

HandX: Scaling Bimanual Motion and Interaction Generation

Supplementary Material

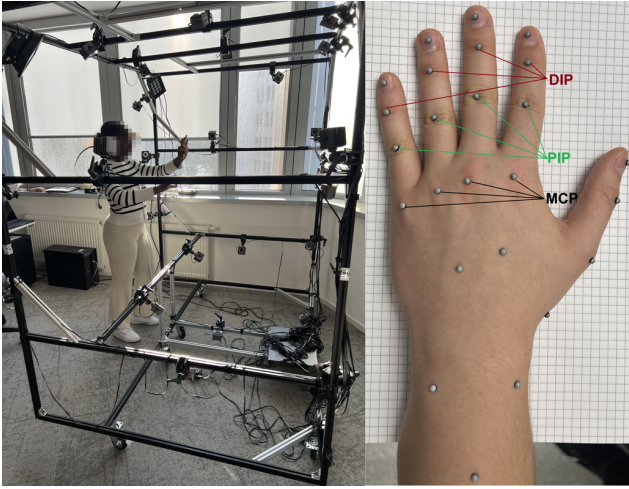


Figure A. Our motion capture configuration: (a) 36-camera OptiTrack studio and (b) placement of 25 markers on each hand.

This supplementary material is organized as follows. Sec. A describes our motion capture system and the construction of HandX from public data, including quality control procedures. Sec. B details our bimanual motion captioning pipeline. Sec. C provides additional details on bimanual motion representations. Sec. D explains how our model supports versatile motion generation. Finally, Sec. E presents additional details on metrics, evaluation results, and the user study.

A. HandX Dataset

A.1. Motion Capture

This subsection details the full data capture pipeline, from motion acquisition to marker-based skeleton reconstruction, and highlights the key design choices that ensure high-quality kinematic data.

Motion Capture Setup. We record all motion data using an OptiTrack motion capture system [39] within a dedicated studio. The capture volume is monitored by 36 high-speed infrared cameras, placed to maximize coverage and minimize marker occlusion during complex two-hand interactions or rapid finger movements, as shown in Figure A.

Marker Placement. Actors have 25 miniature infrared reflective markers (3mm in diameter) glued directly onto the skin of each hand. The placement of these markers (illustrated in Figure A(b)) is meticulously designed to capture the nuanced articulation of the hand, covering the wrist, the dorsal surface of the palm (metacarpals), three key points

on each finger corresponding to the metacarpophalangeal (MCP), proximal interphalangeal (PIP), and distal interphalangeal (DIP) joints, and the fingertip (center of the nail).

Skeleton Reconstruction. The raw output from the OptiTrack system provides 3D coordinates for the 25 surface markers. To estimate the underlying skeleton joint positions from these surface markers, we first compute the anatomical normal direction \vec{n} for each marker, pointing from the skin surface inward towards the bone. For joints along a finger (e.g., PIP, DIP), this normal is derived from the plane formed by the marker and its proximal and distal neighbors (e.g., using MCP, PIP, and DIP markers). For MCP joints, the normal is computed from the plane defined by neighboring MCP markers on the dorsal surface of the hand.

Next, we estimate the depth d from the skin to the joint center along this normal. This depth value is scaled based on the actor’s overall hand size, which is determined during a calibration phase. The final 3D position of a skeleton joint J_p is then calculated by offsetting its corresponding marker position M_p along the computed normal \vec{n} by the estimated depth d :

$$J_p = M_p + \vec{n} \cdot d. \quad (\text{A})$$

Wrist Optimization. While the above process effectively locates the finger joints, the calculations are still affected to soft-tissue artifacts. The most significant non-rigid deformation occurs at the wrist, where skin and soft tissue can stretch and compress substantially, leading to inconsistent distances between, e.g., an MCP marker and a wrist marker.

To compensate for this, we run an iterative optimization process. We assume that the bone lengths, specifically, the distances from each MCP joint to the wrist joint center, remain constant. First, we calculate a reference “bone length” L_{ref_i} for each MCP-to-wrist connection ($i = 1, \dots, 5$) from a static, neutral calibration pose. Then, for every frame in the recording, we iteratively optimize the 3D position of the single wrist joint, J_{wrist} , to find the position J_{wrist}^* that minimizes the squared error between the current calculated distances and these reference lengths:

$$J_{wrist}^* = \underset{J_{wrist}}{\operatorname{argmin}} \sum_{i=1}^5 (\|J_{MCP_i} - J_{wrist}\| - L_{ref_i})^2. \quad (\text{B})$$

This optimization ensures that the wrist joint position is anatomically consistent with the relatively rigid palm, effectively filtering out the artifacts caused by soft tissue deformation.

A.2. Data Consolidation and Unification

As discussed in Sec. 3, an important component of HandX comes from aggregating public datasets, whose licenses are summarized in Sec. F. To ensure consistency across these heterogeneous sources, we unify both the motion representation and the global coordinate system. First, we map all hand motion to a unified 21-joint skeletal topology, ensuring a consistent joint definition and ordering across datasets. We then canonicalize the global coordinates at the sequence level while preserving a right-handed coordinate frame. After this transformation, in the first frame, the positive x -axis points from the left wrist to the right wrist, the positive y -axis points from the wrists toward the fingertips, and the positive z -axis points upward. This two-stage unification establishes a consistent skeletal and spatial reference across all data sources, facilitating subsequent processing and analysis.

A.3. Clip Extraction

During the post-processing stage of HandX, we divide long motion sequences into clips of 60 frames (2 seconds at 30 FPS). This clip length follows the standard setting in prior work on short-term hand motion modeling [67, 81]. It is long enough to capture a complete atomic hand action (*e.g.*, a full pinch-and-release, a grasp, or a sign) while remaining efficient for training. However, the aggregated data may contain defective frames that either do not provide meaningful bimanual pose or motion information or exhibit abrupt changes across consecutive frames. To ensure data quality, we adopt two principles during clip extraction: **(a)** detecting and removing defective frames, and **(b)** using a non-overlapping extraction strategy.

