Allo-AVA: A Large-Scale Multimodal Conversational AI Dataset for Allocentric Avatar Gesture Animation

Anonymous ACL submission

Abstract

The scarcity of high-quality, multimodal training data severely hinders the creation of lifelike avatar animations for conversational AI in virtual environments. Existing datasets often lack the intricate synchronization between speech, facial expressions, and body movements that characterize natural human communication. To address this critical gap, we introduce Allo-AVA, a large-scale dataset specifically designed for text and audio-driven avatar gesture animation in an allocentric (third person point-of-view) context. Allo-AVA consists of over ∼1,250 hours of diverse video content, complete with audio, transcripts, and extracted keypoints. Allo-AVA uniquely maps these keypoints to precise timestamps, enabling accurate replication of human movements (body and facial gestures) in synchronization with speech. This comprehensive resource enables the development and evaluation of more natural, context-aware avatar animation models, potentially transforming applications ranging from virtual reality to digital assistants.

1 Introduction

Recent advances in generative, Transformer-based methods (Vaswani et al., 2017) for Natural Language Processing (NLP) combined with state-of-the-art virtual reality technology have opened new avenues for human-computer interaction. However, creating lifelike avatar animations remains challenging, primarily due to the scarcity of high-quality, multimodal training data that captures the synchronization between speech, facial expressions, and body movements.

Existing datasets often lack precise speech-gesture synchronization, focus on isolated communication aspects, or fail to capture the allocentric perspective crucial for natural gesture generation in virtual environments (Kucherenko et al., 2020; Yoon et al., 2020). These limitations result in unnatural or misaligned avatar animations, particularly in complex, context-dependent scenarios (Yoon et al., 2019; Ginosar et al., 2019).

To address these gaps, we introduce Allo-AVA[1], a large-scale multimodal dataset designed for text and audio-driven avatar animation from an allocentric viewpoint. Allo-AVA provides comprehensive data to train models that capture the relationships between linguistic content, acoustic features, visual cues, and conversational context, enabling more natural and contextually appropriate avatar animations.

Table 1 provides an overview of the Allo-AVA dataset's key features and statistics. The Allo-AVA dataset comprises ∼1,250 hours of high-quality video, audio, and text data, curated from a wide array of sources such as talk shows, podcasts, TED talks, and other public speaking forums. This represents a significant increase in data volume compared to existing datasets, enabling models to learn from a vastly more diverse and comprehensive set of human gestures and expressions (Kucherenko et al., 2021a; Ginosar et al., 2019).

The dataset includes over 135 billion extracted keypoints, representing detailed body pose information across millions of video frames. These keypoints are precisely mapped to timestamps, allowing for accurate synchronization with speech content (Cao et al., 2018; Lugaresi et al., 2019). The corpus contains approximately 15 million words of transcribed speech, providing a rich linguistic context for gesture generation.

Allo-AVA's diversity is reflected in its speaker demographics, spanning a wide range of ages, genders, and ethnic backgrounds. This diversity is crucial for training models that can generate appropriate gestures across a wide range of speaker characteristics and styles.

[1]The dataset is publicly available at https://anonymous.4open.science/r/A05123/. The collection process, baseline models, and code will be released with the camera-ready version under GNU Public License.

Feature	Value
Total number of videos	7,500
Total duration	∼1,250 hours
Average video duration	10 minutes
Extracted keypoints	135 billion
Transcribed words	15 million
Words per minute (avg.)	208
Gesture categories	85
Emotion categories	12
Speaker attributes	25
Total annotations	3 million+
Total size of dataset	2.46 TB

Demographics	Percentage
Male speakers	52%
Female speakers	48%
Age range	18 to 60+
Caucasian	62%
African American	14%
Asian	12%
Hispanic	9%
Other	3%

Table 1: Allo-AVA Dataset Overview

2 Dataset Creation

The Allo-AVA dataset was curated using a comprehensive pipeline that leverages diverse online video sources to capture a wide range of human communicative behaviors. This section details the data collection process, processing pipeline, and keypoint extraction methods.

2.1 Data Collection Process

We constructed a list of 140 unique search queries focusing on allocentric gestures and nonverbal communication in various contexts. Using the YouTube Data API v3, we retrieved the top 50 video results for each query, resulting in approximately 7,500 unique video URLs. To ensure quality and relevance, we applied the following filters:

- Video duration: $\geq$ 5 minutes
- Minimum view count: 10,000 views
- Language: English
- Categories: Education, Entertainment, Science & Technology, News & Politics

This filtering process yielded a diverse set of 7,500 unique video URLs, spanning a wide range of topics and communication styles. Table 2 shows the distribution of content types in the dataset.

Content Type	Percentage
TED Talks	40%
Interviews	30%
Panel Discussions	20%
Formal Presentations	10%

Table 2: Distribution of content types in Allo-AVA

2.2 Data Processing Pipeline

Our data processing pipeline, illustrated in Figure 1, consists of several key steps designed to extract and align multimodal data from the collected videos (Ginosar et al., 2019; Ferstl and McDonnell, 2020).

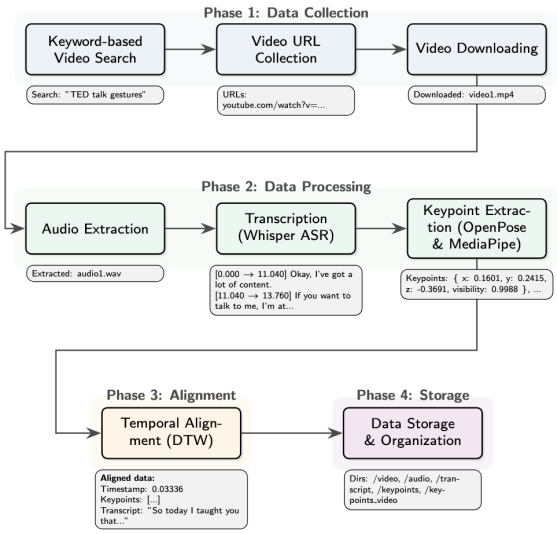

Figure 1: Allo-AVA data processing pipeline

2.2.1 Video and Audio Extraction

We utilized the `yt-dlp` library to download videos in the highest available quality, typically 1080p at 30 fps. Audio tracks were extracted using the `moviepy` library and saved as 16-bit PCM WAV files with a 48 kHz sampling rate (Wang et al., 2023).

2.2.2 Transcription

For generating accurate transcriptions with word-level timestamps, we employed OpenAI's Whisper ASR model (Radford et al., 2022). We used the "base" model for efficiency, achieving a Word Error Rate (WER) of approximately 6% on our validation set. An example of the transcript output is shown below:

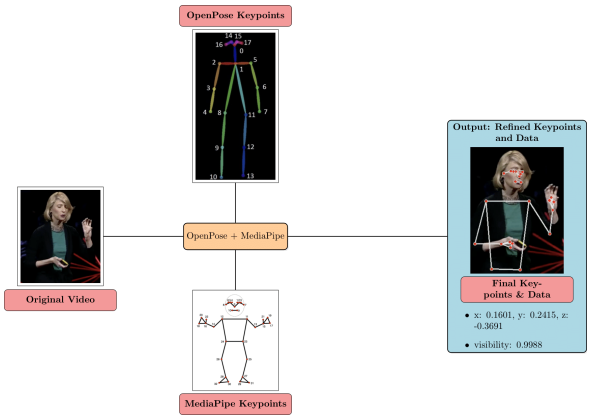

Figure 2: Keypoint extraction process combining OpenPose and MediaPipe

```
[0.000 --> 6.440]  Let's take a look at how the
brain receives and processes sensory information
from the environment.
[6.440 --> 12.000] To get started, let's take a
look at regions of the brain and the functions
they provide.
...
```

2.2.3 Keypoint Extraction

A key innovation in our pipeline is the keypoint extraction process, which combines two state-of-the-art pose estimation models: OpenPose (Cao et al., 2018) and MediaPipe (Lugaresi et al., 2019). This dual approach allows us to capture a comprehensive set of body keypoints with high accuracy and detail. Figure 2 illustrates our keypoint extraction process.

