User Request

Motion Generation Timeline

We utilize a timeline to control motion generation. The format of the timeline is as follows:

```
condition # start time # end time # body part A # body part B …
```

Body Parts

- Head
- Left Arm
- Right Arm
- Spine
- Legs

Guidelines

- Start the timeline at **0 seconds**.
- Ensure all conditions are **asynchronous**, meaning their start or end times do not coincide exactly.
- Audio duration is calculated as:

  ```
  (end Hz - start Hz) / 16000
  ```

- Each text description should last **no less than 5 seconds** and **no more than 5 seconds**.
- Be creative with your text descriptions and don't limit yourself to the examples provided. Textual descriptions should be **concrete actions**, not abstract.
- The audio controls the upper body by default. For example:

```
speak:1_wayne_0_103_103,$9248$99808 # 2.578 # 8.238 # left arm # right arm # head
```

Example Timeline Entries for Semantic Motion

```
a man walks # 0.0 # 8.0 # legs # spine
wave hands # 7.0 # 11.0 # left arm # right arm
speak:1_wayne_0_103_103,$9248$99808 # 2.578 # 8.238 # left arm # right arm # head
speak:1_wayne_0_103_103,$107552$196064 # 9.722 # 15.254 # left arm # right arm # head
```

In this scenario, the speech content involves a greeting. Consequently, we issue the command "wave hands" to control the arm movements from 7.0 to 11.0 seconds. The speech audio spans from 2.578 to 15.254 seconds, creating an overlap between 7.0 and 11.0 seconds during which both the audio and the text command aim to control the arms.

According to our conflict resolution rules, the command with fewer control elements takes precedence. In this case, the "wave hands" command, which specifically governs the arms, prevails during the overlapping period. As a result, users can effectively manage upper body motions using text commands while the speech audio is active.

Text Description Controls Examples

- For a greeting:

  ```
  wave hands # start time # end time # left arm # right arm
  ```

- For expressing anger:

  ```
  punch # start time # end time # left arm # right arm
  ```

- For daily conversation:

  ```
  walking casually # start time # end time # legs # spine
  ```

Emotional Expression

To effectively convey emotions, utilize **full-body language** rather than limiting control to a single body part. Avoid commands like:

```
cry # start time # end time # head
```

Instead, use:

```
cry # start time # end time # left arm # right arm
```

Here are the audio clips:

```
speak:5_stewart_0_87_87,$630304$713696, the speech content is about a man
discussing the weather
speak:5_stewart_0_87_87,$718368$807392, the speech content is about a man
expressing excitement
```

Please generate a timeline that aligns the motion semantically with the speech content.

LLM Responses

GPT-4o-mini

Here’s a timeline that aligns the motion with the provided audio clips, ensuring that all conditions are asynchronous and that the actions are concrete and expressive.

Motion Generation Timeline

```
a man stands and looks around # 0.0 # 5.0 # head # spine
speak:5_stewart_0_87_87,$630304$713696 # 1.0 # 6.0 # head # left arm # right arm
gesture towards the sky # 5.0 # 10.0 # left arm # right arm # head
speak:5_stewart_0_87_87,$718368$807392 # 8.0 # 13.0 # head # left arm # right arm
jump with excitement # 10.0 # 15.0 # legs # spine
```

Breakdown of the Timeline

1. **Action:** "a man stands and looks around"
   - **Time:** 0.0 to 5.0 seconds
   - **Body Parts:** head, spine
   - **Description:** The man is getting ready to start his conversation.
2. **Speech:**
   - **Audio Clip:** Discussing the weather
   - **Time:** 1.0 to 6.0 seconds
   - **Body Parts:** head, left arm, right arm
   - **Description:** The man speaks while using his arms to express thoughts about the weather.
3. **Action:** "gesture towards the sky"
   - **Time:** 5.0 to 10.0 seconds
   - **Body Parts:** left arm, right arm, head
   - **Description:** This gesture emphasizes the discussion about the weather.
4. **Speech:**
   - **Audio Clip:** Expressing excitement
   - **Time:** 8.0 to 13.0 seconds
   - **Body Parts:** head, left arm, right arm
   - **Description:** The man’s excitement is conveyed through animated speech and gestures.
5. **Action:** "jump with excitement"
   - **Time:** 10.0 to 15.0 seconds
   - **Body Parts:** legs, spine
   - **Description:** The man jumps to express his excitement, enhancing the overall emotional impact.