Sequences without defective frames. If no defective frames are detected, we segment the sequence into non-overlapping clips using windows $\mathcal{W}(t)$ of length L :

$$\mathcal{W}(t) = \{t, \dots, t + L - 1\}, \quad (C)$$

$$t \in \{0, L, 2L, \dots\},$$

where $L = 60$ denotes the clip length. This construction ensures that adjacent clips do not overlap.

Sequences with defective frames. If defective frames are present, we first remove their indices and then perform segmentation only on the remaining valid continuous intervals. Let $\mathcal{D} \subset \{0, \dots, T - 1\}$ denote the set of defective frames. We partition the valid index set $\{0, \dots, T - 1\} \setminus \mathcal{D}$ into maximal contiguous intervals:

$$\mathcal{S} = \{[a_k, b_k]\}_{k=1}^{K_s}, \quad [a_k, b_k] \cap \mathcal{D} = \emptyset, \quad b_k - a_k + 1 \geq L.$$

Within each interval $[a_k, b_k]$, we place disjoint windows $\mathcal{W}(t)$ with stride $s = L = 60$. This guarantees that every extracted clip is fully valid and that no overlap exists between adjacent clips.

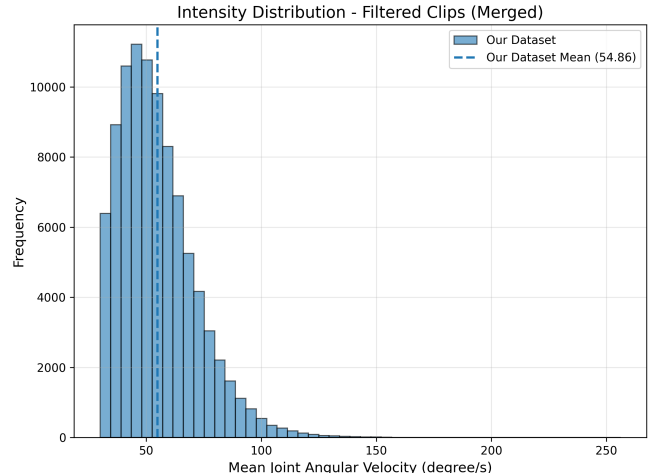


Figure B. **Distribution of filtered clips.** This figure shows the intensity distribution of the filtered clips.

A.4. Clip Filtering

As shown in Table 1(b), analysis of the source datasets reveals a substantial data imbalance in most datasets, particularly in terms of motion intensity and inter-hand interaction, *e.g.*, the hand is idle for most of the time. This imbalance is problematic for training, as it may bias the model toward low-activity states.

To remove static or uninformative segments from the split clips, we apply an *intensity-aware* criterion that identifies and discards static or low-activity segments, retaining only clips with meaningful bimanual dynamics. Specifically, we compute an **action intensity metric** based on joint angular velocity, which is independent of the global motion of the hands. A clip is retained only if both hands exhibit sufficiently strong motion simultaneously, ensuring that the final dataset consists of highly dynamic bimanual interactions. The resulting distribution of filtered clips is shown in Figure B.

Action Intensity Metric. Let \hat{v}_t^{pre} and \hat{v}_t^{next} denote the unit limb directions before and after a joint at time t , respectively. The inter-segment angle is computed as

$$\theta_t = \arccos(\hat{v}_t^{\text{pre}} \cdot \hat{v}_t^{\text{next}}). \quad (D)$$

The angular velocity ω is then computed from the temporal difference of the bending angle for each finger chain.

For each hand $h \in \{\text{left}, \text{right}\}$ with joint set \mathcal{J}_h , we assign a hierarchical importance weight λ_j to each joint $j \in \mathcal{J}_h$. To emphasize structural pose changes, proximal joints are assigned larger weights than distal ones. The hand-level clip intensity over a temporal window \mathcal{W} is computed as the

weighted average of angular velocities:

$$\bar{\omega}_h(\mathcal{W}) = \frac{\sum_{t \in \mathcal{W}} \sum_{j \in \mathcal{J}_h} \lambda_j \omega_t^{(j)}}{\sum_{t \in \mathcal{W}} \sum_{j \in \mathcal{J}_h} \lambda_j}. \quad (\text{E})$$

The bimanual clip intensity is then defined as the average intensity of both hands:

$$\bar{\omega}(\mathcal{W}) = \frac{1}{2} (\bar{\omega}_{\text{left}}(\mathcal{W}) + \bar{\omega}_{\text{right}}(\mathcal{W})). \quad (\text{F})$$

Filtering Rule. We retain only those clips that exhibit sufficiently strong motion on both hands. Specifically, a clip \mathcal{W} is kept if

$$\bar{\omega}_{\text{left}}(\mathcal{W}) \geq \tau_{\text{hand}}, \bar{\omega}_{\text{right}}(\mathcal{W}) \geq \tau_{\text{hand}}, \bar{\omega}(\mathcal{W}) \geq \tau_{\text{avg}},$$

where τ_{hand} and τ_{avg} are fixed thresholds. In practice, we set $\tau_{\text{hand}} = 25$ and $\tau_{\text{avg}} = 30$.

A.5. Dataset Metrics

In Table 1(b), we use several metrics to evaluate the quality of data interaction and action intensity. Specifically, we employ four key indicators. **Contact Ratio** measures the fraction of frames where inter-hand contact occurs, representing the density of interaction. **Contact Duration** denotes the average length of continuous contact spans, reflecting the stability of interactions. **Contact Freq** counts the number of distinct contact events per minute, indicating the complexity and richness of the interaction. **Motion Intensity** quantifies the magnitude of fine-grained finger dynamics, as defined in Sec. A.4.

