

RoMo Hands: A Large Scale Richly Organized Text to Hand Motion Dataset

Yizhak Ben-Shabat¹ Jiahao Zhang^{1,2} Joseph Liu¹ Seonghyeon Moon¹
 Young-Yoon Lee¹ Haomiao Jiang¹ Oren Jacob¹ Mubbasir Kapadia^{1,3}
¹Roblox ²Australian National University ³Rutgers University
 sitzikbs@gmail.com jiahao.zhang@anu.edu.au

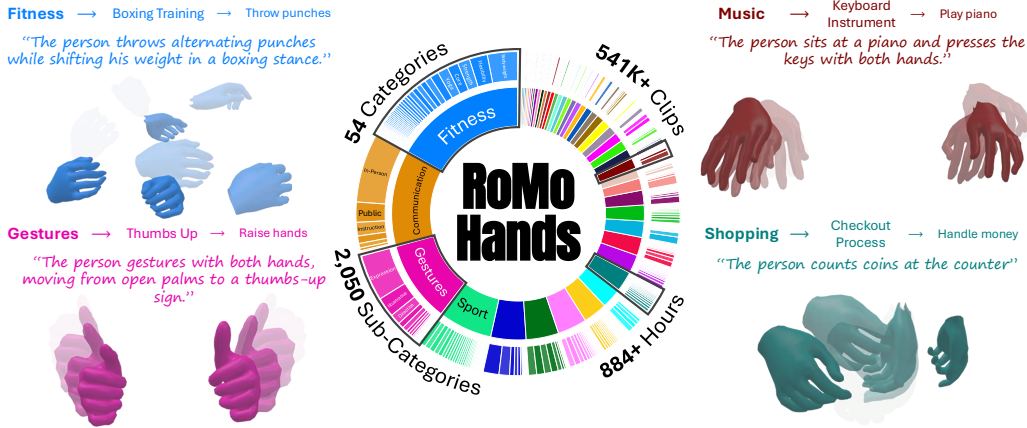


Figure 1: **RoMo Hands scales pure-hand text-to-motion data by an order of magnitude.** $\sim 542K$ in-the-wild bimanual sequences (884 hours, 95.5M frames at 30 fps), each paired with five caption granularities (L0–L4) and a three-level action taxonomy. The dataset enables a benchmark of state-of-the-art diffusion, masked-token, and autoregressive generators adapted to a shared 393-D bimanual hand representation. Interactive dataset statistics, side-by-side benchmark comparisons, dataset explorer, and a 3D motion-grid visualizer are best browsed in the supplementary materials.

Abstract

While text-driven body motion generation has matured, pure-hand synthesis remains bottlenecked by data. Existing datasets often treat hands as rigid end-effectors, lack fine-grained captions, or are too limited in scale to support robust generalization. We introduce **RoMo Hands**, an in-the-wild, text-paired dataset of $\sim 542K$ sequences (884 hours, 95.5M frames at 30 fps) which is an order of magnitude larger than prior work. To ensure high-quality supervision, a stringent dexterity filter discards low-articulation sequences, focusing the distribution on complex bimanual movement. Each clip is paired with five captions at increasing granularities from global semantic tags to atomic finger states, organized under a hierarchical taxonomy. To bridge the gap between body and hand architectures, we propose a novel 393-D bimanual representation that adapts redundant body-motion paradigms to the high-DoF kinematics of hands. Systematic benchmarking of diffusion, masked-token, and autoregressive generators reveals distinct trade-offs under these hand-specific adaptations, including a sharp divergence between semantic alignment and kinematic plausibility. Our dataset, representation, and benchmark establish a rigorous foundation for the next generation of dexterous motion synthesis. Project page: <https://romodataset.github.io/romo-hands/>.

1 Introduction

Text-driven human motion synthesis has matured rapidly in recent years, driven by large-scale datasets paired with natural-language captions that enable models to learn rich semantic correspondences between text and motion [7, 12, 35, 4]. These advances have substantially improved the realism, controllability, and generalization of body-motion generation systems. *Pure hand* motion, however, has not progressed at the same pace. Yet, hands are the primary channel through which humans manipulate the world and convey nuanced intent. High-fidelity finger articulation is therefore essential for applications such as AR/VR avatars, teleoperation, sign-language synthesis, embodied agents, and dexterous robotics. From a kinematic perspective, hands are uniquely challenging: they pack extremely high degrees of freedom into a compact spatial region and exhibit frequent self-contact together with tightly coordinated inter-finger motion. Existing body-centric generators were not designed for this precision-intensive regime, and the lack of sufficiently large text-paired hand datasets has limited the community’s ability to systematically adapt modern generative architectures to dexterous hand motion.

Table 1 and Figure 3 illustrate the sharp trade-off between scale and hand-motion fidelity in current datasets. Large-scale body-centric datasets [12, 38, 35, 4, 40] provide extensive motion diversity and textual supervision, but typically treat hands as rigid end-effectors with generic captions. In contrast, hand-object interaction and egocentric datasets [6, 1, 5, 11, 39] capture accurate contact dynamics and realistic manipulation behavior, yet rarely exceed a few hours or include free-form language annotations. Existing pure-hand datasets [18, 13] similarly lack text supervision, limiting their usefulness for text-conditioned generation. More recent text-paired efforts such as BOTH2Hands [44], HandX [45], and the concurrent CLUTCH [31] begin to address this gap, but remain orders of magnitude smaller than modern body-motion corpora. Compounding this challenge, hand-motion generation demands substantially higher fidelity than full-body synthesis: subtle artifacts such as inter-finger penetration or implausible coordination are immediately perceptible, yet are rarely captured by metrics designed for coarse body dynamics.

To address this bottleneck, we introduce **RoMo Hands**: a pure-hand, text-paired dataset containing approximately 542K in-the-wild motion sequences, corresponding to 884 hours and 95.5M frames at 30 fps, making it more than an order of magnitude larger than the next-largest pure-hand dataset. To ensure high-quality supervision, we design a stringent automated filtering pipeline that removes near-static and low-information sequences, focusing the retained distribution on genuinely dexterous hand activity. Each clip is paired with five fine-grained captions at increasing levels of detail, describing atomic finger states, temporal dynamics, and spatial relations within a hierarchical action taxonomy.

Beyond the dataset itself, we establish a comprehensive benchmark for pure-hand motion generation. Because leading generative architectures were originally developed for coarse-grained body motion, directly applying them to hands introduces substantial representational challenges. To bridge this gap, we propose a novel redundant bimanual representation, $\mathcal{R}_{\text{hand}}$, specifically designed for dexterous hand synthesis. Using this representation, we systematically adapt diffusion, masked-token, and autoregressive generation paradigms within a unified benchmark setting. Under these evaluated adaptations, semantic and physical metrics yield different model rankings: strong text alignment does not guarantee kinematic plausibility or fine-grained physical consistency. These findings suggest that pure-hand generation introduces challenges beyond those typically encountered in conventional body-motion synthesis. Our core contributions are:

- **RoMo Hands Dataset.** A large-scale, in-the-wild pure-hand motion dataset, an order of magnitude beyond existing pure-hand sets.
- **Fine-Grained Captions.** Five hierarchical captions per clip detailing finger articulation, temporal dynamics, and spatial relations, indexed under a pruned RoMo [40] action taxonomy for stratified analysis.
- **Bimanual Motion Representation.** A 393-dimensional hand representation ($\mathcal{R}_{\text{hand}}$) that adapts the redundant body-motion paradigm to hands via a synthetic root and inter-wrist heading.
- **Pure-Hand Benchmark.** A comprehensive evaluation of diffusion, masked-token, and autoregressive architectures under documented hand-specific adaptations, exposing a tension between semantic and physical quality.

Table 1: **RoMo Hands provides an order-of-magnitude leap in hand-motion scale.** It surpasses prior pure-hand datasets by $10\times$ in duration and text volume.

Dataset	Scope	Setting	Duration (h)	Frames (M)	Granularity	#Captions (K)
GRAB [29]	Hands + Object	Studio	3.8	1.6	—	—
OakInk2 [39]	Hands + Object	Studio	37.1	4.01	Action	—
ARCTIC [5]	Hands + Object	Studio	2.02	0.22	Action	—
H2O [11]	Hands + Object	Studio	1.06	0.11	Action	—
HOT3D [1]	Hands + Object	Studio	3.90	0.42	—	—
GigaHands [6]	Hands + Object	Egocentric	34.0	3.7	Coarse	84.0
HoloAssist [34]	Hands + Object	Egocentric	161.2	17.4	Coarse	1.8
HandDiffuse [13]	Pure Hands	Studio	2.0	0.25	—	—
BOTH2Hands [44]	Pure Hands	Studio	8.31	1.8	Coarse	23.5
InterHand2.6M [18]	Pure Hands	Studio	24.0	2.6	—	—
HandX [45]	Pure Hands	Studio	54.2	5.9	Fine-grained	485.7
3D-HIW [31]	Pure Hands	In-the-wild	83.3	12.0	Fine-grained	32.0
RoMo Hands (Ours)	Pure Hands	In-the-wild	884	95.5	Fine-grained	2710

2 Related Work

We categorize prior work along three axes: (1) body-motion scaling paradigms; (2) hand-centric datasets and their scale constraints; and (3) generative architectures adapted to the bimanual domain.

2.1 Large-Scale Body Motion Datasets

Text-to-motion research has transitioned from studio captures like AMASS [16] and HumanML3D [7] to massive, in-the-wild datasets such as Motion-X [12], MotionLib [35], and MotionMillion [4]. While these efforts scale body motion by orders of magnitude, they prioritize gross locomotion and treat hands as rigid end-effectors [12, 4, 40]. RoMo Hands imports the successful scaling recipe of RoMo [40], leveraging in-the-wild video and automated filtering to introduce the first body-scale, text-paired dataset devoted strictly to dexterous articulation.

2.2 Hand and Hand-Object Interaction Datasets

A disjoint line of work focuses on hands (Table 1). Studio mocap datasets like InterHand2.6M [18], GRAB [29], ARCTIC [5], and OakInk2 [39] provide precise MANO [25] parameters but span only a few hours and rarely include free-form text. Egocentric captures like GigaHands [6] and HoloAssist [34] broaden behavioral coverage but produce coarse annotations targeted at assistive agents rather than high-fidelity synthesis.

Pure-hand text-paired datasets are nascent but have made crucial strides. BOTH2Hands [44] records 8.31 hours of motion, while HandX [45] unifies datasets via LLM-based annotation (54.2 hours). Concurrent to our work, CLUTCH [31] introduces an in-the-wild pure-hand dataset of ~ 83 hours. These efforts establish the foundation for text-to-hand generation, yet they remain far below the scale required for the data-hungry architectures dominating the body domain. RoMo Hands provides this missing scaling engine with 884 hours, $\sim 542K$ sequences, and dense five-level captioning.

2.3 Text-Conditioned Motion Generation

Generation has converged on three paradigms. *Diffusion* models (e.g., MotionDiffuse [42], MDM [30], PriorMDM [26], Kimodo [24]) iteratively denoise continuous sequences and lead the high-fidelity regime. *Autoregressive* models (e.g., T2M-GPT [41], MotionMillion [4]) predict discrete tokens over VQ-VAE or FSQ [17] codebooks. Finally, *masked-token* models like MoMask [8] and SnapMoGen [9] utilize bidirectional transformers for joint token generation.

Multiple prior works target text-to-HOI [2, 3, 46], however, they rely on fixed object assets. For pure hands, CLUTCH [31] represents a recent end-to-end system. However, it remains unclear how other generative architectures behave on the unique, high-DoF kinematics of hands. Rather than reinventing architectures, we ask how the existing models can be adapted. By bridging them with our novel representation ($\mathcal{R}_{\text{hand}}$) and benchmarking on RoMo Hands, we isolate which body-centric design choices transfer to dexterous articulation and which require non-trivial adaptation.

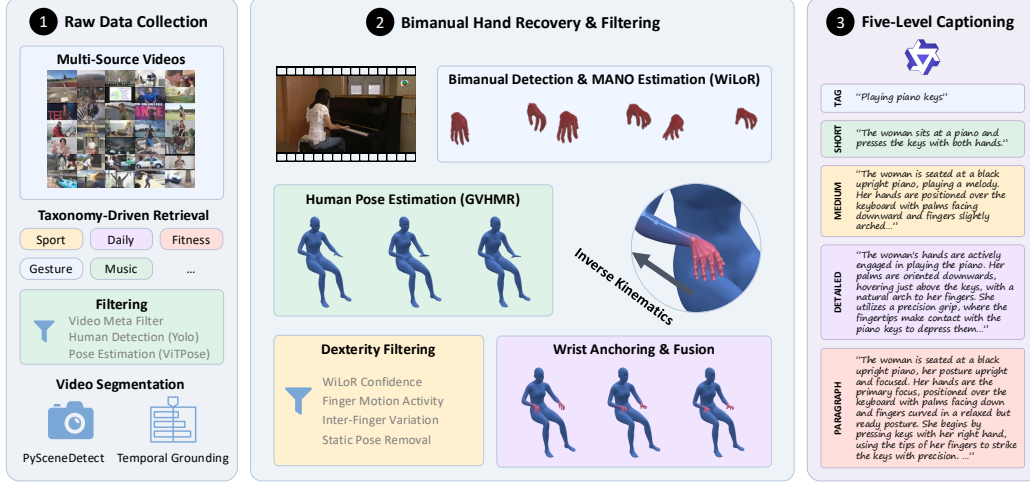


Figure 2: **The RoMo Hands curation pipeline.** It extends the RoMo body-motion stack with three hand-specific stages: hand detection and bimanual MANO pose recovery (WiLoR), a dexterity filter (detection-ratio + finger-articulation thresholds), and a five-level VLM captioner (Qwen3.5).

3 The RoMo Hands Dataset and Representation

3.1 Collection and Captioning

The RoMo Hands collection pipeline. RoMo Hands is built upon the data-collection and curation pipeline introduced by RoMo [40], which we rigorously repurpose for fine-grained pure-hand motion. The RoMo pipeline begins with a multi-source web-video collection assembled via taxonomy-driven queries. They apply metadata filtering, scene detection, and single-human filtering, followed by semantic temporal segmentation against a three-level action taxonomy. They then estimate 3D body motion in SMPL format for every retained clip using GVHMR [27], caption each segment using a vision-language model, and finally prune the dataset with a category-adaptive dynamic-score filter that removes static or low-activity sequences.

Inheriting RoMo’s video pool, we extend this pipeline along three axes specifically targeted at hand motion failure modes. First, we introduce a bimanual hand-pose stage that recovers per-frame MANO parameters using WiLoR [21] and reconciles them against the inherited wrist trajectory. Second, we enforce a strict visibility threshold, discarding sequences where the mean per-frame WiLoR detection confidence falls below a required per-hand minimum, ensuring the retained clips contain consistently visible hands rather than heavily occluded or off-screen artifacts. Third, we utilize Qwen3.5-27B [22] to caption each segment with five text descriptions at increasing levels of specificity, prompting the LLM to emphasize object manipulation context, temporal hand dynamics, and inter-finger spatial relationships. Figure 2 summarizes the end-to-end stack.

Bimanual hand recovery. We recover hand motion using WiLoR [21], a robust monocular bimanual hand estimator that performs per-frame hand detection and MANO [25] parameter regression in a single pass. Each frame yields up to two hand instances, carrying a left/right label, axis-angle wrist orientation, 15 finger-joint rotations, weak-perspective camera parameters $\{t_{cam}, s_{cam}\}$, and a detection score cached for downstream filtering.

Because WiLoR lacks a temporal component and assumes a weak-perspective camera, per-frame wrist position and orientation suffer from frame-by-frame ambiguity: absolute depth is entangled with hand scale, and wrist orientation lacks a shared anchor across frames. We resolve both issues via a per-frame *wrist anchoring* step that reconciles WiLoR’s wrist pose against the temporally consistent SMPL [14] wrist pose recovered by RoMo’s body pipeline. When the two estimates agree within a configurable threshold, we average them. Otherwise, we trust WiLoR’s wrist pose and solve inverse kinematics (IK) on the body’s arm chain to bring the SMPL wrist into agreement, locking both hands into the body’s temporally consistent world frame. For isolated frames where WiLoR fails to detect a hand, we linearly interpolate the MANO parameters in axis-angle space between

the nearest detections. This interpolation is reliable only for short gaps (up to 5 frames); longer gaps trigger a segment split or discard. In aggregate, the hand-specific pipeline (bimanual recovery, plus the visibility and dexterity filters described next) retains ~ 884 h of the $\sim 1,238$ h inherited from RoMo [40], a $\sim 29\%$ reduction.

