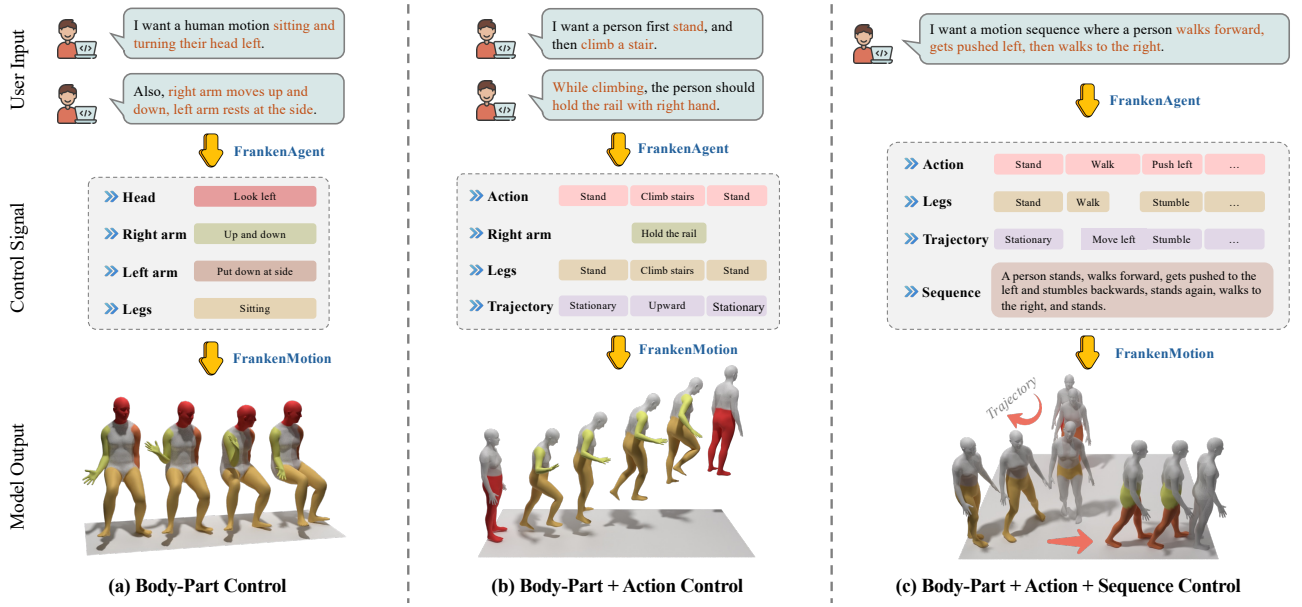
# FrankenMotion: Part-level Human Motion Generation and Composition

Chuqiao Li[1] Xianghui Xie[1,2] Yong Cao[1] Andreas Geiger[1] Gerard Pons-Moll[1,2]

[1]Tübingen AI Center, University of Tübingen, Germany

[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

https://coral79.github.io/frankenmotion/




**(a) Body-Part Control** **(b) Body-Part + Action Control** **(c) Body-Part + Action + Sequence Control**

Figure 1. Overview of our **FrankenMotion** framework. **Left**: Body-Part Control, where users specify fine-grained movements of individual body parts; **Middle**: Body-Part + Action Control, enabling coordinated whole-body actions with part-specific constraints; **Right**: Body-Part + Action + Sequence Control, supporting complex multi-stage motion sequences involving interactions and transitions. In all cases, **FrankenAgent** translates natural-language instructions into structured control signals for precise motion generation.

## Abstract

*Human motion generation from text prompts has made remarkable progress in recent years. However, existing methods primarily rely on either sequence-level or action-level descriptions due to the absence of fine-grained, part-level motion annotations. This limits their controllability over individual body parts. In this work, we construct a high-quality motion dataset with atomic, temporally-aware part-level text annotations, leveraging the reasoning capabilities of large language models (LLMs). Unlike prior datasets that either provide synchronized part captions with fixed time segments or rely solely on global sequence labels, our dataset captures asynchronous and semantically distinct part movements at fine temporal resolution. Based on this dataset, we introduce a diffusion-based part-aware motion generation framework, namely FrankenMotion, where each body part is guided by its own temporally-structured textual prompt. This is, to our knowledge, the first work to provide atomic, temporally-aware part-level motion annotations and have a model that allows motion generation with both spatial (body part) and temporal (atomic action) control. Experiments demonstrate that FrankenMotion outperforms all previous baseline models adapted and retrained for our setting, and our model can compose motions unseen during training. Our code and dataset will be publicly available upon publication.*

## 1. Introduction

Human motion generation is a fundamental task with broad applications in augmented reality (AR) and virtual reality

(VR), gaming, entertainment, and embodied AI [67, 71]. In recent years, significant progress has been achieved in this field, largely driven by the growing availability of motion capture (mocap) datasets and their corresponding textual annotations. [12, 17, 31, 38]. These advancements have enabled motion generation models conditioned on various modalities, including text, scene layouts, and object interactions [8, 19, 20, 52]. Concurrently, various extensions of the base task [4, 64], including motion editing, physics-aware generation, and style transfer, have been introduced.

However, existing methods struggle to achieve both temporal and part-level control, primarily due to limitations in both dataset annotation and model design. For example, models trained on mocap datasets can generate realistic motions [13, 17, 18, 46], yet they still lack fine-grained control over temporal dynamics and body parts. This is because most mocap datasets [30] lack temporally and part-level aligned annotations. To address this, several approaches [24, 62] have incorporated part-level labels to enhance model performance. Despite leveraging such information, these methods still fail to capture temporally coherent part-level features, which are crucial for fine-grained motion generation and controllable editing.

In this paper, we tackle motion generation from the perspective of composition. We argue that complex motions can be decomposed into simpler atomic motion elements. The key to allowing fine-grained control and complex motion generation lies in designing models that can learn these fundamental motion elements and their compositional relationships. We consider body parts as the basic elements in motion composition and design a model that maps prompts of body parts into complete temporal segments, the atoms of motion. We also input the high level action descriptions so that the model understands how different body parts compose into semantically meaningful complex motions.

Training such a model requires detailed per frame body part annotation, which is prohibitive to obtain. There exist datasets with textual annotations yet they lack structured part annotations. HumanML3D [12] and KITML [38] feature high level summary of a full motion sequence. Babel [39] contains action level annotations such as 'walk', 'stand', and 'knock'. Some actions, like 'raising arms', do describe what one part is doing yet the annotations are not structured, and the part movement of most actions, such as 'knock', remains unannotated. However, if a person sits down, the knees are probably bending; if a person is tying their shoes, their spine is bending, and the arms are tying the shoelaces. In other words, we can infer what the body parts are doing from high-level descriptions. Remarkably, we find that LLMs are powerful in inferring such relationships. Thus, we instantiate FrankenAgent, an LLM agent that consumes existing datasets and outputs coherent per-frame body part annotations together with high level annotations. Using this, we automatically annotate the largest part-level human motion dataset to this date, which we call FrankenStein dataset.

Using FrankenStein, we train FrankenMotion, a method that can be controlled at the sequence, action and part level as illustrated in Fig. 1. By constitutionality, we can create novel movements not seen during training, such as a person sitting while raising the left arm. Experiments show that our FrankenAgent is highly reliable, with 93.08% annotations considered correct by human experts. Results on our annotated FrankenStein also show that our model consistently outperforms state-of-the-art methods on part based motion generation in terms of both semantic correctness and realism. In summary, our main contributions are:

- We present FrankenMotion, a text-to-motion model that learns to compose complex motions through hierarchical conditioning on part-, action-, and sequence-level text, enabling fine-grained control over body parts and timing.
- We introduce FrankenStein, a new dataset with structured and temporally aligned body part annotations, leveraging existing datasets and LLM agent namely FrankenAgent.
- With hierarchical conditioning, our model allows flexible controlling at part, action or sequence level and generating novel motion compositions.

## 2. Related Work

**Motion generation with control.** Controllable motion generation has become increasingly important as it enables models to synthesize realistic motions aligned with user intent and environmental context. Among various modalities, text-conditioned motion generation [7, 8, 12, 13, 32, 35, 44, 46, 60, 61, 68] has gained popularity for its flexibility and expressiveness. Other forms of control include inter-person conditioning [29, 42, 43, 48] for modeling human–human interactions, key-pose or trajectory guidance [2, 66] for structural control, and audio-driven synthesis [22, 23, 65] for cross-modal alignment. Human–object interaction modeling [21, 50, 53–55, 57–59, 63] further introduces affordance reasoning and contact-based control, while scene-aware methods [19, 34, 47, 49] incorporate spatial constraints to ensure physical plausibility. Despite these advances, achieving fine-grained spatial and temporal control in motion generation remains a challenging open problem.

**Fine-grained Spatio-Temporal Motion Generation.** Achieving precise control in text-to-motion generation requires modeling both temporal and spatial structures. Existing approaches largely emphasize temporal control, regulating the duration and sequencing of motion segments at the frame level. Methods such as TEACH [1] compose motion segments using text and sequence-level conditioning, PriorMDM [44] refines transitions between short clips via a two-stage inference process, and FlowMDM [3] enhances temporal coherence by modeling smooth local

transitions. DART [68] employs a latent diffusion model for real-time, autoregressive motion generation, while UniMotion [20] introduces hierarchical control through both frame- and sequence-level conditioning. Beyond temporal modeling, spatial control has also gained attention. FineMoGen [62] supports diffusion-based motion generation and editing for fine-grained, per-body-part control, but its stage-based annotations enforce synchronized temporal intervals across parts, limiting flexibility. Besides, STMC [37] achieves spatial composition by assembling body-part motions from pretrained diffusion models, yet it remains a post-hoc method lacking end-to-end spatial–temporal reasoning. While these methods produce impressive results, none provide unified, fine-grained multi-level control over atomic body parts, atomic actions, and sequence-level semantics along the temporal axis.