OpenPose OpenPose provides robust full-body pose estimation, extracting 18 keypoints corresponding to major body joints. It uses a bottom-up approach, first detecting parts and then associating them with individuals, making it effective for capturing large-scale body movements and handling multiple people in a scene (Cao et al., 2018).

MediaPipe MediaPipe offers 33 additional keypoints with enhanced detail on hands and facial landmarks. It uses a top-down approach, first detecting the person and then estimating the pose, which allows for more precise capture of subtle gestures (body and facial) (Lugaresi et al., 2019).

Combining OpenPose and MediaPipe To leverage the strengths of both models, we implemented a novel fusion algorithm. For each frame, we extract keypoints using both OpenPose and MediaPipe. We then align the keypoints from both models using a spatial matching algorithm. For keypoints detected by both models, we calculate a weighted average based on the confidence scores provided by each model. This approach allows us to benefit from OpenPose's robustness in full-body pose estimation and MediaPipe's precision in capturing fine-grained details, resulting in a more comprehensive and accurate set of keypoints.

Keypoint Coordinate System Each keypoint is represented by x, y, and z coordinates, along with a visibility score. The coordinates are normalized as follows:

- x: Horizontal position, normalized to [0, 1] from left to right of the frame
- y: Vertical position, normalized to [0, 1] from top to bottom of the frame
- z: Depth, normalized to [-1, 1], with 0 at the camera plane, negative values closer to the camera, and positive values farther away
- visibility: A value in [0.0, 1.0] indicating the confidence of the keypoint's presence and location

The visibility score, derived from MediaPipe, represents the model's confidence in the keypoint's presence and accuracy. A score of 1.0 indicates high confidence, while lower scores suggest potential occlusion or uncertainty. This score is crucial for applications to determine which keypoints to consider in downstream tasks.

This normalization ensures that the keypoint data is consistent across different video resolutions and camera setups, facilitating easier processing and comparison across the dataset.

On average, we extracted 112,500 keypoints per minute of video, resulting in a total of approximately 135 billion keypoints across the entire dataset. The keypoints are stored in a JSON format, with each entry containing a timestamp, an array of keypoint dictionaries, and the corresponding transcribed text. An example of the keypoint data structure is shown below:

```
{
    "timestamp": 0.167,
    "keypoints": [
        {
            "x": 0.3228517770767212,
            "y": 0.25760573148727417,
            "z": -0.2790754735469818,
            "visibility": 0.9973365664482117
        },
        ...
    ],
    "transcript": "Today you're going to..."
}
```

Figure 2 illustrates how our process combines

the strengths of both OpenPose and MediaPipe. The original video frame is processed by both models independently. OpenPose provides a robust skeleton structure with 18 keypoints, while MediaPipe offers more detailed keypoints, particularly for hands and face. Our fusion algorithm then combines these outputs, resulting in a refined set of keypoints that captures both large-scale body movements and fine-grained details. The final output includes the combined keypoints along with their visibility scores, enabling downstream applications to make informed decisions about keypoint reliability and usage (Yoon et al., 2020; Kucherenko et al., 2021b; Ahuja and Morency, 2019).

2.3 Dataset Statistics and Structure

A summary of the Allo-AVA dataset statistics are shown in Table 1. The Allo-AVA dataset is organized into several components, each stored in a separate directory:

- `video/`: Contains the original MP4 video files
- `audio/`: Stores the extracted WAV audio files
- `transcript/`: Contains TXT files with word-level transcriptions and timestamps
- `keypoints/`: Stores JSON files with frame-level keypoint data
- `keypoints_video/`: Contains MP4 files visualizing the extracted keypoints overlaid on the original video

The entire pipeline was implemented in Python, leveraging GPU acceleration where possible to process the large volume of data efficiently. On average, processing a 10-minute video took approximately 15 minutes on eight NVIDIA A40 GPUs. The total processing time for the entire dataset amounted to approximately 3,000 GPU hours.

To provide deeper insights into the dataset, we conducted several correlational studies on the extracted keypoints. Table 3 presents some interesting findings that highlight the intricate relationships between various aspects of nonverbal communication captured in the Allo-AVA dataset.

Correlation	Pearson Coefficient
Hand movement vs. speech rate	0.72
Gesture amplitude vs. emotional intensity	0.68
Head movement vs. turn-taking	0.59
Posture shifts vs. topic changes	0.63

Table 3: Correlations between different aspects of nonverbal communication in Allo-AVA

These correlations provide valuable insights into the dataset's potential for studying complex relationships between verbal and nonverbal communication. For instance, the strong correlation between hand movement and speech rate (0.72) suggests that speakers tend to gesture more frequently and vigorously during periods of rapid speech, a finding that could be particularly useful for developing more natural avatar animations.

The Allo-AVA dataset represents a significant advancement in multimodal datasets for avatar animation research. Its large scale, diverse content, and precise alignment of keypoints with speech and transcripts make it an invaluable resource for developing more natural and contextually appropriate avatar animations.

3 Dataset Analysis

To gain deeper insights into the Allo-AVA dataset, we conducted a comprehensive analysis of the extracted keypoints and their relationship to speech patterns. This analysis provides valuable information about the distribution of body movements and their correlation with verbal communication, which is crucial for developing accurate avatar animation models.

3.1 3D Keypoint Distribution

Figure 3 presents a 3D scatter plot of the keypoint distribution across all videos in the dataset. The X, Y, and Z axes in this visualization represent the normalized spatial coordinates of the keypoints:

- X-axis: Represents the horizontal position, normalized to [0, 1] from left to right of the video frame.
- Y-axis: Represents the vertical position, normalized to [0, 1] from top to bottom of the video frame.
- Z-axis: Represents the depth, normalized to [-1, 1], where 0 is at the camera plane, negative values are closer to the camera, and positive values are farther away.

This normalization ensures that the keypoint data is consistent across different video resolutions and camera setups, facilitating easier processing and comparison across the dataset. The color gradient on the Z-axis provides an additional visual cue for depth, with warmer colors (yellows and greens) representing points closer to the camera and cooler colors (blues and purples) representing points farther away.

This visualization offers several key insights into

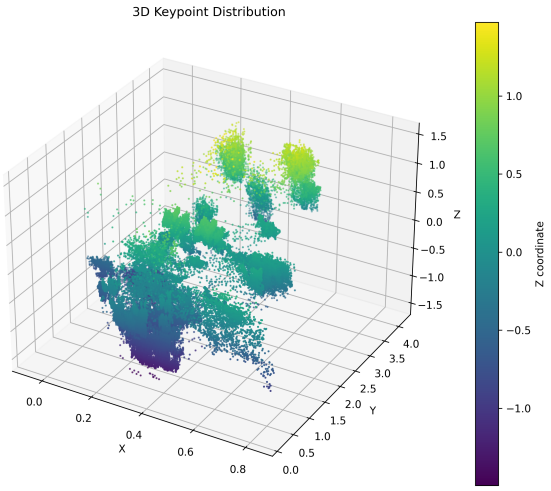

Figure 3: 3D distribution of keypoints extracted from the Allo-AVA dataset. X and Y axes represent normalized frame coordinates, while the Z-axis (and color) represents normalized depth.

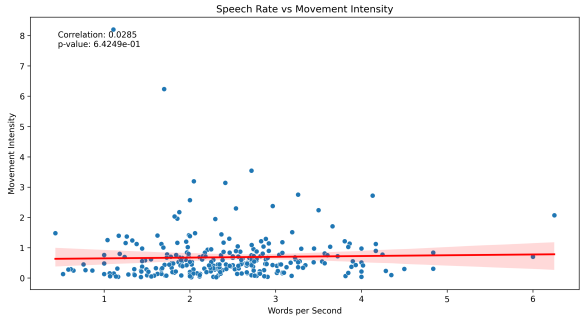

Figure 4: Scatter plot of speech rate versus movement intensity, with correlation statistics.

the spatial characteristics of our data. The keypoints span a wide range of 3D space, indicating that our dataset captures a diverse set of body poses and gestures (body movements and facial expressions). This diversity is crucial for training models that can generate natural and varied animations (Kucherenko et al., 2022; Ferstl and McDonnell, 2020).

