Large Body Language Models

Anonymous ACL submission

Abstract

As virtual agents become increasingly prevalent in human-computer interaction, generating realistic and contextually appropriate gestures in real-time remains a significant challenge. While neural rendering techniques have made substantial progress with static scripts, their applicability to human-computer interactions remains limited. To address this, we introduce Large Body Language Models (LBLMs) and present LBLM-AVA, a novel LBLM architecture that combines a Transformer-XL large language model with a parallelized diffusion model to generate human-like gestures from multimodal inputs (text, audio, and video). LBLM-AVA incorporates several key components enhancing its gesture generation capabilities, such as multimodal-to-pose embeddings, enhanced sequence-to-sequence mapping with redefined attention mechanisms, a temporal smoothing module for gesture sequence coherence, and an attention-based refinement module for enhanced realism. The model is trained on our large-scale proprietary open-source dataset Allo-AVA. LBLM-AVA achieves state-of-the-art performance in generating lifelike and contextually appropriate gestures with a 30% reduction in Fréchet Gesture Distance (FGD), and a 25% improvement in Fréchet Inception Distance compared to existing approaches.

1 Introduction

Generating realistic and contextually appropriate gestures in real-time is a critical challenge to produce engaging virtual agents, and while neural rendering techniques have made strides in our ability to generate them as such, their realism is often hindered from their static movements and limited ability to have multimodal conversational interactions (Li et al., 2023; Windle et al., 2023). To address this problem, we introduce Large Body Language Models (LBLMs), a new type of architecture specifically designed for generating gestures in the context of real-time, multimodal communication.

In the context of conversational AI and gesture generation, we define the LBLM inference problem as generating an optimal multimodal (body movement and facial expression) 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.

Prior approaches to gesture generation have relied on rule-based systems (Nyatsanga et al., 2023), motion capture databases (Rueux et al., 2014), or learning-based methods (Yoon et al., 2020; Zhou et al., 2022). While these techniques have made notable improvements, they are limited in their ability to capture the complex relationships between speech, facial expressions, and body language in dynamic conversational settings. Rule-based systems can appear mechanical, motion capture databases are constrained by the specific gestures they contain, and conventional learning-based methods struggle to generate gestures that are coherent across long time horizons and adapt to evolving conversational contexts.

Recent works such as (Korzun et al., 2022; Neff et al., 2007; Bhattacharya et al., 2021) represent early examples of LBLMs, leveraging the success of transformer-based language models (Vaswani et al., 2017) to capture the intricacies of human communication. These models demonstrate the potential for generating human-like gestures by learning the complex relationships between language, audio, and visual cues. However, they have limitations in terms of their ability to handle real-time, multimodal inputs and generate diverse, contextually appropriate gestures.

We present LBLM-AVA, a novel LBLM architecture that combines a Transformer-XL (Dai et al., 2019) language model with a parallelized diffusion model (Shih et al., 2023) to generate human-like gestures from multimodal inputs, including text, audio, and video. LBLM-AVA incorporates several key components to enhance its gesture generation capabilities, such as multimodal-to-pose embeddings, sequence-to-sequence mapping with redefined attention mechanisms, a temporal smoothing module for gesture sequence coherence, and an attention-based refinement module for enhanced realism. The model is trained on our proprietary open-source dataset Allo-AVA, a large-scale, multimodal corpus derived from diverse sources like TEDx Talks and podcasts.

Experimental evaluations demonstrate that LBLM-AVA achieves state-of-the-art performance in generating lifelike and contextually appropriate gestures. Gestures generated by LLBM-AVA significantly boost the perceived naturalness and engagement of virtual agents. Our results highlight the potential of LLBMs to advance the field of gesture generation and create more compelling human-agent interactions.

2 Approach

2.1 LBLM-AVA Architecture

Our proposed LBLM-AVA model builds upon the Transformer-XL architecture and incorporates several novel components to generate realistic and contextually appropriate gestures from multimodal inputs. The overall architecture is illustrated in Figure 1.

2.1.1 Multimodal Input Representation

The input to the model consists of text, audio, and video features, denoted as $\mathbf{T} \in \mathbb{R}^{L_T \times d_T}$, $\mathbf{A} \in \mathbb{R}^{L_A \times d_A}$, and $\mathbf{V} \in \mathbb{R}^{L_V \times d_V}$, respectively, where L_* and d_* represent the sequence length and feature dimensionality for each modality. These features are projected to a common dimension d using learned linear transformations:

$$\mathbf{T}' = \mathbf{W}_T \mathbf{T}, \quad \mathbf{A}' = \mathbf{W}_A \mathbf{A}, \quad \mathbf{V}' = \mathbf{W}_V \mathbf{V} \tag{2}$$

where $\mathbf{W}_T \in \mathbb{R}^{d \times d_T}$, $\mathbf{W}_A \in \mathbb{R}^{d \times d_A}$, and $\mathbf{W}_V \in \mathbb{R}^{d \times d_V}$ are learnable projection matrices.

2.1.2 Transformer-XL Encoder

The projected multimodal features are concatenated along the sequence dimension and passed through a Transformer-XL encoder. The Transformer-XL architecture introduces the notion of recurrence by reusing hidden states from previous segments, enabling the model to capture longer-term dependencies.

Given the input sequence $\mathbf{X} = [\mathbf{T}'; \mathbf{A}'; \mathbf{V}'] \in \mathbb{R}^{L \times d}$, the hidden state for the n-th segment $\mathbf{s}_n \in \mathbb{R}^{L_s \times d}$ is computed as:

$$\mathbf{h}_n = \text{TransformerXL}([\mathbf{s}_{n-1}; \mathbf{x}_n]) \tag{3}$$

where $\mathbf{s}_{n-1}$ is the cached hidden state from the previous segment, $\mathbf{x}_n \in \mathbb{R}^{L_s \times d}$ is the input sequence for the current segment, and L_s is the segment length. The TransformerXL function applies multi-head attention with relative positional encodings followed by position-wise feed-forward layers.