Detection-ratio and dexterity filtering. Even after segmentation, two failure modes dominate the raw distribution: (a) clips where hands are barely visible due to occlusion or motion blur, and (b) clips where hands are visible but remain near a static rest pose (e.g., resting at the side or statically holding an object). Training a generator on these modes degrades quality by injecting regression noise or biasing the model toward the mean hand pose.

We address this with two complementary filters. First, a *detection-ratio filter* discards clips if the fraction of high-confidence WiLoR frames falls below a per-hand minimum. Second, a *dexterity filter* discards clips failing a threshold that combines finger-articulation activity (mean joint-rotation velocity over the 15 finger joints) and inter-finger pose variation (mean pairwise ℓ_2 distance between fingertip positions over time). Crucially, both filters run *before* captioning to preserve VLM budget and bias the final distribution toward dexterous, well-observed motion.

Captioning at five granularities. Downstream tasks consume hand motion at varying detail levels as retrieval benefits from short headlines, and generation from paragraph-length descriptions. We therefore caption each clip at five progressive granularities via a single structured prompt to Qwen3.5-27B [22], letting downstream models select supervision level at training time and enabling retrieval evaluation as a function of caption complexity:

- **L0 (headline):** A 2–4 word headline naming the action and salient object, e.g., “hold volleyball”.
- **L1 (one-sentence scenario):** A short sentence placing the action in body and scene context, e.g., “The woman holds a volleyball in her left hand while standing on a court.”
- **L2 (multi-sentence description):** A balanced description covering the active hand pose and the role of the passive hand.
- **L3 (half-paragraph description):** Adds finger-level articulation (palm orientation, finger curl) and brief temporal dynamics.
- **L4 (full-paragraph description):** The most verbose level, adding environmental context, fine wrist-and-finger detail, and inferred intent.

Human validation. We audit caption fidelity and recovered-pose plausibility on 250 randomly sampled validation clips (50 per caption level). Twelve adult raters (ages 20–50) with general technical backgrounds (at least a bachelor’s degree), most without animation experience, viewed each source video beside its 3D reconstruction and assigned five-point Likert scores for overall action match, hand-specific caption–motion faithfulness, and reconstruction quality. Participation was uncompensated. The audit collected only technical judgments of the displayed artifacts and no personal or sensitive data; it did not constitute human-subjects research. Across 370 ratings, action match averages 4.16/5 (81% at least 4), caption–motion faithfulness averages 3.79/5 (66% at least 4), and reconstruction quality averages 3.53/5. The per-level breakdown in Table 11 shows that hand-specific caption accuracy remains within 3.70–3.96 across the five levels.

The resulting vocabulary is analyzed in Section 3.2 and visualized in the right panel of Figure 4.

Action taxonomy. We adopt RoMo’s three-level coarse-to-fine taxonomy (*category* $\rightarrow$ *sub-category* $\rightarrow$ *atomic action*) [40] and prune categories emptied by our visibility and dexterity filters; no new categories are introduced. We use the pruned taxonomy as a stratification axis for evaluation, surfacing per-category model behavior that aggregate metrics hide (Figure 5).

3.2 Dataset Statistics and Coverage

RoMo Hands totals ~ 542 K bimanual hand-motion sequences (884 hours, 95.5M frames at 30 fps), with a mean clip duration of 5.9 s and an 85%/10%/5% train/val/test split at the segment level. The left panel of Figure 4 situates this scale against prior pure-hand and hand–object datasets, placing RoMo Hands at $10.6\times$ the duration of CLUTCH and $5.6\times$ the caption count of HandX. To anchor activity coverage, we score GRAB [29] (1.1K seq., 2.86 h) and GigaHands [6] (12.7K seq., 16.40 h) under

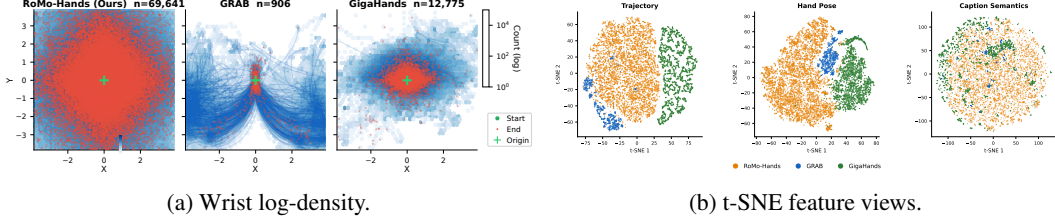


Figure 3: **RoMo Hands provides broader spatial, kinematic, and linguistic coverage than prior bimanual datasets.** (a) Wrist log-density on the workspace plane. (b) t-SNE in trajectory, hand-pose, and caption-feature spaces, where RoMo Hands spans a wider manifold than GRAB and GigaHands.

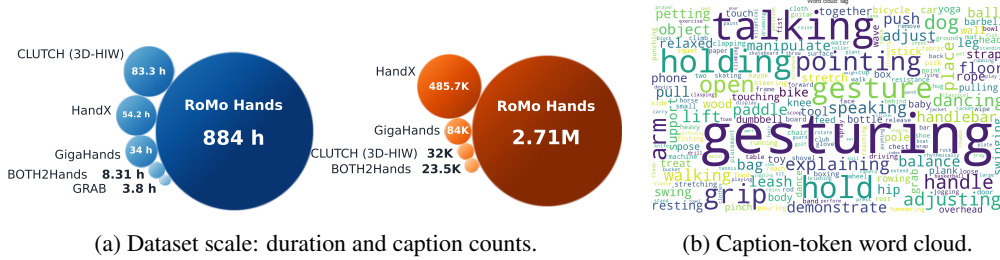


Figure 4: **RoMo Hands surpasses prior hand-motion datasets in both scale and caption specificity.** (left) bubble areas track duration and caption count, with RoMo Hands an order of magnitude beyond the largest prior pure-hand dataset on both axes; GRAB is text-free and omitted from the right bubble. (right) caption-token word cloud, dominated by finger-level verbs (*grip, pinch, flex*) and concrete interaction targets (*phone, cup*) rather than body-motion verbs. Per-finger and per-joint dynamic scores, smoothness, autocorrelation, per-tier caption statistics, and side-by-side breakdowns against GRAB and GigaHands are available in the supplementary materials.

the same dynamic-score function used during curation: RoMo Hands lands at 0.2% static / 53.7% active, versus 20.4%/14.2% for GigaHands and 0.0%/65.6% for GRAB, with a mean dynamic score of 0.33 that sits between GRAB’s choreographed 0.66 and GigaHands’ uncured 0.14.

Figure 3 further characterizes spatial, kinematic, and linguistic coverage: wrist XY log-density (left) shows RoMo Hands filling a substantially larger workspace, and t-SNE embeddings in three feature spaces (wrist-trajectory, hand-pose, caption-semantic) reveal a wider manifold along all three axes. The token-frequency view in Figure 4 (right) is dominated by finger-level verbs (*grip, pinch, flex*) and concrete interaction targets (*phone, cup*) rather than body-motion verbs. Per-finger and per-joint dynamic scores, smoothness, autocorrelation, per-tier caption statistics, and side-by-side breakdowns against GRAB and GigaHands are available in the supplementary materials.

3.3 Bimanual Hand Motion Representation

A useful motion representation must faithfully encode skeletal kinematics while plugging cleanly into generative architectures. For body motion, standard templates utilize redundant joint-aware per-frame channels (position, velocity, and rotation) to improve the training signal across diffusion and autoregressive models. The 263-D HumanML3D representation [7] and the 272-D representation introduced by MotionStreamer [37] (subsequently adopted by MotionMillion [4], with public reference implementation by Xiao [36]) rely on a low-dimensional heading frame followed by per-joint blocks. However, no analogous template has emerged for bimanual hand motion. Existing hand generators like CLUTCH [31] train on a 198-dimensional MANO-derived representation that lacks the redundant velocity and heading-canonicalized channels that make body templates effective.

We propose $\mathcal{R}_{\text{hand}} \in \mathbb{R}^{393}$, which extends MANO into a redundant per-frame template. We wrap the parametric skeleton in explicit position, velocity, and heading-canonicalized rotation channels, specialized to MANO’s 32-joint bimanual topology (16 joints per hand). $\mathcal{R}_{\text{hand}}$ stacks four per-frame channel groups:

$$\mathbf{h}_t = \left[\underbrace{\mathbf{f}_t}_{9, \text{ root frame}} , \underbrace{\{\tilde{\mathbf{p}}_t^j\}_{j=0}^{31}}_{96, \text{ joint positions}} , \underbrace{\{\dot{\mathbf{p}}_t^j\}_{j=0}^{31}}_{96, \text{ joint velocities}} , \underbrace{\{\phi(\mathbf{R}_t^j)\}_{j=0}^{31}}_{192, \text{ 6D rotations}} \right] \in \mathbb{R}^{393}, \quad (1)$$

Table 2: **The redundant $\mathcal{R}_{\text{hand}}$ representation improves text alignment across two generators.** Each block holds the architecture, data, optimizer, training budget, text encoder, evaluator, and normalization fixed while changing only the motion representation. Best within each model block is **bold**.

Generator	Representation	R@1 $\uparrow$	R@3 $\uparrow$	FID $\downarrow$	MM-Dist $\downarrow$
MoMask	$\mathcal{R}_{\text{hand}}\text{-393}$	0.547	0.794	0.153	0.941
	MANO-102	0.395	0.658	0.221	1.049
	CLUTCH-198	0.368	0.617	0.269	1.077
Kimodo M	$\mathcal{R}_{\text{hand}}\text{-393}$	0.665	0.871	0.114	0.868
	MANO-102	0.646	0.868	0.135	0.896
	CLUTCH-198	0.646	0.856	0.106	0.882

where $\mathbf{f}_t$ is the 9D root frame defined below, $\tilde{\mathbf{p}}_t^j$ are joint positions whose xz components are root-relative in the heading frame while the y component is the absolute world height, $\dot{\tilde{\mathbf{p}}}_t^j$ are their first-order finite differences, and $\phi : SO(3) \rightarrow \mathbb{R}^6$ encodes joint rotations via 6D continuous parameterization [47]. The 32 joints follow MANO’s bimanual indexing: joints 0–15 and 16–31 for the left and right hands respectively, with wrists at $j = 0$ and $j = 16$.

Because the bimanual hand has no skeletal joint at its kinematic center, $\mathcal{R}_{\text{hand}}$ introduces two bimanual-specific design choices: **(i) a synthetic root** which is the wrist midpoint $\mathbf{r}_t = \frac{1}{2}(\mathbf{p}_t^L + \mathbf{p}_t^R)$, stored explicitly in $\mathbf{f}_t$ as 2 channels of planar velocity in the heading frame, 1 channel of world- y height, and 6 channels of heading rotation, with the two actual wrist joints ($j \in \{0, 16\}$) retained in the per-joint blocks alongside the fingers; and **(ii) an inter-wrist heading** which is a yaw rotation $\mathbf{R}_y(\psi_t)$ about the world-up axis that aligns the inter-wrist segment with the world x -axis (where $\psi_t = \text{atan2}(-(\mathbf{p}_t^R - \mathbf{p}_t^L)_z, (\mathbf{p}_t^R - \mathbf{p}_t^L)_x)$), applied to per-joint xz positions and the two wrist rotations only, while finger rotations remain parent-relative within each MANO chain. Restricting canonicalization to world-up yaw preserves absolute world height, which carries critical semantic context for two-handed gestures.

Because several channels in $\mathcal{R}_{\text{hand}}$ have near-zero variance, we replace standard per-channel z-score normalization with a type-aware scheme (detailed in the supplementary material). Ultimately, this four-block layout, synthetic root, and normalization scheme yield a robust bimanual representation that slots natively into modern text-to-motion architectures.

Representation ablation. Table 2 isolates the representation while holding each generator’s data, architecture, optimization, text encoder, and frozen evaluator fixed. For MoMask, $\mathcal{R}_{\text{hand}}$ improves every reported metric over raw MANO and CLUTCH. Kimodo M is less representation-sensitive: $\mathcal{R}_{\text{hand}}$ leads R@1, R@3, MM-Dist, and CLIP, while CLUTCH has the best FID. Thus the redundant position, velocity, and heading channels provide a useful generation-side inductive bias across two distinct architectures, although the magnitude and metric-wise effect are architecture dependent. Within each generator all variants share the same data, architecture, and training budget, so differences reflect the representation; the Kimodo M models were retrained for this ablation and are not directly comparable to Table 3. Full results are in Table 12.

4 Experiments

The goal of this benchmark is to evaluate how state-of-the-art text-to-motion architectures that were originally designed for body motion generalize to pure hand motion when trained on a large-scale, text-paired bimanual dataset. We adapt one representative model from each major generative family (diffusion, masked-token, and autoregressive) to our shared $\mathcal{R}_{\text{hand}}$ representation, train them on the RoMo Hands split, and score them under matched semantic and physical metrics.

4.1 Evaluation Protocol

We adopt the two-tier text-to-motion evaluation protocol of HumanML3D [7] and SnapMoGen [9], which was later adapted by CLUTCH [31] for bimanual hands. This protocol comprises *semantic* metrics scored by a learned text–motion evaluator, and *physical* metrics computed directly on the

generated MANO meshes. Full metric definitions and hyperparameters for the TMR [20] evaluator are provided in the supplementary material.

Semantic metrics. Following SnapMoGen [9], we utilize a TMR evaluator trained on the RoMo Hands dataset. To ensure the evaluator scores articulation rather than spatial reconstruction, which would otherwise disproportionately reward VQ-style methods, we strip positions and velocities before normalization. We report: FID [10], measuring the distributional distance between generated and ground-truth (GT) motions; R-Precision ($R@k$) [7] for $k \in \{1, 3, 10\}$, measuring the retrieval accuracy of the correct caption from a pool of 32 captions (1 correct, 31 distractors drawn uniformly from the test set); MM-Dist, the average feature distance between a generated motion and its paired text; CLIP score, the cosine similarity between a generated motion and its paired text in the same TMR joint embedding space; Diversity, the variance across the generated set; and MultiModality, the variance of outputs generated from a single repeated prompt.

Physical metrics. To assess kinematic plausibility, we compute the following directly on the generated MANO meshes: Jerk ($\downarrow$), which quantifies the temporal smoothness of the motion; Inter-hand penetration ($\downarrow$), the percentage of frames where the left and right hand meshes intersect; Maximum self-hand penetration ($\downarrow$), the worst-case collision depth within a single hand (a depth-valued variant of the BimArt [43] self-penetration metric; see the supplementary material); and HF Power Ratio ($\downarrow$), the fraction of velocity energy residing in the high-frequency band. The HF Power Ratio is speed-normalized, isolating high-frequency jitter from the overall motion velocity to complement the absolute Jerk metric.

Test-set protocol. For each generator, we sample one motion per test prompt on the 17,348 RoMo Hands test clips of 40–196 frames, the training length range. Sequence lengths are cropped to the nearest multiple of 4 to accommodate VQ tokenizer temporal downsampling. We apply the same type-aware normalization used during training, extract the essential features, and embed them via the frozen TMR evaluator. During each replication, we draw the conditioning caption uniformly at random across the five granularity levels (L0–L4). Consequently, reported numbers are means marginalized over both replication noise and caption granularity. A per-level breakdown is available in the supplementary materials. Ground-truth motions are processed through the identical pipeline to establish the reference distribution. Because the split is segment-level, we also remove every test clip whose source video appears in train and recompute the benchmark. The four-model ranking is preserved, with $R@1$ shifts between -0.026 and $+0.009$; Table 13 reports the full audit. We will additionally release a strict source-video-disjoint partition with the dataset.

Implementation details. We benchmark Kimodo M [24] and MDM [30] (diffusion), MoMask [8] (residual-VQ masked-token), and MotionMillion [4] (FSQ-tokenized autoregressive). All four were originally proposed for full-body motion; we retrain each on RoMo Hands under the shared $\mathcal{R}_{\text{hand}}$ representation, a frozen T5-v1.1-base text encoder [23], AdamW [15] optimization, and type-aware normalization, so any cross-architecture difference reflects representational fit rather than conditioning. Kimodo M is held at its published medium configuration (the only change is the input/output dimension from SMPL features to $\mathcal{R}_{\text{hand}}$); full per-model hyperparameters and schedules are in the supplementary materials.

4.2 Results

We report semantic and physical results on the RoMo Hands test split in Table 3. Under the documented adaptations, the two metric families yield different rankings: MoMask and Kimodo M lead the semantic metrics, while no single model universally dominates the physical metrics. Because MDM is prefix-conditioned, MotionMillion is evaluated at 150M rather than its published 1B scale, and MoMask Stage 2 had not fully converged, per-model rankings should be read as indicative rather than definitive.