B. Additional Details on Motion Captioning

As Sec. 4 introduces, the objective of kinematic feature extraction is to transform raw hand motion sequences into structured, semantically meaningful representations that can be subsequently interpreted directly by an LLM. We compute six types of kinematic descriptors for bimanual hand motion. When the value of a kinematic descriptor exhibits changes over a period of time, we extract it as an “event.” These events collectively constitute the structured features of the two-hand motion.

Kinematic Descriptors. We define six types of kinematic descriptors to comprehensively characterize the motion: *Finger Flexing*, *Finger Spacing*, *Finger-finger Distance*, *Palm-palm Relation*, *Finger-palm Distance*, and *Wrist Trajectory*. These descriptors aim to comprehensively capture the motion by covering both local, single-hand features and global, bimanual dynamics.

Finger Flexing. For a specific joint of a finger, let $\mathbf{p} \in \mathbb{R}^3$ denote the joint’s coordinates, with \mathbf{p}^{pre} and \mathbf{p}^{next} representing the coordinates of the predecessor and successor joints, respectively. Define $\mathbf{v}^{\text{pre}} = \mathbf{p}^{\text{pre}} - \mathbf{p}$ and $\mathbf{v}^{\text{next}} =$

$\mathbf{p}^{\text{next}} - \mathbf{p}$. We compute the “signed angle” θ of joint \mathbf{p} ’s flexion using the dot and cross products of \mathbf{v}^{pre} and \mathbf{v}^{next} . If $\theta > 0$, the joint exhibits normal flexion; if $\theta < 0$, the joint is in a hyperextended state.

Finger Spacing. The finger spacing of two adjacent fingers is measured by the angle formed between their respective finger directions, computed from the MCP to PIP joints.

Finger-finger Distance. The distance between any two fingertips represents *Finger-finger Distance*. This includes both intra-hand distances between different fingertips of the same hand and inter-hand distances between left and right hand fingertips.

Palm-palm Relation. For a single hand, we first randomly sample 100 points within the convex hull formed by the wrist joint and the five MCP joints, representing the palm’s point cloud $\mathcal{G}^h = \{\mathbf{q}_i^h \in \mathbb{R}^3\}_{i=1}^{100}$ ($h = \text{L, R}$). From all point pairs between \mathcal{G}^{L} and \mathcal{G}^{R} , we select the 30 closest pairs $\{(\mathbf{q}_i^{\text{L}}, \mathbf{q}_i^{\text{R}})\}_{i=1}^{30}$. Computing the differences of their world coordinates yields 30 vectors $\{\mathbf{v}_i\}_{i=1}^{30}$. We then average these vectors to obtain $\bar{\mathbf{v}} \in \mathbb{R}^3$, which serves as the kinematic descriptors for the left-to-right palm relation.

Finger-palm Distance. For a fingertip joint \mathbf{p} of one hand, we identify the 5 closest points $\{\mathbf{q}_i\}_{i=1}^5$ in the other hand’s palm cloud \mathcal{G} . The average distance between \mathbf{p} and these points serves as the this descriptor.

Wrist Trajectory. We use the world coordinate of the wrist joint at the current timestep.

Event Segmentation. In general, if the value of a kinematic descriptor exhibits a detectable change over a time period, we extract it as an “event.” Additionally, if the feature’s value remains essentially constant throughout the entire motion, we also consider this as a distinct type of event.

Table A. Quantitative state intervals for different hand motion features used in event segmentation.

Feature	Range	State
Finger Flexing (°)	$[-180, -20)$	Hyper extend
	$[-20, 30)$	Fully extend
	$[30, 60)$	Partially bent
	$[60, 180)$	Fully bent
Finger Spacing (°)	$[0, 20)$	Closed
	$[20, 180)$	Open
Finger-Finger Distance (m)	$[0, 0.02)$	Contact
	$[0.02, +\infty)$	No Contact
Finger-Palm Distance (m)	$[0, 0.025)$	Contact
	$[0.025, 0.035)$	Near
	$[0.035, +\infty)$	Far

As shown in Table A, we have defined value intervals for certain descriptors, with each interval corresponding to a semantic label. For *Finger Flexing*, *Finger Spacing*, *Finger-finger Distance*, and *Finger-palm Distance*, events are characterized as transitions from one state to another or maintaining a constant state throughout the motion.

For *Wrist Trajectory* and *Palm-palm Relation*, we decompose their vectors along the X, Y, and Z axes of the world coordinate system, where changes along these axes correspond to spatial movements in left-right, forward-backward, and up-down directions, respectively. Figure C shows an example of extracted feature.

Prompt Design. Figure D illustrates our prompt design. By feeding the large language model JSON-formatted features, it can generate five distinct annotations at varying levels of granularity as requested.

C. Hand Motion Representation

We here provide the detailed definition of the representation component, rotational scalar. At frame i , let $v^i \in \mathbb{R}^3$ denote the vector formed from the MCP joint of the little finger to the MCP joint of the index finger. For a given joint, let α be the angle formed by the joint itself, its predecessor joint, and its successor joint along the finger’s kinematic chain. We project angle α onto the plane perpendicular to v^i , and the resulting projected angle magnitude serves as the “rotation scalar” s for this joint. We concatenate the 3D coordinates and rotation scalar for each joint at each frame

$$\mathbf{x}^i = [\mathbf{p}^i; \mathbf{s}^i] \in \mathbb{R}^{2J \times 4}, \quad (\text{G})$$

where $[\cdot; \cdot]$ denotes the concatenation operation. Thus, $\mathbf{x} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^F\} \in \mathbb{R}^{F \times 2J \times 4}$ is our data representation.

D. Versatile Bimanual Motion Generation

D.1. Masked Partial Denoising for Skeleton Control

Our versatile generation module is implemented using a masked partial-denoising strategy inspired by [47]. The goal is to explicitly condition a subset, e.g., keyframes, wrist trajectories, or single-hand constraints, while letting the remaining degrees of freedom be generated.