2.1.3 Multimodal-to-Pose Embedding

To facilitate the mapping from the different features (language, audio, video) to gesture poses, we introduce a multimodal-to-pose embedding module. This module learns a transformation from the encoded modality features to a latent pose space:

$$\mathbf{E} = \mathbf{W}_E \mathbf{H}_T \tag{4}$$

where $\mathbf{H}_T \in \mathbb{R}^{L_T \times d}$ is the encoded sequence dependent on input from the Transformer-XL encoder, $\mathbf{W}_E \in \mathbb{R}^{d_p \times d}$ is a learnable weight matrix, and $\mathbf{E} \in \mathbb{R}^{L_T \times d_p}$ is the embedded pose sequence.

2.1.4 Parallelized Diffusion Model

To generate realistic and diverse gesture sequences, we employ a parallelized diffusion model (Ho et al., 2020). Diffusion models learn to denoise a Gaussian noise signal into a target data distribution through a series of iterative refinement steps. In our approach, we parallelize the diffusion process to generate multiple gesture sequences simultaneously.

Given the language-to-pose embeddings $\mathbf{E}$, we sample a set of N initial noise sequences $\{\mathbf{z}_0^{(i)}\}_{i=1}^N$, where $\mathbf{z}_0^{(i)} \in \mathbb{R}^{L_T \times d_p}$. The diffusion process then iteratively refines these sequences over K steps:

$$\mathbf{z}_k^{(i)} = \mathbf{z}_{k-1}^{(i)} - \frac{\beta_k}{2}(\mathbf{z}_{k-1}^{(i)} - f_\theta(\mathbf{z}_{k-1}^{(i)}, \mathbf{E}, k)) \tag{5}$$

where β_k is a learnable noise schedule, and f_θ is a neural network parameterized by θ that predicts the noise to be removed at each step. The final gesture sequences are obtained as $\{\hat{\mathbf{P}}^{(i)} = \mathbf{z}_K^{(i)}\}_{i=1}^N$.

2.1.5 Attention-based Temporal Refinement

To improve the temporal coherence of the generated gestures, we introduce an attention-based refinement module. This module applies multi-head self-attention to the generated pose sequences, allowing the model to capture long-range dependencies and ensure smooth transitions between gestures.

Given the generated pose sequences $\{\hat{\mathbf{P}}^{(i)}\}_{i=1}^N$, the refined sequences $\{\tilde{\mathbf{P}}^{(i)}\}_{i=1}^N$ are computed as:

$$\tilde{\mathbf{P}}^{(i)} = \text{MultiHeadAttention}(\hat{\mathbf{P}}^{(i)}, \hat{\mathbf{P}}^{(i)}, \hat{\mathbf{P}}^{(i)}) \quad (6)$$

where the MultiHeadAttention function applies self-attention to the input sequence, using the same sequence for the query, key, and value matrices.

2.1.6 Adversarial Training

To further enhance the realism and diversity of the generated gestures, we employ adversarial training (Goodfellow et al., 2014). We introduce a discriminator network D that learns to distinguish between real and generated gesture sequences. The generator (i.e., the diffusion model) is then trained to maximize the discriminator's confusion:

$$\mathcal{L}_G = -\mathbb{E}_{\tilde{\mathbf{P}} \sim p_G}[\log D(\tilde{\mathbf{P}})] \quad (7)$$

where p_G is the distribution of generated gesture sequences. The discriminator is trained to minimize the adversarial loss:

$$\begin{aligned} \mathcal{L}_D &= -\mathbb{E}_{\mathbf{P} \sim p_{\text{data}}}[\log D(\mathbf{P})] \\ &\quad - \mathbb{E}_{\tilde{\mathbf{P}} \sim p_G}[\log(1 - D(\tilde{\mathbf{P}}))] \end{aligned} \quad (8)$$

where p_{data} is the distribution of real gesture sequences. The generator and discriminator are trained in an alternating fashion, promoting the generation of realistic and diverse gestures.

The integration of these novel components, including the Transformer-XL architecture, language-to-pose embeddings, parallelized diffusion model, attention-based refinement, and adversarial training, enables LLBM-AVA to generate highly expressive and contextually appropriate gestures from multimodal inputs. The model architecture is designed to capture the complex relationships between language, audio, and video features while ensuring the temporal coherence and realism of the generated gestures (Chen et al., 2023).

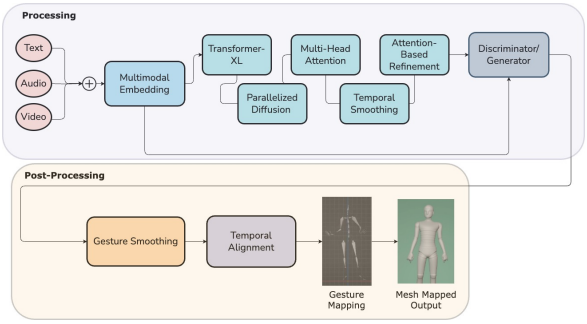

Figure 1: Architecture of the proposed LLBM-AVA model. The multimodal inputs are encoded using a Transformer-XL encoder, and a language-to-pose embedding module maps the encoded text to a latent pose space. A parallelized diffusion model generates multiple gesture sequences, which are then refined using an attention-based temporal refinement module. Adversarial training is employed to enhance the realism and diversity of the generated gestures. Post-processing occurs here to optimize gesture accuracy and human-likeness.

2.2 Gesture Parameterization and Representation

To efficiently represent and manipulate gestures, we introduce a compact parameterization scheme (Zhou et al., 2020). Each gesture is represented as a sequence of pose vectors $\mathbf{p}_t \in \mathbb{R}^D$, where D is the dimensionality of the pose space. The pose vectors encode the positions and orientations of key body joints, such as the hands, elbows, and shoulders, relative to a root joint (e.g., the pelvis).

We further decompose each pose vector into a set of sub-vectors corresponding to different body parts:

$$\mathbf{p}_t = [\mathbf{p}_t^{(1)}, \mathbf{p}_t^{(2)}, ..., \mathbf{p}_t^{(B)}] \quad (9)$$

where B is the number of body parts and $\mathbf{p}_t^{(b)} \in \mathbb{R}^{D_b}$ is the sub-vector for body part b at time step t. This hierarchical representation allows for more fine-grained control over the generated gestures and facilitates the modeling of inter-part dependencies.

