HOI-Blender: A Unifying Blender Add-on for Standardization and Visualization of Diverse Human-Object Interaction Datasets

Anonymous ICCV submission

Paper ID *****

Figure 1. An intersection for interaction: HOI-Blender takes position as a consolidating toolkit, standardizing heterogeneous datasets such that actions may become mutual actors. Coexistence of diverse motions within coherent scenes is facilitated via acquired unified formatting, assisting conjoined visualization, equitable evaluation and aggregated downstream application.

Abstract

Human-Object Interaction (HOI) datasets differ in coordinate frames, formats, data structures, and directory layouts, forcing bespoke loaders and brittle dataset-specific rendering setups. Thus, we present HOI-Blender, a Blender add-on which standardizes HOI data as a unified loader: it normalizes coordinates and scale, harmonizes metadata, and maps motion parameters to skinned bodies and object meshes across congruent scenes. Once standardized, camera rigs, lighting composition and render settings are reapplicable across datasets, enabling one-click import, animation preview, and image rendering. The add-on further decouples motion and appearance, allowing the same motion to drive different human identities—or the same identity to enact different motions. Additionally, to streamline annotation, HOI-Blender includes an auto-captioning module which forwards rendered frames to Vision-Language Models to produce action-object descriptions, supporting rapid weak labeling and dataset curation. We demonstrate support for an initial 15 HOI public datasets and report a comprehensive qualitative evaluation spanning motion fidelity, mesh integrity, hand-object interactions, and inter-mesh collisions. By adapting heterogeneous HOI resources into coherent scenes, HOI-Blender reduces implementation overhead and facilitates efficient visualization, evaluation and culminating application in human-centric research.

1. Introduction

Human-Object Interaction (HOI) describes the interplay between human and object motion. Improvement upon its understanding is core to advancing human-centric vision, as it ties into an extensive range of applications, including assistive robotics [2], scene understanding [9], synthetic data generation [35], and behavioral modeling [12]. More distinctly, character design and animation [7], augmented and virtual reality [22], sports analytics [21], and human-agent interaction [23]. Despite the rapid growth in count of HOI datasets, working with motion data from various sources

within established communal tools remains a significant bottleneck. Issues consistently stem from the heterogeneity of dataset publications, where differences include variability in conventions, data formatting, distribution structures and a multitude of individual oddities. Hence, researchers are often forced to implement dataset-specific importers, manually assemble scenes with unique setup and properties, and deal with undesired complications, which slow progress and restrict accessibility of datasets for downstream usage.

To bridge the gap between HOI datasets and their practicality, we introduce HOI-Blender, a Blender add-on which transforms the aforementioned fragmented, repetitive per-dataset workflow into one seamless, single-click and standardized. HOI-Blender applies dataset-specific preprocessing and affords a multi-purpose unified interface such that all dataset handling is made equal, eliminating any need for one-off scripts. Human and object motion data is standardized into an internal universal structure, enabling ease of application onto skinned bodies and object meshes, together assembled across coherent scenes. This underpins shared camera, lighting and rendering setups across datasets, eliminating need for manual per-dataset adjustments. Further, visualization flexibility is enhanced via decoupled human motion and appearance, allowing cross-combination from diverse sources without additional configuration. Lastly, integration for automatic captioning forwards rendered frames to OpenAI-compatible API for action-object descriptions, facilitating rapid weak labeling and dataset curation.

Beyond support for standardized HOI dataset import, as well as integrated visualization, rendering and motion captioning capability, across an initial 15 HOI public datasets, we present a comprehensive qualitative analysis of key usability metrics, covering motion quality, mesh fidelity, interaction plausibility, and collision robustness. Our evaluation underscores a type of efficiency and broad accessibility provided by HOI-Blender over prior ad-hoc approaches, identifiable in further cumulative applications.

By lowering the technical barriers of accessibility to HOI assets, HOI-Blender supports many diverse avenues of HOI research and workflows. Some notable examples include (i) rapid visualization for inspection, correction, and qualitative evaluation across HOI datasets for benchmarking and reproducible comparison, (ii) integration into Blender, a widely adopted, general-purpose 3D creation environment, enabling its full functionality to be leveraged for interactive assembly, manipulation, and refinement of HOI scenes for development of 3D HOI datasets [33], (iii) extending of prior synthetic rendering pipelines from human-only datasets [4, 26] to high-quality rendering of generated HOI scenes for construction of image-to-HOI datasets and further various image-based synthetic training data, (iv) incorporation with physics-based simulation for development of physically grounded HOI data or robotic policy learning.

Our work culminates upon the following contributions:

- A Blender add-on which standardizes HOI data from an initial 15 datasets into a common universal motion parameter format, enabling one-click import for skinned bodies and object meshes within coherent scenes. High-quality visualization capabilities are enhanced via decoupled motion and appearance for human meshes.
- A streamlined workflow for consistent animation preview, evaluation, and rendering with lightweight scene assembly and cross-dataset reapplicable preset spanning camera rigging, lighting setup and render settings; where frames may be forwarded for automatic action-object captioning.
- A comprehensive qualitative evaluation of standardized HOI datasets for motion fidelity, mesh integrity, interaction plausibility, and collision accuracy, emphasizing manner of support for future HOI workflows and diverse downstream applications.