Semantic results. MoMask (masked-token) and Kimodo M (diffusion) are statistically tied at the top of the semantic rankings: MoMask slightly edges out R-Precision retrieval, while Kimodo M leads on distribution-matching scores (FID, MM-Dist). MotionMillion and MDM (DiP) trail under their evaluated configurations; the latter uses a prefix-completion variant with a restricted 60-frame horizon.

Table 3: **No single generator dominates both metric families.** Semantic (left) and physical (right) results on the RoMo Hands test split. MoMask narrowly leads R-Precision retrieval, while Kimodo M leads the distribution-matching scores and most physical axes. “→” marks metrics whose target is to *match* GT (Diversity, Acceleration, Intra-hand Jerk). The italicized Ground Truth row is the reference distribution.

Method	Semantic						Physical				
	R@1 ↑	R@3 ↑	R@10 ↑	MM-Dist ↓	FID ↓	Div. →	Accel → (m/s ²)	Intra-Jerk → (m/s ³)	Inter-Pen ↓ (%)	Self-Pen ↓ (mm)	HF Pow. ↓ (ratio)
<i>Ground Truth</i>	<i>63.77</i>	<i>82.59</i>	<i>96.13</i>	<i>0.864</i>	<i>0.000</i>	<i>120.41</i>	<i>8.38</i>	<i>14.3</i>	<i>0.64</i>	<i>0.002</i>	<i>0.0081</i>
MDM (DiP) [30]	42.28	66.15	88.92	1.031	0.144	113.44	7.92	71.2	2.19	0.382	0.0132
MoMask [8]	54.00	78.62	93.88	0.945	0.129	115.27	4.99	265.2	1.58	0.017	0.0153
Kimodo M [24]	53.82	76.99	93.76	0.942	0.087	116.34	6.03	22.9	1.06	0.066	0.0077
MotionMillion [4]	31.26	51.55	82.37	1.105	0.314	109.14	5.77	76.1	1.82	0.000	0.0082

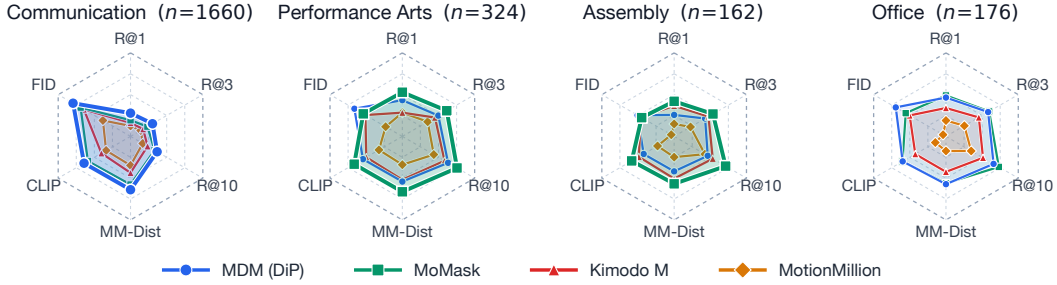


Figure 5: **Taxonomy uncovers per-category performance differences that are hidden in aggregate.** No single generator wins across the taxonomy: the leader rotates both *across* categories and, in *Office*, *within* a single panel across the six semantic axes (R@1, R@3, R@10, MM-Dist, CLIP, FID), normalized to $[0, 1]$. The supplementary covers all 40 categories of the pruned taxonomy.

Physical results. The physical landscape inverts the semantic ranking. Kimodo M is the cleanest overall with the lowest inter-hand penetration, the lowest HF power, and the best intra-hand jerk among baselines. MotionMillion incurs zero detected self-penetration and tracks GT HF power most closely, consistent with temporal regularization from its FSQ codebook. MoMask shows a counterintuitive pattern: its acceleration sits well *below* GT yet its intra-hand jerk reaches $18\times$ GT, a slow-but-jagged signature consistent with abrupt residual-VQ token transitions that semantic metrics do not expose. MDM (DiP) has the highest contact violation rate under its evaluated 60-frame, 10-step configuration.

This split between metric families is a primary finding of the benchmark: *no single evaluated configuration dominates both*. The preferred configuration depends on whether a downstream application prioritizes semantic text alignment or physical plausibility. This direction-level finding does not depend on treating the per-model ranking as definitive: pure-hand generation cannot be evaluated on semantic retrieval alone and requires complementary physical evaluation. The interactive statistics dashboard, side-by-side animated benchmark comparisons of every model on matched test prompts, and the 3D motion-grid visualizer are all available in the supplementary materials. We encourage the reader to inspect the comparisons in particular, since over-smoothing and high-frequency jitter are perceptually obvious in motion but only summarized in numbers here.

Taxonomy-stratified results. Figure 5 shows the semantic metrics over the taxonomy categories and reveals blind spots that the aggregate Table 3 hides. The leader rotates across categories: MDM (DiP) sweeps the *Communication* category that has $n = 1660$ samples; MoMask leads *Performance Arts* ($n=324$) and *Assembly* ($n=162$) on five of the six axes while MDM keeps the FID edge; and within *Office* ($n=176$) the rotation persists across metric families, with MoMask winning the retrieval cluster (R@1, R@3, R@10, CLIP) and MDM winning the distribution-matching axes (MM-Dist, FID). MotionMillion does not lead any category by mean, consistent with its aggregate performance. The full breakdown over all 40 categories, with semantic and physical-metric stratification, is available in the supplementary materials.

5 Conclusion

RoMo Hands is an expansive in-the-wild dataset comprising $\sim 542\text{K}$ text-paired sequences across 884 hours. By pairing this scale with a five-level hierarchical captioning scheme and a coarse-to-fine taxonomy, we address the critical data bottleneck in dexterous motion synthesis. Our benchmark, underpinned by a novel 393-D bimanual representation ($\mathcal{R}_{\text{hand}}$), shows that body-motion generators exhibit distinct trade-offs under hand-specific adaptations. In particular, semantic and physical rankings diverge, and representation fit affects two different generator families. By releasing the dataset, taxonomy, and trained evaluator, we provide a robust foundation for the next generation of text-to-motion research.

Limitations. Sourcing in-the-wild video introduces residual noise from occlusions and motion blur, which we mitigate via filtering and evaluate through a 250-clip human audit. The audit supports average caption fidelity and pose plausibility, but its five-point ratings are coarse and do not replace reconstruction error against ground-truth 3D hand pose. The dexterity filter introduces a selection bias toward well-observed, active motion rather than the raw web-video distribution, and the dataset inherits the demographic and geographic composition of its public source pool. Retrieval metrics use a single TMR evaluator trained on RoMo Hands, so they support relative comparisons under this shared evaluator rather than absolute comparison to body-motion benchmarks. Finally, the benchmark uses documented but non-identical adaptations—prefix conditioning for MDM, a 150M MotionMillion decoder, and a still-improving MoMask Stage 2—so architecture-specific rankings are indicative rather than definitive.

Future Work. Promising directions include scaling autoregressive decoders, such as pushing the MotionMillion [4] decoder to 1B+ parameters, as well as exploring cross-dataset transfer to complementary benchmarks such as HandX [45] and CLUTCH [31]. We also aim to integrate rich latent object context for hand-object interaction synthesis. Ultimately, this work paves the way for a unified multimodal generator capable of jointly synthesizing body, hands, and facial expressions under a shared tokenizer.

Broader Impact. RoMo Hands accelerates research in AR/VR avatars, telepresence, and robotics. To mitigate misuse, we release strictly de-identified motion features and text annotations. However, as generative hand models mature, they could theoretically be combined with synthesized visuals to create misleading gestural content. We urge the community to employ responsible access controls and usage policies when deploying downstream applications.

References

- [1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. HOT3D: Hand and object tracking in 3d from egocentric multi-view videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, 2025. 2, 3, 16
- [2] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3d motion generation for hand-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 2024. 3
- [3] 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. *ACM SIGGRAPH Asia 2024 Conference Papers*, 2024. 3
- [4] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Juntao Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Fan_Go_to_Zero_Towards_Zero-shot_Motion_Generation_with_Million-scale_Data_ICCV_2025_paper.html. 2, 3, 6, 8, 9, 10, 13, 15, 16, 18, 19, 20, 27, 31, 33
- [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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, 2023. 2, 3, 16
- [6] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Giga-Hands: A massive annotated dataset of bimanual hand activities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, 2025. 2, 3, 5, 16, 17, 26, 27
- [7] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5142–5151. IEEE, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Guo_Generating_Diverse_and_Natural_3D_Human_Motions_From_Text_CVPR_2022_paper.html. 2, 3, 6, 7, 8, 15, 16, 18, 20, 23
- [8] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3d human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 2024. 3, 8, 9, 27, 30, 33
- [9] Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. SnapMoGen: Human motion generation from expressive texts. *CoRR*, abs/2507.09122, 2025. URL <https://doi.org/10.48550/arXiv.2507.09122>. 3, 7, 8, 31
- [10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 8, 23
- [11] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 2021. 2, 3, 16
- [12] 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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10-16, 2023*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/700cb96c1c4b8f01b51b41a5b66a16de-Abstract-Datasets_and_Benchmarks.html. 2, 3, 15, 16
- [13] Pei Lin, Sihang Xu, Hongdi Yang, Yiran Liu, Xin Chen, Jingya Wang, Jingyi Yu, and Lan Xu. HandDiffuse: Generative controllers for two-hand interactions via diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025)*, 2025. URL <https://ojs.aaai.org/index.php/AAAI/article/view/32561>. 2, 3, 16
- [14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2015. 4

- [15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 8, 28
- [16] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019. 3
- [17] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations (ICLR)*, 2024. 3, 31
- [18] 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 *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 16
- [19] 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/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 2019. 33
- [20] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 8, 23, 31
- [21] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 4, 33
- [22] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL <https://qwen.ai/blog?id=qwen3.5>. 4, 5, 33
- [23] 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. 8, 28, 33
- [24] Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling controllable human motion generation, 2026. URL <https://arxiv.org/abs/2603.15546>. NVIDIA Research. 3, 8, 9, 27, 29, 33
- [25] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. In *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2017. 3, 4, 18, 33
- [26] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=dTpbEdN9kr>. 3
- [27] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024. 4
- [28] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 31
- [29] 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, 5, 16, 17, 26, 27
- [30] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=SJ1kSy02jwu>. 3, 8, 9, 27, 30, 33
- [31] Balamurugan Thambiraja et al. CLUTCH: Contextualized language model for unlocking text-conditioned hand motion modelling in the wild. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.17770>. 2, 3, 6, 7, 10, 16, 18
- [32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 31

- [33] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 31
- [34] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: An egocentric human interaction dataset for interactive ai assistants in the real world. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 20270–20281. IEEE, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_HoloAssist_an_Egocentric_Human_Interaction_Dataset_for_Interactive_AI_Assistants_ICCV_2023_paper.html. 3, 16
- [35] Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Weishuai Zeng, Qin Jin, and Zongqing Lu. Scaling large motion models with million-level human motions. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL <https://proceedings.mlr.press/v267/wang25fb.html>. 2, 3, 15, 16
- [36] Lixing Xiao. 272-dim Motion Representation. <https://github.com/Li-xingXiao/272-dim-Motion-Representation>, 2024. Public reference implementation of the 272-dimensional per-frame motion feature introduced in MotionStreamer [37] and subsequently adopted by MotionMillion [4]. 6, 18, 19
- [37] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10086–10096, 2025. 6, 13, 18, 19
- [38] Sirui Xu et al. InterAct: Advancing large-scale versatile 3d human-object interaction generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xu_InterAct_Advancing_Large-Scale_Versatile_3D_Human-Object_Interaction_Generation_CVPR_2025_paper.html. 2, 16
- [39] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 2024. 2, 3, 16
- [40] Jiahao Zhang, Joseph Liu, Young-Yoon Lee, Seonghyeon Moon, Victor Zordan, Guy Tevet, C. Karen Liu, Stephen Gould, Oren Jacob, Haomiao Jiang, Mubbasir Kapadia, and Yizhak Ben-Shabat. RoMo: A large-scale, richly organized dataset and semantic taxonomy for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2, 3, 4, 5, 15, 16, 33
- [41] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. *CoRR*, abs/2301.06052, 2023. URL <https://arxiv.org/abs/2301.06052>. 3
- [42] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 3
- [43] Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. BimArt: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27694–27705, 2025. 8, 23, 26, 27
- [44] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_BOTH2Hands_Inferring_3D_Hands_from_Both_Text_Prompts_and_Body_CVPR_2024_paper.html. 2, 3, 16
- [45] Zimu Zhang, Yucheng Zhang, Xiyan Xu, Ziyin Wang, Sirui Xu, Kai Zhou, Bing Zhou, Chuan Guo, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. HandX: Scaling bimanual motion and interaction generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026*, 2026. URL <https://arxiv.org/abs/2603.28766>. 2, 3, 10, 16

- [46] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. GEARs: Local geometry-aware hand-object interaction synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 2024. 3
- [47] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7, 18, 19, 20

Supplementary Material for

RoMo Hands: A Large Scale Richly Organized Text to Hand Motion Dataset

Companion HTML files. A set of HTML files accompanies this PDF and is best viewed in a modern browser; each is self-contained and reproducible from the released dataset. We strongly recommend opening `index.html` first as it is a single-page portal to every artifact below, with short descriptions and direct entry points into the interactive dashboards, the 3D motion grid, and the side-by-side generator comparisons that complement the static figures and tables in this paper.

- `index.html`: entry portal for the supplementary website; one card per artifact below with a short description and direct link.
- `RoMo_Hands_generative_method_comparison.html`: side-by-side animation playback of ground-truth, MotionMillion, Kimodo M, MDM, and MoMask outputs on matched train, validation, and test prompts.
- `Dataset_statistics.html`: interactive statistical report covering durations, activity classes, per-finger and per-joint dynamic scores, inter-hand distance, smoothness, autocorrelation, and per-tier caption statistics, with cross-dataset comparison against GRAB and GigaHands. A curated subset of these views is reproduced as static figures in Appendix A.
- `Hand_motion_explorer.html`: a searchable, filterable card grid over a 20-clip subset of RoMo Hands; each card opens a full MANO viewer with a timeline scrubber, the five-level captions, taxonomy labels, and per-clip quality scores.
- `Hand_motion_example_grid.html`: a single-scene 3D viewer that plays the same 20 representative bimanual sequences in a synchronized grid; clicking a tile focuses it for scrubbing and inspection.
- `RoMo_Hands_generative_method_benchmark.html`: quantitative evaluation of every benchmark model, including taxonomy-stratified semantic retrieval (R@K, MM-Dist, CLIP), distributional metrics (FID, diversity), and physical plausibility.

Scope. This supplement expands the main paper along five axes: extended dataset comparisons including body-only datasets that did not fit in the main table (Appendix A); the full specification of the bimanual hand motion representation $\mathcal{R}_{\text{hand}}$, including channel layout, type-aware normalization, and a comparison to alternative parameterizations (Appendix B); precise mathematical definitions of every semantic and physical metric (Appendix C); the complete result tables with 95% confidence intervals, additional baselines, and the per-caption-level Ground Truth breakdown (Appendix D); and per-model implementation details, training schedules, and scaling decisions (Appendix E). The main paper stands alone; this supplement is optional further reading for reproducibility and depth.

A Full Dataset Comparison

This appendix expands the cross-dataset claims in Section 3.2 along four axes: a wider table that situates RoMo Hands against the full body-motion regime in addition to the pure-hand and hand-object datasets shown in the main paper (Table 4); a like-for-like activity-coverage comparison against GRAB and GigaHands under a shared scoring pipeline (Table 5); per-finger and per-joint dexterity coverage across the three datasets (Figure 6, Figure 7); and a caption-tier depth comparison highlighting RoMo Hands’ five-level hierarchy (Figure 6, right).

Body-scale context. The pure-hand datasets in Table 1 top out at ~ 83 h (CLUTCH); the body-motion regime, by contrast, has scaled by roughly two orders of magnitude in the last three years, from HumanML3D’s 28.6 h workhorse [7] to Motion-X (144 h) [12], MotionLib (1456 h) [35], RoMo (1238 h) [40], and MotionMillion (2000 h) [4]. Table 4 places RoMo Hands directly against this body-scale block: at 884 h, RoMo Hands is the first pure-hand dataset to reach the same band, an order of magnitude beyond the previous pure-hand state of the art and squarely inside the regime that recently unlocked the current generation of body-motion benchmarks. The body datasets themselves are not directly comparable on hand fidelity (they treat hands as rigid SMPL end-effectors), but their

recent scale trajectory is the relevant yardstick for the data jump that RoMo Hands brings to the pure-hand regime.