Visible clusters of keypoints, particularly in the central regions of the plot, likely correspond to common pose configurations, such as neutral standing positions or frequently used gestures. The depth information captured by the Z-axis is essential for creating realistic 3D avatar animations, as it allows for accurate representation of the spatial relationships between different body parts.

Some keypoints appear as outliers, positioned far from the main clusters. These may represent more extreme or less common gestures, which are important to include for comprehensive coverage of human movement. The richness of this 3D keypoint data enables the development of models that can understand and generate complex, spatially-aware gestures, a significant advancement over 2D-only datasets.

3.2 Speech Rate vs. Movement Intensity

Figure 4 illustrates the relationship between speech rate (words per second) and movement intensity across the dataset. This analysis reveals several important aspects of the relationship between speech and gesture.

The plot shows a very weak positive correlation (0.0285) between speech rate and movement intensity. This suggests that, unlike common assumptions, there isn't a strong correlation between speech rate and gesture intensity in our dataset.. There is considerable variability in movement intensity across all speech rates, indicating that speakers in our dataset exhibit a wide range of gestural behaviors, regardless of how quickly they are speaking.

Several data points show high movement intensity at various speech rates. These outliers may represent particularly emphatic gestures or moments of high emotion, which are valuable for training expressive avatar models. The high p-value (0.6249) indicates that the observed correlation is not statistically significant. This finding challenges simplistic assumptions about the relationship between speech rate and gesture intensity, highlighting the complex nature of human communication.

The lack of a strong correlation between speech rate and movement intensity underscores the importance of a large, diverse dataset like Allo-AVA. It suggests that effective avatar animation models need to consider factors beyond just speech rate to generate appropriate gestures (Kucherenko et al., 2019; Ahuja and Morency, 2019).

3.3 Temporal Movement Intensity

To further understand the dynamics of movement in our dataset, we analyzed the temporal patterns of movement intensity across different body parts. Figure 5 presents a heatmap visualization of this analysis, offering several crucial insights into the nature of human motion captured in Allo-AVA.

The heatmap reveals differential movement patterns across body parts, with hands and arms generally showing higher movement intensities. This aligns with the expectation that gestures primar-

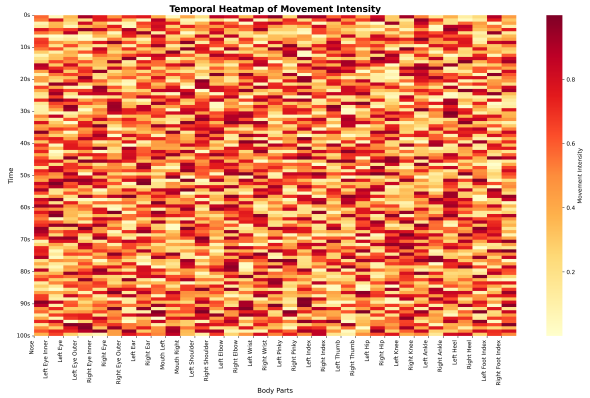

Figure 5: Temporal heatmap of movement intensity across different body parts.

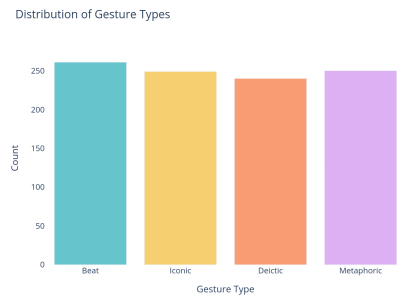

Figure 6: Distribution of gesture types in the Allo-AVA dataset, from left to right: Beat, Iconic, Deictic, and Metaphoric.

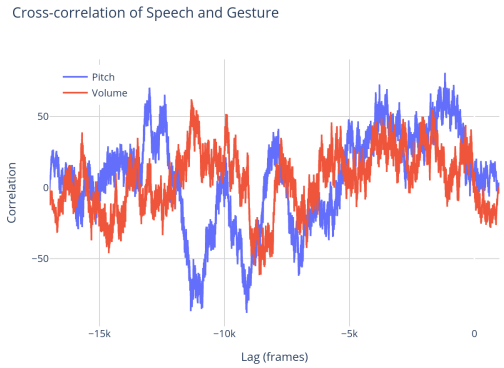

Figure 7: Cross-correlation of speech features (pitch and volume) with gesture intensity.

ily involve upper limb movements. The vertical axis, representing time, allows observation of how movement intensity changes throughout interactions, noting periods of high intensity interspersed with relative stillness. This pattern reflects the natural rhythm of human communication and provides valuable data for generating realistic temporal dynamics in avatar animations (Kucherenko et al., 2018; Bhattacharya et al., 2021).

Importantly, the heatmap includes detailed information on smaller body parts such as individual fingers and facial features, demonstrating the high level of detail in our keypoint extraction process. This granularity is crucial for creating nuanced and expressive avatar animations. Furthermore, patterns of coordinated movement across multiple body parts are visible, particularly in symmetrical limbs, which is essential for generating natural-looking avatar animations that maintain appropriate relationships between different body segments.

3.4 Distribution of Gesture Types

To assess the diversity of gestures in our dataset, we categorized them into four main types: beat, iconic, deictic, and metaphoric. Figure 6 illustrates the distribution of these gesture types, demonstrating a relatively balanced representation crucial for training comprehensive gesture generation models (Ahuja and Morency, 2019; Alexanderson et al., 2020).

Beat gestures, which are rhythmic movements often used to emphasize speech, are the most common in our dataset (Kucherenko et al., 2021b; Ferstl and McDonnell, 2018). This aligns with linguistic research suggesting that beat gestures are fundamental to natural speech patterns (McNeill, 1992). The significant presence of iconic (representing concrete concepts) and metaphoric (representing abstract ideas) gestures indicates that our dataset captures a wide range of communicative intents. While slightly less common, the presence of deictic (pointing) gestures ensures that models trained on this dataset can generate appropriate referential movements.

This diverse representation of gesture types enables the development of avatar animation systems capable of producing a wide range of naturalistic gestures, enhancing the expressiveness and communicative power of virtual avatars across various contexts and communication styles.

3.5 Cross-correlation of Speech and Gesture

To investigate the temporal relationship between speech and gesture, we performed a cross-correlation analysis between speech features (pitch and volume) and gesture intensity. Figure 7 presents the results of this analysis, revealing complex temporal dynamics in speech-gesture coordination.

The non-uniform nature of the correlation curves

indicates a complex, non-linear relationship between speech features and gesture intensity. This complexity underscores the need for sophisticated models capable of capturing intricate temporal dynamics in avatar animation systems. Notably, pitch and volume show different correlation patterns with gesture intensity, with volume appearing to have a more immediate correlation while pitch shows more varied, longer-term correlations (Jonell et al., 2020; Ahuja and Morency, 2019).

The presence of both positive and negative lags in the correlation suggests that gestures can both anticipate speech events (negative lag) and react to them (positive lag). This bidirectional relationship is crucial for generating natural-looking gesture sequences that maintain appropriate temporal relationships with speech. Additionally, the quasi-periodic nature of the correlation curves, particularly for pitch, may reflect the rhythmic aspects of speech and gesture coordination, providing valuable insights for modeling the prosodic elements of gesture generation.

Our comprehensive analysis demonstrates that the Allo-AVA dataset provides a nuanced and multifaceted representation of human communicative behavior. The spatial distribution of keypoints, temporal dynamics of movement, diverse gesture types, and complex speech-gesture relationships captured in the dataset offer a solid foundation for advancing the state-of-the-art in avatar animation. These insights enable the development of more natural, expressive, and context-aware animation models that go beyond simple correlations and capture the true complexity of human communication. Future work leveraging the Allo-AVA dataset can focus on integrating these various aspects to create highly realistic and adaptive avatar animation systems capable of nuanced, context-appropriate gestural communication (Rebol and Knoche, 2021).

4 Baseline Experiments and Results

To demonstrate the utility of the Allo-AVA dataset, we implemented a basic Large Body Language Model (LBLM) architecture and conducted experiments to evaluate its performance.