2. Related Works

2.1. HOI Datasets

HOI datasets demonstrate heterogeneity over a manner of characteristics, one of which being scope of action, ranging from isolated body motion to full-body, human-object or multi-human interaction. Their uniting factor (of those in our interest) is a common foundation on SMPL-Family body models (SMPL, SMPL-H, SMPL-X), however, overlap does not pertain further across conventions, data formats, hierarchical structures or otherwise. Our standardization considers grouping of representative datasets by interaction type, which is delineated alongside supported HOI datasets below, with high-level factors outlined in Table 1.

Isolated Human Motion. AMASS [20] and AIST++ [15] provide large-scale full-body motion sequences without paired objects. AMASS unifies motion capture of multiple sources onto common SMPL-Family parameters, while AIST++ focuses on multi-view dance recordings with synchronized music. These datasets are widely used as motion priors for retargeting, generation, and animation control.

Human-Human Interaction. Hi4D [37] and Duolando [30] capture dyadic motion via purposing of registered SMPL-Family bodies. Hi4D emphasizes close-contact motion with dense 4D segmentation, while Duolando contributes paired leader-follower dance sequences highlighting social coordination. Together they support learning social motion priors, role assignment, and contact reasoning.

Full-Body Human-Object Interaction. BEHAVE [3], InterCap [8], COUCH [40], OMOMO [14], NeuralDome [38], and IMHD^2 [41] provide motion parameters for aligned SMPL-Family human bodies and mid-sized object meshes. They differ in address: BEHAVE and NeuralDome offer multi-object coverage with detailed meshes and contacts, InterCap and IMHD^2 explore monocular or inertial

Table 1. **High-Level Comparison of SMPL-Family HOI Motion Datasets.**

Motion Type	Dataset	# Objects	Object Category	Object Texture	Body Model	Published at
Isolated Human Motion	AIST++ [15]	-	-	-	SMPL	ICCV 2021
	AMASS [20]	-	-	-	SMPL-H/-X	ICCV 2019
Human-Human Interaction	Hi4D [37]	-	-	-	SMPL	CVPR 2023
	Duolando [30]	-	-	-	SMPL-X	ICLR 2024
Full-Body Human-Object Interaction	Behave [3]	20	Mid-Size	✔	SMPL-H	CVPR 2022
	InterCap [8]	10	Mid-Size	✔	SMPL-X	GCPR 2022
	COUCH [40]	4	Chairs	✘	SMPL-H	ECCV 2022
	OMOMO [14]	15	Mid-Size	✘	SMPL-X	SGAsia 2023
	NeuralDome [38]	23	Mid-Size	✔	SMPL-X	CVPR 2023
	IMHD^2 [41]	10	Mid-Size	✔	SMPL-H	CVPR 2024
Hand-Object Interaction	GRAB [31]	51	Small-Size	✘	SMPL-X	ECCV 2020
	Arctic [5]	11	Small-Size	✔	SMPL-X	CVPR 2023
	HIMO[19]	53	Small-Size	✘	SMPL-X	ECCV 2024
Multi-Human Object Interaction	HOI-M3 [39]	90	Large-Size	✔	SMPL	CVPR 2024
	CORE4D[17]	6	Large-Size	✘	SMPL-X	CVPR 2025

capture for tracking, COUCH specializes in seated interactions, and OMOMO emphasizes object-conditioned motion synthesis. These datasets enable studies of contact, affordances, and perception in realistic interaction settings.

Hand-Object Interaction. GRAB [31], ARCTIC [5], and HIMO [19] focus on dexterous manipulation with detailed hand motions upon small objects. GRAB emphasizes full-body grasps across a wide object set, ARCTIC captures bimanual articulated object manipulation, and HIMO extends its regard of objects across diverse scenarios. In conjunction, these datasets are well suited for grasp synthesis, contact modeling, and fine-grained hand pose estimation.

Multi-Human Object Interaction. HOI-M3 [39] and CORE4D [17] extend to collaborative multi-actor interaction scenes with large object coordination in shared environments. They promote analysis of role assignment, group dynamics, and cooperative manipulation.

Despite shared SMPL-Family foundation, divergence in dispense of HOI datasets is prominent across many facets. Structural variation span file packaging (per-frame vs. bundled sequences; differed containers: `.npz`, `.pkl`, `.json`, `.p`), path conventions (absolute vs. relative references to objects, textures), metadata allocation (lack-of, co-located with pose data, scattered into auxiliary files), dictionary key consistency (`trans` vs. `translation`, `poses` vs. `body_pose`), as well as parameter grouping (full-body vectors vs. assorted sets of orientation, body, hands, face). Objects provide irregularities via varied mesh formats (`.obj`, `.fbx`, `.glb`, `.ply`) and ranging motion schemas from static meshes with implicit alignment to per-frame transforms at partial attributes; generally, common denomination is deficient. Representational discrepancies further compound upon structural issues. Coordinate frames differ across up-axis and handedness (Y-up vs. Z-up; left-handed vs. right-handed bases), units in scale (`m`, `cm`, arbitrary), and rotation in encoding (axis-angle, Euler triples, rotation matrices, quaternions, 6D; potentially varied in-dataset). Body root pivots are inconsistent (pelvis-centered, floor-aligned, world origin), as are object pivots (centroid, mesh bottom, CAD default). Object representations range from rigid single-part meshes to multi-part articulated assemblies with divergent parent-child conventions, and texturization fluctuates (vertex color, UV-mapped textures, external references). HOI-Blender precisely bridges all aforementioned gaps between datasets, deterministically resolving incongruities, and yielding an internal standardized structure primed for purpose throughout contexts.