Table 4: **RoMo Hands reaches body-motion scale in the pure-hand regime.** The body-only block (top) shows the recent two-orders-of-magnitude scale-up in body-motion datasets, from HumanML3D’s 28.6 h to MotionMillion’s 2000 h; pure-hand datasets (bottom) had remained below 100 h until now. RoMo Hands’ 884 h closes that gap, sitting an order of magnitude beyond the previous pure-hand state of the art and inside the band that recently unlocked the current generation of body-motion benchmarks. Body datasets are not directly comparable on hand fidelity — they treat hands as rigid SMPL end-effectors — but their scale trajectory is the relevant yardstick. Hours and frame counts are the originally reported totals for each source dataset at its source frame rate; RoMo Hands is at 30 fps.

Dataset	Scope	Duration (h)	Frames (M)	Text Granularity	Text (K)
HumanML3D [7]	Body (no hands)	28.6	2.1	fine-grained	45.0
Motion-X [12]	Body (no hands)	144.2	15.6	coarse	8.1
InterAct [38]	Body (no hands)	30.7	3.3	coarse	48.6
MotionLib [35]	Body (no hands)	1456.4	157.3	fine-grained	2480.0
MotionMillion [4]	Body (no hands)	2000.0	216.0	fine-grained	2000.0
RoMo [40]	Body (no hands)	1237.8	133.7	fine-grained	4100.0
GRAB [29]	Hands + Object	3.8	1.6	—	—
OakInk2 [39]	Hands + Object	37.1	4.01	action	—
ARCTIC [5]	Hands + Object	2.02	0.22	action	—
H2O [11]	Hands + Object	1.06	0.11	action	—
HOT3D [1]	Hands + Object	3.90	0.42	—	—
GigaHands [6]	Hands + Object	34.0	3.7	coarse	84.0
HoloAssist [34]	Hands + Object	161.2	17.4	coarse	1.8
HandDiffuse [13]	Pure Hands	2.0	0.25	—	—
BOTH2Hands [44]	Pure Hands	8.31	1.8	coarse	23.5
InterHand2.6M [18]	Pure Hands	24.0	2.6	—	—
HandX [45]	Pure Hands	54.2	5.9	fine-grained	485.7
CLUTCH (3D-HIW) [31]	Pure Hands	83.3	12.0	fine-grained	32.0
RoMo Hands (Ours)	Pure Hands	884	95.5	fine-grained	2710

Dynamic score. We score per-clip motion activity with a *hybrid dynamic score* δ_j that combines a temporal term and a spatial term in equal weight,

$$\delta_j = \frac{1}{2} \overline{\|v_{t,j}\|_2}^t + \frac{1}{2} \left\| \max_t p_{t,j} - \min_t p_{t,j} \right\|_2,$$

where the first term is the clip-mean speed of joint j in metres per second and the second is the ℓ_2 norm of the per-axis bounding-box extents of joint j over the clip in metres. Both terms are kept in their native physical units (no sigmoid, no per-dataset percentile cap, and no cross-term rescaling), so δ_j is a mixed-unit summary in which the spatial term tends to dominate on typical bimanual hand clips. Joint positions are taken in the bimanual-canonical frame of $\mathcal{R}_{\text{hand}}$ (root-relative xz in the inter-wrist heading frame, world-absolute y ; Section B.3), so δ is invariant to where the subject is in the scene and which way they are facing, and measures motion relative to the bimanual root. Joints aggregate arithmetically: a per-finger score averages the three phalanges of that finger (no fingertips, 5 fingers $\times$ 2 hands = 10 groups), a per-hand score averages its 15 finger joints, and the per-clip “overall” score averages the 30 finger joints across both hands with the two wrists excluded; wrist scores are tracked separately over $\{0, 16\}$. Figure 6 and Figure 7 report the per-finger and per-joint hybrid score respectively (the bar chart includes wrists; the per-finger heatmap and the headline “Mean Dynamic” do not), and Table 5’s Mean Dynamic column is the across-clips mean of each clip’s overall hybrid score.

Activity ratio and buckets. A complementary, simpler statistic is the per-clip *activity ratio*: the fraction of frames whose mean joint velocity exceeds 0.01 m/s. This ratio assigns each clip to one of

three buckets, Static ($< 10\%$), Low ($10\text{--}50\%$), or Active ($> 50\%$), which are the columns reported in Table 5; the Mean Dynamic column on the same table reports the mean of δ across all clips in each dataset.

Activity coverage under shared scoring. Table 5 reports the cross-dataset comparison summarized in Section 3.2: each clip in GRAB, GigaHands, and RoMo Hands is scored by the same dynamic-score function defined above, then bucketed under shared thresholds. We do *not* re-segment GRAB or GigaHands; their clips are taken as released by the original authors and resampled to our 30 fps reference rate, with clips that fail to load discarded, and only the scoring tool is shared, so per-clip scores are directly comparable across the three datasets without altering the reference splits. (The cross-dataset totals in Table 1 and Table 4 preserve each source paper’s originally reported sequence count, hours, and frames at the source frame rate; the activity-coverage analysis below is reported under our 30 fps loader and therefore lists 2.86 h / 16.40 h for GRAB / GigaHands respectively.) Duration histograms are normalized to density so shapes are comparable across very different dataset sizes. Three patterns stand out: **scale**: 541,929 clips (884 h) vs. GRAB’s 1,067 (2.86 h) and GigaHands’ 12,704 (16.4 h), 2–3 orders of magnitude more clips at comparable per-clip lengths; **curation effect**: GigaHands’ uncuration egocentric stream is 85.8% Static+Low (20.4% Static), whereas RoMo Hands collapses Static to 0.2% and recovers a 53.7% Active majority despite being two orders of magnitude larger; **regime placement**: the RoMo Hands mean dynamic score (0.330) sits between GRAB’s tightly choreographed studio capture (0.658) and GigaHands’ uncuration stream (0.136), RoMo Hands trades GRAB’s onset-saturated grasp-cycle regime for broader coverage while still removing the static mean-pose mass that dominates raw web video, and we unpack the per-finger and per-joint structure of these bands below.

Table 5: **Cross-dataset activity coverage under a shared scoring tool.** Each clip in GRAB, GigaHands, and RoMo Hands is scored by the same per-frame dynamic-score function used during RoMo Hands curation. Activity buckets (Static / Low / Active) are assigned per clip from shared thresholds; Mean Dynamic is the average per-clip score in $[0, 1]$.

Dataset	Sequences	Total (h)	Mean (s)	Median (s)	Active (%)	Low (%)	Static (%)	Mean Dynamic
GRAB [29]	1,067	2.86	9.66	8.10	65.6	34.4	0.0	0.6579
GigaHands [6]	12,704	16.40	4.65	3.13	14.2	65.4	20.4	0.1357
RoMo Hands (Ours)	541,929	884.17	5.87	4.00	53.7	46.1	0.2	0.3304

Per-finger and per-joint coverage. Figure 6 reports the mean hybrid dynamic score per finger across the three datasets and Figure 7 reports the same quantity at the per-joint level (32 MANO joints). The three datasets occupy cleanly separated bands with GRAB at 0.57–0.72, RoMo Hands at 0.28–0.35, GigaHands at 0.07–0.17, which we read as diagnostic of *collection setup* rather than as an intrinsic ordering of motion quality: δ is a sum of physical units rather than a normalized score, so a higher value here is not automatically “better.” GRAB sits at the top because every clip is a short studio-recorded grasp + manipulate + release cycle: the rest-to-grasp transition alone sweeps the fingers across most of their workspace, which inflates the spatial bounding-box term every clip even before the manipulation phase, and the absence of idle frames in between keeps the velocity term elevated as well. GigaHands sits at the bottom because its uncuration egocentric stream contains substantial idle / hands-resting frames: with hands at rest, both the per-joint bounding box and the average speed collapse, and a per-clip average over a stream where most frames look like that necessarily lands in the low-dynamic regime. RoMo Hands lands in the middle by design: dexterity filtering at curation removes the static frames that depress GigaHands, while the absence of a forced rest-to-grasp framing avoids the onset-driven inflation that lifts GRAB. Within the RoMo Hands band the figures recover two anatomical patterns one would expect: a proximal-to-distal gradient driven by the bimanual-canonical frame (wrists 0.28–0.29, where motion in the canonical frame collapses to inter-wrist-distance variation plus world- y displacement, vs. distal phalanges 0.32–0.35 where the full articulation workspace is exposed) and slightly less active thumbs than long fingers (thumbs ≈ 0.30 , index–pinky 0.33–0.35), confirming that the score still discriminates major from minor articulators on clip distributions where neither collection-setup artefact dominates. These figures are a static slice of the interactive views in `Dataset_statistics.html`.

Caption-tier word counts. Figure 6 (right) plots the mean number of words per caption across the five RoMo Hands caption tiers (L_0 tag $\rightarrow L_4$ paragraph) against the comparators’ single-caption

baselines. RoMo Hands grows from 3.5 words at L_0 to 149.3 at L_4 (nearly a 43 $\times$ span across tiers) whereas GigaHands ships a single, fixed-granularity caption per clip (~ 12.0 words). GRAB releases no captions; its reference line (~ 6.7 words) shows template sentences we generate from GRAB’s object and action labels (e.g., “a person drinks from a mug”), included for reference only. This explicit hierarchy is what enables the per-tier retrieval breakdown of Table 14 and the uniform-over-levels training mixture used by the TMR evaluator.

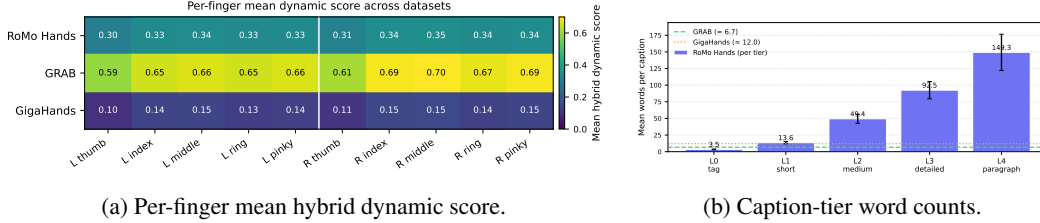


Figure 6: **Finger coverage and caption-tier depth across datasets.** Left: GRAB saturates every finger near 0.65, GigaHands’ uncensored stream sits at 0.10–0.15, and RoMo Hands occupies the middle (0.30–0.35) with thumbs slightly less dynamic than the long fingers (a consistent cross-dataset pattern). Right: RoMo Hands’ five-tier caption hierarchy spans 3.5 to 149.3 words (43 $\times$), while GigaHands ships a single caption per clip and the GRAB line shows our template sentences built from its object–action labels (GRAB itself is text-free).

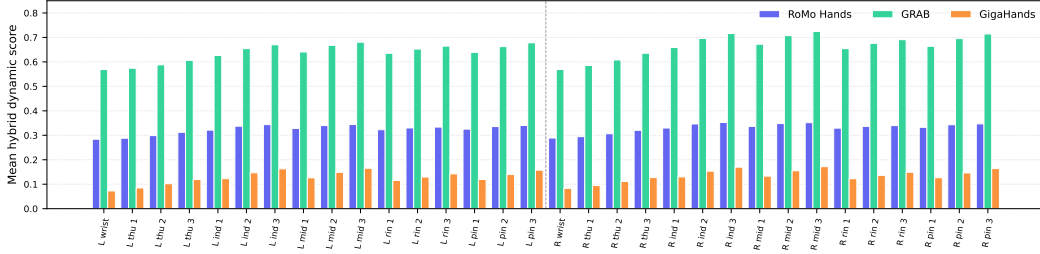


Figure 7: **Per-joint mean hybrid dynamic score** for all 32 MANO joints (16 left, 16 right; vertical guide marks the hand boundary). GRAB is uniformly saturated (≈ 0.65); RoMo Hands shows the expected proximal-to-distal gradient (wrist ≈ 0.28 , distal phalanges ≈ 0.35); GigaHands sits at 0.08–0.17 across the entire skeleton. Joints are ordered wrist $\rightarrow$ thumb $\rightarrow$ index $\rightarrow$ middle $\rightarrow$ ring $\rightarrow$ pinky, with depth indices 1–3 from proximal to distal.

B Bimanual Hand Motion Representation: Full Specification

This appendix expands the compact main-text description in Section 3.3. We start from MANO and the body-motion-generation template, formalize the channel equivalence between $\mathcal{R}_{272}$ [37, 4, 36] and our $\mathcal{R}_{\text{hand}}$, list the per-dimension layout, derive the type-aware normalization scheme, and compare to alternative bimanual parameterizations.

B.1 From MANO to a body-style representation

MANO [25] parameterizes each hand by a global axis-angle rotation $\mathbf{r}^g \in \mathbb{R}^3$, a translation $\mathbf{t}^g \in \mathbb{R}^3$, 15 local joint rotations ($\mathbb{R}^{45}$), and 10 PCA shape coefficients $\beta \in \mathbb{R}^{10}$ (61 per hand, 122 bimanually). Existing hand generators [31] drop β and convert rotations to a 6D continuous form [47], yielding the 99-D CLUTCH parameterization.

Body-motion generators consume a richer template: the HumanML3D 263-D [7] and 272-D [37, 4, 36] representations pack rotations alongside redundant root-relative joint positions, finite-difference velocities, and a 6D heading frame, and this redundancy is what makes diffusion, masked-token, and autoregressive backbones converge well. Our 393-D representation transports this template to MANO’s 32-joint bimanual skeleton (16 joints per hand; a wrist root and five three-phalanx finger

chains, indices 0–15 left, 16–31 right), uses 6D rotations throughout, and drops β since motion is generated at fixed mean shape. Two decisions distinguish it from a naive port:

- (i) **Heading is derived from the inter-wrist axis.** Body representations canonicalize a clip by aligning a pelvis-anchored heading; hands have no natural pelvis. We extract a yaw angle ψ_t from the wrist-to-wrist direction and apply $\mathbf{R}_y(\psi_t)$, a rotation about the world-up axis only. Restricting heading canonicalization to yaw preserves the absolute y -component of every position, exposing world height as a directly observable feature and avoiding the gravity-coupled rotations that would otherwise corrupt joint-position channels for two-handed gestures.
- (ii) **The bimanual root is the wrist midpoint, not a skeleton joint.** We define $\mathbf{r}_t = \frac{1}{2}(\mathbf{p}_t^L + \mathbf{p}_t^R)$ as the root used by the heading frame and the velocity channel. Because $\mathbf{r}_t$ is not one of the 32 skeleton joints, its height is stored as an explicit scalar; without this, root height could not be recovered without integration.

B.2 Structural equivalence to the body 272-D representation

The body 272-D representation introduced by Xiao et al. [37] (adopted by Fan et al. [4], with public reference implementation by Xiao [36]), hereafter $\mathcal{R}_{272}$, lays out five per-frame blocks (root XZ linear velocity, root angular velocity as a 6D continuous rotation, and per-joint positions, velocities, and 6D rotations over the 22-joint SMPL skeleton), whose exact dimension ranges are listed in the body column of Table 6. Its only structural specialization is the absence of an explicit root-height channel: the pelvis is joint 0 of the SMPL skeleton, so its world- y component is carried inside the joint-position block via XZ-centering. Stripping this one implicit-vs-explicit difference, $\mathcal{R}_{272}$ and our $\mathcal{R}_{\text{hand}}$ instantiate the same generic per-frame template $\mathbf{m}_t$ at $J=22$ and $J=32$ respectively, with the hand-specific vector $\mathbf{h}_t$ of (1) recovered as the $J=32$ instance of $\mathbf{m}_t$:

$$\mathbf{m}_t = \left[\underbrace{\mathbf{f}_t}_{\text{root frame}}, \underbrace{\{\tilde{\mathbf{p}}_t^j\}_{j=0}^{J-1}}_{3J, \text{ joint positions}}, \underbrace{\{\dot{\tilde{\mathbf{p}}}_t^j\}_{j=0}^{J-1}}_{3J, \text{ joint velocities}}, \underbrace{\{\phi(\mathbf{R}_t^j)\}_{j=0}^{J-1}}_{6J, \text{ 6D rotations}} \right], \quad (2)$$

with $\dim \mathbf{f}_t = 8$ when the root joint coincides with a skeleton joint (body) and $\dim \mathbf{f}_t = 9$ when the root is off-skeleton (our hand case, where the bimanual root is the wrist midpoint and an explicit height channel is required). This gives $\dim \mathbf{m}_t = 12J + 8$ for the body case (with $J=22$ this recovers $\mathcal{R}_{272}$ exactly: $12 \cdot 22 + 8 = 272$) and $\dim \mathbf{m}_t = 12J + 9$ for the bimanual hand case (with $J=32$: $12 \cdot 32 + 9 = 393$). Table 6 pins this equivalence channel-by-channel against the verified $\mathcal{R}_{272}$ index ranges.