During inference, the diffusion model first denoises the current noisy sample \mathbf{x}_t to obtain a clean prediction \mathbf{x}_0 , and then re-applies the forward noising process to produce the next timestep sample \mathbf{x}_{t-1} . Let $\mathbf{x}_0^{\text{pred}} = \mathcal{G}(\mathbf{x}_t, t, T)$ denote the clean motion predicted by the model given timestep t and text condition T , and let \mathbf{x}_0^{gt} denote the target motion that provides conditions (e.g., keyframe poses or reference trajectories). We index frames by $i \in \{0, \dots, L-1\}$ and joints by $j \in \{0, \dots, 2J-1\}$, and write $\mathbf{x}_0(i, j)$ for the state of joint j at frame i , where L denotes the generation length, and J denotes the number of joints per hand.

Soft Interpolation. To enable smooth temporal control, we apply constraints at the joint level and softly extend them from selected *keyframes to nearby frames*. For each joint $j \in \{0, \dots, 2J-1\}$, we define a set of center frames $\mathcal{C}_j \subseteq \{0, \dots, L-1\}$ where the target motion is most strongly enforced. This formulation applies to any control setting in

which only a subset of frames is specified, e.g., *motion in-betweening*, *keyframe-based generation*, and *long-horizon generation*, detailed in Sec. D.2.

For each center frame $i \in \mathcal{C}_j$, we define a local temporal window

$$\mathcal{T}(i) = \{t \in \{0, \dots, L-1\} \mid |t-i| \leq k_{\text{trans}}\}, \quad (\text{H})$$

where k_{trans} controls the transition range. Within this window, the constraint weight decays linearly with temporal distance:

$$\gamma_{t,j}^{(i)} = p_{\text{hard}} - (p_{\text{hard}} - p_{\text{soft}}) \frac{|t-i|}{k_{\text{trans}}}, \quad (\text{I})$$

where $p_{\text{hard}} = 0.85$, $p_{\text{soft}} = 0.10$, and $k_{\text{trans}} = 5$.

If multiple windows overlap, we use the strongest weight:

$$\gamma_{t,j} = \begin{cases} \max_{i \in \mathcal{C}_j, t \in \mathcal{T}(i)} \gamma_{t,j}^{(i)}, & \text{if } \exists i \in \mathcal{C}_j \text{ s.t. } t \in \mathcal{T}(i), \\ 0, & \text{otherwise.} \end{cases} \quad (\text{J})$$

Thus, $\gamma_{t,j} = 0$ means joint j at frame t is unconstrained.

We then update the clean signal by interpolating between the model prediction and the target:

$$\mathbf{x}_0(t, j) = (1 - \gamma_{t,j}) \mathbf{x}_0^{\text{pred}}(t, j) + \gamma_{t,j} \mathbf{x}_0^{\text{gt}}(t, j). \quad (\text{K})$$

At the center frame $t = i \in \mathcal{C}_j$, the constraint is strongest, with $\gamma_{i,j} = p_{\text{hard}}$.

Finally, we re-noise the modified \mathbf{x}_0 using the standard forward diffusion step:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon, \quad (\text{L})$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_{t-1}$ follows the training noise schedule.

D.2. Task-Specific Mask Construction

Based on the per-joint masking formulation in Sec. D.1, we implement versatile generation tasks by configuring the center frame set \mathcal{C}_j for different groups of joints. For clarity, we define \mathcal{J}_{all} as the set of all joints, and define specific subsets (e.g., $\mathcal{J}_{\text{wrist}}$) for spatial control.

Motion In-betweening. We constrain the start and end of the sequence for all joints. We set the target frame set $T_{\text{target}} = \{0, \dots, K_{\text{inbet}} - 1\} \cup \{L - K_{\text{inbet}}, \dots, L - 1\}$, with $K_{\text{inbet}} = 5$. The joint masks are configured as:

$$\mathcal{C}_j = T_{\text{target}}, \quad \forall j \in \mathcal{J}_{\text{all}}. \quad (\text{M})$$

Keyframe-Based Generation. Given sparse keyframes at indices $T_{\text{key}} = \{t_1, \dots, t_k\}$, we enforce these poses on the full skeleton:

$$\mathcal{C}_j = T_{\text{key}}, \quad \forall j \in \mathcal{J}_{\text{all}}. \quad (\text{N})$$

```

{
  "frame_count": 60,
  "left_hand_events": {
    "finger_flexing": {
      ...,
      "pinky_mcp": [
        {"start": 0, "end": 21, "start_des": "Fully extended", "end_des": "Hyper extend"},
        {"start": 42, "end": 60, "start_des": "Hyper extend", "end_des": "Fully bent"}
      ],
      "pinky_pip": [
        {"start": 0, "end": 60, "constant_des": "Fully bent"}
      ],
      ...,
    },
    ...
  },
  "right_hand_events": {
    ...
  },
  "two_hand_relationship": {
    ...,
    "palm_palm_relative_position": {
      "left-right": [
        {"start": 0, "end": 30, "start_des": "right hand is to the RIGHT of the left hand.", "end_des": "right hand is to the LEFT of the left hand."},
        {"start": 50, "end": 60, "start_des": "right hand is to the LEFT of the left hand.", "end_des": "right hand is to the RIGHT of the left hand."}
      ],
      ...
    }
  }
}

```

Figure C. Structured representation of extracted motion features. The JSON format captures temporal segmentation with frame indices and semantic state descriptions for transition events and constant events.