**Motion understanding and annotation.** Contrary to motion generation, motion understanding with natural language is also an important task for human behavior analysis and motion data annotation. Earlier work PoseScript [9] generates text descriptions from pose parameters based on predefined rules. Follow up work MotionScript [56] extends PoseScript to generate texts for motions programmatically. Despite detailed descriptions, the generated texts are complex and less similar to natural human language. Recent line of works focus more on using LLMs to understand human poses [11, 25] or motions [5, 18, 33, 70]. A common practice is to fine tune the pre-trained language models to align text and motion in the latent space. These methods differ in the formulation of motion understanding as language translation [18] or interactive question answering [5, 33]. Some works can also handle raw videos [10, 51] in addition to 3D motions. In general, rule based annotation [9, 56] lacks human readability, while fine-tuned models [18] lose the generalization ability of original language models. Furthermore, these annotations focus on high-level action descriptions, lacking detailed body part decomposition to learn the essential motion elements. Our proposed annotation pipeline uses existing data and LLMs to generate atomic body-part level text prompts with temporal structure.

# 3. FrankenStein Dataset Construction

The quality and granularity of motion annotations directly constrain the upper bound of learning in motion understanding tasks. Existing datasets typically provide only coarse-grained labels, while lacking fine-grained temporal decomposition and body-part-specific annotations. For example, as shown in Fig. 2a, existing datasets only include sequence and action labels. To address this limitation, we propose a new annotation paradigm that explicitly operates at three levels of granularity: (1) **sequence level**: a global description of the full motion sequence, consistent with existing annotation formats; (2) **action level**: temporally localized coarse atomic actions; and (3) **body-part level**: fine-grained annotations for individual body parts (e.g., head, arms, legs, spine, trajectory) over time. This hierarchical annotation design enables a richer and more structured representation of human motion. As illustrated in Fig. 2b, the entire annotation process is automated, leveraging the reasoning capabilities of LLMs.

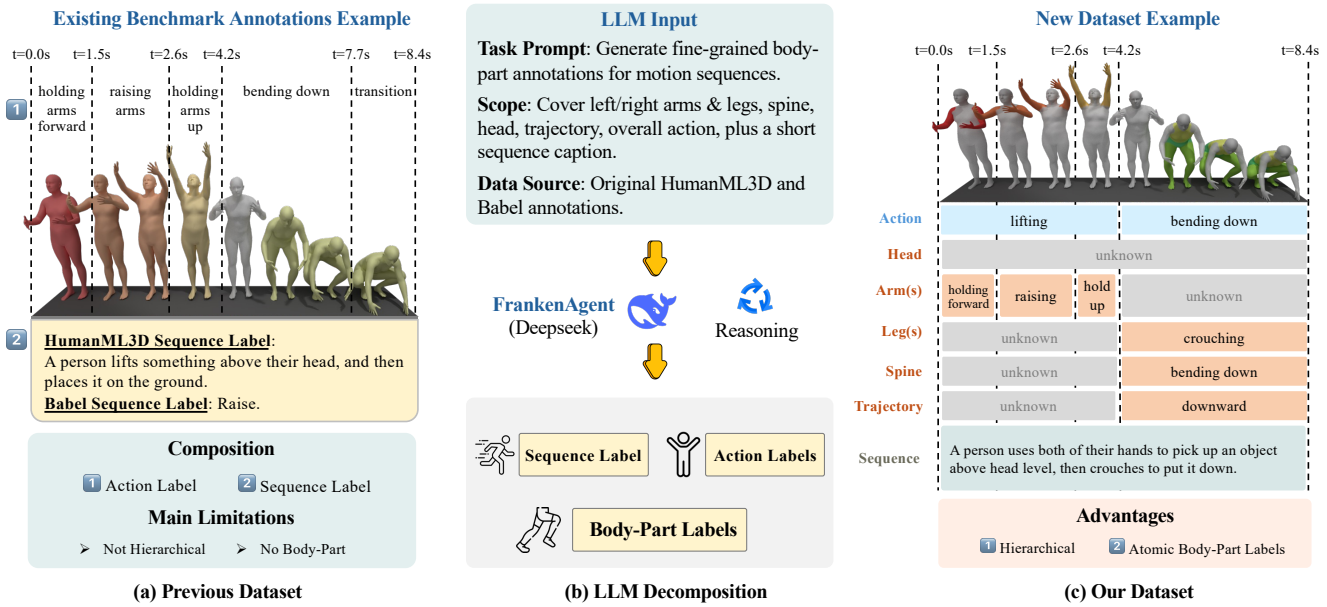
## 3.1. Data Source

Before presenting our automatic annotation pipeline, we first introduce several widely used motion–language datasets to which our approach can be applied: KIT-ML [38], BABEL [39], AMASS [30] and HumanML3D [12]. KIT-ML is one of the earliest datasets linking motion capture data with natural language. It combines sequences from the CMU and KIT mocap datasets, providing motion–text pairs that enabled early research on motion–language alignment. However, it remains small in scale and lacks temporal segmentation or part-level annotations. AMASS later unified motion capture data from 15 publicly available mocap datasets into a consistent parameterization of human motion, forming a large-scale foundation for subsequent motion–language datasets. Building on this resource, BABEL annotates a subset of AMASS with atomic action and coarse sequence-level labels, covering diverse activities but offering mainly categorical and short temporal annotations. Similarly, HumanML3D extends AMASS with crowdsourced natural language descriptions, providing richer semantics but limited temporal resolution and no part-level detail.

Although these datasets have propelled progress at the intersection of motion understanding and natural language, they share two critical shortcomings: (1) **lack of hierarchical structure**: existing annotations provide only coarse action or sequence labels without temporally decomposing motions into sub-actions or body-part levels; and (2) **absence of body-part granularity**: the datasets do not specify which body parts are responsible for each motion, limiting fine-grained understanding.

## 3.2. LLM-based Annotation Framework

To address these limitations, we aim at building a motion dataset with structured text annotations that describe the motion at three *different granularity levels* and are *temporally aligned*. Specifically, given one motion sequence $\mathcal{M}$ starts from $t = 0$ and ends at $t = T$, we define one basic annotation element $a$ using a text label $L$ that describes the motion segment $m$ starts from $t_s$ and ends at $t_e$, formally: $a = (L, t_s, t_e)$. The goal is to obtain a structured collection of annotation elements $\mathcal{A} = \{\mathcal{A}_s, \mathcal{A}_a, \mathcal{A}_p\}$ that covers sequence level annotation $\mathcal{A}_s$, atomic actions $\mathcal{A}_a$, and body part annotations $\mathcal{A}_p$. The sequence level annotation summarizes the full motion sequence in one text hence the annotation is simply one element: $\mathcal{A}_s = \{(L_s, t_s = 0, t_e = T)\}$.




**(a) Previous Dataset** **(b) LLM Decomposition** **(c) Our Dataset**

Figure 2. **LLM-assisted fine-grained motion annotation pipeline, compared with existing dataset.** Given motions with high level action descriptions, we instruct LLM to decompose the actions into part level descriptions and align with corresponding temporal windows. This gives the most important body part text and corresponding motions needed to learn essential motion elements.

Atomic actions are a list of $N$ non-overlapping annotations, where each one summarizes a motion segment covering a temporal window: $\mathcal{A}_a = \{(L^i, t_s^i, t_e^i)\}_{i=1}^N$, with $t_s^1 = 0, t_e^N = T$, and $t_s^i = t_e^{i-1}$. The part annotations, unique in our dataset, contain more fine-grained and structured annotations for $K$ body parts: $\mathcal{A}_p = \{\mathcal{A}_k\}_{k=1}^K$, here $\mathcal{A}_k$, similar to atomic actions $\mathcal{A}_a$, contains atomic motion annotations but with description specific for body part $k$. Formally: $\mathcal{A}_k = \{(L_k^j, t_s^j, t_e^j)\}_{j=1}^M$. Note that annotating every single element in this collection $\mathcal{A}$ is too expensive and unnecessary. Hence we allow the text labels $L$ to be *unknown* for atomic actions $\mathcal{A}_a$ and part annotations $\mathcal{A}_p$, see example annotation in Fig. 2c.

Leveraging the powerful body part reasoning capability of LLMs, we instantiate FrankenAgent to construct body part annotations $\mathcal{A}_p$ from sequence annotation $\hat{\mathcal{A}}_s$ and atomic actions $\hat{\mathcal{A}}_a$ provided by existing datasets. To ensure consistency with part annotations, we also further refine the existing sequence and part annotations:

$$\mathcal{A}_p, \mathcal{A}_a, \mathcal{A}_s = \text{FrankenAgent}(\hat{\mathcal{A}}_a, \hat{\mathcal{A}}_s) \quad (1)$$

We adopt Deepseek-R1 as our primary FrankenAgent due to its strong reasoning ability and robust long-context understanding[1]. We carefully design prompts to ensure high-quality decomposition. The LLMs are instructed to provide temporally aligned annotations with explicit body-part coverage, to decompose complex actions into interpretable segments, and to avoid vague expressions by outputting unknown when uncertain. BABEL serves as the primary reference, with KIT-ML and HumanML3D used for augmentation when available. To avoid hallucination, we explicitly instruct the agent to produce *unknown* for the motion segments that it is unsure.

Detailed prompt templates can be found in Supp.

[1]We use Deepseek-R1-0528: https://www.deepseek.com/.

| **Attributes** | BABEL | HumanML3D | KITML | **Ours** |
|---|---|---|---|---|
| *Annotation Type* | | | | |
| Sequence Label | ✓ | ✓ | ✓ | ✓ |
| Atomic Action Label | ✓ | ✗ | ✗ | ✓ |
| Body-part Label | ✗ | ✗ | ✗ | ✓ |
| *Statistics* | | | | |
| Dataset Size | 43.5h | 28.6h | 11.2h | 39.1h |
| Vocabulary | 2,162 | 6,995 | 1,577 | 4,117 |
| Total Label | 91.4k | 44.9k | 6.3k | 138.5k |
| Unseen Label | | N/A | | 28.8k |
| Part Label | | N/A | | 46.1k |

Table 1. Comparison of source motion–language datasets and our extended dataset. Our dataset builds upon **KIT-ML** [38], **BABEL** [40], and **HumanML3D** [12], using LLM-based reasoning to produce multi-level, part-aware, and unseen annotations.

