SICAGE: Speaker-Independent Culture-Aware Gesture Generation using TED4C-L Dataset

Ariel Gjaci[1]*, Antonio Sgorbissa[2], and Vittorio Murino[1]

[1] Italian Institute of Technology, Genoa, Italy
*ariel.gjaci@iit.it
[2] University of Genoa, Genoa, Italy

Abstract. Recent co-speech gesture generation methods often overlook cultural differences, limiting their effectiveness in human–agent interaction. Moreover, culture-conditioned models are rarely evaluated under speaker-disjoint splits, so apparent "cultural" behavior may be confounded with speaker-specific gesturing style. We introduce SICAGE, a modular framework for culture-aware co-speech gesture generation that conditions motion synthesis models on speaker-independent cultural representations. SICAGE learns these representations from audio and text by treating each speaker as a separate domain while imposing invariance across speakers. This encourages representations to remain culture-discriminative while reducing dependence on speaker identity. The resulting cultural embeddings condition a multimodal generator to produce culturally appropriate gestures. We instantiate this idea with two domain generalization approaches: adversarial learning and Fishr regularization. We further introduce ALaDiT, a real-time diffusion-based gesture generator designed to efficiently incorporate the learned cultural embeddings. To validate our method, we built TED4C-L, a 106-hour multimodal dataset of 764 TED speakers from four cultural groups. Experiments show that SICAGE improves motion realism, diversity, beat synchronization, semantic relevance, and cultural consistency.

Keywords: Co-speech gestures · Culture · Multimodal learning

1 Introduction

Human communication is inherently multimodal, combining speech with co-speech gestures that reinforce verbal messages and convey abstract concepts and emotions. The importance of gestures is well documented in human–human [2, 5, 31, 40], human–agent [7, 18, 30, 37], and social robotics interactions [6, 47]. A critical yet underexplored dimension in gesture generation is culture, as cultural norms shape gesture performance and interpretation [4, 29, 38]. Defining cultural boundaries is itself nontrivial, especially in dynamic and multiethnic societies [49]. As a result, "culture-aware" modeling must clearly specify what aspects of culture are being measured and how evaluation is conducted; otherwise, apparent cultural effects may simply reflect confounding factors.

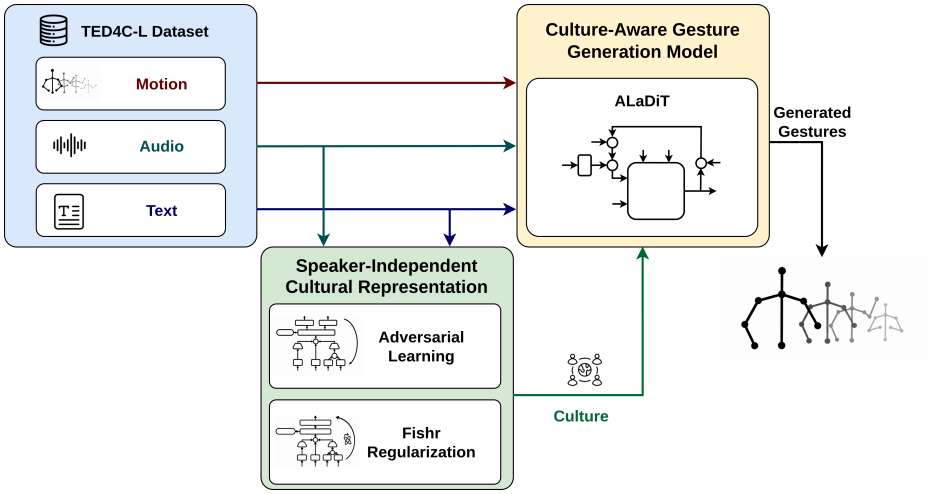

Fig. 1: Overview of our implementation of SICAGE. We extract text, audio, and motion features from TED4C-L (blue), regularize text/audio into speaker-independent cultural embeddings via Fishr or adversarial training (green), then feed these embeddings together with other raw inputs into ALaDiT (yellow) to generate real-time, culture-aware gestures. All components are modular and replaceable.

Existing culture-aware gesture generation methods [19, 48, 54] rely on culturally annotated datasets but often do not rigorously test whether learned cultural patterns transfer across speakers. In particular, evaluations frequently use speaker-dependent protocols in which the same speakers appear in both training and test splits. In such settings, models may appear to capture culture while relying on speaker-specific cues that reflect individual style rather than group-level patterns. A robust evaluation of cultural generalization therefore requires both large-scale data and explicitly speaker-disjoint splits, so that culture must be inferred from patterns shared across speakers rather than memorized from individuals.

In this paper, we address these limitations by proposing **SICAGE**[3] (Speaker-Independent Culture-Aware Gesture gEneration), a modular framework for synthesizing culture-aware gestures while explicitly reducing dependence on speaker identity. Within SICAGE, we adopt the view of culture as a learned social and individual construct affecting both behavior and interpretation [49], and operationally model it as speaker-disjoint, group-level communication patterns shared within cultural groups. Since these patterns are expressed through entangled linguistic, prosodic, semantic, and other behavioral cues [21, 35], our goal is not to causally disentangle culture from these modalities. Instead, we learn conditioning embeddings that are discriminative of cultural groups while reducing their dependence on individual speaker style. We therefore cast cultural representation

[3] Project page, dataset, and source code: `https://arielgjaci.com/sicage`

learning as a domain-generalization problem, treating each speaker as a domain and using speaker-invariant learning objectives to encourage embeddings that generalize to unseen speakers from the same cultural group.

To support robust training and evaluation under this objective, we introduce **TED4C-L**, a large-scale, multimodal, and speaker-balanced dataset with 106 hours of TED talks from 764 speakers across four regions: India (TED Talks in Hindi), Italy, Turkey, and Japan. We select these groups based on abundant native-language TED talks and clear geographic separation to minimize cultural overlap. Unlike English, these languages closely reflect geographic and cultural identity, enhancing annotation reliability. To assess whether cultural patterns are detectable from gesture motion alone, we train a motion-based classifier to predict cultural labels (see supplementary for details). Under speaker-disjoint evaluation, the classifier achieves approximately 45% balanced accuracy on unseen speakers, compared to a 25% random baseline for four classes. This result indicates that gesture motion contains culture-related signals that generalize across speakers, while also highlighting the substantial intra-cultural variability that makes the task challenging.

Within SICAGE, we learn speaker-invariant cultural representations using two domain generalization strategies that treat each speaker as a domain: Fishr regularization [46] and adversarial learning [16], which has proven effective in related multimodal settings [21]. These objectives encourage representations that remain predictive of culture while reducing sensitivity to speaker identity.