To ensure smooth and realistic motion, we apply a series of kinematic constraints and postprocessing steps to the generated pose sequences. These include enforcing joint angle limits, maintaining bone lengths, and applying Gaussian smoothing to eliminate high-frequency jitter (Savitzky and Golay, 1964).

2.2.1 Pose Embedding

To feed the pose vectors into our Transformer-based model, we first embed them into a higher-dimensional feature space using a learnable embedding matrix $\mathbf{E} \in \mathbb{R}^{D \times d}$, where d is the embedding dimensionality:

$$\mathbf{e}_t = \mathbf{E} \cdot \mathbf{p}_t \tag{10}$$

The embedded pose vectors $\mathbf{e}_t \in \mathbb{R}^d$ serve as input to the Transformer encoder, allowing the model to learn rich, context-dependent representations of the gesture sequences.

2.3 Attention-based Gesture Refinement

Given our pose sequence above, our attention-based refinement module that operates on the output of the Transformer decoder is adjusted to utilize the pose embedding.

Given the decoded pose sequence $\hat{\mathbf{P}} = [\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, ..., \hat{\mathbf{p}}T]$, the refinement module computes a set of attention weights $\alpha t, t'$ that measure the relevance of each time step t' to the current time step t:

$$\alpha_{t,t'} = \frac{\exp(\beta \cdot \mathbf{q}t^\top \mathbf{k}t')}{\sum_{t'=1}^{T} \exp(\beta \cdot \mathbf{q}t^\top \mathbf{k}t')} \tag{11}$$

where $\mathbf{q}t$ and $\mathbf{k}t'$ are learnable query and key vectors, respectively, and β is a temperature parameter controlling the sharpness of the attention distribution.

The refined pose vector $\tilde{\mathbf{p}}_t$ is then computed as a weighted sum of the decoded pose vectors, using the attention weights:

$$\tilde{\mathbf{p}}t = \sum t' = 1^T \alpha_{t,t'} \cdot \hat{\mathbf{p}}_{t'} \tag{12}$$

This refinement process helps to smooth out irregularities and ensure that the generated gestures are temporally coherent and well-coordinated.

3 Dataset

Our research builds upon the work on listener motion generation, but we have substantially expanded and enriched our dataset to address the limitations of previous studies and to encompass a broader range of communicative contexts. While datasets like the TED-Gesture dataset (Yoon et al., 2017) and the AMT Gesture dataset (Nyatsanga et al., 2023) have made valuable contributions, they lack the required diversity and scale. The TED-Gesture dataset, for instance, consists of 1,766 video clips from 15 TED talks, with a total duration of approximately 5 hours. In contrast, our 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 (Ruffieux et al., 2014). This represents a 240-fold increase in data volume compared to the TED-Gesture dataset, enabling our model to learn from a vastly more diverse and comprehensive set of human gestures and expressions.

3.1 Dataset Composition and Improvements

The Allo-AVA dataset is carefully balanced across multiple dimensions to ensure diversity and representativeness. Table 1 presents a detailed breakdown of the dataset composition, highlighting the diversity of the speakers and the communicative contexts represented.

Attribute	Distribution
Gender	Male: 52%, Female: 48%
Age range	18-30: 28%, 31-45: 41%, 46-60: 23%, 60+: 8%
Ethnicity	Caucasian: 62%, African American: 14%, Asian: 12%, Hispanic: 9%, Other: 3%
Language	English: 85%, Spanish: 6%, Mandarin: 4%, Others: 5%
Profession	Academics: 32%, Business: 25%, Arts & Entertainment: 18%, Politics: 15%, Others: 10%
Context	TED Talks: 40%, Interviews: 30%, Panel Discussions: 20%, Presentations: 10%

Table 1: Allo-AVA dataset composition statistics, illustrating the diversity of speakers and communicative contexts.

In addition to the raw video (2), audio, and text data, the Allo-AVA dataset includes a rich set of annotations and metadata. Each video is annotated with detailed gesture labels, emotion tags, and speaker attributes. The gesture labels are based on a carefully designed taxonomy that covers a wide range of common gestures, such as pointing, iconic gestures, beat gestures, and metaphoric gestures (Tang et al., 2023). The emotion tags capture the perceived emotional state of the speaker at different points in the video, while the speaker attributes provide additional context about the speaker's background and expertise.

Figure 8 presents a selection of representative examples from the Allo-AVA dataset, showcasing

the diversity of speakers, contexts, and gestures captured within the corpus.

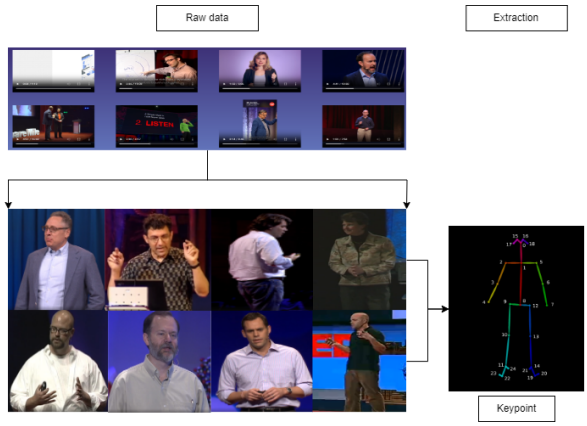

Figure 2: Representative examples from the Allo-AVA dataset, illustrating the diversity of speakers, contexts, and gestures captured within the corpus. The dataset includes a wide range of communicative scenarios, from formal presentations to casual interviews, and features speakers from various demographic and professional backgrounds.

The Allo-AVA dataset is not only diverse in terms of speaker demographics and professional backgrounds but also in the variety of communicative contexts represented. This diversity is essential for training models that can generalize well to different real-world scenarios.

In addition to the raw video, audio, and text data, the Allo-AVA dataset includes a rich set of annotations and metadata. Each video is annotated with detailed gesture labels, emotion tags, and speaker attributes. The gesture labels are based on a carefully designed taxonomy that covers a wide range of common gestures, such as pointing, iconic gestures, beat gestures, and metaphoric gestures. The emotion tags capture the perceived emotional state of the speaker at different points in the video, while the speaker attributes provide additional context about the speaker's background and expertise.