Wrist Trajectories Generation. We constrain only the wrist joints throughout the entire sequence, leaving fingers free. Let $T_{\text{seq}} = \{0, \dots, L - 1\}$. We set:

$$C_j = \begin{cases} T_{\text{seq}}, & \text{if } j \in \mathcal{J}_{\text{wrist}}, \\ \emptyset, & \text{otherwise.} \end{cases} \quad (\text{O})$$

This explicitly controls the wrist positions while setting $\gamma_{t,j} = 0$ for all other joints.

Hand-Reaction Synthesis. One hand (e.g., left hand) is fully constrained as the condition. Let $\mathcal{J}_{\text{left}}$ denote the joints of the left hand. We set:

$$C_j = \begin{cases} T_{\text{seq}}, & \text{if } j \in \mathcal{J}_{\text{left}}, \\ \emptyset, & \text{if } j \in \mathcal{J}_{\text{right}}. \end{cases} \quad (\text{P})$$

Long Horizon Generation. We generate motion autoregressively. We take the last $K_{\text{hor}} + k_{\text{trans}}$ frames of the preceding generated sequence as the ground truth to constrain the start of the current sequence. We configure the center frames as:

$$C_j = \{0, \dots, K_{\text{hor}} - 1\}, \quad \forall j \in \mathcal{J}_{\text{all}}. \quad (\text{Q})$$

where the temporal smoothing radius is set to k_{trans} .

E. Evaluation Details

E.1. Contact Metrics

To comprehensively evaluate the accuracy of two-hand interactions generation, we decompose the contact metrics

Table B. **Ablation study** on diffusion model for analyzing scaling curves. R-Precision is evaluated with a batch size of 16.

Configuration				FLOPs (G)	R-Precision [†]			FID [‡]
L	d	H	FFN		Top 1	Top 2	Top 3	
2	256	4	1,024	3.83	0.117 \pm 0.072	0.210 \pm 0.088	0.295 \pm 0.091	5.342 \pm 0.073
4	256	4	1,024	4.22	0.160 \pm 0.085	0.264 \pm 0.098	0.355 \pm 0.107	2.705 \pm 0.057
4	384	6	1,536	5.22	0.185 \pm 0.087	0.298 \pm 0.059	0.389 \pm 0.092	2.053 \pm 0.033
8	384	6	1,536	6.98	0.262 \pm 0.101	0.407 \pm 0.112	0.501 \pm 0.115	1.959 \pm 0.066
8	512	8	2,048	9.74	0.318 \pm 0.117	0.463 \pm 0.124	0.566 \pm 0.124	1.830 \pm 0.033
12	512	8	2,048	12.87	0.300 \pm 0.109	0.437 \pm 0.119	0.536 \pm 0.116	2.492 \pm 0.030
12	768	12	3,072	24.62	0.411\pm0.127	0.570\pm0.127	0.672\pm0.117	1.695\pm0.039

into two categories: intra-hand and inter-hand contacts. For intra-hand contacts, we assess whether the thumb fingertip makes contact with each of the other four fingertips within a single hand. A contact is registered when the distance between two fingertips falls below a predefined threshold at any frame in the sequence.

For inter-hand contacts, we measure the minimum distance between the closest point pair across the two hands. A contact is detected when this minimum distance is below the threshold.

We extract contact labels for both intra-hand and inter-hand categories from both ground truth and generated motion. We then compute true positives (TP), true negatives (TN), and false positives (FP), and report standard metrics. Intra-hand evaluation is presented in Table 2, the precision of inter-hand evaluation is shown in Table D.

E.2. Evaluator Details

Our evaluator takes as input the global hand joint positions. Following [36, 42], we jointly train the text en-

[Task overview & Goal]:

You are an expert in motion analysis. Your goal is to generate a list of five distinct annotations for a given JSON-formatted two-hand 3D motion sequence, serving as text augmentation. Each annotation must be a JSON object with keys 'left', 'right', and 'two_hands_relation', describing the physical motion of the hands. Focus strictly on joint states, finger movement, and inter-hand spatial relationships—without interpreting gesture meaning or intent. The five annotations should vary in length and detail as specified in the [Output Specification] section.

[Input Format & Structure Description]:

You are given a JSON file representing a 3D two-hand motion sequence. The structure is composed of three top-level keys:

- frame_count
- left_hand_events
- right_hand_events
- two_hand_relationships

Each section includes multiple events, organized by type. An event is recorded as a list of dictionaries, each representing a meaningful temporal change. Each event dictionary includes:

- start: Start frame index of the event
- end: End frame index of the event
- start_des: Description of the state at the start
- end_des: Description of the state at the end
- constant_des: Describes a consistently maintained state throughout the interval without significant change
- v_des: Description of movement speed (e.g., Slow, Medium, Fast)
- direction: (only present in wrist trajectory) Direction of movement over time

Hand Event Types:

Under left_hand_events and right_hand_events, the following categories are defined:

1. *finger_flexing*
 - Keyed by finger joints (e.g., index_mcp, thumb_ip)
 - Describes flexion transitions: bending/straightening
2. *finger_spacing*
 - Keyed by adjacent finger pairs (e.g., "index, middle")
 - Captures dynamic spacing between adjacent fingers.
3. *finger_tip_contact*
 - Keyed by fingertip pairs (e.g., "thumb_tip, index_tip")
 - Describes the contact condition between the thumb and the fingertips of the other four fingers on the same hand.
4. *wrist_trajectory*
 - Axes: left-right, front-back, down-up
 - Describes the wrist's motion along vertical, horizontal, and depth axes.