For gesture synthesis, we adopt a diffusion-based motion generator [50], ALaDiT, conditioned on audio, text, and the learned cultural embeddings. This design enables gestures that remain synchronized with speech while reflecting culture-dependent motion patterns. Figure 1 provides an overview of the SICAGE framework. Our experiments show that enforcing speaker-invariance improves motion quality, diversity, cultural consistency, and alignment with rhythmic and contextual features extracted from text and culture. In addition, our diffusion-based generator outperforms recent baselines on standard gesture generation metrics.

Overall, the contributions of this work are: (i) **SICAGE**, a modular framework for speaker-independent, culture-aware co-speech gesture generation; (ii) **TED4C-L**, a large-scale, multimodal dataset explicitly designed for cultural generalization with disjoint-speaker splits; (iii) the introduction of **domain generalization** methods, instantiated with **Fishr** and **adversarial learning**, to learn cultural cues that generalize across speakers; and (iv) ALaDiT, a diffusion-based gesture generator conditioned on these cues to synthesize high-quality culture-aware motion.

2 Related Work

2.1 Rule-based methods

Early gesture generation relied on rule-based systems mapping linguistic or prosodic cues to predefined gestures. BEAT [9], for example, uses linguistic and

contextual annotations to animate gestures via a knowledge-based engine. Similarly, other approaches combine part-of-speech tagging and speech timing with grammar rules to sample gesture trajectories [41]. More recent works [1] learn mappings from video but suffer scalability and memory issues. Although these methods produce smooth motion, they rely on predefined gesture units, and to our knowledge, only one study [20] explicitly considers cultural variations.

2.2 Data-driven methods

With larger datasets, gesture synthesis has transitioned to data-driven models, favoring scalability over interpretability. Early probabilistic models, such as MDP controllers [43] and parameter-based selection [15], infer gestures from prosodic features. LLM-based approaches use language models for intent extraction to generate gestures [17, 23], yet remain limited by gesture set size. End-to-end generative methods (GANs [22], attention-based [52], diffusion models [24]) produce novel, diverse motions, with diffusion-transformer models achieving state-of-the-art quality due to tight text/audio alignment [3, 10, 36, 55, 61]. However, existing diffusion models lack explicit cultural conditioning. Some works implicitly embed cultural cues via attention [54] or culture-specific GANs [19], but provide limited quality and generalization to unseen speakers. Our approach conditions diffusion models on dedicated, speaker-independent cultural embeddings for robust cross-cultural synthesis.

2.3 Domain Generalization

Domain Generalization (DG) aims to generalize to unseen domains by training on related ones [53]. Common strategies include data augmentation [25, 44], representation learning [16, 39], and advanced training schemes [13, 46]. Following [21], in this work we treat each speaker in the dataset as a domain and condition gesture synthesis on cultural embeddings. We utilize adversarial learning [16], shown to be effective in similar tasks [21], and Fishr [46], known for efficiently matching domain gradient variances. Fishr's scalability and performance make it ideal for embedding cultural representations across numerous domains.

2.4 Datasets

High-quality multimodal data is critical for gesture synthesis. Motion-capture datasets such as CMU Panoptic [28] and Talking With Hands 16.2M [32] provide accurate 3D keypoints but lack diversity; only LISI-HHI [51] provides cultural annotations, but at a small scale. TED-Talk datasets [58, 59], using OpenPose-extracted skeletons [8], GloVe text embeddings [42], and audio features, offer similar size and greater diversity (97 hours, 1,700+ speakers) but omit cultural labels. MCGD [54] annotates culture for 263 speakers, but it is not public and covers only ∼20 speakers per culture. We introduce TED4C-L, comprising approximately 190 speakers per culture across four cultures (106 hours total), explicitly designed for cultural generalization and gesture generation tasks.

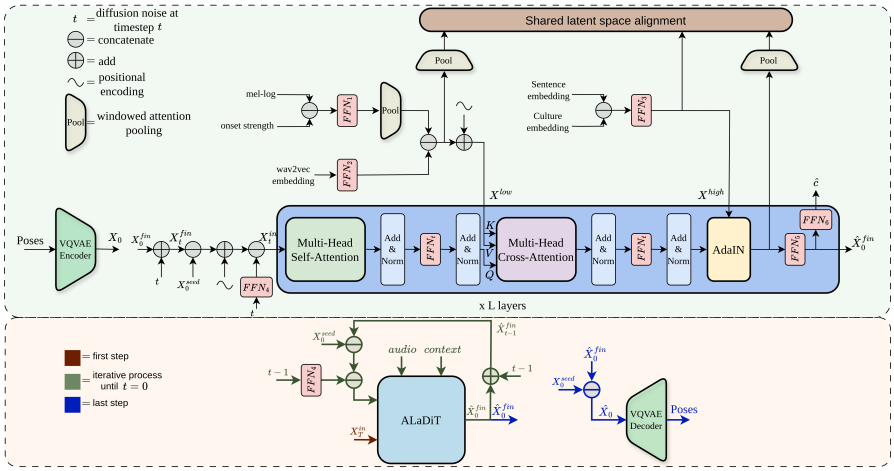

Fig. 2: *Overview of ALaDiT.* **Top (Training):** Motion is encoded via a pretrained VQVAE encoder, split into seed X_0^{seed} and target X_0^{fin}, which is noised to X_t^{fin}. A timestep embedding is concatenated, and motion is processed via self- and cross-attention with audio features (X^{low}), while cultural and textual features (X^{high}) are injected through AdaIN. All features are aligned in a shared space, while cultural classification $\hat{c}$ further enforces cultural consistency. **Bottom (Inference)** Given X_T^{fin}, X_0^{seed}, and context, the model iteratively denoises motion to $\hat{X}_0^{fin}$, then decodes it via the VQVAE decoder.

3 Methodology

We introduce SICAGE, a modular framework for culture-aware co-speech gesture generation comprising: (i) a culturally diverse dataset (TED4C-L in our implementation), (ii) a model for learning speaker-independent cultural representations (via Fishr regularization or adversarial learning in our implementation), and (iii) a motion generator conditioned on culture and other features (ALaDiT in our implementation). Each step can be implemented in various ways; our implementation is detailed in the following sections.

3.1 Data Collection

TED4C-L contains YouTube TED Talks from four coarse country-level groups: India (Hindi TED talks), Italy, Japan, and Turkey. These labels serve as practical group annotations and should not be interpreted as implying cultural homogeneity within countries. Our assumption is only that, despite intra-group variability, these groups may contain detectable group-associated communication patterns that can be studied at scale under speaker-disjoint evaluation. We selected groups with distinct spoken languages and a large number of available TED Talks.