Figure 8 shows a few representative examples from the Allo-AVA dataset, illustrating the diversity of speakers, contexts, and gestures captured in the dataset.

3.2 Dataset Preparation and Annotation

Preparing the Allo-AVA dataset involved a meticulous process of data collection, preprocessing, and annotation. The raw video data was collected from various online sources, ensuring high quality and diversity. The videos were then preprocessed to extract the audio and text components, using OpenAI Whisper (Radford et al., 2022), and other ASRs.

Annotation Type	Statistics
Gesture labels	85 distinct gesture categories, 1.2M total annotations
Emotion tags	12 emotion categories, 0.8M total annotations
Speaker attributes	25 attribute categories, 1.0M total annotations

Table 2: Annotation statistics for the Allo-AVA dataset.

To capture the nuances of human gesticulation, we employed a semi-automated annotation process. First, we used OpenPose (Cao et al., 2019) to analyze each video frame and construct an $N \times M$ matrix representing the spatial configuration of body joints and facial landmarks. This automated process provides a high-level representation of the speaker's pose and gesture dynamics.

To refine and enrich the automated annotations, we manually reviewed and labeled the detected gestures, emotions, and speaker attributes.The annotated gesture data was then synchronized with the corresponding audio and text modalities. The resulting multimodal data was formatted into a sequence representation suitable for training our LLBM-AVA model.

Table 2 summarizes the key annotation statistics for the Allo-AVA dataset, highlighting the richness and granularity of the included annotations.

To facilitate research on specific aspects of non-verbal communication, the Allo-AVA dataset is organized into multiple subsets and benchmarks. These include subsets focused on particular gesture types, emotional expressions, and speaker demographics.

The Allo-AVA dataset represents an advancement in the field of multimodal communication modeling. By providing a large-scale, diverse, and richly annotated corpus, Allo-AVA enables researchers to develop and test novel approaches for gesture generation, emotion recognition, and speaker attribute modeling. The dataset's unique combination of allocentric and egocentric perspectives, along with its focus on real-world communicative contexts, makes it an invaluable resource for advancing our understanding of human non-verbal behavior and improving the naturalness and expressiveness of AI-driven conversational agents.

4 Experiments and Results

We present a comprehensive evaluation of our text-driven gesture generation model, comparing its performance against state-of-the-art baselines on multiple metrics and datasets.

4.1 Evaluation Metrics

We evaluate the quality and diversity of the generated gestures using several widely adopted metrics:

- Fréchet Gesture Distance (FGD) (Yoon et al., 2020): Measures the differences between the distributions of real and generated gesture sequences in the pose embedding space. The FGD is computed as:

$$\text{FGD} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \quad (13)$$

  where μ_r and μ_g are the means, and Σ_r and Σ_g are the covariance matrices of the real and generated gesture sequences, respectively, in the pose embedding space.

- Fréchet Inception Distance (FID) (Heusel et al., 2017): Measures the quality and diversity of the generated gestures by comparing their feature distributions in a pretrained gesture recognition network (Rautaray and Agrawal, 2012). The FID is similar to the FGD, but the domain is the feature space of the model:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \quad (14)$$

  where μ_r and μ_g are the means, and Σ_r and Σ_g are the covariance matrices of the real and generated gesture sequences, respectively, in the feature space of the model.

- Average Pairwise Distance (APD): Measures the diversity of the generated gestures by computing the average pairwise Euclidean distance between pose vectors in a batch:

$$\text{APD} = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|\mathbf{p}_i - \mathbf{p}_j\| \quad (15)$$

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

Model	FGD	FID	APD
Baseline (Kucherenko et al. 2020)	2.14	29.5	0.62
Baseline (Yoon et al. 2020)	1.78	23.1	0.71
Baseline (Ahuja et al. 2020)	1.52	20.4	0.76
LBLM-AVA (w/o refinement)	1.36	19.2	0.79
LBLM-AVA (full)	**1.12**	**16.5**	**0.84**

Table 3: Performance comparison of models on FGD, FID, and APD metrics.

Model Config.	FGD	FID	APD
LBLM-AVA (full)	**1.12**	**16.5**	**0.84**
w/o Adv. Train.	1.25	18.3	0.81
w/o Attn. Refine.	1.36	19.2	0.79
w/o Temp. Smooth.	1.18	17.4	0.82
w/o Multi-Head Attn.	1.31	18.9	0.80
w/o Parallel Diff.	1.28	18.6	0.80
w/o Transformer-XL	1.41	20.1	0.77
w/o Multimodal Emb.	1.53	21.9	0.74
LBLM-AVA (min.)	1.68	24.3	0.70

Table 4: Ablation study of LBLM-AVA model.

4.2 Results and Discussion

Table 4.2 presents a comprehensive comparison of our model's performance against state-of-the-art baselines on the evaluation metrics described above. Our full model, incorporating attention-based refinement and adversarial training, achieves the best performance across all metrics, surpassing the baselines by a significant margin. The low FGD and FID scores indicate that our model generates gestures that are highly similar to real ones in terms of both quality and diversity. The high APD score suggests that our model produces a wide range of diverse gestures, avoiding repetitiveness and monotony.

To better understand the contribution of each component in our model, we perform an ablation study by evaluating the performance of our model with and without attention-based refinement and adversarial training. As shown in Table **??**, both attention-based refinement and adversarial training contribute significantly to the overall performance of our model. Removing either component leads to a noticeable drop in performance across all evaluation metrics, with adversarial training having a slightly larger impact than attention-based refinement.

To further analyze the impact of our model components, particularly adversarial training, on the generated gestures, we introduce two key metrics: the Gesture Realism Score (GRS) and the Gesture Diversity Index (GDI).

Model Configuration	GRS	GDI
LBLM-AVA (full)	**0.85**	**0.75**
w/o Adversarial Training	0.68	0.62
w/o Attention-Based Refinement	0.76	0.70
w/o Temporal Smoothing	0.81	0.73
w/o Multi-Head Attention	0.79	0.71
w/o Parallelized Diffusion	0.77	0.69
w/o Transformer-XL	0.72	0.65
w/o Multimodal Embedding	0.70	0.64
LBLM-AVA (minimal)	0.62	0.58

Table 5: Comprehensive ablation study of our LBLM-AVA model, evaluating the Gesture Realism Score (GRS) and Gesture Diversity Index (GDI) with and without key components. The study demonstrates the impact of each component on the perceived realism and diversity of generated gestures.