### 3.3. Comparison with Existing Datasets

Based on our pipeline, we obtain a large-scale, high-quality dataset that substantially improves over existing datasets,

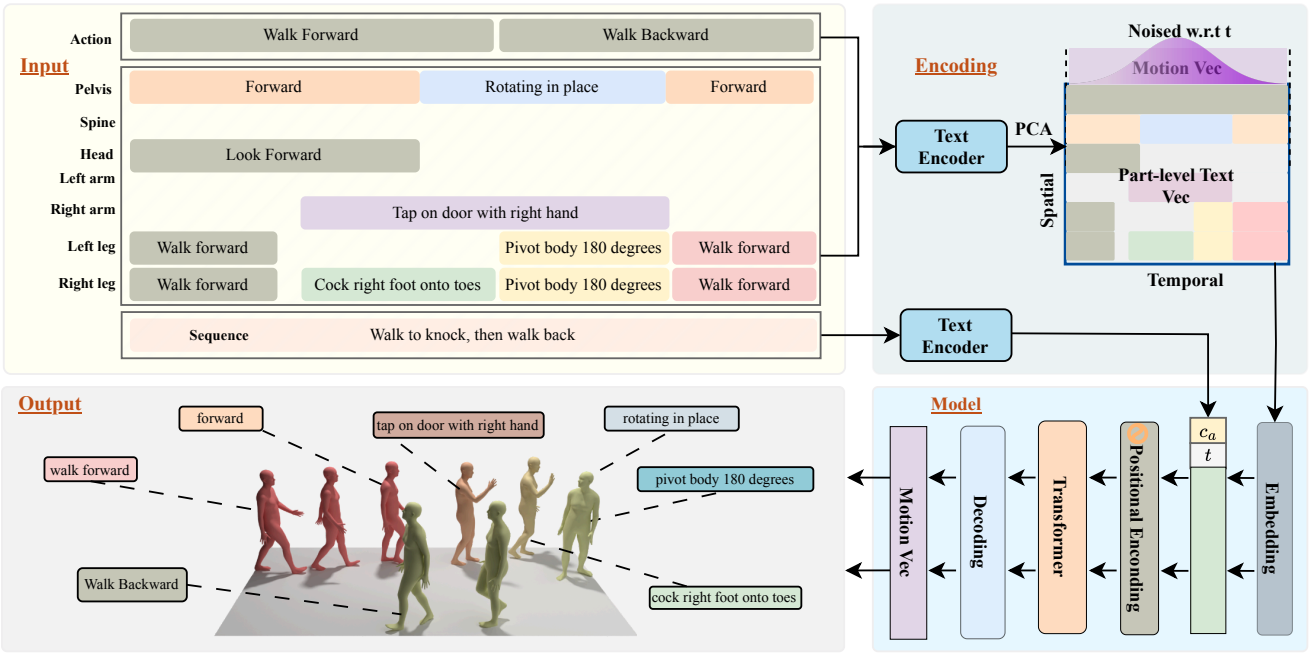



Figure 3. **Overview of our FrankenMotion**. Our model is a transformer-based diffusion model that can be input conditioned on a) sequence level prompt, b) action-level prompt and c) part-level prompt. After training with our paired data of motion and structured multi-granularity text annotations, it learns the essential motion elements and how to compose them into complex motions.

see comparison in Tab. 1. Our dataset offers two major advantages: (1) **Hierarchical**: it provides multi-level annotations that span from the overall sequence to individual body parts; and (2) **Atomic labels**: each body part and atomic actions are annotated.

Statistically, our dataset spans 39 hours of motion data and includes three levels of labels with around 15.7k, 31.5k, and 46.1k annotations respectively, forming approximately 16k sequences and 265k atomic motion segments with an average duration of 4.8 seconds. We split the dataset into 80% training, 10% validation, and 10% testing subsets. Beyond human-provided annotations, our system infers 28.8k new annotations through reasoning based on existing ones. FrankenStein offers precise spatial–temporal grounding at the atomic level, making it both challenging and valuable for advancing fine-grained motion understanding and generation. Further details are provided in the Supp.

# 4. FrankenMotion: Part-Based Spatiotemporal Composition

We aim to build a model that learns the spatial and temporal relationship between different parts and high level action semantics, allowing complex motion composition and fine-grained input control and editing. In the following, we start with problem setup (Sec. 4.1) and then discuss our network details (Sec. 4.2) followed by robust training strategy Sec. 4.3). An overview can be found in Fig. 3.

## 4.1. Preliminaries

**Pose representation.** We follow STMC [37] to represent a single frame pose $\mathbf{x} \in \mathbb{R}^d$ using SMPL [27] pose parameters, joint positions, velocities and angular velocities:

$$\mathbf{x} = [r_z, \dot{r}_x, \dot{r}_y, \dot{\alpha}, \boldsymbol{\theta}, \mathbf{j}], \tag{2}$$

where $r_z$ is the $Z$ ($\uparrow$) coordinate of the pelvis, $\dot{r}_x$ and $\dot{r}_y$ are the linear velocities of the pelvis in the $x$–$y$ plane, and $\dot{\alpha}$ is the angular velocity around the vertical ($Z$) axis. $\boldsymbol{\theta}$ denotes the SMPL pose parameters (encoded using the 6D representation [69]), and $\mathbf{j}$ are the 3D joint positions computed from the SMPL layer, forming a rotation-invariant representation by defining $\mathbf{j}$ in a local coordinate frame aligned with the body. To make $\boldsymbol{\theta}$ local to the body, the $Z$ rotation is removed from the SMPL global orientation.

**Multi-granularity control.** We adopt a transformer-based diffusion model as the framework for our text conditioned motion generation. To generate one motion sequence of $T$ frames, our model allows text control at three different levels of granularity: 1). $\mathbf{L}_s = \{L_s\}$: a single text description for the full sequence. 2). $\mathbf{L}_a = \{L_a^j | j \in \{1...W\}\}$: action labels of $W$ different non-overlapping temporal windows for the full sequence. 3). Body part prompts $\mathbf{L}_p = \{L_k^i | i \in \{1,...T\}, k \in \{1,...,K\}\}$: text prompt for body part $k$ (among $K$ predefined parts) at each frame $i$. This is the most fine-grained input condition and allows part-based motion composition and editing. At inference time, users

can conveniently provide only a sequence-level description, specify sparse part-level prompts, or edit existing control signals to guide the generated motion.

For the output, we adopt the pose representation defined in Eq. (2) aim at generating a sequence of realistic human motion: $\mathbf{x}^{[1\ldots T]}$. We adopt the sample prediction mode in diffusion models. Specifically, let $\mathbf{x}_0^{[1\ldots T]}$ be the clean motion sequence and $\mathbf{x}_\sigma^{[1\ldots T]}$ be the noisy motion at diffusion step $\sigma$, our model $f_\theta$ predicts the clean motion from three levels of text prompts $\mathbf{L}_s, \mathbf{L}_a, \mathbf{L}_p$ and diffusion timestep $\sigma$:

$$\hat{\mathbf{x}}_0^{[1\ldots T]} = f_\theta(\mathbf{x}_\sigma^{[1\ldots T]}, \sigma, \mathbf{L}_s, \mathbf{L}_a, \mathbf{L}_p) \tag{3}$$

To effectively learn the complex structured information of sequence, action and part level input, we design a network that extracts the features from different granularity levels into a joint latent space which we discuss next.

### 4.2. Spatio-temporal Embedding

To effectively learn the atomic text-to-motion mapping and composition, we use a joint embedding for sequence, action, part-level text and motion, see Fig. 3.

**Action-part-motion embedding:** we use CLIP [41] to extract text features for all input prompts. For action and part labels, we apply PCA to reduce the embedding dimension to $D = 50$, yielding $\mathbf{F}_a \in \mathbb{R}^{W\times D}$ and $\mathbf{F}_p \in \mathbb{R}^{T\times(K\times D)}$. Each action embedding in $\mathbf{F}_a$ corresponds to one of the $W$ temporal windows.To align these with the per-frame features, we expand each window embedding to its corresponding frame range, producing $\mathbf{F}_a \in \mathbb{R}^{T\times D}$. Thus, every frame is associated with both detailed part-level and high-level action text features, where $D_{m+t}$ denotes the dimension of the fused motion–text feature space. We then concatenate both matrices, leading to a matrix of text features $\mathbf{F}_{a+p} \in \mathbb{R}^{N\times(K+1)D}$. The combined text feature $\mathbf{F}_{a+p}$ is then further concatenated with the noisy motion input $\mathbf{x}_\sigma^{[1\ldots T]}$ to form the full conditioning input for the model. After passing through an MLP, this fusion produces motion–text embeddings $\mathbf{F}_{a+p+m} \in \mathbb{R}^{T\times D_{m+t}}$.

**Sequence level context:** To incorporate sequence-level context, we encode the sequence text $\mathcal{C}_s$ using the CLIP text encoder followed by an MLP, obtaining a global feature vector $\mathbf{F}_s \in \mathbb{R}^{D_{m+t}}$. We append this global feature as an additional token to the fused sequence representation. The diffusion timestep embedding, after an MLP projection, is also added as a separate token. Together, these yield a final input representation of size $\mathbb{R}^{(T+2)\times D_{m+t}}$, which jointly encodes motion, part-level, action-level, sequence-level, and diffusion timestep information.