To ensure quality, we include only videos in which speakers are clearly visible, standing, speaking their native language (Hindi for India), and not holding objects that could affect gestures. From a total of 106.45h of video data, we extracted 659,454 overlapping five-second samples using 0.5s stride, each with aligned audio, motion, and transcripts. Table 1 compares TED4C-L with the largest available TED-Talk dataset. Although TED4C-L has fewer speakers (764 vs. 1766), it features longer, higher-quality videos and a shorter stride (0.5s vs. 0.67s), resulting in longer duration (106h vs. 97h) and significantly more samples (659k vs. 252k), along with explicit cultural labels, multiple languages, and balanced speaker counts per culture.

Audio is downsampled to 16 kHz, then processed to extract 64 mel-log features (spectral content), onset strength (rhythm), and 1024-dimensional wav2vec features from a model pretrained on 56 languages [11]. For text, we use Language-Agnostic BERT Sentence Embeddings (LaBSE) [14] to represent contextual meaning, including the first and last words that overlap each sample window for better alignment. Since LaBSE mainly aligns sentences by semantic content and may attenuate language-specific idiomatic structure, the audio features provide complementary acoustic, rhythmic, and prosodic cues not captured by text semantics alone. Motion is represented by 9 upper-body 3D keypoints (neck, head, central hip, shoulders, elbows, wrists) at 15 FPS, extracted with pretrained MMPose [12] models. Choosing only 9 keypoints reflects a deliberate reliability and scope trade-off. SICAGE targets macro-level upper-body gesticulation in unconstrained TED videos, where hands and fingers are often blurred, occluded, or out of frame, and the lower body is frequently truncated by camera framing. Under these conditions, monocular 3D finger extraction or full-body tracking would introduce substantial detection noise and sharply reduce the amount of usable data. Nonetheless, this compact configuration retains a robust group-level signal, enabling our speaker-disjoint classifier to achieve approximately 45% balanced accuracy against a 25% random baseline. Following [58], we retain only sequences where the main speaker is continuously detected for at least 5 seconds. To reduce camera-dependent variability, we normalize poses to be yaw-invariant (removing average shoulder rotation around the vertical axis) and pitch-invariant (removing hip-to-neck rotation around the horizontal axis), ensuring speakers are front-facing and upright. Each 75-frame motion sequence is represented in 6D continuous format [60] and encoded by a pretrained VQVAE into 25 tokens (1024-entry VQVAE codebook with 512-dimensional embeddings) via a 1D convolutional encoder with residual blocks and velocity/acceleration losses [56] to reduce jitter and improve reconstruction (see supplementary for details).

3.2 Speaker-Independent cultural representation

Cultural traits should generalize across individuals from the same group. To obtain speaker-independent representations, we split the dataset so that training, validation, and test sets contain disjoint speakers within each culture. We then train a feed-forward network (FFN) to classify culture using either Fishr regular-

	TED Talk	TED4C-L
Speakers	1766	764
Cultures	/	4
Shots of Interest	/	16,326
Motion FPS	15	15
Duration shots of interest (hours)	97h	106.45h
Average Video Length (minutes)	13m	15.41m
Transcript Languages	En	En/Tr/It/Hi/Ja
Languages Spoken	/	Tr/It/Hi/Ja
Sample Duration (frames)	34	75
Stride (seconds)	0.67	0.5
Total Samples	252,109	659,454

Table 1: Comparison between TED Talk [58] and TED4C-L Datasets.

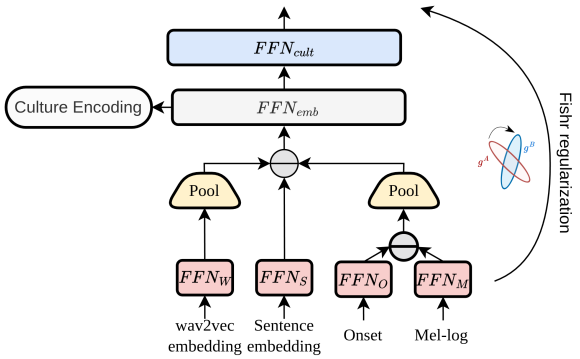

Fig. 3: Feed-forward network trained with Fishr regularization to learn speaker-independent cultural embeddings.

ization or adversarial learning for domain generalization, treating each speaker as a domain and culture as the prediction target.

For Fishr, due to the large number of domains, we randomly sample $k = 64$ speakers per training step. For each selected domain $d \in \mathcal{S}$ (with $\mathcal{S} \subseteq \{1, \ldots, D\}$ and $|\mathcal{S}| = 64$), we form a minibatch of $n_d = 16$ samples.

As illustrated in Figure 3, audio features (wav2vec, mel-log, and onset) and sentence embeddings are processed by separate $FFN_{(*)}$ branches with attention pooling, concatenated, and projected by FFN_{emb}, before being passed to the culture classifier FFN_{cult}. Motion is intentionally not used to learn the cultural embedding: target motion is the output to be generated and is therefore unavailable at inference, while the one-second seed is absent at sequence start and too short to estimate culture reliably.

The overall loss for a training step is:

$$L(\theta) = \frac{1}{N} \sum_{d \in \mathcal{S}} \sum_{i=1}^{n_d} \ell_i^d + \lambda_p \, P(\theta) + \lambda_s \mathcal{L}_{\text{SupCon}}^{(\tau)},$$

$$\text{where} \quad N = \sum_{d \in \mathcal{S}} n_d, \quad \ell_i^d = -\log p\left(c_i^d \mid x_i^d; \theta\right),$$

where λ_p is a penalty weight, c_i^d denotes the cultural label of sample i from speaker d, and θ represents the model parameters.

The term $\mathcal{L}_{\text{SupCon}}^{(\tau)}$ represents the supervised contrastive loss operating on the learned cultural embeddings to enforce speaker invariance. By considering a multi-domain batch from 1 to N, it is formulated as:

$$\mathcal{L}_{\text{SupCon}}^{(\tau)} = \frac{1}{N} \sum_{i=1}^{N} \left[-\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\mathbf{z}_i^\top \mathbf{z}_a / \tau\right)} \right],$$

where τ is the temperature parameter, $A(i) = \{1, \ldots, N\} \setminus \{i\}$ represents the set of all sample indices within the current batch excluding the anchor instance itself, and $P(i) = \{p \in A(i) : c_p = c_i\}$ denotes the set of all positive sample indices sharing the same cultural label c as the anchor.