Table 6: $\mathcal{R}_{\text{hand}}$ is structurally the body $\mathcal{R}_{272}$ template re-indexed by joint count, with one extra channel forced by the off-skeleton bimanual root. Per-channel equivalence between the body 272-D representation $\mathcal{R}_{272}$ [37, 4, 36] and the proposed hand 393-D representation. Both instantiate the template of eq. (2); the bimanual root being off-skeleton forces an explicit root-height channel on the hand side.

Channel group	Channel content	Body ($\mathcal{R}_{272}$, $J=22$)		Hand ($\mathcal{R}_{\text{hand}}$, $J=32$)	
		dim	range	dim	range
Root frame	root XZ linear velocity	2	[0:2)	2	[0:2)
	root height (absolute world y)	–	implicit	1	[2:3)
	root angular velocity (6D rotation)	6	[2:8)	6	[3:9)
Joint positions	root-relative, $J \times 3$	66	[8:74)	96	[9:105)
Joint velocities	finite-difference, $J \times 3$	66	[74:140)	96	[105:201)
Joint rotations	6D continuous [47], $J \times 6$	132	[140:272)	192	[201:393)
Total		272	(12J+8)	393	(12J+9)

B.3 Detailed per-dimension breakdown

We now specify the exact layout of $\mathbf{h}_t \in \mathbb{R}^{393}$. At frame t , let $\mathbf{p}_t^L, \mathbf{p}_t^R \in \mathbb{R}^3$ denote the wrist positions and $\mathbf{r}_t = \frac{1}{2}(\mathbf{p}_t^L + \mathbf{p}_t^R)$ the bimanual root. The heading-frame yaw is

$$\psi_t = \text{atan2}(-(\mathbf{p}_t^R - \mathbf{p}_t^L)_z, (\mathbf{p}_t^R - \mathbf{p}_t^L)_x), \quad \mathbf{R}_t^\psi = \mathbf{R}_y(\psi_t) \in SO(3). \quad (3)$$

Heading degeneracy. The yaw ψ_t is undefined whenever the two wrists project to the same point on the XZ plane, i.e. whenever $(\mathbf{p}_t^R - \mathbf{p}_t^L)_x = (\mathbf{p}_t^R - \mathbf{p}_t^L)_z = 0$; this covers prayer, overhead-clasp, and the limit case of coincident wrists. On each maximal degenerate run we replace ψ_t by linear interpolation between the bracketing non-degenerate frames; if a run touches a clip boundary we hold the nearest non-degenerate value across it. Because linear interpolation in ψ keeps every replaced frame on the yaw subgroup $\{\mathbf{R}_y(\psi)\}_{\psi \in \mathbb{R}}$, the round-trip property of the type-aware normalization (Section B.4) is preserved.

We store the heading rotation as a 6D continuous descriptor $\omega_t = \phi(\mathbf{R}_t^\psi)$, where the operator ϕ returns the first two columns of the rotation matrix flattened to $\mathbb{R}^6$ [47]. By construction this descriptor encodes the absolute heading at the start of each clip and incremental yaw at all subsequent frames. Joint quantities are expressed in the heading-canonical reference: $\tilde{\mathbf{p}}_t^j$ are joint positions made root-relative on the XZ plane (their y -component is absolute world height) and yaw-aligned by $\mathbf{R}_t^\psi$; $\dot{\tilde{\mathbf{p}}}_t^j$ are their first-order finite differences; $\mathbf{R}_t^j \in SO(3)$ are joint rotations, where the two wrist rotations ($j \in \{0, 16\}$) are expressed in the heading-canonical frame and all finger rotations remain parent-relative. Table 7 reports the resulting layout.

Table 7: Per-dimension layout of the proposed hand 393-D representation. The type column drives the type-aware normalization of Section B.4: GAUSSIAN dims are z -scored and clipped to $[-5, +5]\sigma$; CYCLIC dims pass through untouched to preserve the Stiefel-manifold structure; CONSTANT dims are zeroed on forward and restored from μ on inverse.

Dim range	Channel	Width	Type
0–1	Root XZ velocity, heading-aligned ($\dot{\mathbf{r}}_t^{xz}$)	2	GAUSSIAN
2	Root height ($r_{t,y}$, absolute world y)	1	GAUSSIAN
3–8	Heading 6D rotation ($\omega_t = \phi(\mathbf{R}_t^\psi)$)	6	CYCLIC ($\times 2$), CONSTANT ($\times 4$)
9–104	Joint positions, root-relative & heading-aligned ($\{\tilde{\mathbf{p}}_t^j\}, 32 \times 3$)	96	GAUSSIAN
105–200	Joint velocities, finite difference ($\{\dot{\tilde{\mathbf{p}}}_t^j\}, 32 \times 3$)	96	GAUSSIAN
201–392	Joint 6D rotations on $\text{St}(3, 2)$ ($\{\phi(\mathbf{R}_t^j)\}, 32 \times 6$)	192	CYCLIC
Total		393	
of which GAUSSIAN		195	root XZ vel (2) + height (1) + positions (96) + joint vel (96)
of which CYCLIC		194	heading (2 of 6) + joint rotations (192)
of which CONSTANT		4	4 of 6 heading dims (zero by construction)

B.4 Choice of normalization scheme

Two normalization schemes are natural candidates for $\mathbf{h}_t \in \mathbb{R}^{393}$: the **vanilla per-channel z -score** that is the de-facto standard on full-body data [7, 4], and the **type-aware** scheme we adopt. The vanilla recipe is empirically misapplied to $\mathbf{h}_t$ for three independent reasons rooted in geometry, not data quality; the type-aware scheme is a strict generalization that coincides with vanilla on a benign representation but diverges on $\mathbf{h}_t$.

Option A: vanilla per-channel z -score. Fit (μ_i, σ_i) on the train split and apply the same affine map per dimension,

$$\tilde{h}_{t,i} = (h_{t,i} - \mu_i)/\sigma_i, \quad \sigma_i \leftarrow \max(\sigma_i, 1), \quad (4)$$

with the standard $\max(\sigma, 1)$ guard against degenerate channels. This works on 263-D and 272-D body representations because every dimension is approximately Gaussian, well-conditioned, and geometrically unconstrained. None of those preconditions holds on $\mathbf{h}_t$:

- (i) *Manifold-constrained channels.* Each joint’s 6D rotation is not six free numbers but two 3-vectors that must remain unit-norm and mutually orthogonal; rescaling the six coordinates independently breaks both constraints, so the encoder no longer sees rotations as rotations. Concretely, the 6D rotation features (ω_t and $\{\phi(\mathbf{R}_t^j)\}$, $192 + 2 = 194$ active dimensions) live on the Stiefel manifold $\text{St}(3, 2)$ (the first two columns of a rotation matrix), carrying $3 \cdot 32 + 1 = 97$ true degrees of freedom in their 194 ambient dimensions. The redundancy is the price of continuity: Zhou et al. [47] prove that no continuous mapping $\mathbb{R}^n \rightarrow SO(d)$ exists below a critical n , so the constraint $\|\mathbf{a}_1\| = \|\mathbf{a}_2\| = 1, \mathbf{a}_1 \cdot \mathbf{a}_2 = 0$ must be borne in the representation itself. After z -scoring this constraint no longer holds in the input the encoder sees, and the manifold structure has to be re-discovered from training samples rather than being baked into the representation.

- (ii) *Bimodal channels by construction.* Several channels mix two regimes with very different scales (a one-shot draw at the start of each clip versus a near-identity increment everywhere else, or the dual of that for velocities, where rare bursts coexist with long near-stationary runs), so a single (μ, σ) fits neither, and after normalization the rare regime surfaces as a $30\text{--}50\sigma$ “outlier” that is in fact an entirely valid frame. The clearest case is the heading rotation ω_t , which stores the *absolute* heading at $t=0$ (uniformly distributed on S^1 across clips) and the *incremental* change at $t \geq 1$ (sharply peaked near identity, $\cos \Delta\psi \approx 1$, $\sin \Delta\psi \approx 0$). Per-channel statistics average over both regimes: in our training set the cached σ for $\cos \psi$ is 0.055, while the $\text{MAD} \cdot 1.4826$ measured on frames with $t \geq 1$ alone is 4×10^{-5} (approximately $1300\times$ smaller), so frame-zero values normalize as $30\text{--}50\sigma$ “outliers” even though they are perfectly valid orientations. The dual pathology applies to the very-low-variance translation channels (root XZ velocity, joint velocities; ~ 98 dims) on hand data, where most frames are near-stationary and σ is dominated by rare bursts; the result is occasional normalized values exceeding ± 30 on real validation data.
- (iii) *Constant-by-construction channels.* Four of the six heading dimensions carry no information at all (they are pinned to 0 or 1 for every frame of every clip) yet the standard $\sigma \leftarrow \max(\sigma, 1)$ guard rescales them as if they were random and the model has to learn that they are meaningless. Specifically, these are the third row of $\mathbf{R}_y(\psi_t)$ and the second column’s y -component, which the yaw-only restriction pins by construction; their empirical σ collapses to 0, the $\max(\sigma, 1)$ guard kicks in, and the resulting channels become structured noise that the encoder has to explicitly learn to ignore.

None of these failure modes is data-driven: each follows from how $\mathbf{h}_t$ is constructed, and each would persist under perfect mocap.

Option B: type-aware normalization (used). We address all three pathologies with a per-dimension type mask $\tau \in \{\text{GAUSSIAN}, \text{CYCLIC}, \text{CONSTANT}\}^{393}$ (right-most column of Table 7) that selects one of three transforms:

$$\tilde{h}_{t,i} = \begin{cases} \text{clip}((h_{t,i} - \mu_i)/\sigma_i, \pm\sigma_{\text{clip}}), & \tau_i = \text{GAUSSIAN} \\ h_{t,i}, & \tau_i = \text{CYCLIC} \\ 0, & \tau_i = \text{CONSTANT} \end{cases} \quad (5)$$

with $\sigma_{\text{clip}} = 5$. The CYCLIC block covers all 194 active rotation dims and is passed through, preserving the Stiefel-manifold structure (neutralizing (i)). The CONSTANT block covers the four degenerate heading dims; they are zeroed on forward and restored from μ_i on inverse, so the round-trip is lossless (neutralizing (iii)). The 195 GAUSSIAN channels are z -scored and clipped at $\pm\sigma_{\text{clip}}$; the clip leaves the well-behaved position/height channels untouched (their values lie within $\pm 3\sigma$) and acts only on the ~ 98 small- σ velocity channels, absorbing the heavy-tail bursts of (ii). 98.7% of velocity-channel mass lies within $\pm 3\sigma$ on the training set, so the clip is conservative.

Relationship between the two options. Option B reduces to Option A as every dim is tagged GAUSSIAN and $\sigma_{\text{clip}} \rightarrow \infty$: (5) collapses to (4). The two schemes are operationally identical on benign representations like the body 272-D template, but diverge on $\mathbf{h}_t$ along the three axes above. The divergence matters empirically: the TMR retrieval evaluator trained under Option A produces validation features with occasional $\pm 30\sigma$ excursions and degraded R@1, whereas the same architecture trained under Option B converges to the noise floor reported in Section E.4. The transform is invertible per type and lossless on the CYCLIC block, preserving the full bimanual geometry end-to-end through training and evaluation. All hand experiments in this paper including the TMR evaluator are trained under Option B with $\sigma_{\text{clip}} = 5$.

B.5 Comparison to prior bimanual parameterizations

Table 8 compares $\mathcal{R}_{\text{hand}}$ to the principal bimanual alternatives. Our representation is roughly $2\times$ the dimensionality of CLUTCH’s parameterization and $4\times$ that of raw MANO, but this redundancy follows the same principle that makes the body 263-D and 272-D representations effective for full-body generation: making geometric structure explicitly accessible through complementary views of the same underlying pose.

Table 8: $\mathcal{R}_{\text{hand}}$ extends raw MANO and CLUTCH’s parameterization with body-style position, velocity, and heading channels. This roughly doubles the dimensionality but matches the inductive bias of HumanML3D-272. Rows enumerate the design axes that distinguish the three options; the redundancy of $\mathcal{R}_{\text{hand}}$ is the same redundancy that makes body templates effective for full-body generation.

	Raw MANO	CLUTCH	Ours ($\mathcal{R}_{\text{hand}}$)
Rotation repr.	axis-angle	6D continuous	6D continuous
Bimanual root	per-hand $\mathbf{t}^g$	per-hand $\mathbf{t}^g$	wrist midpoint, world y explicit
Heading canon.	$\times$	$\times$	yaw-only, inter-wrist axis
Heading repr.	—	—	6D rotation
Root velocity	$\times$	$\times$	2D XZ, heading-aligned
Joint positions	$\times$	$\times$	$\checkmark(32 \times 3)$
Joint velocities	$\times$	$\times$	$\checkmark(32 \times 3)$
Joint rotations	$\checkmark(32 \times 3)$	$\checkmark(32 \times 6)$	$\checkmark(32 \times 6)$
Shape coefficients	$\beta \in \mathbb{R}^{10}$	$\times$	$\times$
Total dims (bimanual)	122	198	393
Body-template parity	SMPL	SMPL + 6D rot.	extends $\mathcal{R}_{272}$ ($J=22 \rightarrow 32$)

B.6 The 201-D essential subset for the TMR evaluator

The TMR retrieval evaluator (Section E.4) does not consume the full 393-D $\mathcal{R}_{\text{hand}}$ vector. It operates instead on a 201-dimensional *essential* subset $\mathbf{h}_t^{\text{ess}}$ obtained by dropping the joint-position and joint-velocity blocks of $\mathbf{h}_t$ and keeping the root frame together with the per-joint 6D rotations:

$$\mathbf{h}_t^{\text{ess}} = \left[\underbrace{\mathbf{f}_t}_{\substack{9, \text{ root frame}}}, \underbrace{\{\phi(\mathbf{R}_t^j)\}_{j=0}^{31}}_{\substack{192, \text{ joint 6D rotations}}} \right] \in \mathbb{R}^{201}. \quad (6)$$

In the layout of Table 7, $\mathbf{h}_t^{\text{ess}}$ is the dim union $[0:9) \cup [201:393)$ of $\mathbf{h}_t$; the dropped block spans $[9:201)$ and contains the 96 joint positions and the 96 joint velocities. The same type-aware normalization of Section B.4 is applied to the surviving 201 dimensions on both branches of the evaluator.

Why drop positions and velocities. The essential subset is the minimal channel set that determines the bimanual hand pose under MANO forward kinematics: given the root frame and the 32 per-joint 6D rotations, the dropped positions $\{\tilde{\mathbf{p}}_t^j\}$ and their finite differences $\{\dot{\tilde{\mathbf{p}}}_t^j\}$ are recoverable exactly under the fixed mean MANO shape used throughout (Section 3.3). The position and velocity blocks of $\mathbf{h}_t$ are therefore redundant with the rotations as far as identifying a pose is concerned as they exist in $\mathcal{R}_{\text{hand}}$ as an inductive bias for continuous generative architectures, exactly as their analogues do in HumanML3D-263 and -272 for body motion, but they do not carry information that the rotations do not.

The redundancy matters at evaluation time because of quantization. VQ-style generators (MoMask’s residual VQ, MotionMillion’s FSQ) approximate the $\mathbb{R}^{96}$ position and velocity channels through a finite codebook and so reconstruct them less faithfully than the continuous diffusion baselines (Kimodo M, MDM), even when their rotational reconstructions are equally good. An evaluator with access to the position and velocity channels would therefore learn to score motions partly on whether those continuous channels match a continuous reference, systematically penalising VQ generators on a non-articulatory axis. Restricting the TMR evaluator to the rotational 201-D subset removes that confound: every generator is scored on the channels in which it is on equal footing, and the resulting $R@k$ and FID numbers measure articulation fidelity rather than a mixed signal that conflates articulation with continuous-space reconstruction.

C Evaluation Metric Details

Text-conditioned hand motion generation has no per-clip ground truth: many distinct motions can plausibly satisfy the same caption, so a generated clip cannot be scored by direct reference comparison the way one would score a reconstruction or denoising model. As a consequence, no single metric and in particular no single *physical* metric is sufficient to declare a motion “physically plausible.” Fol-

lowing the body-motion literature, we therefore triangulate with a bank of complementary measures that fall into two families.