2.2. 3D Human-Centric Visualization Tools

While our focus lies in unifying HOI datasets, this challenge sits within a broader landscape of human-centric, interactive visualization tools, where prior efforts have addressed related but narrower aspects of body model motion.

Viser [36] is a general-purpose 3D visualization library, which provides an imperative Python API and a web-based viewer, exposing scene primitives and GUI controls. SMPL mesh visualization is supported via pose and shape parameter sliders, enabling rapid debugging and interactive exploration. Its design simplifies embedding in research pipelines and sharing browser-based viewers.

AIT-Viewer [11] spans to broader coverage, supporting SMPL, SMPL-H, SMPL-X, MANO, FLAME, and STAR. Both GUI and headless rendering modes are available, allowing either manual posing or sequence loading. Its applicability for inspection and visualization thus particularly stands out when working with diverse body model types.

Meshcapade's SMPL Blender Add-on [25] enables the import of SMPL, SMPL-H, SMPL-X, and SUPR models, material application, shape and pose parameter adjustment,

as well as loading of AMASS animation sequences. While it brings parametric body manipulation into Blender, its scope remains focused on the body models themselves.

These established toolkits provide essential accessibility for visualization, inspection, and interaction with parametric human bodies, yet their functionality is largely confined to sliders and rudimentary sequence playback. Our work is complementary: HOI-Blender integrates standardized dataset import, which yields coherent scene assembly alongside consistent preferences, cameras, lighting, materials and further prospective desired arrangement. Such consolidation transforms various fragmented resources into interchangeable foundations for diverse downstream applications, acting as a bridge across datasets, eliminating ad-hoc setup and enabling analyses, augmentations, and simulations to be repeated seamlessly upon a one-click workflow.

3. Background

Human-Object Interaction (HOI) can be viewed as the time-varying spatial configuration of articulated human bodies and rigid objects, both represented as 3D meshes. We function upon the SMPL-Family [16, 18, 27, 29] of parametric bodies to model humans, which efficiently encodes identity via shape parameters and motion via joint rotations, yielding skinned, watertight meshes suitable for animation and rendering. Objects are ordinarily provided as mesh assets comprised of geometry and material, occasionally with cohesive part-level kinematics. Such mesh-centric representation aligns both with common datasets and with established frameworks, enabling direct motion control via joint transforms for humans and rigid-body transforms for objects.

3.1. Parametric Body Model

The SMPL-Family is a set of widely adapted and applied parametric body models for representing human shapes and poses. Its included model types begin with a canonical template mesh $\overline{\mathbf{T}} \in \mathbb{R}^{3N}$ at N vertices, and subsequently incorporate both shape blend shapes $B_S : \mathbb{R}^{|\boldsymbol{\beta}|} \mapsto \mathbb{R}^{3N}$ (controlled by shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$) as well as pose blend shapes $B_P : \mathbb{R}^{|\boldsymbol{\theta}|} \mapsto \mathbb{R}^{3N}$ (controlled by pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$, at K joints) as additional vertex offsets:

$$\begin{aligned} T_C(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \overline{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}), \\ B_S(\boldsymbol{\beta}; \mathcal{S}) &= \sum_{n=1}^{|\boldsymbol{\beta}|} \beta_n \mathbf{S}_n, \\ B_P(\boldsymbol{\theta}; \mathcal{P}) &= \sum_{n=1}^{9K} R_n(\boldsymbol{\theta}) \mathbf{P}_n, \end{aligned} \tag{1}$$

where $\mathbf{S}_n \in \mathbb{R}^{3N}$ and $\mathbf{P}_n \in \mathbb{R}^{3N}$ are the respective shape and pose blend shape deformation matrices, and $R_n(\boldsymbol{\theta}) \in \mathbb{R}$ is the matching rotation matrix element of the pose $\boldsymbol{\theta}$. The resulting additive mesh $T_C \in \mathbb{R}^{3N}$ is in canonical space, and is to be transformed into deformation space $T_P \in \mathbb{R}^{3N}$ via Linear Blend Skinning (LBS) with joint rotation:

$$T_P = \left(\sum_{i=1}^{K} w_i \mathbf{G}_i(\boldsymbol{\theta}) \right) T_C, \tag{2}$$

where $\mathbf{G}_i(\boldsymbol{\theta}) \in \mathbb{R}^{4\times4}$ is the transformation matrix of joint i and $w_i \in \mathbb{R}^N$ are the respective skinning weights. Please refer to Sec. 3 and Fig. 3 of SMPL [18] for more details on the parametric blend shapes and skinning.