The Fishr penalty $P(\theta)$ quantifies discrepancies in gradient variance across domains:

$$P(\theta) = \frac{1}{|\mathcal{S}|} \sum_{d \in \mathcal{S}} \|v^d - \bar{v}\|^2$$

$$\text{with} \quad v^d = \frac{1}{n_d} \sum_{i=1}^{n_d} \left(g_i^d - \bar{g}^d\right)^2, \quad \bar{g}^d = \frac{1}{n_d} \sum_{i=1}^{n_d} g_i^d, \qquad g_i^d = \nabla_\theta \ell_i^d$$

where $\bar{v} = |\mathcal{S}|^{-1} \sum_{d \in \mathcal{S}} v^d$ denotes the mean variance across domains. We set $\lambda_p = 0$ for the first 500 updates to stabilize training, and then increase it to $\lambda_p = 1000$ to enforce invariant gradient statistics. The supervised contrastive weight is set to $\lambda_s = 0.2$, and the temperature parameter is $\tau = 0.07$.

For adversarial learning, we use the same architecture as in the Fishr model, with an additional speaker-classification head after a gradient reversal layer (GRL) [16]. The speaker classifier is trained to predict speaker identity, while the gradient reversal layer makes the shared encoder discard speaker-specific information. Let θ_l, θ_c, and θ_s denote the parameters of the shared encoder, culture classifier, and speaker classifier. The encoder is optimized with:

$$\mathcal{L}_{\text{tot}}(\theta_l, \theta_c, \theta_s) = \mathcal{L}_{\text{cult}}(\theta_l, \theta_c) - \lambda_2 \cdot \mathcal{L}_{\text{spk}}(\theta_l, \theta_s) + \lambda_s \mathcal{L}_{\text{SupCon}}^{(\tau)}$$

The adversarial weight λ_2 is gradually increased during training [16]. The losses $\mathcal{L}_{\text{cult}}$ and $\mathcal{L}_{\text{spk}}$ are the cross-entropy terms for culture and speaker classification.

3.3 Culture-aware Gesture Generation model

We propose ALaDiT, a diffusion-based architecture for gesture generation incorporating six modalities: mel-log spectrogram, onset strength, wav2vec embeddings, sentence embeddings, seed motion, and culture embeddings. By using motion embeddings instead of raw poses, our model reduces jitter and focuses

on key motion characteristics. ALaDiT can generate a 4-second motion sequence in under 14 ms, enabling real-time gesture synthesis.

As shown in Figure 2, each sample comprises 25 motion tokens, split into a 1-second seed ($X_0^{seed} \in \mathbb{R}^{T_p \times d}$, $T_p = 5$) and a four-second target ($X_0^{fin} \in \mathbb{R}^{T_r \times d}$, $T_r = 20$). Following Motion Diffusion Model (MDM) [50], Gaussian noise at timestep t is added to X_0^{fin}, yielding X_t^{fin}. Each sample also contains five seconds of aligned features: mel-log spectrogram (X^{mel}), onset strength (X^{on}), wav2vec embeddings (X^{w2v}), sentence embedding (X^{text}), and culture embedding (X^{cu}).

To construct the low-level audio context $X^{low} \in \mathbb{R}^{(T_p+T_r)\times d}$, we concatenate mel-log and onset features and project them to $d/2$ dimensions via FFN_1, while wav2vec features are projected to $d/2$ via FFN_2. We apply windowed attention pooling to the mel-log and onset features to align their temporal dimension (T_{mel}) to that of wav2vec (T_{w2v}). The projected features are then concatenated and further downsampled to $(T_p + T_r)$ tokens using another windowed attention pooling step, yielding X^{low}, such that each audio token corresponds to a motion token. This pooling operation functions like standard attention-pooling but restricts attention to inputs within a fixed window, efficiently downsampling the temporal dimension while preserving local context.

After applying Positional Encoding (PE) to X^{low}, we build the high-level context X^{high} by concatenating sentence and culture embeddings, followed by a projection FFN_3. The seed X_0^{seed} and noisy motion X_t^{fin} are concatenated, and then positional encoding is added to create X^{mot}. The diffusion timestep embedding e_t is computed using sinusoidal encodings followed by a projection FFN_4, and then concatenated with X^{mot} to yield:

$$X^{in} = \mathsf{concat}\Big(\mathbf{e}_t,\ X_{mot}\Big), \quad X^{in} \in \mathbb{R}^{(1+T_p+T_r)\times d}$$

A 10-layer hierarchical Transformer then applies: self-attention to X^{in}, cross-attention to X^{low}, and finally AdaIN conditioning [26] with X^{high}, producing $X^{out} \in \mathbb{R}^{(1+T_p+T_r)\times d}$

A residual connection and layer normalization follow each attention layer. After processing, the first six tokens (e_t and X_0^{seed}) are discarded. The rest are projected by FFN_5 to form the denoised motion $\hat{X}_0^{fin}$. The reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \mathbb{E}_{\substack{X_0 \sim p(X_0|(X^{low}, X^{high})) \\ t\sim[1,T]}} \left[\mathcal{H}\left(X_0^{fin} - \hat{X}_0^{fin}\right)\right]$$

where $\mathcal{H}$ is the Huber Loss [27]. To align the generated motion with both low-level audio and high-level textual/cultural contexts, we project X^{out} and X^{low} into a common space of dimension d using windowed attention pooling, resulting in the embeddings z^o and z^l, while X^{high} remains in $\mathbb{R}^d$. We then define the cosine-alignment losses:

$$\mathcal{L}_{low} = \mathbb{E}[\,1 - \cos(z_o, z_l)],\ \mathcal{L}_{high} = \mathbb{E}[\,1 - \cos(z_o, X^{high})]$$

To avoid trivial solutions, we further incorporate a contrastive loss [45] to pull together matching pairs and separate non-matching ones.

We apply this loss separately for the low-level ($\mathcal{L}^{\text{low}}_{\text{contrastive}}$) and the high-level ($\mathcal{L}^{\text{high}}_{\text{contrastive}}$) contexts, and define the final contrastive loss $\mathcal{L}_{cont}$ as their average.

Following [54], we add a classification head FFN_6 on $\hat{X}_0^{fin}$ to predict the culture label $\hat{c}$, and include a cross-entropy loss $\mathcal{L}_{\text{cult}}$.