Semantic metrics (Section C.1: FID [10], R-Precision@ k [7], Matching Score, Diversity, Multi-Modality) operate in the latent space of the TMR [20] text-motion evaluator (Section E.4, trained on the 201-D essential subset of $\mathcal{R}_{\text{hand}}$, Section B.6). They jointly measure (i) distributional alignment between generated and GT motions (FID), (ii) text-conditional alignment between a motion and its prompt (R-Precision@ k , Matching Score), and (iii) within- and cross-prompt variability (Diversity, MultiModality), so a model cannot win all five simultaneously by collapsing to a single mode or by ignoring the caption. Because every value is evaluator-specific, absolute numbers are not directly comparable to HumanML3D body-motion numbers; the reference ranges we cite from body-motion literature should be read as orientation, not as targets to clear. Throughout, E_m and E_t denote the TMR motion and text encoders.

Physical metrics (Section C.2) are evaluator-free proxies computed directly on per-joint positions and posed MANO meshes. We borrow smoothness measures (Acceleration, Intra-hand Jerk, HF Power Ratio) and contact-validity measures (Inter- and Self-hand penetration variants) from the body-motion literature [7, 43], and add two hand-specific axes that we found necessary in pilots: bone-length violation rate, which catches token-by-token drift in autoregressive generators, and a finger-flexion posture mean, which catches the “flat hand” failure mode. None of these individually proves a motion is physically valid; together they flag the dominant failure modes we observed in pilots.

C.1 Semantic Metrics

The arrow next to each metric ($\downarrow/\uparrow$) indicates whether smaller or larger is better; metrics whose target is to *match* the GT row of the same table rather than minimise or maximise are flagged inline.

FID [10] ($\downarrow$, lower is better). Squared 2-Wasserstein distance between Gaussians fit to motion-encoder features of the real (GT) and generated test sets:

$$\text{FID} = \|\mu - \hat{\mu}\|_2^2 + \text{Tr}(\Sigma + \hat{\Sigma} - 2(\Sigma \hat{\Sigma})^{1/2}), \quad \downarrow,$$

where (μ, Σ) is the empirical mean and covariance of $\{E_m(m_i)\}_{i=1}^N$ over N GT motions, $(\hat{\mu}, \hat{\Sigma})$ is the same for $\{E_m(\hat{m}_j)\}_{j=1}^M$ over M generated motions, $\|\cdot\|_2$ is the Euclidean norm, and $\text{Tr}(\cdot)$ is the matrix trace. For orientation, body-motion generators reported on HumanML3D typically land at 0.05–1.0 for SOTA and ≤ 0.05 on the GT row.

R-Precision@ k [7] ($\uparrow$, higher is better). Let $\mathcal{T} = \{\hat{m}\}$ denote the set of generated motions over the test split. For each $\hat{m} \in \mathcal{T}$, the ground-truth caption t^* is placed in a pool of 32 (1 correct + 31 random distractors drawn uniformly from the test split’s captions); the 32 candidates $\{t_i\}_{i=1}^{32}$ are ranked by ascending cosine distance $d_i = 1 - \cos(E_m(\hat{m}), E_t(t_i))$ to the generated motion, and a hit is counted when t^* falls in the top k :

$$\text{R-Precision@}k = \frac{1}{|\mathcal{T}|} \sum_{\hat{m} \in \mathcal{T}} \mathbb{I}[\text{rank}(t^*) \leq k] \in [0, 1], \quad \uparrow,$$

where $\mathbb{I}[\cdot]$ is the *indicator function* (1 if the condition holds, 0 otherwise), $|\mathcal{T}|$ is the size of the test-generation set, and $\text{rank}(t^*)$ is the position of t^* in the ascending-distance ordering of the 32-candidate pool. Main paper reports $k \in \{1, 3, 10\}$; supplemental tables add $k \in \{2, 5\}$. Chance under random retrieval from the pool of 32 is $k/32$, i.e. 0.031, 0.094, 0.313 at $k = 1, 3, 10$. On HumanML3D the same protocol typically lands SOTA body-motion generators at ~ 0.50 – 0.55 for R@1 and ~ 0.85 – 0.92 for R@10, with the GT row near the same numbers.

Matching Score (MS, a.k.a. MM-Dist) ($\downarrow$, lower is better). Mean paired Euclidean (L_2) distance between each generated motion and its ground-truth caption in the TMR latent space, following the HumanML3D convention [7]. Let $\{(\hat{m}_i, t_i^*)\}_{i=1}^B$ be the $B = |\mathcal{T}|$ matched (motion, GT-caption) pairs over the test-generation set:

$$\text{MS} = \frac{1}{B} \sum_{i=1}^B \|E_m(\hat{m}_i) - E_t(t_i^*)\|_2 \geq 0, \quad \downarrow,$$

with 0 corresponding to perfect text–motion alignment. We additionally report a scale-invariant variant MS_{norm} (Table 14) that L_2 -normalizes each embedding to unit length before the distance, bounded in $[0, 2]$ and satisfying $\text{MS}_{\text{norm}}^2 = 2 - 2 \cos(E_m(\hat{n}), E_t(t^*))$. Because both variants are reported in evaluator-specific units, absolute values do not transfer across evaluators (in particular, HumanML3D body-motion numbers are not directly comparable); we report MS only against our own GT row.

CLIP Score ($\uparrow$, higher is better). Mean paired cosine similarity between each generated motion and its ground-truth caption in the TMR text–motion joint embedding space, the scale-invariant analogue of MS:

$$\text{CLIP score} = \frac{1}{B} \sum_{i=1}^B \cos(E_m(\hat{n}_i), E_t(t_i^*)) \in [-1, 1], \quad \uparrow,$$

with 1 corresponding to perfect text–motion alignment. We name it “CLIP score” by analogy with the image–text CLIP-score convention, although the joint space here is TMR rather than CLIP; this is the headline alignment axis used in the per-category radar plot of Figure 5.

Diversity (target: match GT). Mean Euclidean distance between two random subsets of $S_d = 300$ TMR motion features. Let $\{v_i\}_{i=1}^{S_d}$ and $\{v'_i\}_{i=1}^{S_d}$ be two disjoint random subsets of $\{E_m(\hat{n}) : \hat{n} \in \mathcal{T}\}$, each of size S_d :

$$\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} \|v_i - v'_i\|_2, \quad \text{target: match GT.}$$

Values substantially below the GT row of the same table indicate mode collapse; values substantially above indicate noise. Body-motion literature reports Diversity in the same units as its evaluator-specific MS, so direct numerical comparison is again only meaningful within a single evaluator.

MultiModality (target: match GT). Within-prompt analogue of Diversity over $C = 30$ shared prompts with $S_l = 10$ pairs per prompt. For each prompt $c \in \{1, \dots, C\}$ we draw S_l pairs of generations conditioned on the same caption and let $v_{c,i}, v'_{c,i} = E_m(\hat{n}_{c,i}), E_m(\hat{n}'_{c,i})$ denote the TMR motion-encoder features of the i -th pair:

$$\text{MModality} = \frac{1}{C \cdot S_l} \sum_{c=1}^C \sum_{i=1}^{S_l} \|v_{c,i} - v'_{c,i}\|_2, \quad \text{target: match GT.}$$

Higher indicates more variation under a fixed prompt; excessively high values indicate noise rather than valid alternative interpretations. As with Diversity, the relevant comparison is to the same evaluator’s GT row.

All semantic metrics are averaged over 20 independent evaluation runs with 95% confidence intervals.

C.2 Physical Metrics

Physical metrics are evaluator-free and computed at 30 fps directly from the per-joint positions obtained by decoding $\mathcal{R}_{\text{hand}}$ and posing the left and right MANO templates by forward kinematics. A single shared routine takes a $(T, 32, 3)$ position array in metres and produces every number reported in the physical half of Table 3 of the main paper and Table 10 here, for both generator outputs and the GT rows for RoMo Hands, GRAB, and GigaHands. We split the metrics into two families with very different compute profiles.

Joint-based metrics (Acceleration, Intra-hand Jerk, HF Power Ratio, Bone-length Violations, Finger Flexion Mean) read the $(T, 32, 3)$ position array directly. They are cheap and we evaluate them over the full RoMo Hands test split.

Mesh-based metrics (Inter-hand and Self-hand penetration variants) require posing both ~ 778 -vertex MANO meshes every frame and querying the signed point-to-mesh distance from each query (vertex or joint) to a posed mesh, where the sign denotes whether the query lies outside ($d > 0$) or inside ($d < 0$) the mesh; to make the inside/outside test well-defined, each MANO mesh is closed at the wrist with a small triangular cap before the distance query. These mesh queries are roughly two

orders of magnitude more expensive than joint-array reads, so we estimate the mesh-based metrics on a randomly sampled 256-sequence subset of the test split at frame stride 4. The reported numbers are therefore order-of-magnitude estimates rather than full-split averages, but the subset is large enough for the rankings to be stable across the 20 evaluation replications used for the 95% CIs.

Let $\mathbf{x}(t) \in \mathbb{R}^{J \times 3}$ denote joint positions at frame $t \in \{1, \dots, T\}$ ($J = 32, 16$ per hand; T is the per-clip frame count, with hand $h \in \{L, R\}$ subscripts written as $\mathbf{x}_j^h(t) \in \mathbb{R}^3$ when needed), $\mathbf{V}^L(t), \mathbf{V}^R(t)$ the left/right mesh vertices at frame t , and $\Delta t = 1/30$ s the inter-frame interval. Derivatives use central differences:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \frac{\mathbf{x}(t+1) - \mathbf{x}(t-1)}{2\Delta t}, & \ddot{\mathbf{x}}(t) &= \frac{\mathbf{x}(t+1) - 2\mathbf{x}(t) + \mathbf{x}(t-1)}{\Delta t^2}, \\ \ddot{\mathbf{x}}(t) &= \frac{\ddot{\mathbf{x}}(t+1) - \ddot{\mathbf{x}}(t-1)}{2\Delta t}.\end{aligned}$$

The physical half of Table 3 in the main paper is the headline physical-metrics block; it reports Acceleration, Intra-hand Jerk, Inter-hand Penetration Ratio, Self-hand Penetration MAX, and HF Power Ratio. Supplementary axes such as mean inter-hand penetration depth, the BimArt > 10 mm self-penetration ratio, bone-length violations, and finger flexion mean live in Table 10 here.

Acceleration (target: match GT). Mean Euclidean acceleration magnitude over all frames and joints:

$$\text{Acc} = \frac{1}{TJ} \sum_{t=1}^T \sum_{j=0}^{J-1} \|\ddot{\mathbf{x}}_{t,j}\|_2 \quad (\text{m/s}^2), \quad \text{target: match GT},$$

where $\ddot{\mathbf{x}}_{t,j} \in \mathbb{R}^3$ is the central-difference acceleration of joint j at frame t defined above. The target is to *match* the GT row of the same table, not to minimise: under-acceleration indicates over-smoothing and over-acceleration indicates jitter, so deviation in either direction is a quality flag.

Intra-hand Jerk (target: match GT). Mean Euclidean jerk magnitude after rotating each hand into a per-frame local wrist frame, isolating finger-articulation smoothness from gross arm sweep. For each hand $h \in \{L, R\}$ we build a right-handed orthonormal wrist frame $R^h(t) \in \text{SO}(3)$ via Gram–Schmidt on the index-MCP and pinky-MCP joints relative to the wrist root $\mathbf{w}^h(t) \in \mathbb{R}^3$ (forward axis: wrist→index-MCP; transverse: index-MCP/pinky-MCP), express each joint locally as

$$\mathbf{x}_j^{h,\text{loc}}(t) = R^h(t)^\top (\mathbf{x}_j^h(t) - \mathbf{w}^h(t)) \in \mathbb{R}^3,$$

and apply the same central-difference triple-derivative as above to obtain $\ddot{\mathbf{x}}_{t,j}^{\text{loc}}$ for the joint j in its own hand’s local frame:

$$\text{IntraJerk} = \frac{1}{TJ} \sum_{t=1}^T \sum_{j=0}^{J-1} \|\ddot{\mathbf{x}}_{t,j}^{\text{loc}}\|_2 \quad (\text{m/s}^3; \text{target: match GT}).$$

The wrist derotation removes the gross arm-sweep contributions that otherwise dominate the third derivative on this wide-workspace bimanual dataset (see World-space Jerk below).

Inter-hand Penetration Ratio ($\downarrow$, lower is better). Fraction of frames in which the left and right meshes interpenetrate by at least 1 mm:

$$\text{InterPen} = \frac{1}{T} \sum_{t=1}^T \mathbb{I} \left[\max_{v \in \mathbf{V}^L(t)} (-d_t^R(v)) \geq 1 \text{ mm} \vee \max_{v \in \mathbf{V}^R(t)} (-d_t^L(v)) \geq 1 \text{ mm} \right] \quad (\%, \downarrow),$$

where $d_t^h(v)$ is the signed distance from query vertex v to the hand- h mesh at frame t (positive outside, negative inside; $h \in \{L, R\}$), $\mathbf{V}^h(t)$ is the vertex set of mesh h at frame t , $\vee$ is logical OR, and $\mathbb{I}[\cdot]$ is the indicator function defined in the semantic-metric block above. The wrist-cap closure adds a small (~ 0.5 – 1%) baseline that affects all rows uniformly.

Self-hand Penetration MAX ($\downarrow$, lower is better). Per-sample maximum signed-distance penetration of any joint into its own hand’s mesh, averaged across the S test motions. With $\mathbf{x}_j^h(t)$ as defined above and $d_{t,\text{rest}}^h(\cdot)$ the signed distance to the mesh of hand h at frame t , computed with the local rest neighbourhood of the queried joint j excluded so a joint cannot be flagged as penetrating its own bone:

$$\text{SelfPen}_{\max} = \frac{1}{S} \sum_{s=1}^S \max_{t,h,j} \max(0, -d_{t,\text{rest}}^h(\mathbf{x}_j^h(t))) \quad (\text{mm}, \downarrow),$$

where the inner $\max(0, \cdot)$ keeps only frames/joints that lie inside the mesh ($-d > 0$). This depth-valued variant replaces the BimArt [43] “Pen 1 cm” frame-ratio in the headline Table 3 because every FK-correct generator collapses to 0% on the ratio with overlapping CIs, while the depth varies continuously; the BimArt ratio is retained as a supplementary axis in Table 10.

HF Power Ratio ($\downarrow$, lower is better). Per-joint position power above $f_c = 8$ Hz, normalised by total power and averaged over joints. Let $\hat{\mathbf{x}}_j(f) \in \mathbb{C}^3$ be the temporal Fourier transform of $\mathbf{x}_j(\cdot)$ over the clip’s T frames, evaluated at frequency f in the discrete spectrum $\{0, 1/(T\Delta t), \dots, 1/(2\Delta t)\}$ (Nyquist $1/(2\Delta t) = 15$ Hz at 30 fps), and $|\hat{\mathbf{x}}_j(f)|^2 = \sum_{d=1}^3 |\hat{x}_j^{(d)}(f)|^2$ the squared magnitude summed across the three spatial channels:

$$\text{HFPR} = \frac{1}{J} \sum_{j=0}^{J-1} \frac{\sum_{f \geq f_c} |\hat{\mathbf{x}}_j(f)|^2}{\sum_f |\hat{\mathbf{x}}_j(f)|^2} \in [0, 1], \quad \downarrow.$$

Real hand motion has a roughly $1/f$ spectrum dropping sharply above ~ 5 Hz; excess power above 8 Hz signals noise or token-by-token decoding artefacts. HFPR complements Acceleration and Intra-hand Jerk by being normalised for overall motion energy.

Supplementary axes (Table 10). Table 10 reports four supplementary physical axes that complement the headline numbers from Table 3 without replacing them. *Inter-hand Penetration Depth* (mm; $\downarrow$) is the mean over interpenetrating frames of the maximum left/right interpenetration depth which is a continuous-valued complement to the binary headline Inter-hand Penetration Ratio. *Self-hand Penetration Ratio* (% of frames; $\downarrow$) is the original BimArt [43] “Pen 1 cm” convention (fraction of frames where any joint penetrates its own-hand mesh by > 10 mm), retained here for cross-paper comparability although it discriminates poorly on RoMo Hands. *Bone-length Violations* (% of frames; $\downarrow$) is the fraction of frames where any bone deviates from its MANO template length by more than a per-joint threshold (50/35/30 mm at root/PIP/DIP, calibrated from the empirical RoMo Hands maximum); FK-correct generators score 0% by construction, while discrete-token models drift token-by-token. *Finger Flexion Mean* (m; target: GT) is the time-and-finger mean of fingertip-to-MCP distance, a coarse posture proxy that catches the “flat hand” failure mode.

Reference rows. Both physical tables include italicised GT rows for RoMo Hands, GRAB [29], and GigaHands [6]; the latter two are streamed through the same $\mathcal{R}_{\text{hand}}$ decoder and scoring routine as our generators. GRAB (slow sit-and-grasp on a small workspace) provides a clean lower bound for finger-articulation smoothness, while GigaHands is closer to RoMo Hands in workspace and variety. Reference rows are excluded from per-column “best” bolding among generators and serve only to anchor the absolute scale of the physical metrics, which are unitful and otherwise hard to interpret in isolation.