3.2. Human Appearance Model

As the base SMPL-Family model types provide plain mesh, they well represent a naked human equivalent by specified influence via pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ parameters, however, human appearance such as coloration, hair or clothing are not taken into consideration. Following ClothCap [28], we represent clothed humans by applying adjunct displacement. Notably, cloth blend shapes $B_C \in \mathbb{R}^{3N}$ are appended in canonical space, adding to the canonical naked-body mesh:

$$T_C(\boldsymbol{\beta}, \boldsymbol{\theta}) = \overline{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}) + B_C. \tag{3}$$

To enable coloration, artist-designed UV-spaces $\boldsymbol{U} \in \mathbb{R}^{2N}$ are employed for texturization. Despite possibility of artifacts arising from applying naked-body skinning weights on displacement—as often evident under loose clothing akin to skirts—this low-cost, user-friendly representation well integrates into HOI-Blender, contrasting other potential choices and their attributes.

4. HOI-Blender: A Unifying Visualization Tool

We introduce HOI-Blender (see Figure 2), an add-on which upon one-click (i) standardizes diverse HOI datasets via resolving of incidental incompatibilities (file containers, coordinate conventions, rotation encodings, missing channels, scale ambiguities, and inconsistent object metadata); (ii) assembles attained parameters into Blender-canonical components (armatures, shape keys, color attributes), enabling interactive animation viewing, parallel to appearance compositioning; (iii) provides render-ready setups with established camera placement, lighting arrangement and base material across datasets; and (iv) arranges automatic captioning on rendered frames, with outlook towards downstream applications, extended via Blender-native options, such as auxiliary render passes. The HOI-Blender User Interface is delineated in Figure 3, with demonstrated workflow as part of our Supp. Video.

4.1. Standardization of Heterogeneous Datasets

To reconcile inconsistencies, HOI-Blender performs automated characteristic-based dataset identification and applies

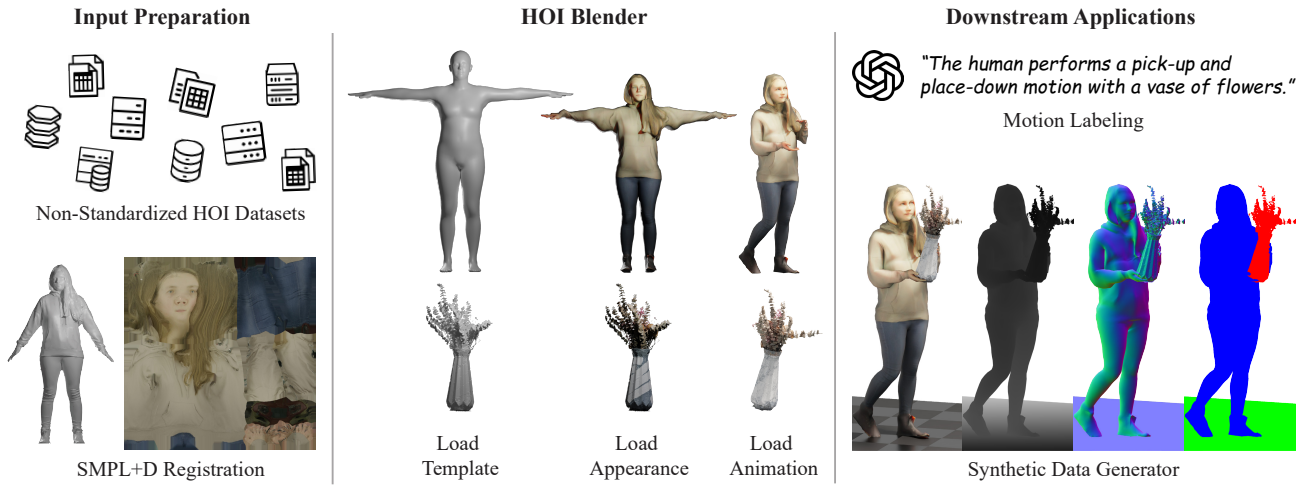


Figure 2. **HOI-Blender Overview.** Our add-on adapts heterogeneous HOI datasets into unified Blender-canonical components, facilitated over three key stages: (1) **Data Standardization:** ingests diverse HOI datasets and normalizes them into consistent representation; (2) **Scene Assembly:** maps SMPL-Family parameters to rigged human bodies, with decoupled motion and appearance, and assembles object meshes within a coherent 3D scene; (3) **Rendering & Applications:** enables scalable rendering with auxiliary render passes, such as depth, normals, and segmentation masks. The unified pipeline supports downstream applications encompassing synthetic supervision generation, VLM-based motion captioning, and cross-dataset evaluation.