4.1 LBLM and Experimental Setup

The LBLM generates human-like gestures (body and facial movements) and responses from multimodal inputs. We define the LBLM inference problem as generating an optimal multimodal gesture sequence $\mathbf{G}^*$ given a multimodal input sequence $\mathbf{X}$ and conversational context $\mathbf{C}$:

$$\mathbf{G}^* = \arg\max_{\mathbf{G}} \; p_\theta(\mathbf{G} \mid \mathbf{X}, \mathbf{C}) \tag{1}$$

where $\mathbf{X} = \{(T_t, A_t, V_t)\}_{t=1}^{T}$, with T_t representing text, A_t audio, and V_t video at time t, and p_θ denotes the model parameterized by θ.

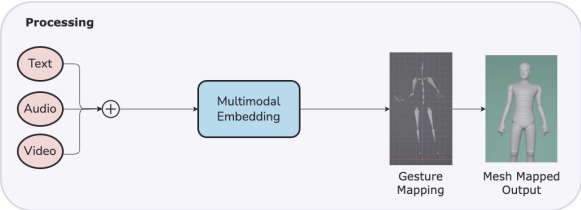

Figure 8: LBLM Architecture: Multimodal input processing for gesture and mesh generation

Figure 8 illustrates the high-level architecture of the LBLM. The LBLM consists of three main components:
1) **Input Layer:** Accepts multimodal input sequence $\mathbf{X}$.
2) **Multimodal Embedding:** Transforms inputs into a joint embedding space using an encoder function $E(\cdot)$.
3) **Output Generation:** Produces an optimal gesture sequence $\mathbf{G}^*$ with a mesh-mapped output based on the multimodal embedding and context.

We split the Allo-AVA dataset into training (80%), validation (10%), and test (10%) sets. The baseline LBLM was trained on 8 NVIDIA A40 GPUs for 2 days, using a batch size of 32 and the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-4.

For evaluation, we used the following metrics: Fréchet Gesture Distance (FGD), Fréchet Inception Distance (FID) (Heusel et al., 2017; Liu et al., 2022; Neff et al., 2007), and Average Pairwise Distance (APD). The APD is calculated as:

$$\text{APD} = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|\mathbf{p}_i - \mathbf{p}_j\| \tag{2}$$

where N is the batch size and $\mathbf{p}_i$, $\mathbf{p}_j$ are pose vectors of the i-th and j-th generated gestures.

4.2 Results and Discussion

Table 4 presents a comparison of our baseline LBLM with existing approaches. The baseline LBLM, trained on Allo-AVA, outperforms existing approaches across all metrics. We observed

Model	FGD↓	FID↓	APD↑
Gesticulator	18.41	145.2	0.62
Speech2Gesture	15.71	132.6	0.71
Baseline LBLM (Ours)	**10.20**	**98.7**	**0.84**

Table 4: Comparison of gesture generation models based on FGD, FID, and APD metrics.

Model	Naturalness	Appropriateness
Speech2Gesture	3.2 ± 0.4	3.4 ± 0.3
Baseline LBLM (Ours)	$\mathbf{4.1 \pm 0.3}$	$\mathbf{4.3 \pm 0.2}$

Table 5: Human evaluation results for naturalness and appropriateness of generated gestures.

a 35% reduction in FGD and a 25.6% improvement in FID compared to the best previous model (Speech2Gesture). The APD of 0.84 indicates an 18.3% improvement in gesture diversity.

To assess perceptual quality, we conducted a human evaluation study with 50 participants rating the naturalness and appropriateness of gestures on a scale of 1 to 5. Results are shown in Table 5.

To analyze gesture coherence, we introduced a Temporal Coherence Score (TCS):

$$\text{TCS} = \frac{1}{T-1} \sum_{t=1}^{T-1} \cos(\mathbf{g}_t, \mathbf{g}_{t+1}) \tag{3}$$

where T is the total frames and $\mathbf{g}_t$ is the gesture vector at time t. Our model achieved a TCS of 0.89, compared to 0.81 for Speech2Gesture (Wang et al., 2023) and 0.76 for Gesticulator (Kucherenko et al., 2020), indicating smoother and more natural-looking animations.

These results underscore the importance of large-scale, diverse datasets like Allo-AVA in training models for complex tasks such as gesture generation. The comprehensive coverage of different speakers, contexts, and gesture types enables models to learn more nuanced relationships between speech content and corresponding gestures, paving the way for more natural and engaging virtual human interactions (Zhou et al., 2020; Tang et al., 2023).

5 Conclusion and Future Work

We presented Allo-AVA, a large-scale multimodal dataset for avatar animation, which significantly advances the field by providing researchers with comprehensive, temporally-aligned data for developing and evaluating models. Our baseline experiments demonstrate the dataset's utility in improving gesture generation across multiple metrics.

Future work on Allo-AVA will focus on:

- Expanding linguistic and cultural diversity to enable cross-cultural studies and globally adaptive systems.
- Enhancing annotations with fine-grained labels for gestures, emotions, and semantic meanings (Kucherenko et al., 2019).
- Incorporating multiview recordings and interaction scenarios to support 3D reconstruction and study of interactive behaviors.
- Improving multimodal synchronization for capturing subtle expressions and movements (Ferstl and McDonnell, 2020).
- Developing domain-specific subsets to facilitate targeted research in various contexts.

These enhancements will further our understanding of human communication and nonverbal behavior, supporting the development of more natural and engaging embodied conversational AI. We invite researchers from diverse fields to collaborate and leverage Allo-AVA in pushing the boundaries of avatar animation and human-computer interaction.

6 Limitations

While Allo-AVA and our baseline LBLM model demonstrate significant improvements in gesture generation, several limitations must be acknowledged. The dataset's primary focus on English-language content from Western cultures may restrict the model's ability to generate culturally appropriate gestures for non-Western contexts, potentially impacting its global applicability. The model may also struggle with highly complex or abstract gestures requiring deep contextual understanding. The significant computational resources required for training models on Allo-AVA could limit accessibility for researchers with constrained computing capabilities. Additionally, the current implementation may not be optimized for real-time performance, potentially limiting its immediate applicability in interactive systems requiring low-latency gesture generation. Lastly, while our model shows improved short-term gesture coherence, maintaining consistency over extended periods remains challenging and requires further investigation.

7 Ethical Considerations

The development and use of large-scale datasets and AI models for human behavior synthesis raise important ethical considerations. Privacy and data protection are paramount; although we have implemented strict anonymization and data handling protocols, the risk of potential re-identification in large-scale datasets cannot be entirely eliminated. There is also a potential for the model to learn and reproduce gestural stereotypes, necessitating ongoing analysis and implementation of debiasing techniques. The use of publicly available data for AI model training raises questions about informed consent, requiring clear communication about the potential uses of such content in AI research. As with any technology capable of generating human-like behavior, there is a potential for misuse in creating deceptive content. We strongly advocate for responsible use and the development of safeguards against such misuse, including watermarking techniques and detection tools for AI-generated content. Transparency and accountability are crucial; we are committed to maintaining openness about our models' capabilities and limitations, and we support efforts to establish industry-wide standards for ethical use of avatar animation technologies.

References

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. *2019 International Conference on 3D Vision (3DV)*, pages 719–728.

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum*, 39(2):487–496.

Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1998–2007.

Gary Bradski. 2000. The opencv library. *Dr. Dobb's Journal of Software Tools*, 25(11):120–126.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*. John Wiley & Sons.

Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 93–98.

Ylva Ferstl and Rachel McDonnell. 2020. Expressivity in generated co-speech gestures. *Computer Animation and Virtual Worlds*, 31(6):e1986.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. In *To appear in Proceedings of the 7th International Conference on Natural Language Processing*.

Ian T. Jolliffe and Jorge Cadima. 2016. *Principal Component Analysis*. Springer.

Patrik Jonell, Taras Kucherenko, Erik Ekstedt, and Jonas Beskow. 2020. Let's face it: Probabilistic multimodal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Taras Kucherenko, Jonas Beskow, and Hedvig Kjellström. 2018. Data-driven speech animation using convolutional neural networks. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 33–36.

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104.