Two-Hand Relationships:

The two_hand_relationships section captures inter-hand spatial dynamics:

1. *finger_tip_contact*
 - Contact state between fingers across hands
2. *finger_palm_distance*
 - Fingertip-to-opposite-palm distances across hands
3. *palm_palm_distance*
 - Scalar distance between palms over time
4. *palm_palm_relative_position*
 - Axial description of spatial positioning between palms
 - Includes: 'left-right', 'front-back', 'up-down'

[Output Specification]:

Your output must be a single JSON list containing **exactly five** JSON objects. Each object represents one annotation and must adhere to the following structure:

```
{
  "left": "...",
  "right": "...",
  "two_hands_relation": "..."
}
```

The five annotations must vary in detail and length to provide diverse descriptions of the same motion:

- **Annotation 1 & 2:** Use concise and succinct language. Focus only on the most significant movements or state changes.
- **Annotation 3 & 4:** Provide a moderately detailed description. Include primary movements and some secondary but notable details. The length should be balanced.
- **Annotation 5:** Generate a highly detailed and comprehensive description. Cover all significant events, including subtle changes, speed variations, and precise temporal markers. This should be the longest and most fine-grained annotation.

[Motion Description Rules]:

1. **Critical Event Reporting:**

- If there are events such as **contact** or **hyperextension**, you must always include them in the annotation and never omit them. These are high-priority features.
- Contact overrides all other spatial relationships. If a fingertip makes contact with another part, describe the contact event only and do not describe other proximity or distance relations involving that fingertip.
Good: The left index fingertip made contact with the right palm center.
Bad: The left index fingertip moved closer to the right palm and made contact, while staying near the right middle finger.

2. **Temporal Context:**

- Always indicate the temporal position of an event within the sequence. Use phrases like "initially," "at the beginning," "mid-sequence," "towards the end," or describe events that "recur" or are "maintained throughout."

3. **Left/Right Hand Motion Guidelines:**

- Highlight notable features: extreme state changes, high-speed motions, significant spacing shifts, or extreme flexion states.
Good: The left index finger quickly moves from fully extended to fully bent.
Good: The spacing between the right index and middle fingers changes from touching to far apart at a medium speed.
- Finger-Level Flexion Description: Always describe finger flexing at the level of the **whole finger**. Do not break it down into joint-level details. Combine joint data to describe each finger's overall bend state.
Exception: If a finger shows hyperextension at the base joint (MCP) but bending at the distal joints (PIP/DIP), you must describe the joints separately and clearly to capture this complex posture.
Good: The left index finger remains in a fully bent state.
Bad: The MCP joint of the left index finger remains in a fully bent state; the PIP joint of the left index finger remains in a fully bent state...
- Condense similar finger behaviors for clarity and conciseness.
Good: The right index and middle fingers move from partially bent to fully extended.
- Summarize repeated patterns of motion.
Good: The left index finger and the right palm repeatedly make contact and then move apart.

4. **Two-Hand Relationship Guidelines:**

- Emphasize meaningful interactions: contact, high-speed approach/separation, or significant changes in relative positioning.
Good: The right index fingertip quickly moved from a distant position to make contact with the left palm.
Good: Initially, the right hand was positioned above the left hand, but it gradually moved to a position below it.
- Avoid trivial spatial details (e.g., maintaining a large, unchanging distance) unless it is part of a significant transition.

[Style]:

- Maintain clarity and fluency; use direct, confident, and unambiguous language.
- Avoid speculation or inferring intent (e.g., "might," "seems," "possibly").
- Summarize only the motion—do not include explanations, metadata, or any text outside of the specified JSON output format.
- Avoid repetition within a single annotation. Prioritize precision and tempo in the narrative.

Figure D. Our prompt design to facilitate LLMs to summarize kinematic features. This is for bimanual motion captioning.

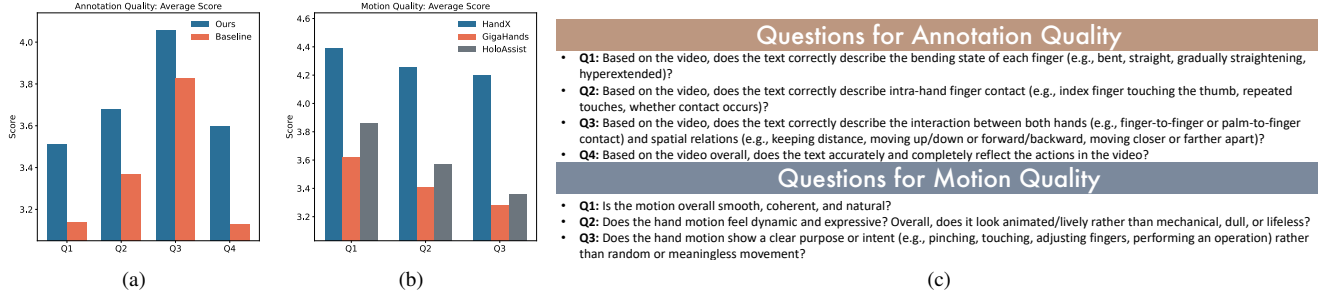


Figure E. **User study on data quality.** Our HandX exhibits high-quality motion and annotations. For annotation quality evaluation, we compare our method with a baseline that utilizes Gemini 3 Pro [51] to directly caption rendered motion videos.

Table C. Detailed configurations of the diffusion and autoregressive models. Parameter counts exclude the frozen text encoder. Results from Figure 4 are extracted from this table.

Model	Layers	Latent dim	FF size	Trainable params
Diffusion	4	256	1,024	4.63M
	8	512	1,024	26.33M
	12	512	1,024	38.95M
	16	768	3,072	260.97M
Autoregressive	8	512	512	29.63M
	12	768	768	92.27M
	16	1,024	1,024	215.31M

Table D. **Ablation study on inter-hand interaction.** We report the inter-hand contact precision (C_{prec}) on diffusion model across different dataset scales and model depths.

Dataset Ratio	Layers	Inter-hand $C_{\text{prec}}^{\uparrow}$
0.05	4	0.7310
0.05	8	0.7381
0.05	12	0.7111
0.2	8	0.7066
0.2	12	0.7971
1.0	4	0.7551
1.0	8	0.7838
1.0	12	0.8593

coder and the hand motion encoder. Unlike traditional approaches that rely on classification-based training [21], we adopt a sequence-level contrastive learning objective based on the InfoNCE loss [40]. Given the text prompts $T = (T_L, T_R, T_I)$, we concatenate them with tokenizer-defined separator tokens to form the input sentence. The text encoder uses T5 [44] as its backbone.