### 4.3. Robust FrankenMotion Training

**Masking strategy.** We set zero to our text feature when text label is *unknown*. This creates sparse input text condition as our body part annotations are sparse. To improve robustness, we introduce random masking with stochastic masking probability for each labelled text condition. We adopt Beta distribution to randomly decide the zero out probability $p$ of a body part text label $L_k^i$: $p \sim \mathrm{Beta}(5r, 5(1-r))$, where $r$ is the desired masking rate. At each training step, we sample different $p$ for body part labels $L_k^i$ that are not *unknown*. This stochastic masking enhances robustness to incomplete conditioning and improves generalization under sparse supervision [26].

**Training loss.** We train our diffusion model $f_\theta$ parametrized by $\theta$ using the standard DDPM objective [16]:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0^{[1\ldots T]},\,\sigma,\epsilon}\left[\|f_\theta(\mathbf{x}_\sigma^{[1\ldots T]}, \sigma, \mathbf{c}) - \mathbf{x}_0\|_2^2\right] \tag{4}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is the diffusion noise and $\mathbf{c} = (\mathbf{L}_s, \mathbf{L}_a, \mathbf{L}_p)$ is our hierarchical text condition.

**Implementation details.** Following [36], we employ a cosine noise schedule with 100 diffusion steps, as introduced by [6]. We use the AdamW optimizer [28] with a learning rate of $2\times10^{-4}$ and a batch size of 32. For text encoding, we adopt the frozen text encoder from CLIP (ViT-B/32) [41]. Our model trains for approximately 47.5 hours on a single NVIDIA H100 GPU and each evaluation model trains for around 16 hours on a single NVIDIA A100 GPU.

## 5. Experiments

In this section, we first evaluate our newly introduced motion dataset, and then compare our model with baselines for hierarchical text to motion generation. We further ablate the design choices of our model and showcase various applications due to the flexibility of our model design. Results validate the high quality of our data and demonstrate that our model outperforms all previous methods.

### 5.1. Dataset Quality

To evaluate the quality of our FrankenAgent, we conduct a human evaluation on the annotated FrankenStein dataset. We randomly sampled 50 motion sequences and asked three human experts to assess whether the generated part, action, and sequence labels are consistent with the corresponding motion segments. Each expert assigns a binary correctness score for every label, and we compute the average as the annotation accuracy. The overall accuracy of our FrankenAgent generated annotation is 93.08%, demonstrating the high accuracy of our method. We also report inter-annotator agreement using Gwet's AC1 coefficient ($AC_1$) [15] to assess the reliability of human evaluation, where we obtain a score of $AC_1$ = 0.91 ($0.8 \leq AC_1 \leq 1.0$), indicating high reliability. Please refer to the Supp for more details.

### 5.2. Fine-grained Motion Generation

Our FrankenMotion allows motion generation from hierarchical texts of sequence, atomic action, and body-part la-

| Method | Avg-part semantic correctness | | | Per-action semantic correctness | | | Per-seq semantic correctness | | | Per-action realism | | Per-seq realism | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 ↑ | R@3 ↑ | M2T ↑ | R@1 ↑ | R@3 ↑ | M2T ↑ | R@1 ↑ | R@3 ↑ | M2T ↑ | FID ↓ | Div. → | FID ↓ | Div. → |
| GT | $52.04^{\pm0.16}$ | $64.88^{\pm0.18}$ | $0.71^{\pm0.00}$ | $54.83^{\pm0.12}$ | $72.42^{\pm0.11}$ | $0.77^{\pm0.00}$ | $72.66^{\pm0.19}$ | $91.47^{\pm0.17}$ | $0.78^{\pm0.00}$ | $0.00^{\pm0.00}$ | $53.56^{\pm0.05}$ | $0.00^{\pm0.00}$ | $48.81^{\pm0.04}$ |
| STMC | $40.67^{\pm0.29}$ | $51.38^{\pm0.30}$ | $0.66^{\pm0.00}$ | $40.96^{\pm0.61}$ | $56.32^{\pm0.66}$ | $0.70^{\pm0.00}$ | $43.58^{\pm0.35}$ | $62.32^{\pm0.45}$ | $0.67^{\pm0.00}$ | $0.10^{\pm0.00}$ | $51.79^{\pm0.10}$ | $0.20^{\pm0.00}$ | $46.82^{\pm0.17}$ |
| DartControl | $38.67^{\pm0.70}$ | $50.23^{\pm0.46}$ | $0.65^{\pm0.00}$ | $39.77^{\pm0.77}$ | $57.55^{\pm0.46}$ | $0.69^{\pm0.00}$ | $54.28^{\pm0.60}$ | $76.95^{\pm0.68}$ | $0.70^{\pm0.00}$ | $0.14^{\pm0.00}$ | $52.66^{\pm0.21}$ | $0.28^{\pm0.00}$ | $46.58^{\pm0.03}$ |
| UniMotion | $45.72^{\pm0.24}$ | $57.36^{\pm0.20}$ | $\mathbf{0.69}^{\pm0.00}$ | $47.58^{\pm0.12}$ | $65.62^{\pm0.33}$ | $\mathbf{0.75}^{\pm0.00}$ | $62.66^{\pm0.30}$ | $82.08^{\pm0.35}$ | $0.74^{\pm0.00}$ | $0.05^{\pm0.00}$ | $53.12^{\pm0.23}$ | $0.08^{\pm0.00}$ | $48.36^{\pm0.03}$ |
| FrankenMotion | $\mathbf{47.21}^{\pm0.19}$ | $\mathbf{58.97}^{\pm0.18}$ | $\mathbf{0.69}^{\pm0.00}$ | $\mathbf{48.10}^{\pm0.13}$ | $\mathbf{65.79}^{\pm0.14}$ | $\mathbf{0.75}^{\pm0.00}$ | $\mathbf{65.27}^{\pm0.22}$ | $\mathbf{85.62}^{\pm0.18}$ | $\mathbf{0.76}^{\pm0.00}$ | $\mathbf{0.04}^{\pm0.00}$ | $\mathbf{53.82}^{\pm0.10}$ | $\mathbf{0.06}^{\pm0.00}$ | $\mathbf{48.60}^{\pm0.05}$ |

Table 2. **Evaluating text to motion generation.** We report the semantic correctness and realism of parts (averaged), action and sequence level motion, with 95% confidence interval (±) after 20 repeated evaluations. Across all settings, our FrankenMotion achieves the best performance, outperforming all prior baselines in both correctness and realism.

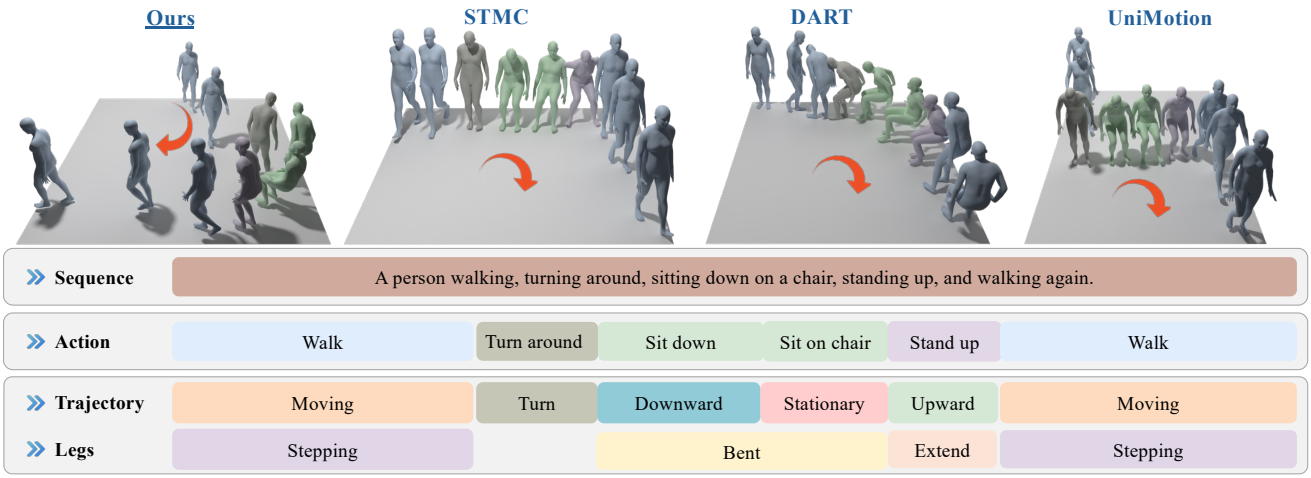



Figure 4. **Qualitative comparison with baselines**. Prior methods cannot compose parts into realistic motion (STMC [37]), generates repetitive motions (DART [68]), or do not follow the intricate details like "turn around" (UniMotion [20]). Our method faithfully composes the complex parts into one realistic motion while also following precisely the detailed body part prompts and high level semantics.

bels. To the best of our knowledge, none of the prior methods are able to accomplish this complex task. To be able to compare, we adapt prior methods to include part control and train on our proposed FrankenStein dataset. We outline the baseline setup below and provide the details in Supp.

**Baselines.** We adapt UniMotion [20], STMC [37], and DART [68] to our setup as the baselines. For fair comparison, all methods are modified to use the same pose state representation (Eq. (2)) defined in STMC [37] that includes SMPL [27] pose parameters, velocity, and joint positions. **1). STMC** is a post-hoc, test-time composition method that stitches body-part motions predicted by MDM [46] at each frame. We retrain the base MDM on all possible motion-text pairs from our FrankenStein dataset to ensure that MDM understands our text labels. At inference time, STMC receives part labels as conditioning and stitches the part motions generated by the retrained MDM. For the frames where no part labels are given, we use atomic action labels or sequence level text. **2). UniMotion** is a hierarchical model that supports sequence-level and frame-level (atomic action) text conditioning with temporal alignment but lacks body part control. We merge all part and action labels into one long text for each frame as the frame-level text input for UniMotion. **3). DART** is an autoregressive model that predicts future motion frames conditioned on motion history and a high-level text prompt. We merge our three levels of text into one long text for each frame and train DART with these pairs of text-motion data.