Taras Kucherenko, Patrik Jonell, and Gustav Eje Henter. 2022. Multimodal co-speech gesture generation with unsupervised text representation learning. *IEEE Transactions on Affective Computing*.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. *Proceedings of the International Conference on Multimodal Interaction*, pages 242–250.

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021a. A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. *26th International Conference on Intelligent User Interfaces*, pages 11–21.

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021b. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 32–39.

Zhenguang Liu, Shuang Qian, Kaixuan Huang, Zhigang Xu, Dingding Cao, and Xiaohai Wen. 2022. Unifiedgesture: A unified gesture synthesis model for multiple skeletons. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2776–2784.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Norbert Marwan, Maria Carmen Romano, Marco Thiel, and Jürgen Kurths. 2007. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2007. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)*, 26(2):5–es.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Michael Rebol and Holger Knoche. 2021. Casual conversations: A dataset for measuring fairness in ai. *arXiv preprint arXiv:2104.02821*.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

David Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1177–1178. ACM.

Tenglong Tang, Ye Liu, Linchao Liu, Shuwen Xie, and Ping Zhang. 2023. Gesture generation with diffusion models. *arXiv preprint arXiv:2302.00332*.

Lloyd N. Trefethen and David Bau III. 1997. *Numerical Linear Algebra*. SIAM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Cheng Wang, Sicheng Zhao, Fei Gao, and Chenliang Xu. 2023. Speech2gesture: Learning individual styles of conversational gesture. *arXiv preprint arXiv:2308.13413*.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6):1–16.

Yuliang Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. 2020. Towards 4d human body capture in the wild. *arXiv preprint arXiv:2007.11431*.

A Dataset Collection and Processing Details

A.1 Keyword Selection

The Allo-AVA dataset was curated using a comprehensive list of 140 keywords and phrases related to allocentric gestures and nonverbal communication. These keywords were carefully selected to capture a wide range of communication scenarios and contexts. The full list of keywords used in our data collection process is as follows:

- TED talk gestures
- TED talk body language
- TED talk hand movements
- TED talk nonverbal communication
- podcast gestures
- podcast body language
- podcast hand movements
- podcast nonverbal communication
- comedy show gestures
- comedy show body language
- comedy show hand movements
- comedy show nonverbal communication
- interview gestures
- interview body language
- interview hand movements
- interview nonverbal communication
- news anchor gestures
- news anchor body language
- news anchor hand movements
- news anchor nonverbal communication
- public speaking gestures
- public speaking body language
- public speaking hand movements
- public speaking nonverbal communication
- presentation gestures
- presentation body language
- presentation hand movements
- presentation nonverbal communication
- lecture gestures
- lecture body language
- lecture hand movements
- lecture nonverbal communication
- vlog gestures
- vlog body language
- vlog hand movements
- vlog nonverbal communication
- interview show gestures
- interview show body language
- interview show hand movements
- interview show nonverbal communication
- talk show gestures
- talk show body language
- talk show hand movements
- talk show nonverbal communication
- debate gestures
- debate body language
- debate hand movements
- debate nonverbal communication
- panel discussion gestures
- panel discussion body language
- panel discussion hand movements
- panel discussion nonverbal communication
- conference gestures
- conference body language
- conference hand movements
- conference nonverbal communication
- seminar gestures
- seminar body language
- seminar hand movements
- seminar nonverbal communication
- webinar gestures
- webinar body language
- webinar hand movements
- webinar nonverbal communication

- workshop gestures
- workshop body language
- workshop hand movements
- workshop nonverbal communication
- training session gestures
- training session body language
- training session hand movements
- training session nonverbal communication
- motivational speech gestures
- motivational speech body language
- motivational speech hand movements
- motivational speech nonverbal communication
- political speech gestures
- political speech body language
- political speech hand movements
- political speech nonverbal communication
- ceremony gestures
- ceremony body language
- ceremony hand movements
- ceremony nonverbal communication
- award show gestures
- award show body language
- award show hand movements
- award show nonverbal communication
- reality show gestures
- reality show body language
- reality show hand movements
- reality show nonverbal communication
- game show gestures
- game show body language
- game show hand movements
- game show nonverbal communication
- sports commentary gestures
- sports commentary body language
- sports commentary hand movements
- sports commentary nonverbal communication
- fitness video gestures
- fitness video body language
- fitness video hand movements
- fitness video nonverbal communication
- dance tutorial gestures
- dance tutorial body language
- dance tutorial hand movements
- dance tutorial nonverbal communication
- music video gestures
- music video body language
- music video hand movements
- music video nonverbal communication
- cooking show gestures
- cooking show body language
- cooking show hand movements
- cooking show nonverbal communication
- DIY tutorial gestures
- DIY tutorial body language
- DIY tutorial hand movements
- DIY tutorial nonverbal communication
- makeup tutorial gestures
- makeup tutorial body language
- makeup tutorial hand movements
- makeup tutorial nonverbal communication
- fashion show gestures
- fashion show body language
- fashion show hand movements
- fashion show nonverbal communication
- product review gestures
- product review body language
- product review hand movements
- product review nonverbal communication
- unboxing video gestures
- unboxing video body language
- unboxing video hand movements
- unboxing video nonverbal communication
- travel vlog gestures
- travel vlog body language
- travel vlog hand movements
- travel vlog nonverbal communication

- reaction video gestures
- reaction video body language
- reaction video hand movements
- reaction video nonverbal communication
- prank video gestures
- prank video body language
- prank video hand movements
- prank video nonverbal communication
- challenge video gestures
- challenge video body language
- challenge video hand movements
- challenge video nonverbal communication
- allocentric gestures
- allocentric body language
- allocentric hand movements
- allocentric nonverbal communication
- allocentric perspective gestures
- allocentric perspective body language
- allocentric perspective hand movements
- allocentric perspective nonverbal communication

A.2 Video Filtering Criteria

To ensure the quality and relevance of the collected videos, we applied the following filtering criteria:

- Minimum duration: 5 minutes
- Minimum view count: 10,000 views
- Language: English
- Categories: Education, Entertainment, Science & Technology, News & Politics
- Video quality: At least 720p resolution
- Content focus: Clear visibility of speaker's full body

A.3 Extra Data Processing Pipeline Information

Our data processing pipeline consists of several key steps designed to extract and align multimodal data from the collected videos. Figure 9 provides a detailed overview of this pipeline.

The pipeline includes the following major components:

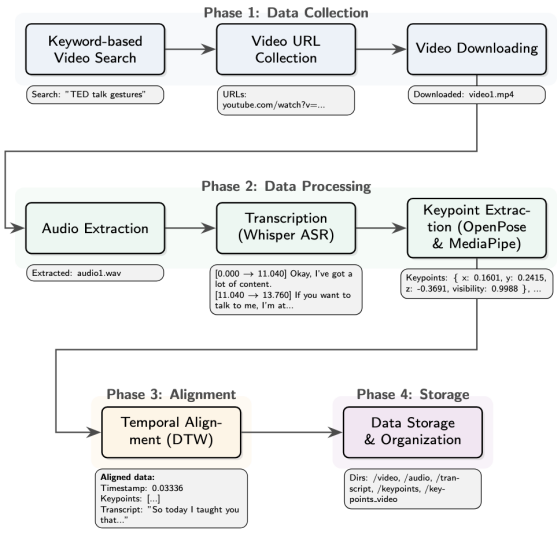

Figure 9: Detailed data processing pipeline for AlloAVA

A.3.1 Video and Audio Extraction

We utilized the `yt-dlp` library to download videos in the highest available quality, typically 1080p at 30 fps. Audio tracks were extracted using the `moviepy` library and saved as 16-bit PCM WAV files with a 48 kHz sampling rate.

A.3.2 Transcription

For generating accurate transcriptions with word-level timestamps, we employed OpenAI's Whisper ASR model. We used the "base" model for efficiency, achieving a Word Error Rate (WER) of approximately 6% on our validation set.

A.3.3 Keypoint Extraction

Our keypoint extraction process combines two state-of-the-art pose estimation models: OpenPose and MediaPipe. This dual approach allows us to capture a comprehensive set of body keypoints with high accuracy and detail.