D Full Experimental Results

Table 9 extends the semantic half of Table 3 with the full retrieval grid (R@1, R@2, R@3, R@5, R@10) plus MM-Dist, FID, and Diversity, all with 95% CIs. Table 10 adds the supplementary physical axes (inter-hand penetration depth, BimArt > 10 mm self-penetration ratio, bone-length violations, finger flexion mean) that are useful for cross-checking the headline numbers and for cross-paper comparability. The per-caption-level GT breakdown referenced in Section 4.2 follows. Unless noted, all semantic evaluations use the TMR evaluator trained with type-aware normalization and the SnapMoGen pool-of-32 protocol with 20 replications.

Table 9: **Full semantic scorecard** (all $R@k$, MM-Dist, FID, Diversity) for the generators we benchmark in the paper body. Means $\pm 95\%$ CI. MotionMillion = v6, CFG 5.0 (matches Table 3). Diversity is reported with “ $\rightarrow$ ” to indicate that closer to the ground-truth value is better.

Method	R@1 $\uparrow$	R@2 $\uparrow$	R@3 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	MM-Dist $\downarrow$	FID $\downarrow$	Diversity $\rightarrow$
Ground Truth	63.77 $\pm$ 0.34	76.25 $\pm$ 0.31	82.59 $\pm$ 0.29	89.49 $\pm$ 0.23	96.13 $\pm$ 0.11	69.04 $\pm$ 0.02	0.000 $\pm$ 0.000	120.41 $\pm$ 0.41
MDM (DiP) [30]	42.28 $\pm$ 0.21	57.30 $\pm$ 0.42	66.15 $\pm$ 0.43	76.95 $\pm$ 0.28	88.92 $\pm$ 0.15	80.19 $\pm$ 0.02	0.144 $\pm$ 0.000	113.44 $\pm$ 0.40
MoMask [8]	54.00 $\pm$ 0.41	70.68 $\pm$ 0.36	78.62 $\pm$ 0.24	86.07 $\pm$ 0.24	93.88 $\pm$ 0.14	74.52 $\pm$ 0.02	0.129 $\pm$ 0.000	115.27 $\pm$ 0.34
Kimodo M [24]	53.82 $\pm$ 0.37	69.04 $\pm$ 0.29	76.99 $\pm$ 0.28	85.56 $\pm$ 0.16	93.76 $\pm$ 0.13	73.67 $\pm$ 0.02	0.087 $\pm$ 0.000	116.34 $\pm$ 0.34
MotionMillion [4]	31.26 $\pm$ 0.46	43.97 $\pm$ 0.46	51.55 $\pm$ 0.43	63.40 $\pm$ 0.42	82.37 $\pm$ 0.46	87.10 $\pm$ 0.000	0.314 $\pm$ 0.000	109.14 $\pm$ 0.49

Table 10: **Supplementary physical metrics on the RoMo Hands test set.** Complements the physical half of the headline Table 3 with the mean inter-hand penetration depth, the original BimArt [43] > 10 mm self-hand penetration frame ratio (retained for cross-paper comparability), bone-length violation rate (token-by-token AR-drift signal), and a finger-flexion posture proxy that catches the “flat hand” failure mode. Means $\pm 95\%$ CI over 20 replications on a 256-sequence subset (stride 4).

Method	Inter-hand Pen Depth (mean, mm; $\downarrow$)	Self-hand Pen Ratio (% frames > 10 mm; $\downarrow$)	Bone-length Viol. (% frames; $\downarrow$)	Finger Flexion (mean, m; target: GT)
Ground Truth	0.081 $\pm$ 0.052	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.1120 $\pm$ 0.0010
GRAB GT (ref.) [29]	0.000 $\pm$ 0.000	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.1217 $\pm$ 0.0007
GigaHands GT (ref.) [6]	0.000 $\pm$ 0.000	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.1251 $\pm$ 0.0005
MDM (DiP) [30]	0.243 $\pm$ 0.112	1.05 $\pm$ 0.84	1.93 $\pm$ 0.28	0.1035 $\pm$ 0.0014
MoMask [8]	0.190 $\pm$ 0.170	0.07 $\pm$ 0.14	6.34 $\pm$ 0.56	0.1136 $\pm$ 0.0016
Kimodo M [24]	0.150 $\pm$ 0.093	0.00 $\pm$ 0.00	1.54 $\pm$ 0.19	0.1127 $\pm$ 0.0014
MotionMillion [4]	0.200 $\pm$ 0.114	0.00 $\pm$ 0.00	0.08 $\pm$ 0.02	0.0906 $\pm$ 0.0012

Representation-ablation protocol. Table 12 reports the generation-side ablation of $\mathcal{R}_{\text{hand}}$ against raw MANO-102 and the CLUTCH-198 representation. The Kimodo M variants share one sweep protocol, each reported at its own best validation checkpoint. This differs from the original Kimodo M run behind Table 3, hence the higher absolute numbers ($R@1$ 0.665 versus 0.538); comparisons within the table are unaffected, and Table 3 keeps the original runs so that all four generators are reported under one shared protocol. A further Kimodo M variant isolates the redundant channels: removing the velocity and heading channels (289-D) lowers $R@1$ to 0.639, level with the rotation-only baselines, so the velocity and heading channels, rather than the joint positions, carry the improvement.

Reading the caption-level breakdown. Table 14 shows a non-monotone shape across $L_0 \rightarrow L_4$: GT retrieval is weakest at L_0 (single-action verbs too coarse to separate 31 distractors), peaks at L_2 – L_3 (atomic finger states with relations are the most discriminative granularity), and decreases at L_4 where verbose paragraph paraphrases reintroduce non-discriminative filler tokens. The “Random” column is the uniform-over-levels mixture used to train the TMR evaluator and to report numbers in the main paper; it sits below the per-level peak by design, conservatively averaging over all five levels and matching the realistic deployment regime where a downstream user does not know which granularity the system was trained against.

E Implementation Details

This appendix expands Section 4.1 with the architectures, optimization, and training schedules of every model in this work. We document *algorithmic* implementation detail (model size, batch size, LR, schedule, duration, design decisions vs. published defaults); compute infrastructure is intentionally omitted.

E.1 Common Setup

Dataset. All RoMo Hands models train on the released LitData shards at 30 fps (an 85%/10%/5% train/val/test split of the 541,929 clips), with clips longer than 196 frames (6.5 s) randomly cropped to 196 frames and clips shorter than 40 frames (1.3 s) zero-padded to 40, 32 bimanual joints, and 5 paraphrased captions per clip sampled uniformly during training.

Table 11: **Human ratings support caption fidelity across all five detail levels.** Twelve raters produced 370 ratings over 250 randomly sampled validation clips. Questions (i) and (ii) assess overall action match and hand-specific caption–motion faithfulness on a five-point Likert scale; the final two columns report the fraction rated at least 4.

Caption level	N	(i) Mean	(i) ≥ 4	(ii) Mean	(ii) ≥ 4
Tag	74	4.51	89%	3.96	69%
Short	74	4.19	85%	3.81	70%
Medium	74	4.15	82%	3.76	64%
Detailed	74	4.03	78%	3.70	64%
Paragraph	74	3.92	72%	3.74	65%
All	370	4.16	81%	3.79	66%

Table 12: **Full representation-ablation results.** Each model is trained under a representation-matched sweep with all non-representation choices fixed within its block. Diversity should approach the corresponding GT value (MoMask: 120.41; Kimodo M: 120.95). CLIP was not reported for the MoMask sweep. Best within each model block is **bold**.

Generator	Representation	R@1 $\uparrow$	R@3 $\uparrow$	R@10 $\uparrow$	FID $\downarrow$	Diversity $\rightarrow$	MM-Dist $\downarrow$	CLIP $\uparrow$
MoMask	$\mathcal{R}_{\text{hand}}\text{-}393$	0.547	0.794	0.948	0.153	114.99	0.941	–
	MANO-102	0.395	0.658	0.888	0.221	111.37	1.049	–
	CLUTCH-198	0.368	0.617	0.862	0.269	109.95	1.077	–
Kimodo M	$\mathcal{R}_{\text{hand}}\text{-}393$	0.665	0.871	0.971	0.114	116.09	0.868	0.601
	MANO-102	0.646	0.868	0.980	0.135	115.15	0.896	0.581
	CLUTCH-198	0.646	0.856	0.974	0.106	117.17	0.882	0.590

Motion representation and normalization. Every model consumes the same $\mathcal{R}_{\text{hand}}$ feature (195 GAUSSIAN + 194 CYCLIC + 4 CONSTANT = 393 dims) with the type-aware normalization of Section B.4 applied at the dataloader stage. Cached (μ_i, σ_i) are computed on the train split and frozen across all downstream eval; the TMR evaluator is fit under the same normalization so retrieval-based metrics live in a representation-consistent latent space.

Text encoding. A frozen T5-v1.1-base [23] encoder runs inline on GPU (bf16 buffer, fp32 params), tokenized to max 256 tokens (180 for MotionMillion, capped to its FSQ token budget). Kimodo M consumes the pooled output (768-D); all other models cross-attend to the full token sequence. For each clip we hold 5 paraphrased captions of escalating verbosity, and at training time sample uniformly across a 5-bucket length schedule (`max_lengths = [16, 32, 96, 192, 256]`); val cycles through buckets round-robin for length balance.

Optimization defaults (overrides noted per-model). AdamW [15], $\beta = (0.9, 0.99)$, weight decay 0 for diffusion / tokenizer heads and 0.01 for LM-style heads (TMR, MotionMillion LLaMA, Kimodo M). bf16-mixed precision; gradient clipping at 0.5 (MDM, TMR) or 1.0 (others). EMA $\beta = 0.999$ every 10 steps on diffusion models only. `torch.compile` enabled by default; disabled on MoMask Stages 1/2 due to recompile thrash from BERT-style mask-ratio sampling. Validation every 5 epochs (every 1 epoch for TMR).

E.2 Per-model footprint at convergence

Table 15 summarizes the training footprint of every component at validation convergence (bf16-mixed throughout; “effective batch” is per-GPU batch $\times$ #GPUs with no gradient accumulation). All four generators are sized inside a single capacity decade (40M–150M trainable text-conditional parameters), so the cross-architecture comparison in Section 4.2 reflects representational fit rather than raw scale. MotionMillion is the only entry that required a material change from its published configuration: its 1B-parameter LLaMA-style decoder would have been an order of magnitude larger than the next-biggest generator and would have confounded the comparison, so we trained the 150M variant which was the smallest published scale at which the autoregressive recipe still converges

Table 13: **Removing source-video overlap preserves the benchmark ranking.** Metrics are recomputed after removing every test clip whose source video appears in the training set. Parentheses give the change from the corresponding segment-level result in Table 3; R@1 is shown as a fraction here.

Method	R@1 $\uparrow$	FID $\downarrow$	MM-Dist $\downarrow$	Diversity $\rightarrow$
MoMask	0.549 (+0.009)	0.158 (+0.029)	0.922 (−0.023)	113.9 (−1.4)
Kimodo M	0.545 (+0.007)	0.116 (+0.029)	0.920 (−0.022)	114.2 (−2.1)
MDM (DiP)	0.405 (−0.018)	0.171 (+0.027)	1.030 (−0.001)	114.0 (+0.6)
MotionMillion	0.287 (−0.026)	0.322 (+0.008)	1.130 (+0.025)	107.7 (−1.4)
<i>Ground Truth</i>	<i>0.597 (−0.041)</i>	<i>0.000 (+0.000)</i>	<i>0.880 (+0.016)</i>	<i>119.1 (−1.3)</i>

Table 14: **Mid-length captions (L2–L3) retrieve best;** the verbose L4 full-paragraph captions hurt slightly versus mid-length descriptions. Levels L0–L4 run from short headline $\rightarrow$ full paragraph; *Random* matches Table 3 (uniform level draw per replication).

Metric	L0	L1	L2	L3	L4	Random (uniform)
R@1 $\uparrow$	53.02 $\pm$ 0.41	65.04 $\pm$ 0.40	68.43 $\pm$ 0.40	66.61 $\pm$ 0.38	65.67 $\pm$ 0.26	63.77 $\pm$ 0.34
R@2 $\uparrow$	67.52 $\pm$ 0.34	77.21 $\pm$ 0.32	80.08 $\pm$ 0.28	79.03 $\pm$ 0.24	78.27 $\pm$ 0.31	76.25 $\pm$ 0.31
R@3 $\uparrow$	75.22 $\pm$ 0.26	83.16 $\pm$ 0.29	85.50 $\pm$ 0.32	84.75 $\pm$ 0.26	84.23 $\pm$ 0.26	82.59 $\pm$ 0.29
R@5 $\uparrow$	83.69 $\pm$ 0.24	89.78 $\pm$ 0.23	91.57 $\pm$ 0.23	91.27 $\pm$ 0.26	90.91 $\pm$ 0.20	89.49 $\pm$ 0.23
R@10 $\uparrow$	93.53 $\pm$ 0.18	96.02 $\pm$ 0.15	96.51 $\pm$ 0.10	96.79 $\pm$ 0.10	96.70 $\pm$ 0.11	96.13 $\pm$ 0.11
MM-Dist $\downarrow$	73.32	68.24	66.67	67.86	68.72	69.04
CLIP score $\uparrow$	0.544	0.609	0.628	0.615	0.606	0.600
MM-Dist _{norm} $\downarrow$	0.924	0.852	0.831	0.848	0.860	0.864
FID $\downarrow$	0.000	0.000	0.000	0.000	0.000	0.000
Diversity	120.41 $\pm$ 0.41 (level-independent: pooled across all generated motions)					

to non-trivial accuracy on a 15K-token FSQ codebook. Appendix E.5 unpacks each per-generator adaptation; detailed architectures and schedules follow.

Table 15: Per-model training footprint at validation convergence. Cross-references point to the full per-model subsections below.

Model	Family	Params	Eff. batch	LR ($\times 10^{-4}$)	Val loss / acc.
Kimodo M (§E.3.1)	Diffusion (DDPM/DDIM)	80M	1024	0.20	0.0810
MDM DiP (§E.3.2)	Diffusion (prefix)	40M	512	3.27	0.0469
MoMask Stage 0 RVQ (§E.3.3)	VQ tokenizer (6×512)	10M	2048	2.83	0.026 (rec.)
MoMask Stage 1 mask (§E.3.3)	Masked transformer	12M	1536	3.46	3.79 / 25.9%
MoMask Stage 2 res. (§E.3.3)	Residual transformer	12M	1536	3.46	4.50 / 10.4%
MotionMillion FSQ (§E.3.4)	FSQ tokenizer (15.4K)	12M	512	2.00	0.0171 (rec.)
MotionMillion LLaMA (§E.3.4)	Causal LM (RoPE)	150M	1024	2.45	7.36 / 6.4%
TMR evaluator (§E.4)	Dual encoder (InfoNCE+KL)	80M	768	2.50	6.38

E.3 Generative Models

E.3.1 Kimodo M (text-to-motion diffusion, paper-M target)

Architecture. Single-stack transformer denoiser (8 layers, 8 heads, latent 1024, dropout 0.1, max length 512 tokens, 49 learnable conditioning tokens prepended). Text condition is the pooled T5-v1.1-base vector (768-D) projected and concatenated with the conditioning block. Attention is SDPA (PyTorch 2 fused).

Diffusion. 1000-step training schedule; 100-step DDIM at inference; CFG dropout 0.1, CFG weight 2.5. Variance fixed.

Optimization. AdamW ($\beta = (0.9, 0.99)$, no decay), LR 2×10^{-5} held flat (Kimodo [24] does not LR-scale with batch). EMA $\beta = 0.995$, update every 10 steps. Effective batch 1024 (256×4 GPUs), matching the published “medium” config. The registered checkpoint is at epoch 1919 of the 2300-epoch budget (val/loss 0.0810); val/loss reached its 0.082 floor ~ 700 epochs earlier and the tail was within the noise band.

E.3.2 MDM (Motion Diffusion Model)

We adopt the DiP (Diffusion-in-Prefix) variant of MDM [30], based on the `trans_dec` architecture (transformer *decoder* denoiser, conditioning concatenated with the noisy motion sequence as a prefix token) rather than MDM’s original encoder-style architecture, which we found underperformed in earlier sweeps. The model is conditioned on a 20-frame prefix and predicts the next 40 frames rather than generating the full 196-frame clip unconditionally; this collapses most of the multi-modal posterior and lets a much shorter sampling schedule suffice.