E.3. User Study on Data Quality

To evaluate the quality of our data, we conduct a comprehensive user study focusing on two primary dimensions: **annotation quality** and **motion quality**. Regarding annotation quality, we compare our decoupled strategy against a baseline where Gemini 3 Pro [51] is prompted to directly

caption rendered videos. For motion quality, HandX is evaluated against GigaHands [19] and HoloAssist [55].

We recruit 20 participants for the study. Sequences are randomly permuted and distributed to participants. We ensure that each sequence receives ratings from at least three independent participants to maintain consistency. Following the question design shown in Figure Ec, participants rated the samples on a scale of 1 to 5, where higher scores indicate better quality. Figures Ea and Eb demonstrate that our approach significantly surpasses direct motion captioning in annotation quality and existing bimanual datasets in motion quality.

E.4. User Study on Scaling Trend

To complement the quantitative scaling trends reported in the main paper, we conduct a perceptual user study to evaluate the visual quality and semantic consistency of the generated motion across different data scales.

Experimental Setup. We randomly sample 10 distinct text prompts from the test set to ensure an unbiased evaluation of the model’s performance. For each prompt, we generate videos using the diffusion models trained on three different subsets of the training dataset: 5%, 20%, and 100%, consistent with the quantitative ablation settings. All motions are rendered as 3D meshes. Crucially, to ensure clear visibility of the fine-grained finger interactions and spatial relations, we unify and optimize the camera viewing angles for each sequence.

Procedure. We recruit 10 participants to evaluate the generation quality; all of them are graduate or undergraduate students without previous knowledge on hand motion generation. For each of the 10 prompts, participants are presented with the text description and three generated videos displayed in a randomized order to prevent bias. Participants are allowed to replay the videos an unlimited number of times to inspect motion details carefully. We instruct participants to consider both motion naturalness and semantic alignment when making their choice.

Results. The user study shows a clear preference for the model trained on the full dataset. Specifically, the model

trained on 100% of the data receives 48% of the votes, compared with 33% for the model trained on 5% of the data and 19% for the model trained on 20% of the data. This suggests that increasing the data scale leads to perceptually better motion quality and semantic alignment.

F. License

Users must review and follow the original licenses for each sub-dataset utilized in HandX. Please find the licenses of corresponding assets in the code directories, and below is a summary of the licenses for the assets we have used:

1. GigaHands [19] uses the Creative Commons Attribution-NonCommercial 4.0 International License.
2. HOT3D [5] uses the HOT3D Dataset License Agreement.
3. ARCTIC [14] uses the Data & Software Copyright License for non-commercial scientific research purposes.
4. H2O [26] uses custom Terms of Use restricted to academic and non-commercial purposes.
5. HoloAssist [55] uses the CDLA v2 Permissive License.

G. Discussion

Limitations. Despite the comprehensive benchmark established in this work, several limitations remain. First, although HandX significantly scales up bimanual motion data with fine-grained annotations, the dataset is still finite in both volume and diversity. Consequently, it cannot exhaustively cover the full spectrum of human dexterity or every possible interaction scenario found in the real world. Second, a portion of our training corpus is aggregated from existing public datasets. While we employ rigorous filtering and interpolation techniques to standardize such data, inherent quality issues in the raw sources, such as minor jitter or kinematic implausibility, cannot be completely eliminated.

Potential Negative Societal Impact. Our work focuses on generating realistic human hand motion, which has various positive applications. However, like other high-fidelity generative models, there is a risk of misuse in creating deep-fakes or misleading content. The ability to synthesize dexterous hand motion from text could be used to fabricate videos of individuals performing actions they did not carry out. To mitigate this, we release our code and data under licenses that restrict usage to research and non-commercial purposes. Additionally, we address privacy concerns regarding the subjects involved in our data collection. We receive participants consent, and ensure the released dataset is strictly limited to skeletal motion representations, excluding any personally identifiable features to preserve anonymity.

References

[1] Louis Abel, Vincent Colotte, and Slim Ouni. Towards interpretable co-speech gestures synthesis using stargate. In

- International Conference on Multimodal Interaction*, 2024. 2
- [2] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwa Oh. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, 2018. 2
- [3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, 2019. 2
- [4] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *CVPR*, 2024. 3
- [5] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *CVPR*, 2025. 3, 4, 8
- [6] Léore Bensabath, Mathis Petrovich, and Gül Varol. Text-driven 3d hand motion generation from sign language data. *CVPR*, 2026. 3, 4
- [7] Y. Bilge, R. Cinbis, and N. Ikizler-Cinbis. Towards zero-shot sign language recognition. *TPAMI*, 2023. 3
- [8] Yunus Can Bilge, Nazli Ikizler-Cinbis, and Ramazan Gokberk Cinbis. Zero-shot sign language recognition: Can textual data uncover sign languages? In *BMVC*, 2019. 3
- [9] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *CVPR*, 2024. 3, 4
- [10] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffshg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *CVPR*, 2024. 2
- [11] 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. 3
- [12] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *ECCV*, 2022. 3
- [13] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *ICCV*, 2025. 5, 6
- [14] 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 *CVPR*, 2023. 3, 4, 8
- [15] Zicong Fan, Maria Pirelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *CVPR*, 2024. 3
- [16] Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. Signllm: Sign language production large language models. In *ICCV*, 2025. 3
- [17] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. In *ACM SIG-*