A.4 Keypoint Extraction and Fusion

The keypoint extraction and fusion process is a critical component of our pipeline. We leverage the strengths of both OpenPose (Cao et al., 2018) and MediaPipe (Lugaresi et al., 2019) to obtain a rich set of keypoints. Algorithm 3 details the fusion process.

This algorithm ensures that we leverage the strengths of both models, resulting in a more comprehensive and accurate set of keypoints.

Algorithm 1 Keypoint Fusion Algorithm

Require: OpenPose keypoints K_{OP}, MediaPipe keypoints K_{MP}
Ensure: Fused keypoints K_{fused}

$K_{fused} \leftarrow \{\}$
for each joint j in skeleton **do**
  if j in K_{OP} and j in K_{MP} **then**
    $w_{OP} \leftarrow \text{confidence}(K_{OP}[j])$
    $w_{MP} \leftarrow \text{confidence}(K_{MP}[j])$
    $K_{fused}[j] \leftarrow \frac{w_{OP} \cdot K_{OP}[j] + w_{MP} \cdot K_{MP}[j]}{w_{OP} + w_{MP}}$
  else if j in K_{OP} **then**
    $K_{fused}[j] \leftarrow K_{OP}[j]$
  else if j in K_{MP} **then**
    $K_{fused}[j] \leftarrow K_{MP}[j]$
  end if
end forreturn K_{fused}

A.5 Temporal Alignment

To ensure precise synchronization between keypoints, transcriptions, and audio, we developed a custom temporal alignment algorithm based on Dynamic Time Warping (DTW) (Sakoe and Chiba, 1978). Algorithm 2 outlines this process.

Algorithm 2 Temporal Alignment Algorithm

Require: Keypoint sequence K, Transcript T, Audio features A
Ensure: Aligned data $D_{aligned}$

$D_{aligned} \leftarrow \{\}$, $W_K \leftarrow \text{ExtractMotionFeatures}(K)$, $W_A \leftarrow \text{ExtractAudioFeatures}(A)$
$DTW_{matrix} \leftarrow \text{ComputeDTWMatrix}(W_K, W_A)$
$path \leftarrow \text{BacktrackOptimalPath}(DTW_{matrix})$
for (i, j) in $path$ **do**
  $t_{aligned} \leftarrow \text{InterpolateTimestamp}(i, j)$
  $k_{aligned}, a_{aligned} \leftarrow K[i], A[j]$
  $t_{aligned} \leftarrow \text{FindClosestTranscript}(T, t_{aligned})$
  $D_{aligned}.\text{append}(t_{aligned}, k_{aligned}, a_{aligned})$
end forreturn $D_{aligned}$

This algorithm ensures that keypoints, audio features, and transcripts are precisely aligned, accounting for potential variations in timing across modalities.

B Gesture Classification

B.1 Gesture Taxonomy

To categorize the diverse range of gestures in the Allo-AVA dataset, we adopted the gesture taxonomy (McNeill, 1992) and extended it to include additional categories relevant to our dataset. Table 6 presents an overview of this classification system.

Category	Subcategories
Deictic	Pointing, Indicating, Referential
Iconic	Object-mimicking, Action-mimicking, Spatial
Metaphoric	Abstract concept, Emotional state, Magnitude
Beat	Rhythmic, Emphasis, Pacing
Emblematic	Cultural-specific, Universal
Adaptors	Self-touching, Object-manipulation

Table 6: Allo-AVA Gesture Taxonomy

C Speech-Gesture Correlation Analysis

C.1 Cross-modal Correlation Metrics

To analyze the relationship between speech and gesture, we employed several established cross-modal correlation metrics. Table 7 provides an overview of these metrics and their interpretations.

These metrics provide a comprehensive view of the complex relationships between speech and gesture in our dataset (Cover and Thomas, 2006).

D Advanced Keypoint Analysis

D.1 Principal Component Analysis of Keypoints

To better understand the underlying structure of our keypoint data, we performed Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016) on the extracted keypoints. Figure 10 shows the distribution of keypoints projected onto the first two principal components.

This visualization reveals several key insights:

- The wide spread of points indicates a high degree of variability in the captured poses, suggesting that our dataset encompasses a diverse range of gestures and body positions.
- The presence of distinct clusters (e.g., the dense regions in the center and the sparser regions at the edges) may correspond to common pose types or gesture categories.

Metric	Description
Mutual Information (MI)	Measures the mutual dependence between speech and gesture features.
Canonical Correlation Analysis (CCA)	Identifies linear combinations of speech and gesture features that have maximum correlation.
Dynamic Time Warping (DTW)	Computes the optimal alignment between speech and gesture sequences.
Cross-recurrence Quantification Analysis (CRQA) (Marwan et al., 2007)	Quantifies the dynamics of recurrent structures in speech-gesture interactions.

Table 7: Cross-modal correlation metrics

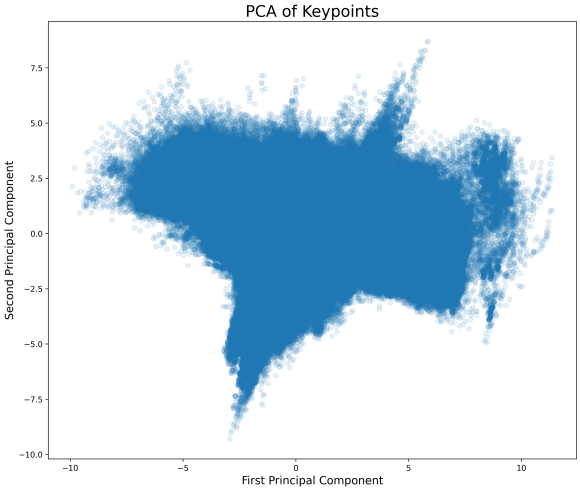

Figure 10: PCA of keypoints extracted from the Allo-AVA dataset

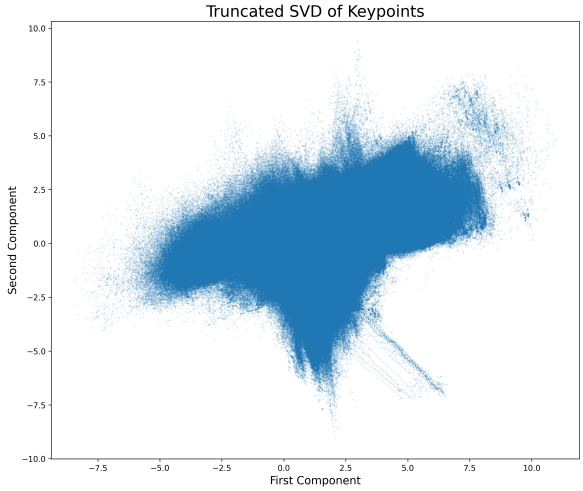

Figure 11: Truncated SVD of keypoints from the Allo-AVA dataset

- The continuous nature of the distribution suggests that our dataset captures smooth transitions between different poses, which is crucial for generating natural-looking animations.

D.2 Truncated Singular Value Decomposition of Keypoints

To further analyze the dimensionality of our keypoint data, we performed a truncated Singular Value Decomposition (SVD) (Trefethen and Bau III, 1997). Figure 11 displays the results of this analysis.

The truncated SVD visualization provides additional insights:

- The overall shape of the distribution is similar to the PCA plot, confirming the consistency of our dimensional reduction techniques.
- The more pronounced "tails" in the SVD plot may indicate the presence of rarer, more extreme poses or gestures in our dataset.
- The density variations across the plot suggest that certain pose configurations are more common than others, which aligns with expected patterns in natural human movement.

D.3 MiniBatch K-means Clustering of Keypoints

To further investigate underlying structure and patterns within our keypoint data, we employed MiniBatch K-means clustering (Sculley, 2010), a variant of the K-means algorithm optimized for large datasets. Figure 12 presents the results of this clustering analysis.