**Evaluation metrics.** Following [20], we evaluate generated motions along two axes: **semantic correctness**, consisting of R-Precision [12] and M2T [37], and **realism**, consisting of Fréchet Inception Distance (FID) and Diversity [12]. All these metrics require a pretrained text to motion generation model to measure the text-motion alignment. To assess the performance in all three levels of control, we train separate evaluation models for each input modality following TMR [36]: seven for each body part, one for actions, and one for full sequences. Each model is trained with paired data of text and corresponding full body motion. For semantic correctness metrics, we follow [36] and use MPNet [45] embeddings to remove false negative pairs due to paraphrased text labels (e.g., "hold arm lateral" vs. "hold arm horizontally"). Due to space constraints, we report semantic alignment scores for sequence-, action-, and av-

| Method | Inputs | | | Avg-part semantic correctness | | | Per-action semantic correctness | | | Per-seq semantic correctness | | | Realism | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Part | Atomic | Seq. | R@3 ↑ | M2T ↑ | M2M ↑ | R@3 ↑ | M2T ↑ | M2M ↑ | R@3 ↑ | M2T ↑ | M2M ↑ | FID ↓ | Div. → |
| GT | ✓ | ✓ | ✓ | $64.88^{\pm 0.18}$ | $0.71^{\pm 0.00}$ | $1.00^{\pm 0.00}$ | $72.42^{\pm 0.11}$ | $0.77^{\pm 0.00}$ | $1.00^{\pm 0.00}$ | $91.47^{\pm 0.17}$ | $0.78^{\pm 0.00}$ | $1.00^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $46.84^{\pm 0.03}$ |
| Ours | ✓ | ✗ | ✗ | $56.34^{\pm 0.42}$ | $0.69^{\pm 0.00}$ | $0.72^{\pm 0.00}$ | −− | −− | −− | −− | −− | −− | $0.08^{\pm 0.00}$ | $46.09^{\pm 0.09}$ |
| | ✓ | ✓ | ✗ | $57.74^{\pm 0.42}$ | $0.69^{\pm 0.00}$ | $0.73^{\pm 0.00}$ | $65.39^{\pm 0.17}$ | $0.75^{\pm 0.00}$ | $0.75^{\pm 0.00}$ | −− | −− | −− | $0.07^{\pm 0.00}$ | $46.19^{\pm 0.07}$ |
| | ✓ | ✓ | ✓ | **58.97**$^{\pm 0.18}$ | $0.69^{\pm 0.00}$ | **0.75**$^{\pm 0.00}$ | **65.79**$^{\pm 0.14}$ | $0.75^{\pm 0.00}$ | **0.76**$^{\pm 0.00}$ | **85.62**$^{\pm 0.18}$ | **0.76**$^{\pm 0.00}$ | **0.75**$^{\pm 0.00}$ | **0.05**$^{\pm 0.00}$ | **46.53**$^{\pm 0.03}$ |

Table 3. **Importance of hierarchical input condition.** We train models that consume input of part only, or additionally with atomic action, or all part, action and sequence level (Seq.) texts. Even with part-level inputs alone, our model attains strong performance, achieving an M2T score close to the upper-bound GT reference. Incorporating action and sequence texts introduces high-level semantics for the desired motion, which further enhances both the correctness and realism of part-level motion generation.

eraged part-level evaluations, while realism metrics are reported for all action- and sequence-level crops. Additional numbers for part-level realism can be found in Supp. For all metrics, we repeat the evaluation 20 times and report the 95% confidence interval, indicated by the variance score.

**Results.** We compare our method against baselines quantitatively in Tab. 2 and qualitatively in Fig. 4. STMC [37] follows part instructions thanks to the well trained MDM model, but it struggles to compose different parts into a realistic motion. This often leads to unsmooth transitions between atomic actions or fine-grained control, such as "turning around", being ignored, as can be seen in Fig. 4 column 2. UniMotion [20] generates more realistic motion due to its frame-level control, as can be seen from the FID score in Tab. 2. However, it lacks an explicit structure of the body part features, leading to less precise text control, and the tiny detailed motions are ignored, as shown in the R-1 score (Tab. 3) and Fig. 4, where it does not turn around. DART primarily follows the sequence level text; however, it cannot control motion precisely at each frame. Due to its autoregressive design, error can accumulate, and it often produces repeated motion segments representing only partial input prompts (see repeated sitting and standing in Fig. 4).

In contrast, our model has structured body part conditioning, together with hierarchical control of atomic action and sequence level text. It generates fine grained motions that are controlled precisely by body parts while also maintaining coherence with the high level semantics introduced by atomic action and sequence level text. Notice in Fig. 4 column one how precisely our model follows the text prompt of body parts and actions at every frame. In Tab. 2, our method clearly outperforms all baselines in both motion quality and consistency with the input text.

### 5.3. Ablation Study

We propose a model that consumes text at three levels of granularity, as we find that these different texts are important for generating coherent motions. To evaluate this, we train separate models that take only part text, part and atomic action, and all text as input conditions, and we evaluate the motion generation performance in Tab. 3. We also show the scores of our evaluation model, which serve as an upper bound. Notably, the model conditioned only on part texts already achieves competitive scores, demonstrating the strong capability of our model for understanding fine-grained part texts. With additional high level semantics of atomic action and sequence level text, our model generates motions that more precisely follow the part labels. Furthermore, such motions are more meaningful due to the guidance of high level text semantics, illustrating the benefits of our hierarchical design. We show some qualitative examples in Supp.

### 5.4. Flexible Input Control Applications

Thanks to its modular design and the sparse structure of our dataset, our model supports flexible conditioning during inference. Users can control the motion at different granularities—such as a dominant body part, an action-level phrase, or a single sequence-level description—allowing adaptive control depending on the available text or user preference. We demonstrate the flexibility of our input in Fig. 1 and Fig. 4. Additional qualitative examples and more application cases are provided in the supplementary video.

## 6. Conclusion and Limitation

We introduced multi-level spatiotemporal motion conditioning, enabling controllable generation at the sequence, atomic action, and atomic part levels. To support this task, we built FrankenStein, a fine-grained, temporally aligned dataset with atomic part-level annotations derived via LLM reasoning. Based on it, FrankenMotion learns to compose motion from atomic elements, achieving fine-grained and flexible spatial–temporal control. Experiments show that it outperforms adapted baselines and establishes a strong foundation for compositional motion generation.

**Limitation** While FrankenMotion enables fine-grained spatiotemporal text-to-motion generation, it cannot yet generate minute-long motion sequences within a single pass. Extending its ability to model long-term temporal structure will be an important direction for future work.

**Acknowledgments:** Special thanks RVH and AVG members for the help and discussion. Prof. Gerard Pons-Moll and Prof. Andreas Geiger are members of the Machine Learning Cluster of Excellence, EXC number 2064/1 - Project number 390727645. Gerard Pons-moll is endowed by the Carl Zeiss Foundation. Andreas Geiger was supported by the ERC Starting Grant LEGO-3D (850533).

## References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, 2022. 2

[2] Jinseok Bae, Inwoo Hwang, Young-Yoon Lee, Ziyu Guo, Joseph Liu, Yizhak Ben-Shabat, Young Min Kim, and Mubbasir Kapadia. Less is more: Improving motion diffusion models with sparse keyframes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11069–11078, 2025. 2

[3] German Barquero, Sergio Escalera, and Cristina Palmero. Flowmdm: Seamless human motion composition with blended positional encodings. *arXiv preprint arXiv:2402.15509*, 2024. 2

[4] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 464–473. IEEE, 2024. 2

[5] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arxiv:2405.20340*, 2024. 3

[6] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023. 6

[7] Jungbin Cho, Junwan Kim, Jisoo Kim, Minseo Kim, Mingu Kang, Sungeun Hong, Tae-Hyun Oh, and Youngjae Yu. Discord: Discrete tokens to continuous motion via rectified flow decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14602–14612, 2025. 2

[8] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*, 2024. 2

[9] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pages 346–362. Springer, 2022. 3

[10] Qihang Fang, Chengcheng Tang, Bugra Tekin, Shugao Ma, and Yanchao Yang. Humocon: Concept discovery for human motion understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[11] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024. 3

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 7, 1

[13] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 2

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1

[15] Kilem Gwet. Handbook of inter-rater reliability. *Gaithersburg, MD: STATAXIS Publishing Company*, pages 223–246, 2001. 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 6

[17] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing, 2024. 2

[18] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[19] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 2

[20] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. *arXiv preprint arXiv:2409.15904*, 2024. 2, 3, 7, 8

[21] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2

[22] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: Generative models for human motion synthesis. *arXiv preprint arXiv:2505.01425*, 2025. 2

[23] Xiaojie Li, Ronghui Li, Shukai Fang, Shuzhao Xie, Xiaoyang Guo, Jiaqing Zhou, Junkun Peng, and Zhi Wang. Music-aligned holistic 3d dance generation via hierarchical motion modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14420–14430, 2025. 2

[24] 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. *Advances in Neural Information Processing Systems*, 2023. 2

[25] Jing Lin, Yao Feng, Weiyang Liu, and Michael J. Black. ChatHuman: Chatting about 3d humans with tools. In *CVPR*, 2025. 3

[26] Lei Liu, Yuhao Luo, Xu Shen, Mingzhai Sun, and Bin Li. $\beta$-dropout: A unified dropout. *IEEE Access*, 7:36140–36153, 2019. 6

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34, 2015. 5, 7

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[29] Yiyi Ma, Yuanzhi Liang, Xiu Li, Chi Zhang, and Xuelong Li. Intersyn: Interleaved learning for dynamic motion synthesis in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12832–12841, 2025. 2

[30] 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*, 2019. 2, 3

[31] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2

[32] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575*, 2024. 2

[33] Jiawei Mo, Yixuan Chen, Rifen Lin, Yongkang Ni, Min Zeng, Xiping Hu, and Min Li. Mochat: Joints-grouped spatio-temporal grounding llm for multi-turn motion comprehension and description, 2024. 3

[34] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5379–5391, 2025. 2

[35] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[36] 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. 6, 7

[37] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Stmc: Multi-track timeline control for text-driven 3d human motion generation. *arXiv preprint arXiv:2401.08559*, 2024. 3, 5, 7, 8

[38] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 3, 4, 1

[39] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2021. 2, 3, 1

[40] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. 4

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[42] Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2in: Leveraging individual information to generate human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1941–1951, 2024. 2

[43] Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. Mixermdm: Learnable composition of human motion diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12380–12390, 2025. 2

[44] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Priormdm: Human motion diffusion as a generative prior. In *ICLR*, 2023. 2

[45] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020. 7

[46] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 7, 3

[47] Yuan Wang, Yali Li, Xiang Li, and Shengjin Wang. Hsi-gpt: A general-purpose large scene-motion-language model for human scene interaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7147–7157, 2025. 2

[48] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7169–7178, 2025. 2

[49] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. 2

[50] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *Pro-*

*ceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11176–11186, 2025. 2

[51] Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Motionbank: A large-scale video motion benchmark with disentangled rule-based annotations. *arXiv preprint arXiv:2410.13790*, 2024. 3

[52] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 2

[53] Sirui Xu, Yu-Xiong Wang, Liangyan Gui, et al. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37:52858–52890, 2024. 2

[54] Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7048–7060, 2025.