The final loss is a weighted sum of these loss components:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{cu}\,\mathcal{L}_{cult} + \lambda_l\,\mathcal{L}_{low} + \lambda_h\,\mathcal{L}_{high} + \lambda_c\,\mathcal{L}_{cont}$$

where λ_{cu}, λ_l, and λ_h are set to 0.1, and $\lambda_c = 0.01$.

4 Experimental Setup

ALaDiT variants. We train three main ALaDiT variants differing only in cultural conditioning:(i) *ALaDiT NC*, where culture must be inferred implicitly from the audio/text while an auxiliary culture-classification head is used during training; (ii) *ALaDiT FI*, which is conditioned on speaker-independent cultural embeddings learned with Fishr regularization; and (iii) *ALaDiT ADV*, which is conditioned on embeddings learned via adversarial domain generalization. These variants isolate the effect of explicit culture embeddings and different domain generalization strategies. To isolate possible confounds, we also evaluate three additional ALaDiT ablations: *OneHot*, which conditions on the discrete group label after projecting it with a two-layer MLP and injecting it through the same conditioning pathway; *NoDG*, which uses the same audio/text embedding architecture as FI but removes speaker-domain regularization; and *NoAlign*, which uses Fishr embeddings but removes ALaDiT's multimodal alignment losses.

Baselines comparison. We adapt Motion Diffusion Model (MDM) [50] and DiffuseStyleGesture+ (DSG+) [57] using the same TED4C-L features as ALaDiT whenever supported by the architecture, while matching optimization settings and diffusion steps ($T = 50$). We train NC, FI, and ADV variants for each baseline. For DSG+, we additionally evaluate *DSG+/FI+Align*, which adds an explicit multimodal alignment loss to test whether DSG+ can better exploit Fishr embeddings when alignment is provided. Details on how culture embeddings are injected are provided in the supplementary material.

Objective evaluation. We evaluate models on the test split using (i) Fréchet Gesture Distance (FGD) [58], (ii) Semantic Relevance Gesture Recall (SRGR) [34], (iii) Beat Alignment Score (BAS) [33], and (iv) Diversity (mean ℓ_1 distance between randomly sampled generated motion codebook sequences). Metrics are averaged over 10 evaluations, each computed on a random subsample of 3000 test instances drawn without replacement; statistical significance is assessed with paired t-tests ($p < 0.01$). To quantify cultural consistency, we apply a speaker-disjoint motion-based cultural classifier to generated gestures and report weighted F1-score on Cultural Expressivity (CE F1) (see supplementary). SRGR is computed on temporally aligned motion pairs using the original PCK

Model	FGD ↓	CE F1(%) ↑	BAS(%) ↑	SRGR(%) ↑	Diversity ↑
(§) OneHot	1.63 ± 0.23	$43.73 \pm 1.13^{*}$	22.51 ± 0.17	67.63 ± 0.25	$\mathbf{111.79 \pm 0.58}^{\ddagger,\dagger,\circ}$
(¶) NoDG	1.56 ± 0.22	43.18 ± 1.20	22.51 ± 0.23	$67.76 \pm 0.23^{*,\S}$	$111.60 \pm 0.71^{\ddagger,\dagger,\circ}$
(○) NoAlign	$1.36 \pm 0.16^{\ddagger,*,\S,\P}$	43.37 ± 0.91	22.58 ± 0.17	$\mathbf{68.17 \pm 0.23}^{\ddagger,*,\S,\P}$	$111.10 \pm 0.77^{\ddagger,\dagger}$
(‡) NC	1.60 ± 0.18	43.41 ± 1.10	22.51 ± 0.15	67.72 ± 0.23	109.50 ± 0.68
(∗) ADV	1.53 ± 0.17	42.71 ± 0.95	22.45 ± 0.17	67.57 ± 0.27	$111.75 \pm 0.71^{\ddagger,\dagger,\circ}$
(†) FI	$\mathbf{1.03 \pm 0.15}^{\ddagger,*,\S,\P,\circ}$	$\mathbf{44.61 \pm 0.95}^{\ddagger,*,\S,\P,\circ}$	$\mathbf{22.63 \pm 0.22}$	$68.09 \pm 0.25^{\ddagger,*,\S,\P}$	$110.27 \pm 0.70^{\ddagger}$

Table 2: ALaDiT ablations, reported as mean ± std over 10 test runs. FI uses Fishr regularization; ADV uses adversarial learning; NC removes explicit cultural conditioning; OneHot uses one hot cultural labels for conditioning; NoDG uses the same audio/text embedding architecture as FI, but without speaker-domain regularization; NoAlign removes ALaDiT alignment losses while using Fishr regularization. Bold marks the best value within the ALaDiT block. Superscripts denote values significantly better than the rows indicated by the corresponding symbols under paired two-sided t-tests ($p < 0.01$).

formulation with $\delta = 0.05$, BAS uses $\sigma = 3$, and FGD embeddings are extracted using the pretrained VQVAE.

Qualitative comparisons. To visualize culture-specific differences, we translate the sentence *"This example helps explain the idea of cultural styles"* into each dataset language (Google Translate) and synthesize speech with Bark. Using the same fixed ground-truth seed pose, we generate motions for each culture and compare ALaDiT-FI, ALaDiT-ADV, and ALaDiT-NC.

User study. We conduct a user study with $N = 20$ participants. For each culture and condition (Real, NC, FI, ADV), we generate two 30-second clips (32 clips total). Clips are presented in random order with original audio and English subtitles, and each participant rates all clips. After each clip, participants evaluate six questions from [19] (speech coherence, appropriateness, fluency, timing, amount of gesticulation, naturalness) plus an additional question on cultural fit using an 11-point Likert scale (0–10). Before the study, participants viewed two short TED Talk examples per culture to familiarize themselves with the evaluated gesturing styles.

5 Results

5.1 Quantitative analysis

We evaluate SICAGE through both ablation studies and comparisons with strong diffusion-based gesture generation baselines.