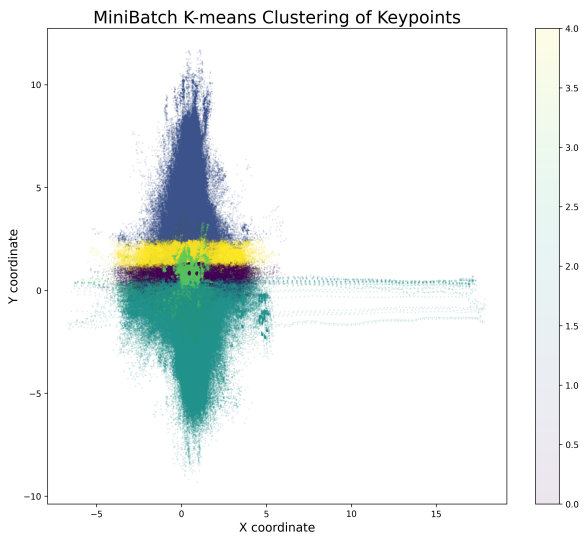

Figure 12: MiniBatch K-means Clustering of keypoints extracted from the Allo-AVA dataset

This visualization offers several significant insights into the spatial distribution and clustering tendencies of our keypoint data:

The clustering results reveal distinct regions of keypoint concentration, each represented by different colors. The central yellow cluster corresponds to neutral or resting positions, serving as a common reference point across various gestures and poses. The expansive blue region extending upwards suggests a continuum of upper body movements, potentially capturing a wide range of arm and hand gestures crucial for expressive communication.

Of particular interest is the dense, elongated cyan cluster extending downwards, which may represent lower body positions and movements. The clear separation between upper and lower body clusters underscores the importance of capturing full-body dynamics in avatar animation, as it highlights the potential for independent yet coordinated movements between these regions.

The presence of smaller, isolated clusters (e.g., the green points scattered throughout) indicates the capture of less frequent but potentially significant pose configurations. The outliers represent the full spectrum of human gestures, including more extreme or specialized movements that may be critical for certain communicative contexts.

The asymmetry observed in the distribution, particularly the extension towards the right side of the plot, suggests a bias towards right-handed gestures in our dataset. This observation aligns with the predominance of right-handedness in the general population and highlights the need for careful consideration of handedness in gesture generation models.

The gradual color transitions between clusters indicate smooth interpolations between different pose categories, a factor essential for generating fluid and natural-looking animations.

D.4 Keypoint Extraction and Fusion

Our keypoint extraction process leverages two state-of-the-art pose estimation models: OpenPose and MediaPipe. Figure 13 shows the keypoint structure used by OpenPose, while Figure 14 illustrates the MediaPipe keypoint model.

D.4.1 OpenPose

OpenPose provides a robust framework for multi-person keypoint detection, employing a bottom-up approach that first detects body parts and then associates them with individuals. This method allows

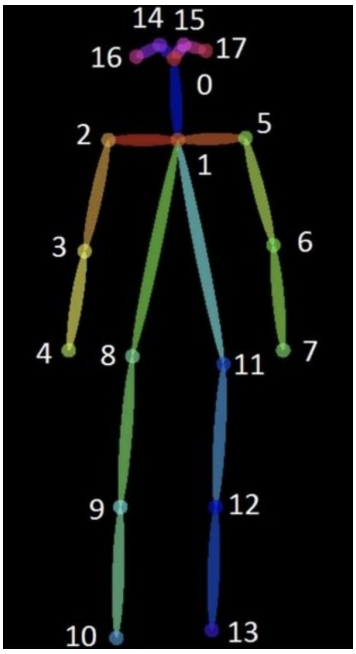

Figure 13: OpenPose keypoint structure

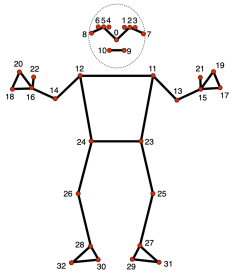

Figure 14: MediaPipe keypoint structure

for efficient handling of multiple subjects within a single frame, a crucial feature for analyzing complex scenes. OpenPose detects 18 key body points, including major joints and select facial landmarks.

The 18 keypoints detected by OpenPose are as follows:

```
 0 - Nose
 1 - Neck
 2 - Right Shoulder
 3 - Right Elbow
 4 - Right Wrist
 5 - Left Shoulder
 6 - Left Elbow
 7 - Left Wrist
 8 - Right Hip
 9 - Right Knee
10 - Right Ankle
11 - Left Hip
12 - Left Knee
13 - Left Ankle
```

```
14 - Right Eye
15 - Left Eye
16 - Right Ear
17 - Left Ear
```

This configuration provides a comprehensive skeletal representation that forms the foundation for our gesture analysis.

D.4.2 MediaPipe

MediaPipe complements OpenPose by offering a more detailed keypoint model, particularly excelling in hand and facial feature detection. We specifically utilized MediaPipe's Pose Landmarker model, which detects 33 body keypoints. This expanded set of points allows for finer tracking of subtle movements and expressions, crucial for capturing the nuances of human communication.

The MediaPipe Pose Landmarker model identifies the following 33 keypoints:

```
 0 - Nose
 1 - Left Eye (Inner)
 2 - Left Eye
 3 - Left Eye (Outer)
 4 - Right Eye (Inner)
 5 - Right Eye
 6 - Right Eye (Outer)
 7 - Left Ear
 8 - Right Ear
 9 - Mouth (Left)
10 - Mouth (Right)
11 - Left Shoulder
12 - Right Shoulder
13 - Left Elbow
14 - Right Elbow
15 - Left Wrist
16 - Right Wrist
17 - Left Pinky
18 - Right Pinky
19 - Left Index
20 - Right Index
21 - Left Thumb
22 - Right Thumb
23 - Left Hip
24 - Right Hip
25 - Left Knee
26 - Right Knee
27 - Left Ankle
28 - Right Ankle
29 - Left Heel
30 - Right Heel
31 - Left Foot Index
32 - Right Foot Index
```

This comprehensive set of keypoints enables precise tracking of facial expressions, hand gestures, and full-body movements.

D.4.3 Fusion Algorithm

To leverage the strengths of both models, we developed a novel fusion algorithm that combines the outputs of OpenPose and MediaPipe. Algorithm 3 outlines this process.

Algorithm 3 Enhanced Keypoint Fusion Algorithm

Require: OpenPose keypoints K_{OP}, MediaPipe keypoints K_{MP}
Ensure: Fused keypoints K_{fused}

$K_{fused} \leftarrow \{\}$
for each joint j in combined skeleton **do**
  if j in K_{OP} and j in K_{MP} **then**
    $w_{OP} \leftarrow \text{ConfidenceScore}(K_{OP}[j])$
    $w_{MP} \leftarrow \text{ConfidenceScore}(K_{MP}[j])$
    $w_{total} \leftarrow w_{OP} + w_{MP}$
    $K_{fused}[j].pos \leftarrow (w_{OP} \cdot K_{OP}[j].pos + w_{MP} \cdot K_{MP}[j].pos)/w_{total}$
    $K_{fused}[j].conf \leftarrow \max(K_{OP}[j].conf, K_{MP}[j].conf)$
  else if j in K_{OP} **then**
    $K_{fused}[j] \leftarrow K_{OP}[j]$
  else if j in K_{MP} **then**
    $K_{fused}[j] \leftarrow K_{MP}[j]$
  end if
  if j is a hand or face keypoint **then**
    $K_{fused}[j] \leftarrow \text{RefineWithMediaPipe}(K_{fused}[j])$
  end if
end for
$K_{fused} \leftarrow \text{TemporalSmoothing}(K_{fused})$
$K_{fused} \leftarrow \text{AnatomicalConstraints}(K_{fused})$
return K_{fused}

This enhanced algorithm includes several key steps:

- Weighted averaging of keypoints detected by both models, based on their confidence scores
- Preference for MediaPipe's detection for hand and face keypoints due to its higher granularity in these areas
- Temporal smoothing to reduce jitter and ensure consistency across frames
- Application of anatomical constraints to prevent physically impossible poses

D.5 Additional Analyses

To further elucidate the intricate relationships between speech, gesture, and context captured in the Allo-AVA dataset, we conducted several in-depth analyses that provide valuable insights for the development of sophisticated avatar animation models.

