

FrankenMotion: Part-level Human Motion Generation and Composition

Supplementary Material

In the following, we start with our project page 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. Project Page

We provide a project page at <https://coral79.github.io/frankenmotion/> with animated motion results, offering a clearer comparison across various tasks and baselines. The page also includes a visualization gallery of our FrankenStein dataset annotations.

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 instruc-

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.

tions.

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 AC_1 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 AC_1 . 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

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±0.18	0.71±0.00	1.00±0.00	72.42±0.11	0.77±0.00	1.00±0.00	91.47±0.17	0.78±0.00	1.00±0.00	0.00±0.00	46.84±0.03
MDM	✓	✓	✓	57.92±0.82	0.69±0.00	0.67±0.00	60.37±0.39	0.72±0.00	0.69±0.00	66.91±0.64	0.69±0.00	0.68±0.00	0.10±0.00	45.47±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.

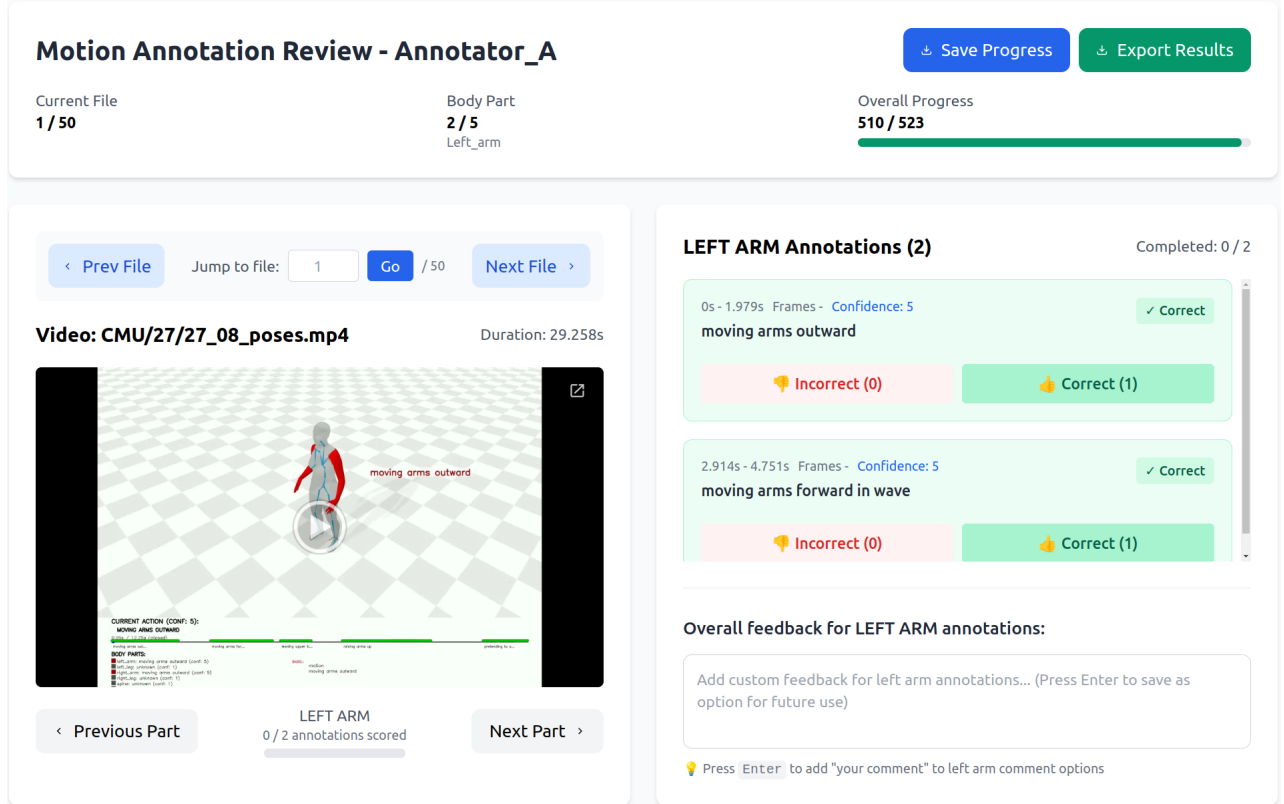


Figure 5. Interface used in the human evaluation for inter-annotator agreement (IAA). Annotators inspect motion clips and judge the alignment of part-level labels with the underlying motion.

label. For each item, let

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

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

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

The expected chance agreement is

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

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.