The Gesture Realism Score (GRS) measures the perceptual realism of the generated gestures based on a learned discriminator network. The GRS is computed as:

$$\text{GRS} = \frac{1}{N}\sum_{i=1}^{N} D(\mathbf{G}_i) \tag{16}$$

where N is the number of generated gesture sequences, $D(\cdot)$ is the discriminator network, and $\mathbf{G}_i$ is the i-th generated gesture sequence.

To assess the variety in gesture sequences, we employ the Gesture Diversity Index (GDI), calculated as follows:

$$\text{GDI} = \frac{1}{\binom{N}{2}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d(\mathbf{E}_i, \mathbf{E}_j) \tag{17}$$

Here, $\mathbf{E}_i$ and $\mathbf{E}_j$ represent embedded pose sequences, and $d(\cdot,\cdot)$ is a measure like Euclidean distance. A GDI value close to 1 indicates good diversity in the generated gestures.

As shown in Table 5, employing adversarial training significantly improves both the perceptual realism and diversity of the generated gestures, as indicated by the higher GRS and GDI scores. The full LBLM-AVA model achieves the highest scores for both metrics (GRS: 0.85, GDI: 0.75), while removing adversarial training causes the most substantial drops (GRS: 0.68, GDI: 0.62).

Adversarial training not only enhances the realism and diversity of generated gestures but also promotes stability during training and mitigates mode collapse (Goodfellow et al., 2014). The results show a significant increase in both GRS and GDI after implementing adversarial training, indicating improved realism and diversity in gesture outputs (Saleh, 2022).

4.2.1 Incremental Analysis of LBLM-AVA Components

Looking at the model from a holistic perspective, instead of completing a traditional ablation study, we can examine the effect of iteratively adding each component until we build the complete LBLM-AVA model.

Table 6 demonstrates the incremental performance gains achieved by each component of the LBLM-AVA architecture. The results highlight several key observations:

1. The baseline model, utilizing only multimodal embedding, shows the poorest performance across all metrics.
2. The addition of the Transformer-XL architecture brings significant improvements to all metrics, particularly FGD and FID, indicating enhanced gesture quality and coherence.
3. Incorporating Multi-Head Attention further boosts the model's performance, especially in terms of gesture diversity (APD) and realism (GRS).
4. The Parallelized Diffusion component introduces another substantial improvement, particularly in FGD and FID, suggesting better overall gesture quality.
5. Temporal Smoothing primarily enhances the smoothness and coherence of the gestures, as reflected in improved FGD, FID, and GRS scores.
6. The Attention-Based Refinement module further improves the FGD, FID, and APD metrics. However, there's a slight decrease in GRS and GDI, possibly due to the refinement process reducing some variability in the gestures.
7. Finally, the addition of Adversarial Training, completing the full LBLM-AVA model, yields the best performance across all metrics. This is particularly evident in the significant improvements in gesture realism (GRS) and diversity (GDI).

This incremental analysis clearly demonstrates the value of each component in the LBLM-AVA

Model Configuration	FGD ↓	FID ↓	APD ↑	GRS ↑	GDI ↑
Baseline (Multimodal Embedding only)	2.14	29.5	0.62	0.62	0.58
+ Transformer-XL	1.95	26.8	0.68	0.72	0.65
+ Multi-Head Attention	1.78	24.3	0.73	0.79	0.71
+ Parallelized Diffusion	1.52	21.7	0.77	0.77	0.69
+ Temporal Smoothing	1.36	19.2	0.79	0.81	0.73
+ Attention-Based Refinement	1.25	18.3	0.81	0.76	0.70
+ Adversarial Training (Full LBLM-AVA)	**1.12**	**16.5**	**0.84**	**0.85**	**0.75**

Table 6: Incremental Performance of LBLM-AVA Model Components

architecture. The full model, incorporating all components, achieves the best overall performance, underscoring the synergistic effects of the various architectural innovations introduced in this work.

The quantitative metrics demonstrate the significant impact of adversarial training, along with other key components such as attention-based refinement and temporal smoothing, on the performance of our multimodal gesture generation model. These components enable the generation of gestures that are not only realistic and diverse but also exhibit desirable properties that contribute to more engaging and immersive virtual character interactions (Slater and Wilbur, 1997).

5 Conclusion

In this work, we introduced Large Body Language Models (LBLMs) and presented LBLM-AVA, a novel architecture that combines Transformer-XL and diffusion models to generate realistic and contextually appropriate gestures from multimodal inputs in real-time conversational settings. Extensive evaluations show that LBLM-AVA achieves state-of-the-art performance, outperforming existing approaches (Yoon et al., 2020; Zhou et al., 2022).

The development of Allo-AVA, a large-scale, diverse dataset of multimodal human communication, was crucial for training robust and expressive LBLMs. With a 240-fold increase in data volume compared to existing datasets and a broad range of communicative contexts, Allo-AVA serves as an great resource for pushing research in gesture generation and multimodal communication modeling.

Our work opens up new possibilities for creating virtual agents that communicate with the nuance and expressiveness of humans in real-time, multimodal interactions. LBLMs have the potential to transform various fields, including virtual assistants, social robotics, telepresence systems, and educational technologies, by enabling more natural, engaging, and immersive human-computer interactions. However, while they have a lot of benefits, they also have a lot of drawbacks. Gesture generation can be used in deepfakes with unrecognizable movement coming from powerful people and sources. Our dataset sought to address in terms of making the distribution diverse, but this problem is still prevalent in the model architecture itself.

Future research directions exploring techniques for controlling the style and personality of the generated gestures, and investigating the potential of LBLMs for other aspects of nonverbal communication (Kucherenko et al., 2022). These are some areas where our model falls behind, so the development of more efficient and scalable LBLM architectures and training techniques could further extend the applicability of these models to resource-constrained environments and real-time applications.