[55] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12266–12277, 2025. 2

[56] Payam Jome Yazdian, Eric Liu, Rachel Lagasse, Hamid Mohammadi, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv preprint arXiv:2312.12634*, 2024. 3

[57] Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12358–12369, 2025. 2

[58] Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12358–12369, 2025.

[59] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024. 2

[60] Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17592–17602, 2025. 2

[61] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[62] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *NeurIPS*, 36, 2024. 2, 3

[63] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[64] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. Force: Dataset and method for intuitive physics guided human-object interaction. *CoRR*, 2024. 2

[65] Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Ziqiang Dang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13761–13771, 2025. 2

[66] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. *arXiv preprint arXiv:2306.10900*, 2023. 2

[67] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. *arXiv preprint arXiv:2403.07487*, 2024. 2

[68] Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 2, 3, 7

[69] 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*, pages 5745–5753, 2019. 5

[70] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding, planning, generation and beyond. *arXiv preprint arXiv:2311.16468*, 2023. 3

[71] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2023. 2

# FrankenMotion: Part-level Human Motion Generation and Composition

## Supplementary Material

In the following, we start with the supplementary website in Sec. 7 and discuss the details of the training dataset in Sec. 8. Then, we present the details of our evaluation setup in Sec. 10 and additional results in Sec. 11.

## 7. Website with Qualitative Results

We provide a website to further present the results with animated motions, showing a clearer comparison across various tasks and against other baselines. Additionally, we show a visualization of our dataset.

## 8. Annotation Process and Dataset Statistics

**FrankenAgent Implementation Details** FrankenMotion is trained on a subset of AMASS, and we obtain text annotations at three granularities—sequence, action, and body-part level—using our FrankenAgent agent. The agent consolidates annotations from HumanML3D [12], BABEL [39], and KIT-ML [38], producing temporally aligned and consistent descriptions across all sources.

This task requires temporal alignment, decomposition of whole-body descriptions into part-specific actions, and the ability to resolve ambiguous or overlapping annotations—capabilities that smaller or instruction-only LLMs often struggle with. Models with explicit reasoning ability consistently produce more coherent and temporally faithful part-level labels. Therefore, we adopt DeepSeek-R1 [14] as our annotation agent due to its strong structured-reasoning performance and stability under long, multi-step instructions.

| Part | Duration | Unique Labels | Segments | Conf. |
|---|---|---|---|---|
| head | 2.22h | 584 | 1769 | 4.89 |
| left arm | 12.37h | 6060 | 8935 | 4.87 |
| right arm | 14.05h | 6953 | 10238 | 4.88 |
| left leg | 22.09h | 3380 | 15556 | 4.91 |
| right leg | 22.33h | 3529 | 15815 | 4.91 |
| spine | 5.27h | 1106 | 3990 | 4.72 |
| trajectory | 20.97h | 2599 | 13732 | 4.82 |
| action | 33.56h | 7946 | 23771 | 4.96 |
| sequence | 37.83h | 13941 | 7378 | 4.95 |

Table 4. Statistics of the FrankenStein annotations across all three granularities (sequence-, action-, and part-level). We report the annotated duration per category, the number of unique textual labels, the number of segments, and the average confidence score assigned by FrankenAgent. Durations exceed the dataset length because each category spans the full sequence.

The complete prompt template used by the agent is provided at the end of this supplementary document.

**Multi-Granularity Annotation Statistics** We provide visual examples of our annotations in the supplementary video. FrankenStein contains 39.07 hours of unique mocap data annotated at three granularities: sequence-level, action-level, and part-level. Tab. 4 reports detailed statistics for all categories across these granularities, including annotated duration, number of unique labels, number of segments, and average confidence.

"Unique labels" denotes distinct textual descriptions generated for each category. Durations exceed the dataset length because each body-part category spans the full temporal extent of every sequence.

## 9. Human Evaluation and Inter-Annotator Agreement

To complement the results in the main paper, we provide additional details on the human evaluation protocol and the computation of inter-annotator agreement (IAA) for FrankenStein.

**Inter-Annotator Agreement (IAA) Setup.** Following the evaluation protocol in the main paper, we randomly sampled 50 motion sequences and asked three human experts to judge whether each generated part, action, and sequence label is consistent with the corresponding motion clip. Each annotator gave a binary correctness score, resulting in three independent ratings per label.

These ratings serve two purposes: (1) computing annotation accuracy, and (2) measuring annotator consistency. As reported in the main paper, our annotations achieve an accuracy of 93.08%. To ensure that this accuracy is trustworthy, it is important that human annotators agree with each other. Therefore, we compute Gwet's AC1 coefficient ($AC_1$), a chance-corrected measure of inter-annotator agreement that is robust to class imbalance. Our evaluation yields $AC_1 = 0.91$, which indicates that the human judgments are consistent and that the reported accuracy reliably reflects annotation quality.

**Computing Gwet's AC1.** Let $N$ be the number of items and $R$ the number of annotators. Each annotator assigns a binary label $y_{ir} \in \{0, 1\}$ to item $i$, where 1 denotes a correct label. For each item, let

$$n_{1,i} = \sum_{r=1}^{R} y_{ir}, \qquad n_{0,i} = R - n_{1,i}.$$

| Method | Inputs | | | Avg-part semantic correctness | | | Per-action semantic correctness | | | Per-seq semantic correctness | | | Realism | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Part | Atomic | Seq. | R@3 ↑ | M2T ↑ | M2M ↑ | R@3 ↑ | M2T ↑ | M2M ↑ | R@3 ↑ | M2T ↑ | M2M ↑ | FID ↓ | Div. → |
| GT | ✓ | ✓ | ✓ | $64.88^{\pm 0.18}$ | $0.71^{\pm 0.00}$ | $1.00^{\pm 0.00}$ | $72.42^{\pm 0.11}$ | $0.77^{\pm 0.00}$ | $1.00^{\pm 0.00}$ | $91.47^{\pm 0.17}$ | $0.78^{\pm 0.00}$ | $1.00^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $46.84^{\pm 0.03}$ |
| MDM | ✓ | ✓ | ✓ | $57.92^{\pm 0.82}$ | $0.69^{\pm 0.00}$ | $0.67^{\pm 0.00}$ | $60.37^{\pm 0.39}$ | $0.72^{\pm 0.00}$ | $0.69^{\pm 0.00}$ | $66.91^{\pm 0.64}$ | $0.69^{\pm 0.00}$ | $0.68^{\pm 0.00}$ | $0.10^{\pm 0.00}$ | $45.47^{\pm 0.12}$ |

Table 5. Performance of the retrained MDM base model used by STMC. MDM achieves reasonably good semantic alignment across part-, action-, and sequence-level evaluations, supporting its use as the text-to-motion backbone for STMC.