- GRAPH Conference on Motion, Interaction and Games*, 2019. 2
- [18] Hongming Fu, Wenjia Wang, Xiaozhen Qiao, Shuo Yang, Zheng Liu, and Bo Zhao. Egograsp: World-space hand-object interaction estimation from egocentric videos. *arXiv preprint arXiv:2601.01050*, 2026. 3
- [19] 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 *CVPR*, 2025. 2, 3, 4, 7, 8
- [20] Ariel Gjaci, Carmine Tommaso Recchiuto, and Antonio Sgorbissa. Towards culture-aware co-speech gestures for social robots. *International Journal of Social Robotics*, 14(6): 1493–1506, 2022. 2
- [21] 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 *CVPR*, 2022. 2, 6, 7
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 5
- [23] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 3
- [24] Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, et al. Hoigt: Learning long-sequence hand-object interaction with language models. In *CVPR*, 2025. 3
- [25] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2023. 2
- [26] Taemin Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 2, 3, 4, 8
- [27] Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *CVPR*, 2025. 3
- [28] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *NeurIPS*, 2023. 3, 4
- [29] Pei Lin. Handdiffuse: generative controllers for two-hand interactions via diffusion models. In *AAAI*, 2025. 2, 3, 4
- [30] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *CVPR*, 2024. 2
- [31] Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *ICCV*, 2025. 2
- [32] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022. 3
- [33] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, 2022. 3
- [34] 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 *CVPR*, 2024. 3
- [35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 2
- [36] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Human-TOMATO: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 5
- [37] Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. In *CVPR*, 2025. 2, 5, 6
- [38] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020. 2, 3, 4
- [39] NaturalPoint, Inc. Optitrack motion capture system. <https://optitrack.com>, 2025. 1
- [40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 7
- [41] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 2
- [42] Mathis Petrovich, Michael J Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, 2023. 5
- [43] Aditya Prakash, David Forsyth, and Saurabh Gupta. Bimanual 3d hand motion and articulation forecasting in everyday images. *arXiv preprint arXiv:2510.06145*, 2025. 3
- [44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 5, 7
- [45] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 4
- [46] Khaled Saleh. Hybrid seq2seq architecture for 3d co-speech gesture generation. In *International Conference on Multimodal Interaction*, 2022. 2
- [47] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *ICLR*, 2024. 2, 4
- [48] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. *ACM Transactions on Graphics*, 43(6), 2024. 2

- [49] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 3
- [50] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using spatial cues and latent consistency. In *3DV*, 2024. 3
- [51] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7
- [52] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 5
- [53] Balamurugan Thambiraja, Omid Taheri, Radek Danecek, Giorgio Becherini, Gerard Pons-Moll, and Justus Thies. CLUTCH: Contextualized language model for unlocking text-conditioned hand motion modelling in the wild. In *ICLR*, 2026. 3, 4
- [54] Ruocheng Wang, Pei Xu, Haochen Shi, Elizabeth Schumann, and C Karen Liu. Fürelise: Capturing and physically synthesizing hand motion of piano performance. In *SIGGRAPH Asia*, 2024. 3
- [55] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *ICCV*, 2023. 3, 4, 7, 8
- [56] Ziyin Wang, Sirui Xu, Chuan Guo, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, and Liangyan Gui. Unleashing guidance without classifiers for human-object interaction animation. In *ICLR*, 2026. 2
- [57] Yilin Wen, Hao Pan, Takehiko Ohkawa, Lei Yang, Jia Pan, Yoichi Sato, Taku Komura, and Wenping Wang. Generative hierarchical temporal transformer for hand pose and action modeling. In *ECCV*, 2024. 2
- [58] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. *arXiv preprint arXiv:2503.15451*, 2025. 2
- [59] Pei Xu and Ruocheng Wang. Synchronize dual hands for physics-based dexterous guitar playing. In *SIGGRAPH Asia*, 2024. 3
- [60] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 2, 5
- [61] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *ICLR*, 2023. 2
- [62] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. InterDreamer: Zero-shot text to 3d dynamic human-object interaction. In *NeurIPS*, 2024. 2
- [63] Sirui Xu, Yu-Wei Chao, Liuyu Bian, Arsalan Mousavian, Yu-Xiong Wang, Liangyan Gui, and Wei Yang. Dexplore: Scalable neural control for dexterous manipulation from reference scoped exploration. In *CoRL*, 2025. 3
- [64] 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 *CVPR*, 2025. 2, 3, 4
- [65] Sirui Xu, Samuel Schuler, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liangyan Gui. InterPrior: Scaling generative control for physics-based human-object interactions. *arXiv preprint arXiv:2602.06035*, 2026. 2
- [66] Xiyan Xu, Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. MoReact: Generating reactive motion from textual descriptions. *arXiv preprint arXiv:2509.23911*, 2025. 2
- [67] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919*, 2023. 2
- [68] Payam Jome Yazdian, Eric Liu, Rachel Lagasse, Hamid Mohammadi, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv preprint arXiv:2312.12634*, 2023. 3
- [69] Yufei Ye, Jiaman Li, Ryan Rong, and C Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *arXiv preprint arXiv:2602.22209*, 2026. 3
- [70] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. DYNAMR: Recovering 4d interacting hand motion from a dynamic camera. In *CVPR*, 2025. 3
- [71] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 2
- [72] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *CVPR*, 2025. 3
- [73] Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *TPAMI*, 2025. 3
- [74] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiandiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2
- [75] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *CVPR*, 2023. 2
- [76] Wenqian Zhang, Molin Huang, Yuxuan Zhou, Juzhe Zhang, Jingyi Yu, Jingya Wang, and Lan Xu. Both2hands: Inferring 3d hands from both text prompts and body dynamics. In *CVPR*, 2024. 2, 3, 4
- [77] 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 *CVPR*, 2025. 3

- [78] Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. In *NeurIPS*, 2025. 3
- [79] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *ECCV*, 2022. 3
- [80] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *CVPR*, 2024. 3
- [81] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, 2023. 2
- [82] 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 *ECCV*, 2024. 2
- [83] Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *ICCV*, 2025. 3