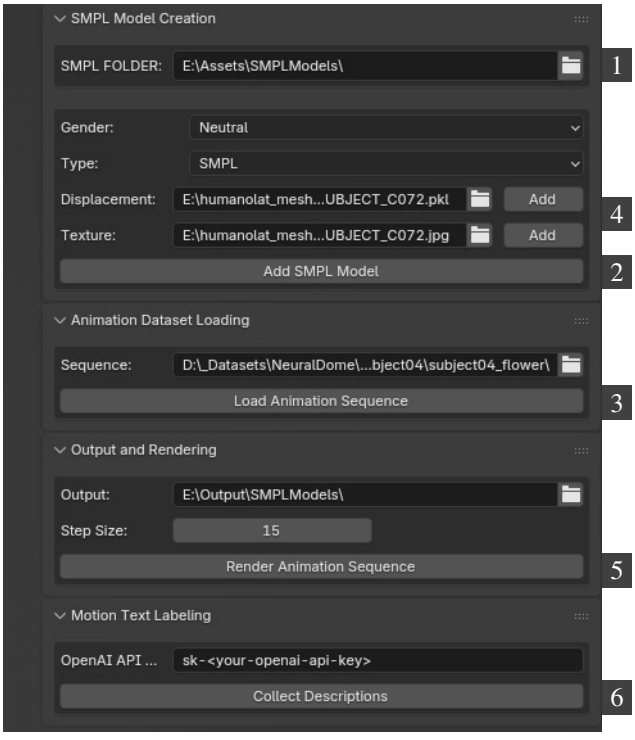


Figure 3. **HOI-Blender User Interface**. After imparting an asset path to SMPL-Family models (1), users can either import a specified model type (2) or load individual sequences directly (3). Both texture and displacement may be optionally provided (4). The select sequence can be directly rendered at specified step-size (5), with output frames electively passed for automatic captioning (6).

tailored loaders, such that a compact per-sequence dictionary of standardized parameters is acquired. Blender compatibility is attained via resolved sequence and asset structures, key mapping, parsed containers, canonicalizated conventions, aligned coordinate frames, unified rotations, and accounting for dataset-specific idiosyncrasies and exceptions. Invisibility of dataset dissensions to all later processing modules via standardization is hence achieved.

4.2. 4D HOI Representation

As generating individual meshes per-frame is highly expensive and inconsistent, we instead leverage the native rigging tools provided by our chosen framework for advantage.

Topology of canonical template mesh $\overline{\mathbf{T}}$ is explicitly represented in Blender, with blendshapes B_S, B_P exposed as shape keys (additive per-vertex offsets), and joint transformations mapped to an armature (hierarchical skeleton) via per-joint vertex groups i representing per-vertex weights w_i, facilitating Blender's native skinning for vertex position updates. Keyframes for global mesh location, shape key values and bone rotations are inserted respectively as read from standardized translation, shape and pose data—enabling efficient and consistent human animation playback.

Appearance is decoupled from motion, and may be realized in two manners (see Figure 4). Available texture assets are integrated into a principled material node network, with UV-mapping U carried over from the SMPL-Family artist-designed template. Provided per-vertex displacement B_C are exposed as shape keys, whose inherent additive nature allows arbitrary layered blending. This separation bears

Figure 4. **MotionAppearance Decoupling.** Top row: same identity enacting various motion sequences. Bottom row: same motion enacted by various identities.

flexibility for dynamic asset swapping and enables diverse visual variations without altering the underlying animation.

Objects are typically considered rigid per-sequence assets. After pivot configuration, keyframes for standardized translation, rotation, and scale are applied. Material assignment adheres to perdataset convention, supporting texture and vertexcolor basis.

Our standardization facilitates flexible import of HOI animations across heterogeneous datasets, laying groundwork for scalable rendering, controlled experimentation, and diverse downstream applications.

4.3. Automated Sequence Rendering

Rendering in HOI-Blender is taken care of by dataset-aware framing, automated instantiation of cameras and lighting, and preference configuration. Sequence mesh translation is aggregated over to compute a scene bounding box, guiding placement of tracked cameras and relative lighting. Frames are configurably rendered and optionally post-processed to achieve compact, centered crops. In addition, object-centric stills may also be produced. Collectively, our standardized cross-dataset basis enables consistent and customizable visual output—potentially adorned by auxiliary render passes or further modifications—serving a multiplicity of foundations for cumulative downstream applications.

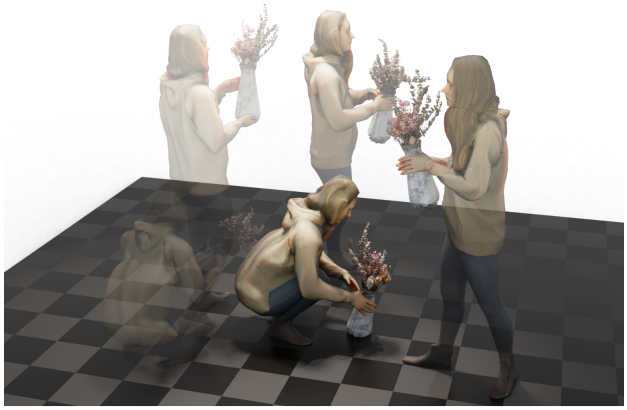


Figure 5. **VLM-Based Motion Captioning.** HOI-Blender renders image sequences of standardized HOI dataset animations and forwards them to VLMs for automated captioning.

4.4. Downstream Applications

Motion Text Description. HOIBlender carries a captioning module, that collects HOIcentric crops and forwards to VLMs under a structured prompt (see Figure 5). Achieved are concise yet semantically rich motion-text pairs describing motion phases, contact patterns, and interaction semantics at scale. Our pipeline ensures standardized visual input with potential for systematic study of visual factors on generated text. Acquired captions not only facilitate rapid weak labeling and dataset curation, but they also serve as training data for downstream tasks such as motion captioning, interaction retrieval, and instructionfollowing applications.