The AC1 coefficient is finally computed as

$$AC1 = \frac{\bar{P} - P_e^{AC1}}{1 - P_e^{AC1}}. \quad (5)$$

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

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 [71] 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 per-

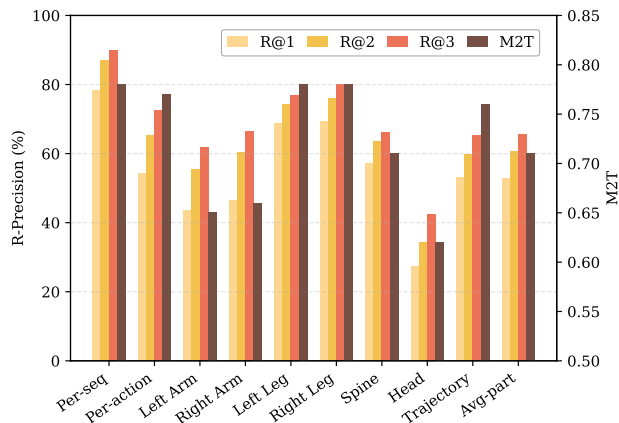


Figure 6. **Ground-truth text–motion alignment metrics.** R-Precision (R@1–3) and M2T retrieval performance for all control levels on FrankenStein, including sequence-, action-, and part-level evaluations. Individual body parts are shown separately, with “Avg-part” indicating the average over all seven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

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

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 \pm 0.51	0.00 \pm 0.00	63.90 \pm 0.62	0.00 \pm 0.00	80.82 \pm 0.47	0.00 \pm 0.00	81.82 \pm 0.37	0.00 \pm 0.00	40.00 \pm 1.37	0.00 \pm 0.00	65.40 \pm 0.93	0.00 \pm 0.00	62.12 \pm 0.40	0.00 \pm 0.00
STMC	46.74 \pm 0.67	0.12 \pm 0.00	46.17 \pm 0.41	0.12 \pm 0.00	68.67 \pm 0.22	0.07 \pm 0.00	69.18 \pm 0.57	0.07 \pm 0.00	31.03 \pm 1.87	0.16 \pm 0.00	50.01 \pm 2.37	0.10 \pm 0.01	47.12 \pm 1.11	0.06 \pm 0.00
DART	40.46 \pm 0.55	0.28 \pm 0.00	43.32 \pm 0.96	0.27 \pm 0.00	69.96 \pm 0.65	0.10 \pm 0.00	69.05 \pm 0.45	0.11 \pm 0.00	25.63 \pm 2.62	0.35 \pm 0.01	52.59 \pm 1.36	0.20 \pm 0.00	50.58 \pm 0.75	0.13 \pm 0.00
UniMotion	51.78 \pm 0.69	0.09 \pm 0.00	53.24 \pm 1.29	0.08 \pm 0.00	74.68 \pm 0.40	0.04 \pm 0.00	75.06 \pm 0.31	0.05 \pm 0.00	33.00 \pm 2.03	0.11 \pm 0.00	56.74 \pm 0.89	0.08 \pm 0.00	56.81 \pm 0.62	0.04 \pm 0.00
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.