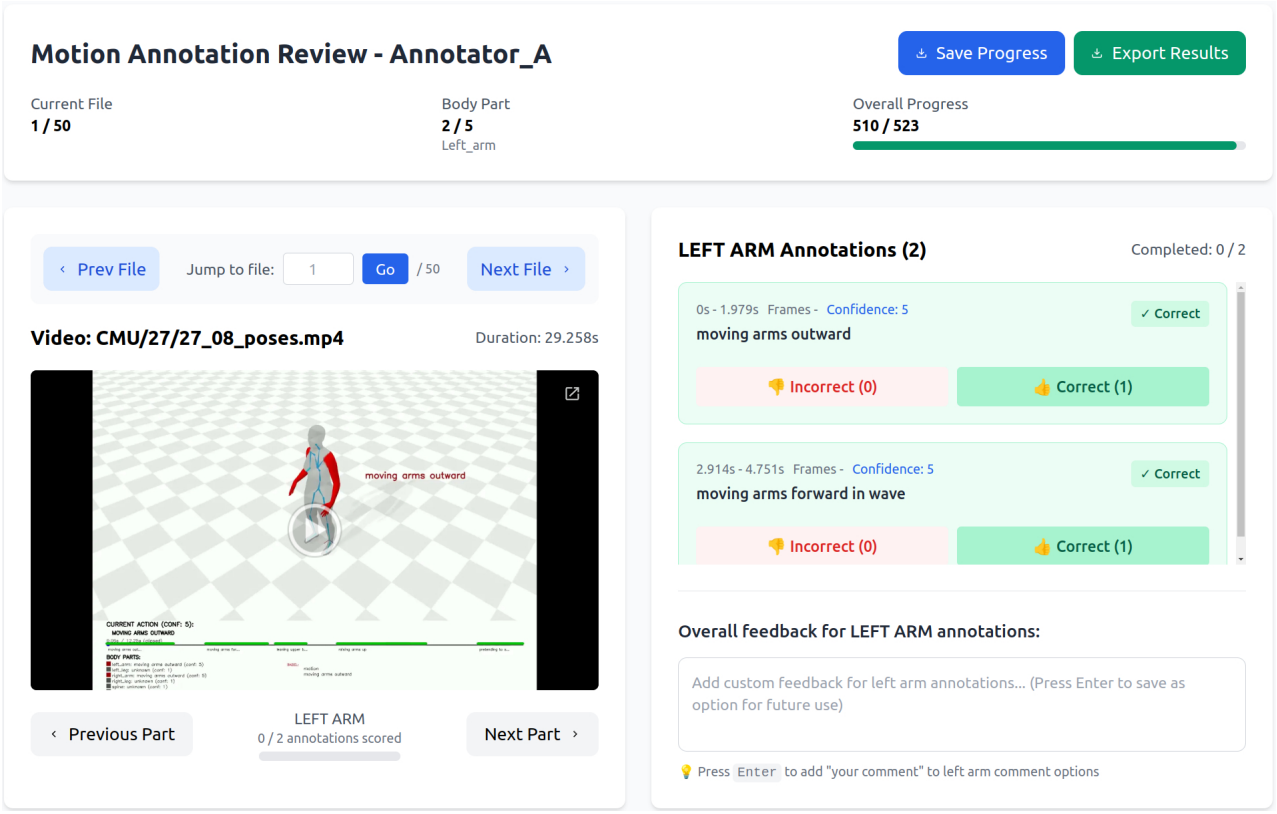



Figure 5. Interface used in the human evaluation for inter-annotator agreement (IAA). Annotators inspect motion clips and judge the alignment of part-level labels with the underlying motion.

The per-item agreement is

$$P_i = \frac{n_{1,i}(n_{1,i}-1) + n_{0,i}(n_{0,i}-1)}{R(R-1)},$$

and the observed agreement is

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i.$$

Let the prevalence of the positive label be

$$p_1 = \frac{1}{NR}\sum_{i,r} y_{ir}, \qquad p_0 = 1 - p_1.$$

The expected chance agreement is

$$P_e^{\mathrm{AC1}} = 2p_0p_1.$$

The AC1 coefficient is finally computed as

$$\mathrm{AC1} = \frac{\bar{P} - P_e^{\mathrm{AC1}}}{1 - P_e^{\mathrm{AC1}}}. \tag{5}$$

A higher AC1 indicates more reliable and consistent annotations. We report both overall and per-category AC1 in Table 6.

| Body Part | # Items | Accuracy (%) | AC1 |
|---|---|---|---|
| Action | 115 | 93.91 | 0.902 |
| Head | 5 | 40.00 | -0.282 |
| Left Arm | 50 | 92.00 | 0.906 |
| Left Leg | 74 | 95.05 | 0.970 |
| Right Arm | 57 | 91.23 | 0.875 |
| Right Leg | 76 | 93.42 | 0.920 |
| Sequence Caption | 48 | 95.83 | 0.925 |
| Spine | 21 | 90.48 | 0.923 |
| Trajectory | 55 | 94.55 | 0.892 |
| **Overall** | 501 | **93.08** | **0.907** |

Table 6. Per-part inter-annotator agreement results. We report accuracy and Gwet's AC1, the primary reliability metric. The overall AC1 of 0.907 indicates strong cross-rater consistency.

## 10. Evaluation Setup

In the main paper (Sec. Sec. 5.2), we evaluate sequence-, action-, and part-level control by training separate text–motion retrieval models for each granularity. Here, we provide additional details and report the full per-part retrieval scores.

Each evaluation model is trained on the corresponding split of FrankenStein (training set only) and tested on the held-out test set. For part-level retrieval, we construct a dedicated dataset for each body part by extracting only the time segments where that part-level annotation is present. This yields seven independent part-level evaluation sets (left-/right arms, left/right legs, head, spine, trajectory), plus action-level and sequence-level evaluation sets.

Fig. Fig. 6 presents the detailed retrieval performance of these evaluation models, reporting R-Precision (R@1–3) and M2T for all three granularities. As in the main paper, we merge paraphrased text labels using a 0.85 sentence embedding similarity threshold to ensure consistent scoring for part-level retrieval.

**Baseline Implementation Details.** We provide the full implementation details for adapting STMC [37], UniMotion [20], and DART [68] to our multi-level spatiotemporal control setting. For each method, we describe how the original text-conditioning interfaces are extended to support our three annotation levels (sequence-, action-, and part-level) and how each training sample is reformatted accordingly.

**STMC.** STMC is a post-hoc test-time composition framework and therefore requires a text-to-motion base model that can reliably interpret the atomic and part-level labels used during stitching. To support this, we retrain the MDM [46] model on our dataset using all aligned text-motion pairs from the three annotation levels. The performance of this retrained MDM is reported in Table Tab. 5. It achieves *reasonably good* semantic alignment given the diversity of our labels, indicating that STMC's lower performance in the main paper primarily arises from its test-time stitching mechanism rather than limitations of the underlying MDM model. During inference, STMC receives all level labels as conditioning signals and composes the corresponding part-specific motion fragments generated by MDM.

**UniMotion and DART.** Both UniMotion and DART are

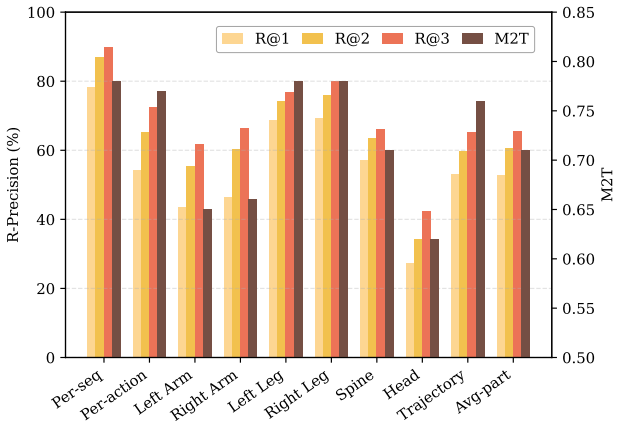



Figure 6. **Ground-truth text–motion alignment metrics.** R-Precision (R@1–3) and M2T retrieval performance for all control levels on FrankenStein, including sequence-, action-, and part-level evaluations. Individual body parts are shown separately, with "Avg-part" indicating the average over all seven part-level evaluation sets.

| UniMotion input | DART input |
|---|---|
| **Sequence-level (global):**<br>`A person walks using a hand rail while moving diagonally to the left.` | **Frame-level (merged):**<br>`A person walks using a hand rail while moving diagonally to the left.` |
| **Frame-level (local):**<br>`Action: walk.`<br>`Left Arm: pull with hand. Right Arm: pull with hand.`<br>`Left Leg: walking.`<br>`Right Leg: walking.`<br>`Trajectory: diagonally left and forward.` | `Action: walk.`<br>`Left Arm: pull with hand. Right Arm: pull with hand.`<br>`Left Leg: walking.`<br>`Right Leg: walking.`<br>`Trajectory: diagonally left and forward.` |

Table 7. Conditioning formats for UniMotion and DART on the 'walk with right support' example. UniMotion uses a two-level hierarchy (sequence-level + frame-level), while DART receives a single merged frame-level description.

| Method | Left Arm | | Right Arm | | Left Leg | | Right Leg | | Head | | Spine | | Trajectory | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@3 | FID | R@3 | FID | R@3 | FID | R@3 | FID | R@3 | FID | R@3 | FID | R@3 | FID |
| GT | $60.13^{\pm0.51}$ | $0.00^{\pm0.00}$ | $63.90^{\pm0.62}$ | $0.00^{\pm0.00}$ | $80.82^{\pm0.47}$ | $0.00^{\pm0.00}$ | $81.82^{\pm0.37}$ | $0.00^{\pm0.00}$ | $40.00^{\pm1.37}$ | $0.00^{\pm0.00}$ | $65.40^{\pm0.93}$ | $0.00^{\pm0.00}$ | $62.12^{\pm0.40}$ | $0.00^{\pm0.00}$ |
| STMC | $46.74^{\pm0.67}$ | $0.12^{\pm0.00}$ | $46.17^{\pm0.41}$ | $0.12^{\pm0.00}$ | $68.67^{\pm0.22}$ | $0.07^{\pm0.00}$ | $69.18^{\pm0.57}$ | $0.07^{\pm0.00}$ | $31.03^{\pm1.87}$ | $0.16^{\pm0.00}$ | $50.01^{\pm2.37}$ | $0.10^{\pm0.01}$ | $47.12^{\pm1.11}$ | $0.06^{\pm0.00}$ |
| DART | $40.46^{\pm0.55}$ | $0.28^{\pm0.00}$ | $43.32^{\pm0.96}$ | $0.27^{\pm0.00}$ | $69.96^{\pm0.65}$ | $0.10^{\pm0.00}$ | $69.05^{\pm0.45}$ | $0.11^{\pm0.00}$ | $25.63^{\pm2.62}$ | $0.35^{\pm0.01}$ | $52.59^{\pm1.36}$ | $0.20^{\pm0.00}$ | $50.58^{\pm0.75}$ | $0.13^{\pm0.00}$ |
| UniMotion | $51.78^{\pm0.69}$ | $0.09^{\pm0.00}$ | $53.24^{\pm1.29}$ | $0.08^{\pm0.00}$ | $74.68^{\pm0.40}$ | $0.04^{\pm0.00}$ | $75.06^{\pm0.31}$ | $0.05^{\pm0.00}$ | $33.00^{\pm2.03}$ | $0.11^{\pm0.00}$ | $56.74^{\pm0.89}$ | $0.08^{\pm0.00}$ | $56.81^{\pm0.62}$ | $0.04^{\pm0.00}$ |
| Ours | **53.92**$^{\pm0.53}$ | **0.06**$^{\pm0.00}$ | **55.16**$^{\pm0.74}$ | **0.05**$^{\pm0.00}$ | **74.76**$^{\pm0.53}$ | **0.03**$^{\pm0.00}$ | **75.49**$^{\pm0.40}$ | **0.04**$^{\pm0.00}$ | **35.63**$^{\pm1.40}$ | **0.09**$^{\pm0.00}$ | **60.19**$^{\pm1.11}$ | **0.07**$^{\pm0.00}$ | **57.64**$^{\pm0.50}$ | **0.03**$^{\pm0.00}$ |

Table 8. **Part-level evaluation across seven control dimensions.** We report R@3 (semantic correctness; higher is better) and FID (realism; lower is better) for each part: left/right arms, left/right legs, head, spine, and trajectory.

adapted to accept temporally aligned, fine-grained part-level conditioning, but their conditioning interfaces differ.

