at batch size one unless stated.

| | **w/o G** | **w/ G** |
|---|---|---|
| *Per window (ms)* | | |
| total | 157 | 1,140 |
| denoising, 100 steps | 138 | 1,122 |
| decoding to meshes | 5 | 5 |
| *Per sequence (s)* | | |
| 119 frames, 1 window | 0.15 | 1.08 |
| 299 frames, 4 windows | 0.63 | 4.56 |
| Batched (s per seq., batch 128) | 0.088 | 0.136 |
| Peak GPU memory (GiB) | 0.78 | 0.97 |

### D.9 Long sequences

Consecutive windows of 120 frames overlap by 60 frames. The human pose and the object states predicted for the first frame of a window define its canonical frame and the reference orientation and hinge angle of every object. In the overlap, the two predictions are blended with linear weights, rotations projected back onto $SO(3)$ and hinge angles averaged as $(\cos, \sin)$ pairs.

### D.10 CONTACT-CONSISTENT GUIDANCE

The two guidance losses of Section 4 invert the contact labels. The score of finger $f$ and object $\mathcal{O}_n$ in $\boldsymbol{c}^t_{\mathcal{O}}$, which is the soft indicator defined above, and the floor-contact score of foot joint $j$ in $\boldsymbol{c}^t_{\mathcal{H}}$ are

$$\begin{aligned} c^t_{\mathcal{O}_n,f} &= \sigma\big(-100\,(d^t_{\mathcal{O}_n,f} - \delta)\big), \\ c^t_{\mathcal{H},j} &= \sigma\big(-100\,(y^t_j - \delta_y)\big)\,\sigma\big(-800\,(v^t_j - \delta_v)\big), \end{aligned} \tag{8}$$

where $d^t_{\mathcal{O}_n,f}$ is the distance between the finger and the object surface, $y^t_j$ the height of the foot joint above the floor, and $v^t_j = \|\boldsymbol{p}^{t+1}_j - \boldsymbol{p}^t_j\|$ its displacement to the next frame, all in meters. The slopes of 100 and 800 per meter set the width of the transition around corresponding $\delta$ thresholds. Both labels are monotone, so a score $c$ implies the bounds

$$d \le d^*(c) = \delta - \tfrac{1}{100}\,\mathrm{logit}(c), \qquad v \le v^*(c) = \delta_v - \tfrac{1}{800}\,\mathrm{logit}(c), \tag{9}$$

which hold exactly on the training data. For every score that predicts a contact, $c > 0.5$, the losses penalize the violation of its bound,

$$\mathcal{L}_{\text{contact}} = \sum_{n,f,t} c^t_{\mathcal{O}_n,f}\,\big[d^t_{\mathcal{O}_n,f} - d^*\big(c^t_{\mathcal{O}_n,f}\big)\big]_+, \qquad \mathcal{L}_{\text{slide}} = \sum_{j,t} c^t_{\mathcal{H},j}\,\big[v^t_j - v^*\big(c^t_{\mathcal{H},j}\big)\big]_+, \tag{10}$$

where $[\cdot]_+ = \max(\cdot, 0)$, and we minimize $\mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{slide}}$. The guidance runs at the last ten of the 100 sampling steps, with five Adam iterations at a learning rate of $5{\cdot}10^{-4}$ on the predicted clean sample at each of them and 100 at the final step, whose prediction is the output. Within every iteration, the edit is blurred along time with a Gaussian kernel of one frame, to reduce jitter.