D.5.1 Temporal Gesture Evolution

Figure 15 presents a detailed analysis of the temporal evolution of gesture intensity in relation to speech features. This visualization reveals a complex interplay between gestural movements and speech characteristics. The graph demonstrates a clear relationship between gesture intensity and speech features such as volume and pitch. Notably, peaks in gesture intensity often coincide with or slightly lag behind peaks in speech volume and pitch, suggesting that gestures frequently serve to emphasize or complement spoken content.

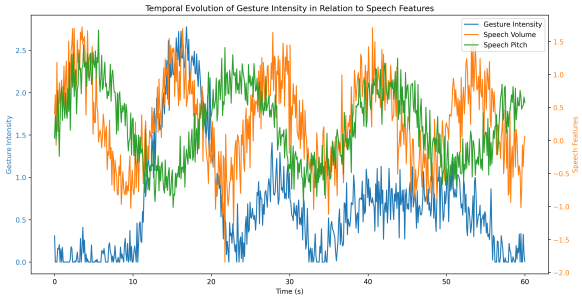

Figure 15: Temporal evolution of gesture intensity in relation to speech features

Interestingly, the data also reveals instances of gesture anticipation, where increases in gesture intensity precede changes in speech volume or pitch. This phenomenon indicates that gestures can function as preparatory movements for upcoming speech events, highlighting the predictive nature of nonverbal communication. The relationship between speech and gesture exhibits significant variability throughout the sequence, with varying degrees of temporal alignment. This variability underscores the dynamic and context-dependent nature of speech-gesture coordination, a crucial consideration for developing naturalistic avatar animation systems.

The observed range of gesture intensity, spanning from near-zero to peaks exceeding 2.5, captures a wide spectrum of nonverbal behaviors, from subtle movements to pronounced gestures. This diversity is essential for training models capable of generating appropriately nuanced gestures across various communicative contexts.

D.5.2 Gesture Trajectory Analysis

The 3D trajectories of hand movements during different types of gestures, as visualized in Figure 16, offer profound insights into the spatial characteristics of various gesture categories. Each gesture type—Deictic, Iconic, Metaphoric, and Beat—exhibits distinct trajectory patterns that reflect its communicative function.

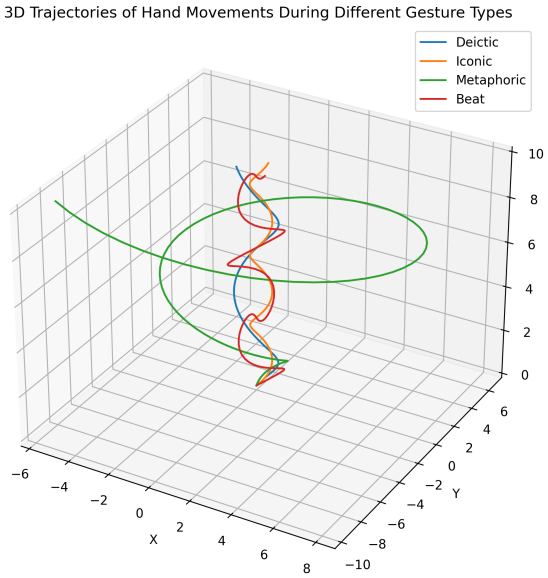

Figure 16: 3D trajectories of hand movements during different gesture types

Deictic gestures, primarily used for pointing, display more linear and directed movements, while beat gestures manifest as repetitive, rhythmic patterns that punctuate speech. Iconic gestures, which often depict concrete objects or actions, present more intricate paths, reflecting the complexity of the concepts they represent. Metaphoric gestures, used to convey abstract ideas, tend to occupy a larger gestural space with sweeping movements, illustrating the expansive nature of the concepts they embody.

The spatial distribution of these trajectories demonstrates how different gesture types utilize the speaker's gesture space, a crucial factor in developing realistic and varied gesture generation models. Moreover, the clear three-dimensional nature of these gestures underscores the importance of capturing depth information in our dataset, as it is essential for creating truly lifelike avatar animations.

D.5.3 Pose Similarity Heatmap

Figure 17 presents a heatmap visualizing the similarity between different poses across the dataset, revealing the underlying structure of the gestural space captured in Allo-AVA. The diagonal blocks of high similarity indicate clusters of related poses, likely corresponding to common gesture types or resting positions that occur frequently in natural communication.

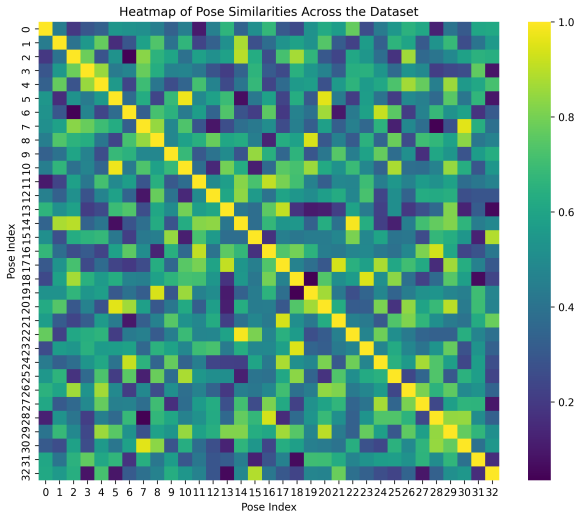

Figure 17: Heatmap of pose similarities across the dataset

Of particular interest are the off-diagonal patterns, which reveal relationships between different pose clusters. These patterns potentially represent common gesture transitions, providing crucial information for generating smooth and natural sequences of gestures. The presence of both high and low similarity regions in the heatmap demonstrates the rich diversity of poses captured in our dataset, a key factor in training models capable of generating a wide range of gestures.

Furthermore, the heatmap suggests a hierarchical structure in the pose space, with some poses being more closely related than others. This hierarchical nature could inform the development of sophisticated, hierarchical models for gesture generation, capable of capturing both broad gestural categories and fine-grained variations within those categories.

E Ethical Considerations and Data Privacy

E.1 Consent and Usage Rights

All videos in the Allo-AVA dataset were collected from publicly available sources on YouTube. To ensure ethical use of the data, we implemented the following measures:

- Adherence to YouTube's Terms of Service and API usage guidelines.
- Exclusion of content marked as "unlisted" or "private" by content creators.
- Implementation of a takedown procedure for content creators who wish to have their data removed from the dataset.
- Creation of usage guidelines that prohibit the use of the dataset for individual identification or tracking purposes.

E.2 Anonymization Techniques

To protect the privacy of individuals in the dataset, we applied several anonymization techniques:

- Face blurring in keypoint visualization videos using OpenCV's (Bradski, 2000) face detection and blurring algorithms.
- Voice pitch modification in audio files using the librosa library (McFee et al., 2015), with random pitch shifts within a range that preserves intelligibility but alters voice characteristics.
- Replacement of proper nouns in transcripts with generic placeholders using Named Entity Recognition (NER) techniques from the spaCy library (Honnibal and Montani, 2017).

E.3 Bias Mitigation

To address potential biases in the dataset, we implemented the following strategies:

- Diversity tracking to ensure representation across demographics, using metadata from video descriptions and automatic classification techniques.
- Regular audits of the dataset composition to identify and address underrepresented groups or contexts.
- Documentation of known biases and limitations in the dataset documentation to inform users and encourage responsible use.

F Computational Resources and Optimization

F.1 Hardware Specifications

The Allo-AVA dataset was processed using the following hardware setup:

- 8 NVIDIA A40 GPUs (48GB VRAM each)
- 2 AMD EPYC 7742 64-Core Processors
- 1 TB RAM
- 200 TB NVMe SSD storage

F.2 Optimization Techniques

To efficiently process the large volume of data, we employed several optimization techniques:

- Parallel processing using Python's multiprocessing library to utilize all available CPU cores.
- GPU acceleration for keypoint extraction and video processing tasks using CUDA-enabled libraries.
- Efficient data loading and preprocessing using PyTorch's DataLoader with custom collate functions.
- Caching of intermediate results to avoid redundant computations.

These optimizations allowed us to process the entire dataset in approximately 3,000 GPU hours.