*UniMotion* supports hierarchical text conditioning and therefore receives:

- **Sequence-level description:** provided once per motion sequence, taken directly from our annotations.
- **Frame-level description:** updated whenever the atomic action or any body-part label changes.

For each segmented time interval, we construct the frame-level description by concatenating the current atomic action with all active part-level annotations. This provides UniMotion with both long-range global context and fine-grained, time-varying control.

*DART* accepts a single text prompt per segment (history: 2 frames; future: 8 frames). Therefore, for each segment, we merge the sequence-level description, the atomic action label, and all active part-level labels into one unified **frame-level prompt**. Whenever any part-level annotation changes, we create a new segment boundary and generate a new merged prompt. This ensures that DART receives temporally aligned conditioning consistent with our annotation structure.

*Example.* Table Tab. 7 shows how a representative training sample is formatted for UniMotion and DART. UniMotion uses a two-level hierarchy (sequence-level + frame-level), while DART receives a single merged frame-level description.

## 11. More Experiment Results

**Part-level results for baseline comparison.** In the main paper, we report only the averaged part-level semantic correctness, and part-level realism metrics are included only in Table Tab. 3 for the ablation study. To provide a complete picture, Table Tab. 8 presents the full seven-part evaluation, covering the left/right arms, left/right legs, head, spine, and trajectory. Each part is evaluated using R@3 (semantic correctness) and FID (realism). Across all individual parts and both metrics, our method consistently outperforms all baselines, demonstrating that the improvements shown in the main paper hold uniformly at the detailed part level.

**Body Part Motion Annotation Task and Example Annotations (Part 1)**

## Goal

Create detailed body part-level annotations for the entire motion sequence using the provided BABEL and HumanML3D/KITML annotations. This task breaks down whole-body motion descriptions into specific body part movements.

## What to Annotate

- **Left arm:** Movements specific to the left arm, hand, and shoulder (e.g., wave arm)
- **Right arm:** Movements specific to the right arm, hand, and shoulder (e.g., wave arm)
- **Left leg:** Movements specific to the left leg, foot, and hip (e.g., stand, kick, jump)
- **Right leg:** Movements specific to the right leg, foot, and hip (e.g., stand, kick, jump)
- **Spine:** Trunk flex/extend, side-bend, twist, posture.
- **Head:** Look left/right/down, turn left/right, nod, tilt
- **Trajectory:** Center of mass movement through space (e.g., forward/backward, left/right, circular path, zigzag, vertical up and down, downward/rise up, stationary on ground, turn left/right)
- **Action:** Atomic, Segment-level overall body activities and movements (e.g., "jumping jacks", "waltz dance", "playing golf", "cartwheel")
- **Sequence caption:** Produce a concise caption that describes the entire process across the clip. Follow the style shown in examples; don't add details beyond the source.

## Core Principles

- Accuracy over completeness: It is better to use `"unknown"` than to guess incorrectly.
- Do **NOT** add details that aren't explicitly inferred from the source annotations.
- Never infer specific directions (forward/backward/sideways) unless explicitly stated.
- Preserve all information: Include all actions mentioned for each body part.
- Effector verbs are explicit: If a verb clearly implies the acting body part (e.g., push/press → arms; step/squat → legs; bend/twist → spine; look/nod → head; turn/rise/down → trajectory), label that part with confidence 5.

## Process

1. **Primary** – BABEL annotations (`babel_seg_X`) for timestamps.
2. **Secondary** – HumanML3D/KITML for added context.
3. When no KITML/HumanML3D is available: rely on BABEL and mark `"unknown"` when uncertain.
4. For gaps: if a movement cannot be reasonably inferred, use `"unknown"` with `"confidence": 1`.

## Important Rules

- Cover the entire duration with **no gaps** for each body part.
- Extract the relevant portion of text for each body part when annotations describe multiple body parts.
- Never use `"transition"` — describe what each body part is actually doing or use `"unknown"`.
- For complex activities (dances, sports, exercises), separate specific limb actions from the overall **Action** category.
- Always include **ALL** actions when multiple are mentioned for the same body part (e.g., "lean forward" **and** "head bobbing").
- Use specific descriptions rather than vague terms (e.g., "spine upright" not "dance stance").

## Bilateral Limb Annotation Guidelines

1. **Explicitly Different Actions:**
   - If left vs. right differ, assign the specific description to each.
2. **Ambiguous Limb References:**
   - If side is not specified, apply the same label to **both** limbs.
3. **Whole Body Movements:**
   - Decompose into part actions; put the overall description in **Action**; use exact BABEL wording when available.
4. **Single Active Limb:**
   - If only one limb is mentioned, annotate it; set the other limb to `"unknown"`.

## Confidence Scoring

- **5:** Explicitly mentioned in source annotations with clear description.
- **4:** Strongly implied by source or confirmed by multiple sources.
- **If confidence would be ¡ 4:** set `text: "unknown"` and `confidence: 1`.

**Body Part Motion Annotation Task and Example Annotations (Part 2)**

## Special Cases and Common Mistakes to Avoid

1. **Static postures:** Describe specific positions (e.g., “arm hanging by side”), not “static”.
2. **Never add details not in source:**
   - No added directions (don’t change “walking” to “walking forward”).
   - No guessing which limb performs an action.
   - No guessing timing when not specified.
3. **Be specific and complete:**
   - Use precise terms (e.g., “spine upright”).
   - Include **ALL** actions mentioned for each body part.
   - For repetitive movements, summarize clearly.
4. **Classify correctly:**
   - Assign turning/standing movements primarily to **legs** (unless explicitly head/torso).
   - For whole-body poses (e.g., “touching ground”, “bridge pose”), describe each involved body part rather than assigning to one category.
   - If only “kick” is given without side, assign to **both legs** (see Ambiguous Limb References).

## Output Format

```
{
    "annotations": {
    "sequence_caption": [
      {
        "text": "a person laying face down on the ground and then slowly crawling backwards",
        "start": 0.0,
        "end": 4.367,
        "confidence": 5,
        "reasoning": "Whole-clip caption per examples",
        "source": "summary_from_sources"
      }
    ],
    "action": [
      {
        "text": "crawl",
        "start": 3.278,
        "end": 4.367,
        "confidence": 5,
        "reasoning": "Atomic segment-level action",
        "source": "babel_seg_id_or_equivalent"
      }
    ],
      "left_arm": [
        {
          "text": "moving arms outward",
          "start": 0.0,
          "end": 1.979,
          "confidence": 5,
          "reasoning": "From BABEL annotation",
          "source": "babel_06475_seg_3"
        }
      ]
      // Repeat for all body parts("right_arm","trajectory","left_leg","right_leg","spine","head")
    }
}
```

**Note:**

1. Maintain the original file information structure with file ID and duration.
2. Each **body part** and **Action** must cover the full duration with no gaps.
3. Follow bilateral limb guidelines.
4. Use `"unknown"` with `"confidence": 1` when movements cannot be reasonably inferred.
5. Review **Special Cases and Common Mistakes to Avoid** to prevent similar errors.

If full annotation is complete add `"ANNOTATION COMPLETE"` at the end of the output, outside of JSON format.

## Body Part Motion Annotation Task and Example Annotations (Part 3)

### Input:


```
{
  "ACCAD/Female1General_c3d/A6 - lift box_poses": {
    "duration": 8.383,
    "annotations": [
      {
        "seg_id": "babel_00629_seg_0",
        "text": "holding arms forward",
        "start": 0.0,
        "end": 1.586
      },
      {
        "seg_id": "babel_00629_seg_2",
        "text": "raising arms",
        "start": 1.752,
        "end": 2.606
      },
      {
        "seg_id": "babel_00629_seg_6",
        "text": "bending down",
        "start": 4.877,
        "end": 7.690
      },
      {
        "seg_id": "humanml3d_002139_1",
        "text": "a person uses both hands to pick up an object above head level, then crouches to put it down, then stands up straight.",
        "start": 0.0,
        "end": 8.383
      },
      ...
    ]
  }
}
```


### Generation Output:


```
{
  "annotations": {
    "sequence_caption": [
      {
        "text": "a person uses both hands to pick up an object above head level, then crouches to put it down, then stands up straight.",
        "start": 0.0,
        "end": 8.383,
        "confidence": 5,
        "reasoning": "From HumanML3D",
        "source": "humanml3d_002139_1"
      }
    ],
    "action": [
      {
        "text": "lifting",
        "start": 0.0,
        "end": 4.877,
        "confidence": 4,
        "reasoning": "From BABEL sequence and HumanML3D",
        "source": "babel_00629_seq_0, humanml3d_002139_1"
      },
      ...
    ],
    "left_arm": [
      {
        "text": "holding arms forward",
        "start": 0.0,
        "end": 1.586,
        "confidence": 5,
        "reasoning": "From BABEL",
        "source": "babel_00629_seg_0"
      },
      ...
    ],
    "right_arm": [ ... ],
    "left_leg": [ ... ],
    "right_leg": [ ... ],
    "spine": [ ... ],
    "head": [ ... ],
    "trajectory": [ ... ]
  }
}
```


ANNOTATION COMPLETE