Human-Centric World Models aim to learn how humans act and interact with objects to predict, simulate, and plan actions in realistic context settings [24]. Their training requires large, coherent HOI datasets with precise, temporally aligned camera parameters, interaction poses, and contact information. In practiced RGBD captures, signals are often degraded by sync errors, rollingshutter artifacts, calibration drift, incomplete object geometry, and missing contact labels, making frametoframe correspondence and contactrich dynamics hard to obtain.

HOI-Blender's standardized render-ready format enables generating largescale synthetic HOI data with perfect correspondences, metric depth, calibrated cameras, and dense labels (see Figure 6). This approach mirrors a broader trend in worldmodeling, where synthetic data—offering dense, noise-free ground truth—boosts generalization and is thus predominantly trained upon [6, 10, 13, 32].

3D & 4D HOI Reconstruction provides the structured, temporally-coherent (geometry, pose, contact) supervision needed for humanoid robot training [1]. Accuracy is hinged for related methods at aforementioned issues, as practiced

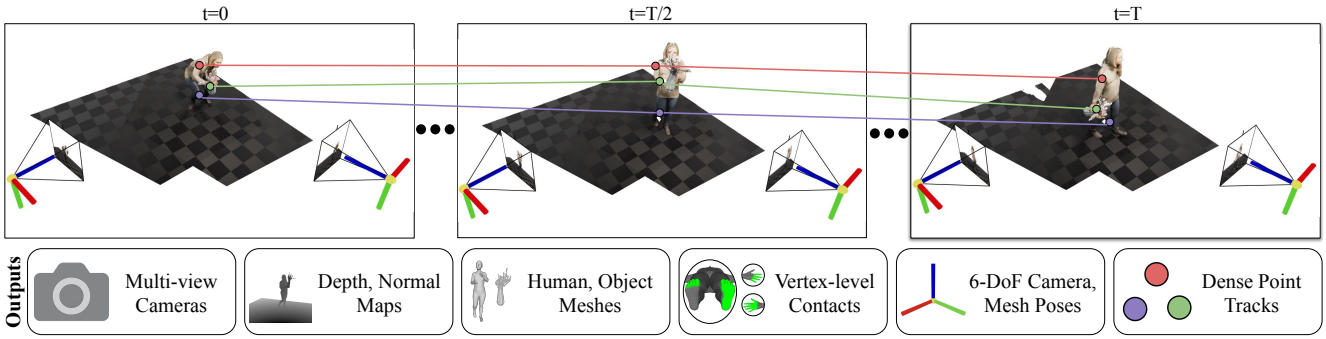


Figure 6. **Generated Data Modalities.** HOI-Blender may assist in generating multi-view synthetic HOI renders with auxiliary render passes and annotations—including depth, normal and segmentation maps, human and object mesh exports, as well as interaction contact points. Additional attainable annotations include 6-DoF camera poses, motion parameters, and temporally consistent dense point tracks; defining comprehensive resources for training models on HOI dynamics and temporal reasoning.

signals become rarely available at high fidelity [34, 35], and datasets commonly lack reliable contact or object geometry due to emphasis on humans over objects, thus hampering interaction modeling [4, 26]. In contrast, HOIBlender facilitates scalable synthetic supervision via multiview RGB, ground-truth cameras, auxiliary render passes, and temporally consistent contact meshes. Decoupled motion and appearance enables controlled ablations and randomized renders, yielding dense, perfectly labeled data for reconstruction and policy learning—without fragile initialization or heavy optimization.

Dataset Evaluation is accommodated by HOI-Blender via direct, controlled comparison of HOI datasets facilitated through standardized, scalable import and one-click rendering at shared camera views. This streamlines verification of alignment, contact plausibility, geometry gaps, mesh penetrations, and reprojection accuracy, as well as disentangling of content from capture factors via decoupled motion from appearance under specified lighting. Stratified sampling and gallery renders help identify outliers, guiding dataset splits and cross-dataset ablations.

We further demonstrate application of HOI-Blender for dataset evaluation by applying its provided tools for qualitative analysis of our initial 15 HOI supported datasets.

5. Results: Dataset Evaluation

HOI-Blender's standardized import and one-click rendering enable controlled, equatable comparisons across heterogeneous HOI datasets. To demonstrate use, we render motion clips across all supported datasets with automated sequence import, camera framing, and lighting—then utilize acquired renders as basis for dataset-level analysis.

We qualitatively evaluate 50 homogenized motion clips per-dataset via following axes across a four-point scale ($4 =$ strong performance, $1 =$ weak performance, see Figure 7):

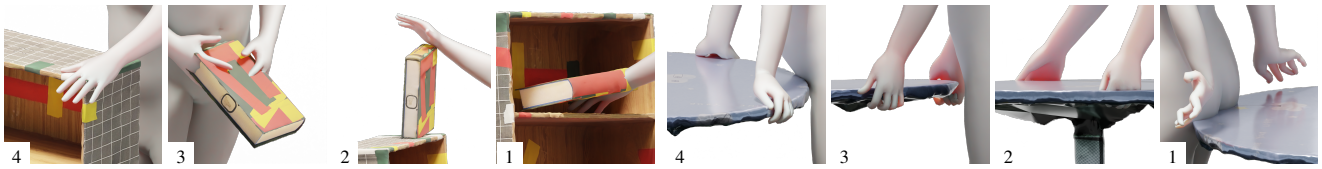
- **Animation Smoothness**: Motion continuity, void of temporal jitter and abrupt frame-to-frame transitions;
- **Motion Realism**: Perceived naturalness of motion;
- **Mesh Quality**: Absence of deformation artifacts as well as topological failures during animation playback;
- **HOI Contact Quality**: Tenability of hand-object contact;
- **Hand Grip Quality**: Plausibility of grip-action and finger placement during grasping;
- **Inter-Mesh Collision**: Lack of irregular interpenetration.