This timeline effectively integrates speech and motion, ensuring a coherent and expressive representation of the man's actions and emotions related to the content of the audio clips.

GPT-4o

To create a timeline that aligns motion with the provided speech content, we need to consider the semantic content of the audio clips and ensure that the motions reflect the emotions and actions being described. Here, we have two audio clips: one discussing the weather and another expressing excitement. We'll craft motions that fit these themes while adhering to the guidelines.

Timeline:

1. **Discussing the Weather:**
    - **Audio Clip:** `speak:5_stewart_0_87_87,$630304$713696`
    - **Duration Calculation:** ((713696 - 630304) / 16000 = 5.222) seconds

    ```
    speak:5_stewart_0_87_87,$630304$713696 # 0.0 # 5.222 # left arm # right arm # head
    gesture thoughtfully # 0.5 # 5.5 # left arm # right arm # head
    ```

    Here, the "gesture thoughtfully" command is used to reflect the contemplative nature of discussing the weather. The slight overlap at the beginning and end ensures asynchronous timing.

2. **Expressing Excitement:**
    - **Audio Clip:** `speak:5_stewart_0_87_87,$718368$807392`
    - **Duration Calculation:** ((807392 - 718368) / 16000 = 5.563) seconds

    ```
    speak:5_stewart_0_87_87,$718368$807392 # 6.0 # 11.563 # left arm # right arm # head
    jump with joy # 6.5 # 12.0 # legs # spine # left arm # right arm
    ```

    The "jump with joy" command captures the excitement expressed in the speech. The motion starts shortly after the audio begins, ensuring there is no exact overlap with the start of the clip.

Explanation:

- **Thoughtful Gestures:** For the weather discussion, the thoughtful gestures involve subtle arm and head movements, aligning with the nature of the conversation.
- **Joyful Jumping:** For expressing excitement, a full-body action like jumping conveys the emotion effectively, using both arms and legs.

This timeline ensures that each motion is semantically aligned with the speech content, and all conditions are asynchronous, adhering to the guidelines provided.

GPT-o1-mini

```
speak:5_stewart_0_87_87,$630304$713696, the speech content is about a man discussing the weather # 0.0 # 5.212 # left arm # right arm # head
nod head and gesture with arms # 0.1 # 5.1 # left arm # right arm # head

speak:5_stewart_0_87_87,$718368$807392, the speech content is about a man expressing excitement # 5.3 # 10.864 # left arm # right arm # head
jump up and down with arms raised # 5.4 # 10.4 # left arm # right arm # legs # spine
```

Timeline Breakdown

1. **First Speech Segment**
    - **Audio**: Discussing the weather
        - **Start Time**: 0.0 seconds
        - **End Time**: 5.212 seconds
        - **Controlled Body Parts**: Left Arm, Right Arm, Head
    - **Motion Command**: Nod head and gesture with arms
        - **Start Time**: 0.1 seconds
        - **End Time**: 5.1 seconds
        - **Controlled Body Parts**: Left Arm, Right Arm, Head
2. **Second Speech Segment**
    - **Audio**: Expressing excitement
        - **Start Time**: 5.3 seconds
        - **End Time**: 10.864 seconds
        - **Controlled Body Parts**: Left Arm, Right Arm, Head
    - **Motion Command**: Jump up and down with arms raised
        - **Start Time**: 5.4 seconds
        - **End Time**: 10.4 seconds
        - **Controlled Body Parts**: Left Arm, Right Arm, Legs, Spine

Notes

- **Asynchronous Conditions**: The start and end times of motion commands are offset by 0.1 seconds from the corresponding audio segments to ensure they are asynchronous.
- **Duration Compliance**: Each motion command lasts exactly 5 seconds, adhering to the guidelines.
- **Full-Body Language**: Especially for expressing excitement, the motion command engages multiple body parts to convey the emotion effectively.
- **Conflict Resolution**: During overlapping periods, motion commands take precedence over audio controls for their specified body parts, ensuring clear and intentional motion behavior.