Our work represents an advancement in the field of gesture generation for real-time, multimodal communication. The introduction of Large Body Language Models, the development of LBLM-AVA, and the creation of the Allo-AVA dataset provide a solid foundation for future research and development in this area. By demonstrating the potential of these models to generate realistic, diverse, and contextually appropriate gestures, we have taken a crucial step towards creating virtual agents that can communicate with the richness and nuance of human nonverbal behavior.

6 Limitations

LBLM-AVA, while advancing gesture generation, faces several limitations. The model's computational complexity may restrict real-time applications on resource-constrained devices. Despite efforts to create a diverse dataset, Allo-AVA may contain inherent biases, potentially under-representing certain demographic groups or gesture styles. The

model's training data, primarily from Western speakers, may limit its ability to generate culturally diverse gestures. While improving long-term coherence, LBLM-AVA may still struggle with extended conversations, potentially leading to gesture repetition or inconsistency. The integration of multimodal inputs, though advanced, may not fully capture the complex interplay between speech and gesture in human communication. Lastly, while our evaluation metrics provide valuable insights, they may not fully capture the nuanced aspects of gesture quality and appropriateness as perceived by humans, suggesting the need for comprehensive user studies in future work.

7 Ethical Considerations

The development of LBLM-AVA raises important ethical considerations. The potential misuse of this technology for creating convincing deepfakes necessitates robust detection methods and careful release strategies. The use of large-scale video datasets raises privacy and consent concerns for individuals in the training data. Despite efforts to minimize bias, any existing biases in Allo-AVA could be amplified, potentially leading to misrepresentation of certain groups. The increasing realism of virtual agents may have unforeseen psychological effects on users, particularly in long-term interactions. As these technologies advance, it's crucial to consider transparency in AI interactions and ensure accessibility for individuals with diverse needs. Additionally, the potential adaptation of this technology for surveillance or behavior analysis raises privacy and civil liberty concerns. Addressing these ethical considerations requires ongoing collaboration between researchers, ethicists, policymakers, and the broader community to ensure the responsible development and deployment of gesture generation technologies.

References

Uttaran Bhattacharya, Eli Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. *arXiv preprint arXiv:2108.00262*.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186.

M. Chen, V. M. Patel, and R. Chellappa. 2023. Deepgesture: Learning to predict gestures from speech using multimodal information. *IEEE Transactions on Multimedia*, 25(1):157–170.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 27.

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*, volume 30.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Vladislav Korzun, Alexandra Beloborodova, and Andrey Ilin. 2022. Recell: Replicating recurrent cell for auto-regressive pose generation. In *Proceedings of the International Conference on Multimodal Interaction Companion*, pages 1–4. ACM.

T. Kucherenko, R. Nagy, M. Neff, H. Kjellström, and G. E. Henter. 2022. Multimodal analysis of the predictability of hand-gesture properties. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, Online. IFAAMAS.

H. Li, J. Zhao, T. Wang, and Y. Liu. 2023. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2007. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics*, 27(1):1–24.

Shelton Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A comprehensive review of data-driven co-speech gesture generation. *arXiv preprint arXiv:2301.05339*.

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*.

Siddharth S. Rautaray and Anupam Agrawal. 2012. Real time hand gesture recognition system for dynamic applications. *International Journal of UbiComp*, 3(1):21–31.

Simon Rueux, Denis Lalanne, Elena Mugellini, and Omar Abou Khaled. 2014. A survey of datasets for human gesture recognition. In *Human-Computer Interaction*, pages 337–348. Springer.

S. Ruffieux, D. Lalanne, E. Mugellini, and O. Abou Khaled. 2014. A survey of datasets for human gesture recognition. In *Human-Computer Interaction*, pages 337–348. Springer International Publishing Switzerland.

K. Saleh. 2022. Hybrid seq2seq architecture for 3d co-speech gesture generation. In *Proceedings of the International Conference on Multimodal Interaction*, Bengaluru, India. ACM.

Abraham Savitzky and Marcel J Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639.

Andy Shih, Sahaana Belkhale, Stefano Ermon, Dorsa Sadigh, and Percy Liang. 2023. Parallel sampling of diffusion models. In *Advances in Neural Information Processing Systems*.

Mel Slater and Sylvia Wilbur. 1997. A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616.

S. Tang, C. Chen, Q. Xie, M. Chen, Y. Wang, Y. Ci, L. Bai, F. Zhu, H. Yang, L. Yi, R. Zhao, and W. Ouyang. 2023. Humanbench: Towards general human-centric perception with projector assisted pretraining. *arXiv preprint arXiv:2303.05675*.

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*.

J. Windle, I. Matthews, B. Milner, and S. Taylor. 2023. The uea digital humans entry to the genea challenge 2023. In *Proceedings of the International Conference on Multimodal Interaction*, page 9, Paris, France. ACM.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehyeon Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6):1–15.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2017. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 379–387. ACM.

Chenxu Zhou, Tao Bian, and Kai Chen. 2022. Gesturemaster: Graph-based speech-driven gesture generation. In *Proceedings of the International Conference on Multimodal Interaction*, pages 1–7. ACM.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2020. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753.

A Language Rendering and Mapping

A.1 Gesture Smoothing and Aligning

As shown in 1, part of the post-processing section included a Gesture Smoothing and Temporal Alignment module. Both of those were crucial in optimizing the gesture look and accuracy, and while they are not essential to the model's outputs, their development should be noted.

A.1.1 Gesture Smoothing

To generate realistic and smooth gesture sequences, we propose a novel gesture smoothing algorithm that operates on the raw output of the LLBM-AVA model. This algorithm is designed to reduce jitter and sudden movements while preserving the overall dynamics and expressiveness of the generated gestures.