Architecture. 8 layers, 8 attention heads, hidden dim 512, feed-forward 2048, dropout 0.1, GELU activations, SDPA attention. The text condition is the full T5-v1.1-base token sequence cross-attended (via prefix tokens) into every layer. Per-channel loss weights match the representation: position dims (9–104) get weight 2; root XZ velocity and height (0–2) get weight 10 (under-represented but high-impact for trajectory).

Sequence shape. Fixed window of 20 frames context + 40 frames prediction = 60 total, with both ends pinned (`min_seq_length` = `max_seq_length` = 60). Inference requires a prefix tensor (in practice: the ground-truth prefix from the eval pipeline).

Diffusion. Cosine 50-step DDPM training schedule with variance fixed-small, CFG dropout 0.1, and CFG weight 2.5 at sampling. At validation we reduce to 10 DDPM steps: prefix-conditioned generation empirically needs an order of magnitude fewer denoising steps than the full-sequence variant.

Optimization. AdamW, learning rate 3.27×10^{-4} (the batch-size-scaled paper baseline), gradient clip 0.5, EMA on ($\beta = 0.999$, update every 10 steps). Per-GPU batch 256 across 2 GPUs (effective batch 512). 200 epochs. Best val/loss = 0.04689 at epoch 74.

Scale-down rationale. The $196 \rightarrow 60$ -frame reduction is a deliberate representation simplification, not a compute compromise: when the model predicts only 40 frames given 20 frames of context, the DDPM posterior is far less multimodal and 10 denoising steps suffice.

E.3.3 MoMask (residual VQ + masked transformer)

MoMask [8] is a 3-stage pipeline: (Stage 0) train a residual VQ tokenizer that maps motion to a discrete sequence; (Stage 1) train a masked transformer that fills in masked tokens conditional on text; (Stage 2) train a residual transformer that refines the higher-residual codes given the Stage 1 output. Stage 1 and Stage 2 are independent given a frozen Stage 0 tokenizer and were trained in parallel on independent jobs.

Stage 0: RVQ tokenizer. Convolutional encoder (`down_t` = 2, stride 2 $\rightarrow$ token rate 7.5 Hz), 6-quantizer residual VQ with 512 codes per quantizer (per-quantizer codebooks, not shared), code dim 512, EMA codebook updates ($\mu = 0.99$), 20% code dropout. 64-frame training window. Text-independent; AdamW, LR 2.83×10^{-4} , no decay/EMA, step LR at iters {150K, 250K} with $\gamma = 0.05$. 200 epochs at single-GPU batch 2048. Reconstruction val/loss 0.026, codebook perplexity 156 (moderate utilization typical for first-stage VQs at this dataset scale).

Stage 1: Masked transformer. Operates on quantizer 1. At each step a cosine-scheduled fraction of tokens is replaced by [MASK] and the model predicts originals given remaining context and the text caption. *Architecture:* 8 layers, 6 heads, latent 384, FFN 1024, dropout 0.2, max length 55 tokens. Text is the full T5 sequence cross-attended into every layer; caption-drop 0.1. *Optimization:* AdamW, LR 3.46×10^{-4} , step at iters {72K, 100K} with $\gamma = 0.1$, 1K linear warmup; 120 epochs at effective batch 1536 (768×2 GPUs). Val/loss 3.79, accuracy 25.85%. Batch was scaled 1.5 $\times$ over the H200 baseline with $\sqrt{1.5}$ -scaled LR; dropout was bumped from MoMask’s 0.1 to 0.2 because the hand dataset is $\sim 5\times$ smaller than HumanML3D (+1.5 pp val accuracy). `torch.compile` was disabled after observing recompile thrash from BERT-style mask-ratio sampling.

Stage 2: Residual transformer. Operates on quantizers 2–6 conditional on the Stage-1 prediction. Same architecture as Stage 1 except: conditioning input is the Stage-1 token sequence (text still

cross-attended), `cond_drop_prob`=0.2, weights shared across the 5 residual heads. *Optimization*: identical schedule to Stage 1; val/loss 4.50, accuracy 10.39%. Stage 2 was still on a slow downward slope at epoch 120 but the schedule was kept identical for parallel-launch simplicity; extending to 200 epochs would likely move accuracy to 13–15%. Stages 1 and 2 share no weights and were trained in parallel on disjoint 2-GPU shards once the Stage-0 checkpoint was available.

E.3.4 MotionMillion (FSQ tokenizer + LLaMA AR decoder)

A two-stage tokenizer + autoregressive language model, modeled after MotionMillion [4]. We trained the small (150M-parameter) variant.

Stage 1: FSQ tokenizer. Finite Scalar Quantization [17] replaces MoMask’s residual VQ; FSQ partitions a low-dim latent into discrete bins per axis, sidestepping codebook collapse. *Architecture*: the same convolutional encoder as MoMask’s Stage 0 (`down_t` = 2, token rate 7.5 Hz, width 512, depth 3, 64-frame window), with 5-axis FSQ at levels [8, 8, 8, 6, 5] for an implicit codebook of 15,360 tokens ($\approx 30\times$ MoMask’s 512). *Optimization*: AdamW, LR 2×10^{-4} , step at iters {150K, 250K} with $\gamma=0.05$; 600 epochs at single-GPU batch 512. Val/loss 0.0171 at epoch 569, about $4\times$ lower reconstruction loss than the MoMask Stage-0 RVQ, as expected for a $30\times$ codebook.

Stage 2: LLaMA AR decoder. Causal LLaMA-style [32] model predicting the next FSQ token given (text, previous tokens). *Architecture*: 12 layers, 12 heads (no GQA), hidden 768, intermediate 2048, RoPE [28] $\theta=10K$, RMSNorm, SDPA. Max motion-token length 64 (≈ 8.5 s). Text condition is the full T5 sequence (180-token cap) cross-attended rather than concatenated, keeping the AR prefix short. CFG drop 0.1, CFG scale 2.5. *Optimization*: AdamW, LR 2.45×10^{-4} ($\sqrt{1.5}$ -scaled from the H200 baseline), weight decay 0.01, gradient clip 1.0, step at iters {30K, 60K} with $\gamma=0.1$, 2K warmup; effective batch 1024 (128×8 GPUs). `torch.compile` enabled on the inner LlamaModel only. *Outcome*: best val/loss 7.36 (accuracy 6.36%) at epoch 79 of a 250-epoch budget with early stopping (patience 15 on val/loss); stop fired at epoch 154. Train accuracy at the best checkpoint was $\sim 58\%$, indicating the 150M variant overfits this dataset. The original MotionMillion uses a 1B-parameter decoder; we started with 150M to de-risk the FSQ integration, and the train/val gap is the green light for capacity scale-up. 1B and 3B configs are checked in but untrained.

E.4 Evaluation: TMR retrieval evaluator

The TMR model [20] is the *evaluator*, not a generator: once trained it is the canonical scorer for every text-to-motion model in this work via $R@k$ retrieval and FID.

Architecture. Dual encoder with a shared transformer backbone, consuming the 201-D essential subset of $\mathcal{R}_{\text{hand}}$ (root frame + per-joint 6D rotations only; Section B.6). Both branches encode into a 768-D joint latent space. 8 layers, 8 heads, FFN 3072, dropout 0.1, SDPA; T5-v1.1-base inline (256-token cap, bf16 buffer).

Loss. Standard TMR combination: reconstruction ($\lambda_{\text{rec}} = 1.0$), motion-latent KL ($\lambda_{\text{KL}} = 10^{-5}$), encoder regularization ($\lambda_E = 10^{-5}$), text–motion latent alignment ($\lambda_{\text{latent}} = 10^{-5}$), and InfoNCE [33] contrastive across the batch ($\lambda_{\text{NCE}} = 0.1$, temperature 0.1). Unlike SnapMoGen [9], we do not use the `threshold_selfsim=0.80` paraphrase mask: in hyperparameter sweeps, removing the mask and using AdamW weight decay 0.01 (vs. SnapMoGen’s 10^{-5}) gave the best $R@1$ by $\sim 3\text{--}4$ pp on the held-out val set.

Optimization. AdamW ($\beta = (0.9, 0.99)$), LR 2.5×10^{-4} , weight decay 0.01, gradient clip 0.5 (tighter than diffusion models because contrastive loss is sensitive to spikes from hard negatives), 1500-step linear warmup, step LR at epochs {40, 60} with $\gamma=0.1$. 100 epochs at effective batch 768 (384×2 GPUs); `torch.compile` enabled. Validation every epoch (rather than every 5) so the best-by- $R@1$ checkpoint is reliably captured. Val/loss 6.38 at epoch 98; the top-6 checkpoints across epochs 93–99 are within 0.004 val/loss of each other. *All retrieval-based metrics ($R@k$, FID, Diversity, MModality) in this paper are computed against this evaluator.*

E.5 Matched-budget scaling across architectures

The cross-architecture comparison in Section 4.2 is a comparison of *representational fit* between four generator families on the same data, not a comparison of compute. To keep that comparison fair we sized every generator under two principles: (i) *matched to the data*: depth, batch, and schedule are tuned to RoMo Hands’ 85% train split rather than carried over from each method’s body-motion default; and (ii) *matched to one another*: generator parameter counts are kept inside a single decade ($\sim 40\text{M}$ – 150M) so any ranking we report reflects how each architecture handles the bimanual representation, not raw capacity. Tokenizers (MoMask RVQ, MotionMillion FSQ) are exempt from (ii) because their objective is reconstruction quality and codebook coverage, not text-conditional generation; we keep them at their published sizes and let the codebook do the work.

The headline consequence is the MotionMillion entry: its as-published 1B-parameter decoder would have been $\sim 7\times$ larger than the next biggest generator and would have confounded the cross-architecture comparison, so we trained the 150M variant in this work instead. Table 16 summarizes the resulting band.

Table 16: **Generator capacities are matched to RoMo Hands’ size and to one another within a single decade ($\sim 40\text{M}$ – 150M).** “Generator params” counts only the trainable text-conditional model; tokenizers (RVQ, FSQ) are excluded because they are not part of the cross-architecture comparison. The only material capacity change versus the published configurations is MotionMillion’s 1B $\rightarrow$ 150M scale-down.

Generator	Generator params	Published params	Adaptation to RoMo Hands
Kimodo M	80M	80M	Paper “medium” config held; DDPM/DDIM schedule unchanged.
MDM (DiP)	40M	40M	Sequence shortened ($196 \rightarrow 60$), DDPM steps $50 \rightarrow 10$; prefix conditioning collapses the posterior.
MoMask Stage 1	12M	12M	Batch $\times 1.5$ with $\sqrt{1.5}$ -scaled LR; dropout $0.1 \rightarrow 0.2$ for the smaller dataset.
MoMask Stage 2	12M	12M	Same schedule as Stage 1; weights shared across the 5 residual heads.
MotionMillion	150M	1B	Decoder shrunk $\sim 7\times$ to keep the comparison inside the same parameter band.
TMR (evaluator)	80M	80M	Paraphrase mask removed; weight decay 0.01 (+3–4 pp R@1 on val).

Two clarifications. First, MoMask is reported as two rows because Stages 1 and 2 are independent transformers given the frozen Stage 0 tokenizer; they share the architecture and the schedule. The Stage 0 RVQ (10M params, 6×512 codes) and the MotionMillion FSQ (12M params, 15.4K-token codebook) are kept at their published sizes for the reasons given above. Second, the MotionMillion train/val accuracy gap reported in Section E.3.4 ($\sim 58\%/6\%$) is the green light for revisiting the 1B and 3B configurations once the rest of the pipeline is locked, but doing so under a fair-comparison protocol would also require pushing the other generators’ capacities up by a comparable factor; we leave that scale-up to future work.

E.6 Reproducibility

The RoMo Hands dataset, the action taxonomy, the five-tier captions, the TMR retrieval evaluator, and the per-baseline training configurations and checkpoints reported in this paper will be publicly released through the project page under the licenses listed in Section F. Together with the architectures, optimizers, and training budgets documented in Section E, this release will allow any single result in Table 3 to be reproduced from the same starting point.

F Licenses for External Assets

The third-party assets used in this work are listed below. Each is cited in the main paper and used under its published license.

- **RoMo** [40]: source video dataset and the curation pipeline we extend. Non-commercial research.
- **WiLoR** [21]: monocular bimanual hand detector and pose estimator used in the same pipeline. Non-commercial research.
- **MANO** [25]: parametric hand model that defines the joint hierarchy and blendshapes underlying our motion representation. MPI non-commercial research license.
- **SMPL-X** [19]: full-body parametric human model that we use to source the SMPL-X-aligned wrist trajectories in the wrist-anchoring step. MPI non-commercial research license.
- **Qwen3.5-27B** [22]: VLM used to generate the five-level captions. Apache License 2.0 accompanying the released checkpoint.
- **T5-v1.1-base** [23]: text encoder inside the TMR retrieval evaluator. Apache License 2.0.
- **MDM** [30]: motion-diffusion-model reference implementation, used with the DiP variant adaptations. MIT License.
- **MoMask** [8]: momask-codes reference implementation, used with our per-baseline adaptations. MIT License.
- **MotionMillion** [4]: MotionMillion-Codes reference implementation by Shanghai AI Laboratory, used with the 150M-variant adaptations. Apache License 2.0. (Note: the released MotionMillion *dataset* is CC BY-NC-SA 4.0, but we do not use it; we train on RoMo Hands.)
- **Kimodo** [24]: NVIDIA’s nv-tlabs/kimodo reference implementation, used with the $\mathcal{R}_{\text{hand}}$ I/O dimension change. Apache License 2.0 for the codebase; the separately licensed pretrained checkpoints (NVIDIA Open Model License) are not used, since we train from scratch on RoMo Hands.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The contributions stated in the abstract and at the end of Section 1 (the dataset, the captions and taxonomy, and the benchmark) match the artifacts and results reported in Section 3 and Section 4.

Guidelines:

- The answer [\[N/A\]](#) means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [\[No\]](#) or [\[N/A\]](#) answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: See Section 5, which covers in-the-wild noise, the capacity-bound MotionMillion run, the lack of hand-object context, and the absence of body and face.

Guidelines:

- The answer [\[N/A\]](#) means that the paper has no limitation while the answer [\[No\]](#) means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper has no formal theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The evaluation protocol and metric definitions are in Section 4.1, and per-model architectures, optimizers, and training budgets are documented in Section E. The dataset, the trained checkpoints, and the per-baseline training configurations will be released through the project page (Section E.6) so that any single result can be reproduced from the same starting point.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset, taxonomy, captions, TMR evaluator, and training configurations are released through the project page (<https://romodataset.github.io/romo-hands/>). The data pipeline (Section 3.1) and the per-model training recipes (Section E) are documented in enough detail to reproduce every reported number.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Splits, evaluator setup, and metric definitions are in Section 4.1. Per-model hyperparameters (optimizer, learning rate, EMA, batch size, training budget) are in Section E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Semantic metrics are means over 20 replications with 95% CIs (Section 4.2, Section C). The supplemental tables in Section D include the CIs as separate columns.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Per-model GPU count, batch size, precision, and training budget are reported in Section E, which is enough to reproduce each run on any equivalent accelerator. The exact cluster and GPU SKU are intentionally omitted, as noted at the start of that section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We follow the NeurIPS Code of Ethics. Source data is publicly available video, and the released artifact contains only MANO motion features and text (no biometrics or identifying content). The human rating exercise in Section 3.1 collected only non-sensitive technical judgments of displayed artifacts, did not constitute human-subjects research, and is documented below.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See the “Broader impact” paragraph at the end of Section 5, which covers both intended applications and the dual-use considerations of motion generation, and notes the de-identified release format we use to limit misuse.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We release MANO motion features and text only (no faces, identities, audio, or raw video), so the misuse risk is low enough that we do not gate access.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Every third-party asset we use is cited in the main paper and listed with its license in Section F. The RoMo Hands release will ship with a license file that respects the upstream non-commercial terms of MANO, SMPL-X, and the source-video pipeline.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The dataset, taxonomy, captions, and TMR evaluator are documented in Section 3, with the full motion representation and metric definitions in Section B and Section C; the public release plan is described in Section E.6. Intended use and scope are covered in Section 1 and Section 5.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The rating protocol and its three five-point questions are described in Section 3.1, with the per-level results in Table 11. Twelve adult raters ages 20–50, each with at least a bachelor’s degree and most without animation experience, provided uncompensated technical judgments of caption fidelity and reconstruction quality. The exercise was not crowdsourced and collected no personal or sensitive data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The rating exercise collected only non-sensitive technical judgments of displayed motion and caption artifacts, with no intervention or collection of personal data, and did not constitute human-subjects research. IRB review therefore did not apply.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use Qwen3.5-27B for captioning and to seed the action taxonomy (Section 3.1). LLMs are not used in the motion representation, the evaluator, or any of the generative baselines.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.