Both Human-only and SMPL-type datasets—or rather, those generally in lack of hand parameters—may not facilitate impression for occasional evaluation criteria, hence are marked via '-'.

Discussion. Table 2 enables dataset comparison across qualitative metrics, and thus the observation of occasional trends. Isolated Human Motion datasets score high for kinematic metrics, likely aided by a lack of dispersed focus between human and object during capture; AMASS outscores AIST++ as dance may perform actions of considerable flexibility and complexity (AIST++: AS 3.52, MR 3.78, MQ 3.54; AMASS: AS 3.94, MR 3.98, MQ 3.96). Human-Human Interaction is characterized by natural motion and clean geometry (Hi4D: MR 4.00, MQ 3.88, IMC 3.96; Duolando : MR 3.98, MQ 3.92, IMC 3.98). Full-Body Human-Object Interaction often sees deformation fault at occlusions (Behave: MQ 3.10; NeuralDome : MQ 2.92; IMHD^2: MQ 3.10) or animation jitter (InterCap: AS 2.00; COUCH: AS 2.86; IMHD^2: AS 3.02), however, cases of particularly well captured interaction are also distinguishable (OMOMO: CQ 3.98; InterCap: GQ 3.68; NeuralDome: GQ 3.74; IMHD^2: GQ 3.88). Hand-Object Interaction datasets majorly present restricted interaction scenarios, thus offer absence of deformation (GRAB: MQ 3.98; Arctic: MQ 4.00; HIMO: MQ 4.00) with high grip plausibility (GRAB: GQ 3.94; Arctic: GQ 3.90; HIMO: GQ 3.53). Multi-Human Object Interaction proffers contrasting perspectives on deformations and interpenetration (HOI-M3: MQ 4.00; CORE4D: MQ 3.24; vs. HOI-M3: IMC 2.96; CORE4D: IMC 3.90), with CORE4D at higher overall scoring impression (CORE4D: AS 3.26; MR 3.41; GQ 3.69).

Table 2. **Evaluation of Human-Centric Motion Datasets with Automatic Visualization via HOI-Blender.**

Dataset	Animation Smoothness	Motion Realism	Mesh Quality	HOI Contact Quality	Hand Grip Quality	Inter-Mesh Collision
AIST++ [15]	3.52	3.78	3.54	-	-	-
AMASS [20]	3.94	3.98	3.96	-	-	-
Hi4D [37]	3.24	4.00	3.88	-	-	3.96
Duolando [30]	3.34	3.98	3.92	-	-	3.98
Behave [3]	2.62	3.61	3.10	2.96	-	2.92
InterCap [8]	2.00	2.56	3.94	2.90	3.68	3.38
COUCH [40]	2.86	2.96	3.67	2.73	-	3.25
OMOMO [14]	3.94	3.94	3.92	3.98	-	3.36
NeuralDome [38]	3.70	3.40	2.92	3.60	3.74	3.29
IMHD^2 [41]	3.02	3.12	3.10	3.31	3.88	3.44
GRAB [31]	3.91	3.52	3.98	3.43	3.94	3.69
Arctic [5]	3.98	3.52	4.00	3.34	3.90	3.46
HIMO[19]	3.82	2.84	4.00	2.90	3.53	3.47
HOI-M3 [39]	3.04	3.26	4.00	3.25	-	2.96
CORE4D[17]	3.26	3.41	3.24	3.71	3.69	3.90



Accurate Contact | Plausible Contact | Dubious Contact | Erroneous Contact | Accurate Grip | Plausible Grip | Dubious Grip | Erroneous Grip

(a) Visualization of adopted metric for HOI-Contact plausibility.

(b) Visualization of adopted metric for Hand Grip plausibility.

Figure 7. **Demonstration of Plausibility Ordering for both Contact and Grip Quality.**

HOI-Blender streamlined our dataset-level comparisons by standardizing import and rendering for equitable contrast across datasets. Consequently, we identified isolated human motion excel in kinematic fidelity, multi-human proficiency in lifelike captures, and hand-object prevail at constrained, refined grasping. Human-object interaction is often impeded by occlusion-driven artifacts, however, carries datasets of exceptional contact point performance as well. Large-scale contrast between datasets can guide the direction of upcoming developmental advancement, downstream application, and focused intervention, overall leading to enhanced performance across a multitude of HOI tasks.

6. Outlook and Conclusion

We introduced HOI-Blender, a unifying loader Blender addon that turns heterogeneous HOI datasets into a consistent, render-ready format. It resolves conflicting conventions and metadata, congregates coherent scenes with shared camera rigs, lighting composition and render configuration, decouples motion from appearance for flexible recombination, and supports one-click import, animation preview, as well as automated rendering. In practice, this makes visualization, cross-dataset comparison, and dataset-driven supervision strikingly simpler and more reliable.