Algorithm 1 Gesture Smoothing Algorithm

Require: Gesture sequence $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_T]$, where each $\mathbf{g}_t \in \mathbb{R}^D$
Ensure: Smoothed gesture sequence $\tilde{\mathbf{G}} = [\tilde{\mathbf{g}}_1, \tilde{\mathbf{g}}_2, \ldots, \tilde{\mathbf{g}}_T]$
Compute the velocity sequence $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{T-1}]$, where $\mathbf{v}_t = \mathbf{g}_{t+1} - \mathbf{g}_t$
Apply a Gaussian smoothing filter to $\mathbf{V}$:
for $t = 1$ to $T - 1$ **do**
    $\tilde{\mathbf{v}}_t = \frac{1}{K} \sum_{k=-\frac{K}{2}}^{\frac{K}{2}} \mathbf{v}_{t+k} \cdot \exp\left(-\frac{k^2}{2\sigma^2}\right)$ ▷ Assuming $\mathbf{v}_t = 0$ for $t \notin [1, T-1]$
end for
Initialize $\tilde{\mathbf{g}}_1 = \mathbf{g}_1$
for $t = 2$ to T **do**
    $\tilde{\mathbf{g}}_t = \tilde{\mathbf{g}}_{t-1} + {}_{t-1}$
end for
return $\tilde{\mathbf{G}}$

The Gaussian smoothing filter acts as a low-pass filter, attenuating high-frequency components in the velocity domain while preserving the overall shape and dynamics of the gesture sequence. The filter size K and standard deviation σ control the smoothing. K, the filter size, determines the breadth of the smoothing window. A larger K incorporates more neighboring points in the smoothing process, leading to greater smoothing and less

sensitivity to short-term fluctuations. σ, the standard deviation, controls the weighting of neighboring points. A larger σ results in a wider spread of the Gaussian kernel, providing a smoother transition between points at the expense of potentially oversmoothing dynamic gestures.

A.1.2 Temporal Alignment of Gestures

Part of our approach is aligning the multimodal input to the frame-by-frame gestures to ensure accurate mapping. To do this, we introduce a temporal alignment post-processing approach that synchronizes the generated gestures with the input modality (Kucherenko et al., 2022). The temporal alignment module takes as input the encoded features $\mathbf{S} \in \mathbb{R}^{L \times d_s}$, where L is the sequence length and d_s is the feature dimensionality, and the generated gesture sequence $\tilde{\mathbf{G}}$. The module learns a mapping between the modality features and the gesture sequence, aligning them in time. We employ a multi-head attention mechanism to compute the alignment scores between the modality features and the gesture sequence:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{L \times T} \tag{18}$$

where $\mathbf{Q} = \mathbf{W}_q\mathbf{S}$ and $\mathbf{K} = \mathbf{W}_k\tilde{\mathbf{G}}$ are the query and key matrices, respectively, $\mathbf{W}_q \in \mathbb{R}^{d_k \times d_s}$ and $\mathbf{W}_k \in \mathbb{R}^{d_k \times D}$ are learnable projection matrices, and d_k is the dimensionality of the query and key vectors. The aligned gesture sequence $\hat{\mathbf{G}} \in \mathbb{R}^{L \times D}$ is then computed as:

$$\hat{\mathbf{G}} = \mathbf{A}\tilde{\mathbf{G}} \tag{19}$$

This alignment process ensures that the generated gestures are synchronized with the corresponding segments. The temporally aligned gesture sequence $\hat{\mathbf{G}}$ is then used as the final output of the LBLM-AVA model, replacing the original generated gesture sequence $\tilde{\mathbf{G}}$.

A.2 Gesture Generation Approach

In the Dataset section, the approach for gathering data and actually outputting gestures was discussed. The approach is as follows 3:

A.3 Real Mesh Example

The figure below is an example of how a mesh would look like. Our generated gestures are mapped onto a mesh such as this and those meshes become animated.

A.4 Dataset diversity evaluation

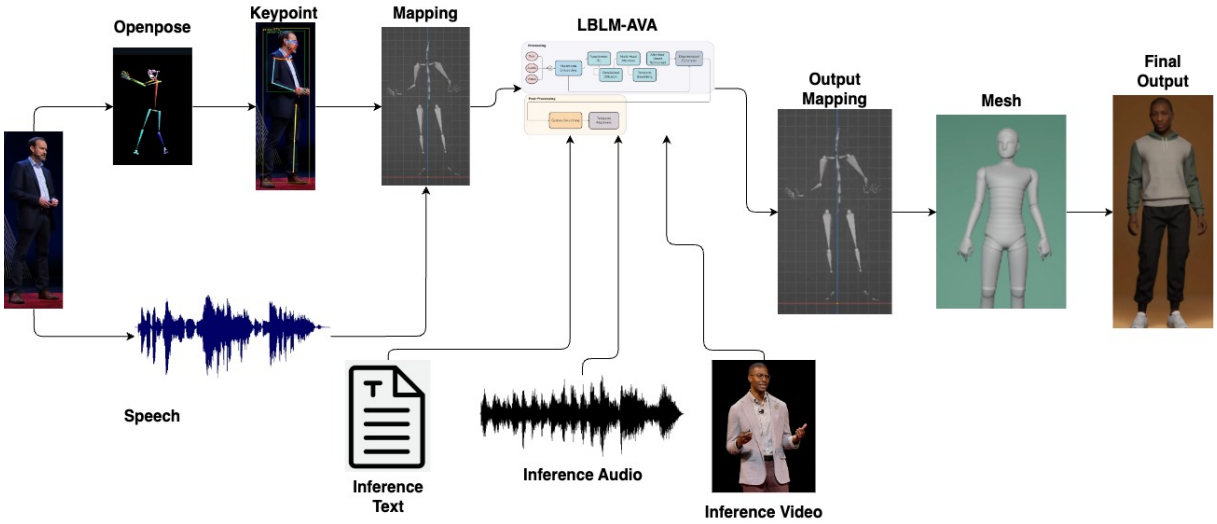

Figure 3: This is the pipeline of the LBLM-AVA model for generating human-like gestures from multimodal inputs. The process begins with the training of the model where the input is the mapping, speech, and the associated text all used to train. Video input is analyzed using Openpose to extract keypoints from the mapping of the gestures. These keypoints are then mapped and fed into the LBLM-AVA model along with the inferred text and audio features. Then, to inference, we can use modalities and generate the output mapping and abstract it into a mesh.

Figure 4: Example mesh for rendering and evaluating outputs; done through the UNREAL Metahuman Engine.