Effect of speaker-independent cultural embeddings. Table 2 reports an ablation study of the proposed cultural representation within the same ALaDiT generator. Introducing explicit cultural representations improves most metrics. Fishr conditioning (FI) achieves the best motion realism (lowest FGD) and the highest cultural classification accuracy (CE F1), indicating that speaker-invariant embeddings learned with Fishr capture culturally relevant motion patterns more effectively. Adversarial learning (ADV) remains closer to NC on most metrics, except for Diversity, suggesting that the adversarial objective provides

Model	FGD ↓	CE F1(%) ↑	BAS(%) ↑	SRGR(%) ↑	Diversity ↑
		DSG+ variants and ablations			
(△) DSG+/NC	$2.76 \pm 0.31^{\diamond,\bullet}$	$41.51 \pm 0.78^{\star}$	22.48 ± 0.11	$68.17 \pm 0.24^{\diamond,\star}$	$108.85 \pm 0.63^{\circ}$
(⋄) DSG+/ADV	4.81 ± 0.43	$\mathbf{42.21 \pm 0.96}^{\triangle,\bullet,\star}$	22.58 ± 0.19	$66.78 \pm 0.32^{\star}$	107.78 ± 0.71
(•) DSG+/FI	4.89 ± 0.40	40.67 ± 1.31	22.48 ± 0.15	$\mathbf{68.46 \pm 0.29}^{\triangle,\diamond,\star}$	$108.85 \pm 1.14^{\circ}$
(⋆) DSG+/FI+Align	$\mathbf{2.52 \pm 0.21}^{\diamond,\bullet}$	39.80 ± 0.56	$\mathbf{22.67 \pm 0.16}^{\triangle}$	65.17 ± 0.24	$\mathbf{111.13 \pm 1.07}^{\triangle,\diamond,\bullet}$
		MDM variants			
(§) MDM/NC	15.58 ± 1.43	38.57 ± 0.80	22.52 ± 0.14	51.62 ± 0.22	$107.62 \pm 1.08^{\P}$
(¶) MDM/ADV	$13.67 \pm 1.17^{\S}$	38.92 ± 0.93	22.59 ± 0.13	$\mathbf{52.25 \pm 0.25}^{\S,\circ}$	105.92 ± 0.84
(○) MDM/FI	$\mathbf{7.59 \pm 0.59}^{\S,\P}$	$\mathbf{47.09 \pm 0.79}^{\S,\P}$	$\mathbf{22.59 \pm 0.17}$	51.86 ± 0.24	$\mathbf{109.37 \pm 0.74}^{\S,\P}$
		ALaDiT variants			
(‡) ALaDiT/NC	1.60 ± 0.18	43.41 ± 1.10	22.51 ± 0.15	67.72 ± 0.23	109.50 ± 0.68
(∗) ALaDiT/ADV	1.53 ± 0.17	42.71 ± 0.95	22.45 ± 0.17	67.57 ± 0.27	$\mathbf{111.75 \pm 0.71}^{\ddagger,\dagger}$
(†) **ALaDiT/FI (ours)**	$\mathbf{1.03 \pm 0.15}^{\ddagger,*}$	$\mathbf{44.61 \pm 0.95}^{\ddagger,*}$	$\mathbf{22.63 \pm 0.22}$	$\mathbf{68.09 \pm 0.25}^{\ddagger,*}$	$110.27 \pm 0.70^{\ddagger}$

Table 3: Comparison across DSG+, MDM, and ALaDiT variants (FI, ADV, NC), reported as mean ± std over 10 matched test runs. Bold marks the best value within each architecture family, not necessarily the best global value. Superscripts denote significantly better results than the marked row within the same family under paired two-sided t-tests ($p < 0.01$).

weaker perceptual and objective gains in this setting. The OneHot ablation shows that a discrete cultural label alone is not sufficient: although it can increase Diversity, it does not match FI in FGD or CE F1. NoDG further shows that using the same audio/text embedding architecture without speaker-domain regularization is also insufficient. Finally, NoAlign confirms that ALaDiT's alignment losses help exploit the learned embeddings, but do not by themselves explain the FI gains: removing alignment worsens FGD and CE F1 despite a small increase in SRGR. Overall, the ablations show that the strongest results come from combining Fishr-based speaker-independent embeddings with an architecture able to align multimodal conditioning.

Comparison with existing models. Table 3 compares ALaDiT with DiffuseStyleGesture+ (DSG+) and Motion Diffusion Model (MDM), each trained with the same features and diffusion settings. Overall, ALaDiT provides the strongest trade-off across the evaluated metrics. In particular, ALaDiT achieves substantially lower FGD, indicating closer similarity to real motion distributions, while maintaining competitive or superior CE F1, BAS, Diversity, and SRGR scores. Introducing cultural embeddings generally improves performance for the baseline models. For MDM, both FI and ADV improve several metrics, particularly FGD and CE F1, with FI consistently outperforming ADV, following the same trend observed for ALaDiT. In contrast, improvements for DSG+ are more limited, suggesting that cultural embeddings are most effective when combined with architectures that explicitly model multimodal alignment. This is further supported by the DSG+/FI+Align ablation, where adding explicit alignment substantially improves FGD, Diversity, and BAS over DSG+/FI, although performance still remains below ALaDiT/FI in terms of motion realism and cultural consistency. Taken together, these results show that (i) speaker-independent cultural embeddings significantly improve gesture generation when effectively integrated within the generator architecture, (ii) Fishr provides the strongest cul-

tural representation among the evaluated domain-generalization strategies, and (iii) the proposed ALaDiT architecture achieves the best overall performance when combined with these embeddings.

5.2 Qualitative analysis

We visually compare ALaDiT-FI and ALaDiT-ADV with the no-culture variant (NC) under matched seed pose and semantic content ("This example helps explain the idea of cultural styles"), translated into each language and synthesized with Bark. Figure 4 shows Japanese and Italian examples, where each column summarizes one second of motion at 15 fps, while the final column shows the remaining frames of the sequence; darker regions indicate joint locations that are occupied more frequently within that second. For Japanese, all models produce relatively compact upper-body motion throughout the utterance. However, NC exhibits higher spatial dispersion at the wrists and elbows, particularly toward the end of the clip when speech activity decreases. This residual motion during silence suggests lower motion realism, consistent with the quantitative results. In contrast, FI and ADV show more stable motion trajectories and terminate gestures more cleanly as the speech ends. For Italian, FI, and ADV generate broader spatial gestures and larger arm excursions than NC, with movement sustained across multiple seconds. In contrast, NC remains more constrained and shows a limited range of motion. Among the culture-conditioned variants, FI distributes motion more globally across the upper body, whereas ADV concentrates motion more strongly on the arms (notably the left arm in this example). Overall, these qualitative trends support the benefit of explicit culture embeddings in producing motion that is both better aligned to speech dynamics and more consistent with culture-specific patterns captured by our motion-based classifier. Additional qualitative examples and rendered videos for all cultures, including comparisons to ground-truth motion, are provided in the supplementary material.