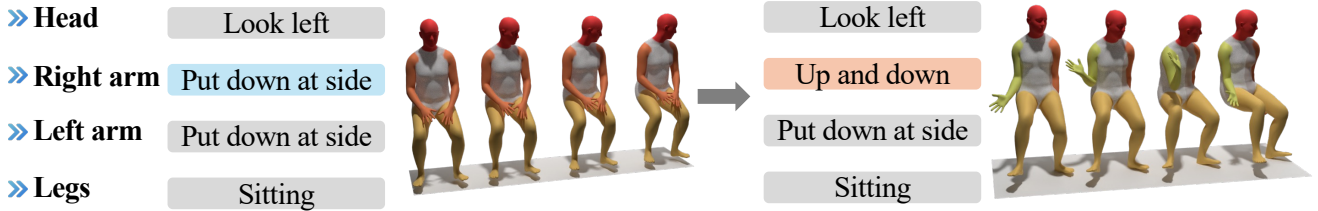


Figure 7. Motion editing example. Changing the right arm prompt from **Put down at side** to **Up and down** while keeping all other body-part prompts unchanged produces a coherent edited motion.

formance 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 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.

Comparison with CoMo and FineMoGen. CoMo [17] and FineMoGen [64] accept body-part text as input, offering partial spatial control, but lack temporal decomposition — their part-level conditioning is applied uniformly across the full sequence without temporal segmentation. As they do not support our temporal multi-level control setting, we report them separately here rather than in the main comparison. We adapt both methods to our setting and evaluate on FrankenStein in Table 9. FrankenMotion outperforms both baselines across all metrics. We also report a sensitivity analysis by injecting annotation noise: we replace the Right Leg text prompt with a random label at varying ratios (0.1–0.4) to simulate annotation errors. As shown in Table 9 (Ours-noise), our model maintains high generation quality across all noise levels, demonstrating robustness of our

Method	Avg-part semantic correctness		Per-action semantic correctness		Per-action realism	
	R@1 \uparrow	M2T \uparrow	R@1 \uparrow	M2T \uparrow	FID \downarrow	Div. \rightarrow
	GT	52.04	0.71	54.83	0.77	0.00
CoMo	42.01	0.66	43.70	0.71	0.04	52.08
FineMoGen	38.00	0.64	37.50	0.66	0.19	48.96
Ours-noise 0.1	45.36	0.69	47.39	0.74	0.04	54.17
Ours-noise 0.2	44.71	0.68	47.33	0.74	0.05	54.20
Ours-noise 0.3	45.30	0.68	46.78	0.73	0.04	53.95
Ours-noise 0.4	46.34	0.69	47.05	0.74	0.04	53.76
Ours-Qwen	45.95	0.68	46.62	0.74	0.05	53.39
Ours-CLIP	47.21	0.69	48.10	0.75	0.04	53.82

Table 9. **Evaluating text-to-motion generation.** We report semantic correctness and realism at the part and action level.

stochastic masking strategy which prevents over-reliance on any single part-level signal.

Text Encoder Analysis. We compare our default CLIP text encoder against Qwen-0.6B in Table 9 (Ours-Qwen vs. Ours-CLIP). The two variants yield similar results, confirming that our performance gains stem from the hierarchical conditioning design rather than text encoder capacity. We use CLIP for fair comparison with all baselines, and will release code supporting modular text encoders to facilitate future scaling.

User Study. We conduct two user studies. First, comparing FrankenMotion against CoMo [17] and FineMoGen [64] across 20 randomly sampled prompts with 20 participants, our method was preferred in 70% of cases over CoMo (20%) and FineMoGen (10%). Second, comparing against UniMotion [20] and DART [71] across 20 sequences with 28 participants, our method was preferred in 75% of cases, strongly corroborating the quantitative results.

Motion Editing Example. Thanks to disentangled part-level conditioning, FrankenMotion supports precise motion editing: users can modify a single body-part prompt (e.g., right arm: “put down at side” \mapsto “up and down”) while preserving all other part motions. Figure 7 shows that the model integrates the new instruction without disrupting temporal coherence.

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_seg_0, humanml3d_002139_1"
      },
      {
        "text": "holding arms forward",
        "start": 0.0,
        "end": 1.586,
        "confidence": 5,
        "reasoning": "From BABEL",
        "source": "babel_00629_seg_0"
      }
    ],
    "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