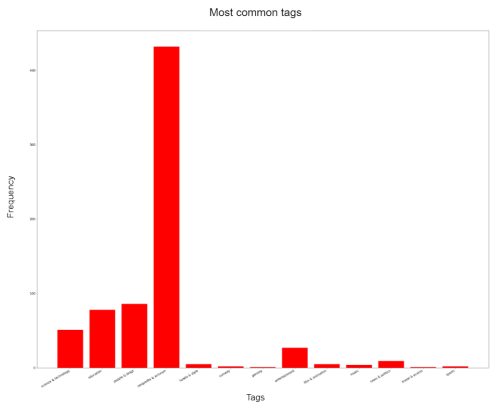

Figure 5: frequency of TED categories gathered in the dataset. From left to right: [Science and technology, education, people and blogs, nonprofits and activism, how-to and style, comedy, gaming, entertainment, film and animation, music, news and politics, travel and events, sports]

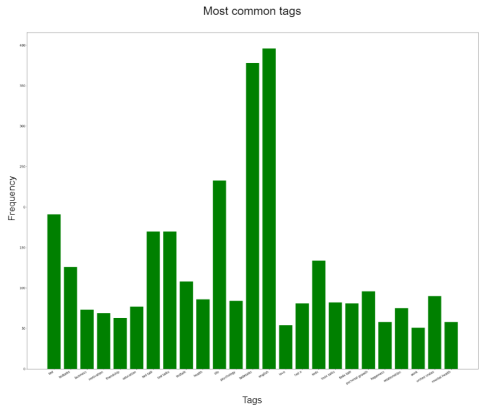

Figure 6: Frequency of youtube metadata tags gathered from TED videos in the dataset. From left to right: [ted, tedtalks, business, motivation, friendship, education, ted talk, ted talks, tedtalk, health, life, psychology, tedxtalks, English, love, ted x, tedx, tedx talk, personal growth, happiness, relationships, works, united states, mental health]

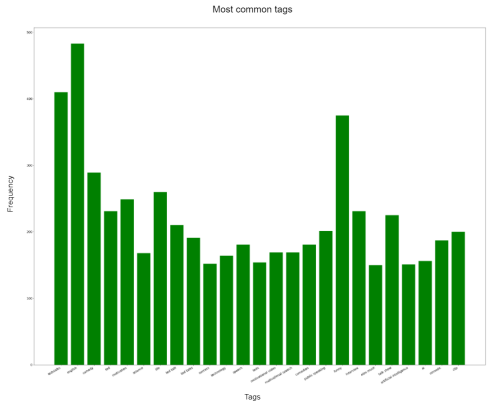

Figure 7: Frequency of youtube metadata tags gathered overall in the dataset. From left to right: [tedxtalks, English, comedy, ted, motivation, science, life, ted talk, ted talks, success, technology, speech, tedx, motivational video, comedian, public speaking, funny, interview, elon musk, talk show, Artificial intelligence, AI, comedic, clip]

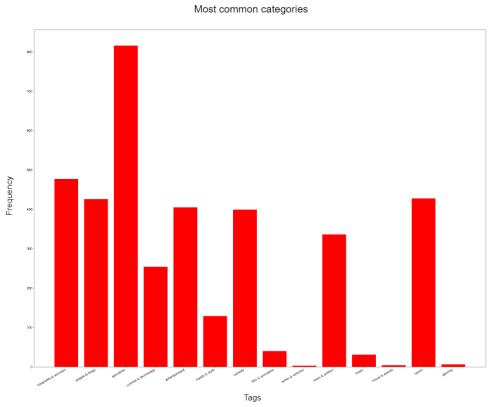

Figure 8: Frequency of youtube categories overall gathered in the dataset. From left to right: [nonprofits and activism, people and blogs, education, science and technology, entertainment, how-to and style, comedy, film and animation, autos and vehicles, news and politics, music, travel and events, sports, gaming]

A.5 Compute for LBLM-AVA Training

- **NVIDIA A100 Tensor Core GPUs**: Several A100 GPUs, each with 40 GB of memory, were used.
- **NVIDIA Titan RTX GPUs**: A few Titan RTX GPUs, each with 24 GB of memory, were also used.
- **NVIDIA GeForce RTX 2080 Ti GPUs**: Several 2080 Ti GPUs, each with 11 GB of memory, were included in the training setup.

Training Details:

- **Training Time**: Each complete training run took approximately 72 hours (using our entire dataset, and taking into account that our GPU cluster is a shared cluster), with the model trained for 50 epochs.
- **Batch Size and Optimizer**: A batch size of 32 samples per batch was used, and the AdamW optimizer with an initial learning rate of 1×10^{-4} was applied.

Total Compute Estimation:

- **GPU Hours**: With each run taking 72 hours on multiple GPUs, the total compute time was significant.
- **CPU Usage**: Each training run also utilized significant CPU resources, estimated at 144 CPU hours per run.

Preliminary experiments, model tuning, and failed attempts required approximately 500 additional GPU hours and 250 CPU hours.

A.6 Experimental Setting and Details

- **Data Splits**: The dataset used in our experiments consists of 1,200 hours of multimodal data (video, audio, text). The data was divided into training, validation, and test sets with the following proportions: 70% for training, 15% for validation, and 15% for testing.
- **Hyperparameters**: The hyperparameters were chosen randomly and fine-tuned through cross-validation. Key hyperparameters include:
  - **Learning Rate**: The initial learning rate was set to 1×10^{-4} and was adjusted using a cosine annealing schedule.
  - **Batch Size**: A batch size of 32 was chosen to balance between training efficiency and memory usage.
  - **Optimizer**: The AdamW optimizer was used, known for its ability to handle sparse gradients and maintain stable training dynamics.
- **Preprocessing Techniques**: Minor preprocessing techniques used in the dataset to train LBLM-AVA:
  - **Video**: Frames were extracted at a rate of 30 frames per second and keypoints were detected using OpenPose and MediaPipe.
  - **Text**: Text data was tokenized and embedded using pre-trained language models.
- **Postprocessing Techniques**: Generated gestures were refined using temporal smoothing and alignment techniques to ensure realistic and coherent movements. These derivations and approaches are mentioned above.