5.3 User study

The user study results shown in Figure 5, conducted with 20 participants from different cultural backgrounds, further support the findings of the objective and qualitative analyses. Overall differences between generated models are relatively small, which is expected given the subjective nature of the task and the limited number of raters. Nevertheless, FI obtains the highest average score among generated models and is significantly preferred over ADV overall (6.06 vs. 5.65, $p = 0.033$). FI is also significantly preferred over NC on Cultural Match (6.16 vs. 5.81, $p = 0.038$), suggesting that Fishr-based cultural embeddings produce perceptible improvements in culture-associated gesture style. NC and ADV receive comparable ratings, indicating that the adversarial domain-generalization objective yields less consistent perceptual gains. Real motion remains the highest-rated condition, confirming that generated gestures are still perceptually distinguishable from ground-truth motion. These findings are consistent with the

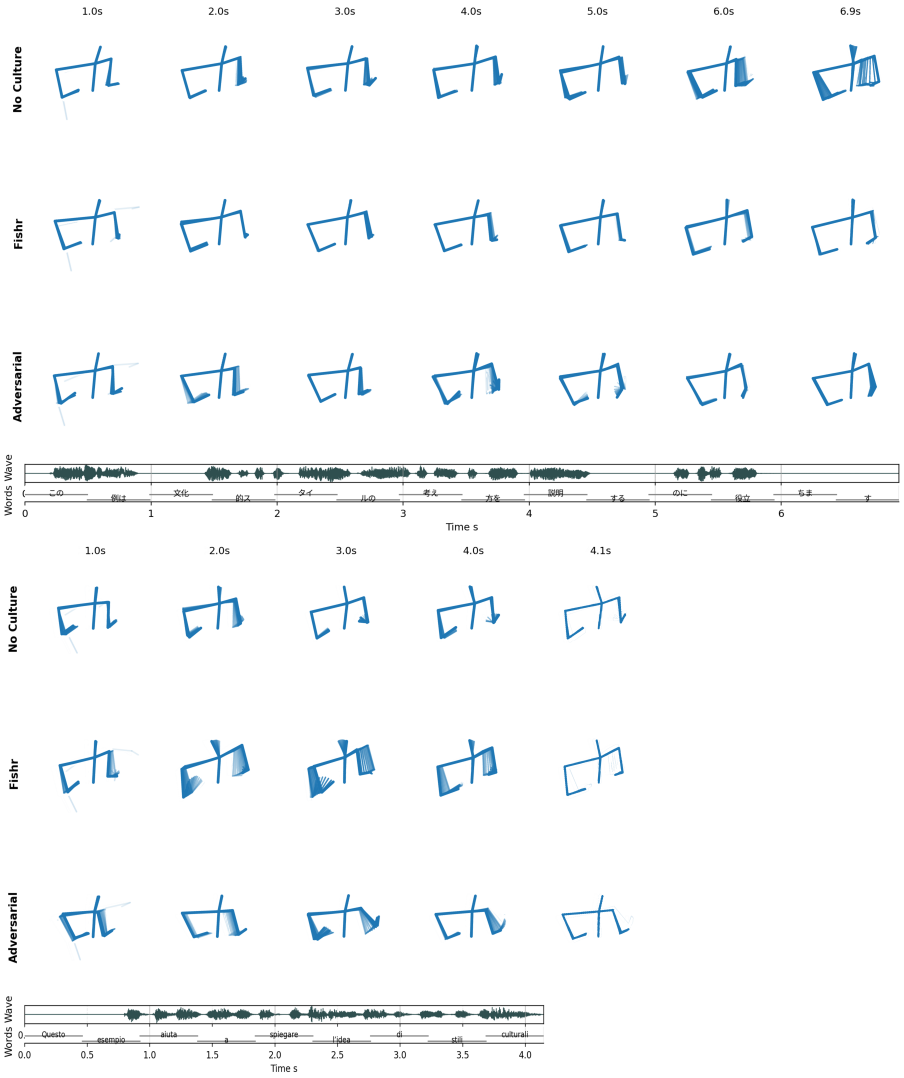

Fig. 4: Motion generated by the No Culture, Fishr and Adversarial models for Japanese (top rows) and Italian (bottom rows) cultures given the sentence "This example helps explain the idea of cultural styles". Each image represents one second of motion, except the last one, which represents the last frames.

objective results, where FI provides the strongest balance between motion realism and cultural consistency. We also observe variation across cultures; for Turkish samples, NC is slightly preferred over FI and ADV, but this difference is not statistically significant. A more detailed per-culture analysis is provided in the supplementary material.

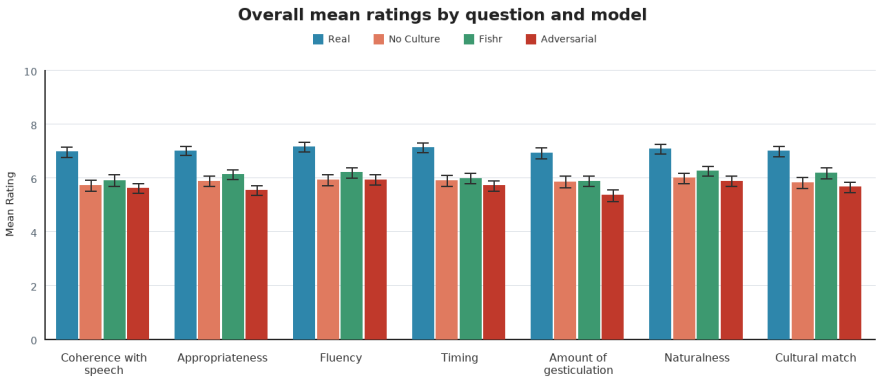

Fig. 5: User-study scores across cultures and conditions. Bars show mean ratings across participants (N = 20); error bars show standard deviation.

6 Conclusion

We introduced SICAGE, a modular framework for culture-aware co-speech gesture generation that learns speaker-independent cultural embeddings by treating speakers as domains. We also introduced TED4C-L, a large-scale multicultural dataset with speaker-disjoint splits for evaluating cultural generalization. Experiments show that explicitly modeling culture improves gesture synthesis. Fishr-based embeddings provide the most consistent gains in motion realism, cultural consistency, and speech–gesture alignment. Combined with these embeddings, our ALaDiT diffusion model achieves state-of-the-art performance compared to recent diffusion-based baselines. Quantitative results, qualitative analysis, and a user study indicate that differences between models are perceptible to human observers, with the Fishr-based model receiving the highest ratings in most evaluation criteria. Future work should consider finer-grained cultural annotations, reliable hand/finger tracking in in-the-wild videos, and representation learning methods that better disentangle culture-associated regularities from language, prosody, semantics, and coarse country-level grouping effects. User studies involving participants more familiar with the evaluated cultures could also provide more reliable perceptual assessments.