This timeline ensures that the motions are semantically aligned with the speech content, providing a coherent and expressive experience.

GPT-o1-preview

```
Gesture about the weather # 1.0 # 6.0 # left arm
Walk forward slowly # 1.1 # 6.1 # legs # spine
speak:5_stewart_0_87_87,$630304$713696 # 0.5 # 5.712 # left arm # right arm # head
Raise hands in excitement # 6.7 #11.7 # left arm # right arm
Walk energetically # 6.4 #11.4 # legs # spine
speak:5_stewart_0_87_87,$718368$807392 #6.3 #11.864 # left arm # right arm # head
```

Claude3.5-sonnet

I'll help create a timeline that matches the semantic content of the speech, considering the weather discussion followed by excitement. I'll calculate the audio durations and create appropriate motion sequences.

First audio: (713696 - 630304) / 16000 = 5.212 seconds
Second audio: (807392 - 718368) / 16000 = 5.564 seconds

Here's the timeline with semantic motions:

```
standing casually # 0.0 # 5.0 # legs # spine
look at sky # 0.2 # 5.2 # head
gesture at clouds # 2.0 # 7.0 # left arm # right arm
speak:5_stewart_0_87_87,$630304$713696 # 0.0 # 5.212 # left arm # right arm #
head
jump excitedly # 5.3 # 10.3 # legs # spine
raise arms enthusiastically # 7.2 # 12.2 # left arm # right arm
nod vigorously # 7.4 # 12.4 # head
speak:5_stewart_0_87_87,$718368$807392 # 5.564 # 11.128 # left arm # right arm #
head
```

This timeline creates a natural progression:

1. First segment (weather discussion):
    - Starts with a casual standing pose
    - Looking up at the sky
    - Making gestures while talking about weather
2. Second segment (excitement):
    - Transitions to excited jumping
    - Enthusiastic arm movements
    - Energetic head nodding

The motions are asynchronous and each text description lasts between 5-5.5 seconds. The full body is engaged in expressing both the weather discussion and excitement, creating natural and expressive movements.

Gemini-1.5 pro

```
# Man discussing the weather

walking casually # 0.0 # 10.0 # legs # spine
speak:5_stewart_0_87_87,$630304$713696 # 4.451 # 8.903 # left arm # right arm # head
glance at the sky # 6.0 # 7.5 # head
shrug shoulders # 9.0 # 10.5 # left arm # right arm # head

# Man expressing excitement

speak:5_stewart_0_87_87,$718368$807392 # 11.0 # 15.453 # left arm # right arm # head
clap hands # 13.0 # 14.0 # left arm # right arm
jump excitedly # 14.5 # 16.0 # legs # spine
point forward # 15.5 # 16.5 # right arm
```

Explanation:

- **Walking casually** is maintained throughout the first part to depict a natural walking motion while the man discusses the weather.
- **Glance at the sky** is timed to coincide with the potential mention of weather elements like clouds or sun in the speech.
- **Shrug shoulders** is used to express uncertainty or indifference, aligning with the possible nature of a weather discussion.
- **Clap hands** is a natural expression of excitement and is timed to occur during the second speech segment.
- **Jump excitedly** further emphasizes the excitement conveyed in the speech.
- **Point forward** suggests the man is excited about something specific and directs attention towards it.

Note:

- This timeline utilizes full-body language for emotional expressions like shrugging and jumping.
- The timings are asynchronous to ensure smooth transitions and avoid robotic movements.
- The duration of each action falls within the 5-second limit.
- Text descriptions are concrete actions, clearly defining the desired motion.