Our findings from Section 5 guide downstream data generation and amendment: alongside HOI-Blenders standardized importing and coherent scenes, motion characteristics can compose into high-quality sequences (AMASS motion, OMOMO contacts, GRAB hands). Smooth fullbody kinematics can be combined with object placements and contact cues, as well as retargeted hand poses to yield natural motion, plausible contacts, and robust grasps. Such compositional strategies would enable exploration of large design spaces and generation of targeted datasets for both training and evaluation.

Looking ahead, HOI-Blender provides a practical foundation for extensions which broaden its significance across abundant synthesis, benchmarking, and embodied learning. By leveraging standardized scene coherence alongside equitable renders, evaluations can grow beyond qualitative inspection into automated, quantitative diagnostics of mesh contact, reprojection, and interpenetration. Through mix and matching of sequences with differing characteristics, those non-interactive can be enriched by interactive components and low-fidelity grasps can be upgraded with higher-quality counterparts. Lightweight post-processing—finger retargeting, pose priors, and minor optimization passes—can reduce hand artifacts and improve contact plausibility, while native scripting support can streamline large-scale or batched generation of synthetic corpora. Finally, stronger physics and contact realism could enable consistent retargeting and contact correction, as well as provide synthetic, force-aware supervision for robotics and control.

References

[1] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025. 6

[2] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009. 1

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and Method for Tracking Human Object Interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 2, 3, 8

[4] 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 (CVPR)*, 2023. 2, 7

[5] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 8

[6] Adam W. Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Pavel Tokmakov, Suya You, Rares Ambrus, Katerina Fragkiadaki, and Leonidas J. Guibas. AllTracker: Efficient Dense Point Tracking at High Resolution, 2025. 6

[7] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing Physical Character-Scene Interactions. New York, NY, USA, 2023. Association for Computing Machinery. 1

[8] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction. In *German Conference on Pattern Recognition (GCPR)*, pages 281–299. Springer, 2022. 2, 3, 8

[9] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs, 2019. 1

[10] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4D: Leveraging Video Generators for Geometric 4D Scene Reconstruction, 2025. 6

[11] Manuel Kaufmann, Velko Vechev, and Dario Mylonopoulos. aitviewer, 2022. 3

[12] Hema S. Koppula and Ashutosh Saxena. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, 2016. 1

[13] Rosario Leonardi, Antonino Furnari, Francesco Ragusa, and Giovanni Maria Farinella. Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection? In *ECCV*, 2024. 6

[14] Jiaman Li, Jiajun Wu, and C Karen Liu. Object Motion Guided Human Motion Synthesis. *ACM Trans. Graph.*, 42 (6), 2023. 2, 3, 8

[15] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation, 2021. 2, 3, 8

[16] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, 2017. 4

[17] 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 Computer Vision and Pattern Recognition Conference*, pages 1769–1782, 2025. 3, 8

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 4

[19] 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, 2024. 3, 8

[20] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 3, 8

[21] Fahad Majeed, Maria Nazir, Kamilla Swart, Marco Agus, and Jens Schneider. Real-time analysis of soccer ballplayer interactions using graph convolutional networks for enhanced game insights. *Scientific Reports*, 15(1), 2025. 1

[22] Madhur Mangalam, Sanjay Oruganti, Gavin Buckingham, and Christoph W. Borst. Enhancing hand-object interactions in virtual reality for precision manual tasks. *Virtual Reality*, 28(4), 2024. 1

[23] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Assistive roBOTs. In *7th Annual Conference on Robot Learning*, 2023. 1

[24] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured World Models from Human Videos. *arXiv preprint arXiv:2308.10901*, 2023. 6

[25] Meshcapade. SMPL_blender_addon [blender add-on]. https://github.com/Meshcapade/SMPL_blender_addon, 2021. Accessed: August 22, 2025. 3

[26] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in Geography Optimized for Regression Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 2, 7

[27] 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 *IEEE Conference*

on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 4

[28] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D Clothing Capture and Retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally. 4

[29] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6):245:1–245:17, 2017. 4

[30] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower GPT with Off-Policy Reinforcement Learning for Dance Accompaniment. In *ICLR*, 2024. 2, 3, 8

[31] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 8

[32] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow, 2020. 6

[33] Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong, Yulong Chen, Lizhuang Ma, and Yong-Lu Li. Reconstructing In-the-Wild Open-Vocabulary Human-Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[34] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, Human and Object REconstruction from a single RGB image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 7

[35] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template Free Reconstruction of Human-object Interaction with Procedural Interaction Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 7

[36] Brent Yi, Chung Min Kim, Justin Kerr, Gina Wu, Rebecca Feng, Anthony Zhang, Jonas Kulhanek, Hongsuk Choi, Yi Ma, Matthew Tancik, and Angjoo Kanazawa. Viser: Imperative, Web-based 3D Visualization in Python, 2025. 3

[37] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D Instance Segmentation of Close Human Interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 8

[38] 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 *CVPR*, 2023. 2, 3, 8

[39] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M3: Capture Multiple Humans and Objects Interaction within Contextual Environment. *arXiv preprint arXiv:2404.00299*, 2024. 3, 8

[40] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: Towards Controllable Human-Chair Interactions. 2022. 2, 3, 8

[41] 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*, pages 729–741, 2024. 2, 3, 8