References

1. Ali, G., Lee, M., Hwang, J.I.: Automatic text-to-gesture rule generation for embodied conversational agents. Computer Animation and Virtual Worlds **31**(4-5), e1944 (2020)
2. Alibali, M.W., Kita, S., Young, A.J.: Gesture and the process of speech production: We think, therefore we gesture. Language and Cognitive Processes **15**(6), 593–613 (2000)
3. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. ACM Transactions on Graphics (TOG) **42**(4), 1–18 (2023)
4. Archer, D.: Unspoken diversity: Cultural differences in gestures. Qualitative sociology **20**, 79–105 (1997)
5. Beyan, C., Vinciarelli, A., Bue, A.D.: Co-located human–human interaction analysis using nonverbal cues: A survey. ACM Comput. Surv. **56**(5) (Nov 2023)
6. Bremner, P., Pipe, A.G., Melhuish, C., Fraser, M., Subramanian, S.: The effects of robot-performed co-verbal gesture on listener behaviour. In: 2011 11th IEEE-RAS International Conference on Humanoid Robots. pp. 458–465. IEEE (2011)
7. Buisine, S., Abrilian, S., Martin, J.C.: Evaluation of multimodal behaviour of embodied agents: Cooperation between speech and gestures. From brows to trust: evaluating embodied conversational agents pp. 217–238 (2004)
8. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
9. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: Beat: the behavior expression animation toolkit. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 477–486 (2001)
10. Chemburkar, A., Lu, S., Feng, A.: Discrete diffusion for co-speech gesture synthesis. In: Companion Publication of the 25th International Conference on Multimodal Interaction. pp. 186–192 (2023)
11. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. Interspeech 2021 (2021)
12. Contributors, M.: Openmmlab pose estimation toolbox and benchmark (2020)
13. Du, Y., Xu, J., Xiong, H., Qiu, Q., Zhen, X., Snoek, C.G., Shao, L.: Learning to learn with variational information bottleneck for domain generalization. In: European conference on computer vision. pp. 200–216. Springer (2020)
14. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic bert sentence embedding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2022)
15. Ferstl, Y., Neff, M., McDonnell, R.: Expressgesture: Expressive gesture generation from speech through database matching. Computer Animation and Virtual Worlds **32**(3-4), e2016 (2021)
16. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
17. Gao, N., Zhao, Z., Zeng, Z., Zhang, S., Weng, D., Bao, Y.: Gesgpt: Speech gesture synthesis with text parsing from chatgpt. IEEE Robotics and Automation Letters (2024)
18. Gjaci, A., Oneto, L., Recchiuto, C.T., Sgorbissa, A.: Culture awareness in intelligent systems. In: Proceedings of the 9th Italian Workshop on Artificial Intelligence

and Robotics, AIRO 2022. CEUR Workshop Proceedings, vol. 3417, pp. 25–30. CEUR-WS.org (2022)
19. Gjaci, A., Recchiuto, C.T., Sgorbissa, A.: Towards culture-aware co-speech gestures for social robots. International Journal of Social Robotics **14**(6), 1493–1506 (2022)
20. Gjaci, A., Recchiuto, C.T., Sgorbissa, A.: Labeling sentences with symbolic and deictic gestures via semantic similarity. In: 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN). pp. 477–484. IEEE (2024)
21. Gjaci, A., Schmuck, V., Sgorbissa, A., Celiktutan, O.: Exploring cultural cues in multimodal interactions. In: 2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). pp. 306–313. IEEE (2024)
22. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
23. Hensel, L.B., Yongsatianchot, N., Torshizi, P., Minucci, E., Marsella, S.: Large language models in textual analysis for gesture selection. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 378–387 (2023)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
25. Huang, J., Guan, D., Xiao, A., Lu, S.: Fsdr: Frequency space domain randomization for domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6891–6902 (2021)
26. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
27. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics: Methodology and distribution, pp. 492–518. Springer (1992)
28. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T.S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
29. Kita, S.: Cross-cultural variation of speech-accompanying gesture: A review. Language and cognitive processes **24**(2), 145–167 (2009)
30. Krämer, N.C., Simons, N., Kopp, S.: The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry. In: Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7. pp. 238–251. Springer (2007)
31. Krauss, R.M., Chen, Y., Chawla, P.: Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In: Advances in experimental social psychology, vol. 28, pp. 389–450. Elsevier (1996)
32. Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S.S., Sheikh, Y.: Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 763–772 (2019)
33. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13401–13412 (2021)
34. Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversa-

tional gestures synthesis. In: European conference on computer vision. pp. 612–630. Springer (2022)
35. Ma, W., Datta, S., Wang, L., Vosoughi, S.: EnCBP: A new benchmark dataset for finer-grained cultural background prediction in English. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 2811–2823. Association for Computational Linguistics, Dublin, Ireland (May 2022)
36. Mao, X., Jiang, Z., Wang, Q., Fu, C., Zhang, J., Wu, J., Wang, Y., Wang, C., Li, W., Chi, M.: Mdt-a2g: Exploring masked diffusion transformers for co-speech gesture generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3266–3274 (2024)
37. Martin, J.C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C.: Multimodal complex emotions: Gesture expressivity and blended facial expressions. International Journal of Humanoid Robotics **3**(03), 269–291 (2006)
38. Matsumoto, D., Hwang, H.C.: Cultural similarities and differences in emblematic gestures. Journal of Nonverbal Behavior **37**, 1–27 (2013)
39. Matsuura, T., Harada, T.: Domain generalization using a mixture of multiple latent domains. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11749–11756 (2020)
40. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago press (1992)
41. Ng-Thow-Hing, V., Luo, P., Okita, S.: Synchronized gesture and speech production for humanoid robots. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4617–4624. IEEE (2010)
42. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
43. Puterman, M.L.: Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons (2014)
44. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12556–12565 (2020)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
46. Rame, A., Dancette, C., Cord, M.: Fishr: Invariant gradient variances for out-of-distribution generalization. In: International Conference on Machine Learning. pp. 18347–18377. PMLR (2022)
47. Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., Joublin, F.: To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. International Journal of Social Robotics **5**, 313–323 (2013)
48. Shen, Y., Johal, W.: Ted-culture: culturally inclusive co-speech gesture generation for embodied social agents. Frontiers in Robotics and AI **12**, 1546765 (2025)
49. Spencer-Oatey, H., Franklin, P.: What is culture. A compilation of quotations. GlobalPAD Core Concepts **1**(22), 1–21 (2012)
50. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023)

51. Tuyen, N.T.V., Georgescu, A.L., Di Giulio, I., Celiktutan, O.: A multimodal dataset for robot learning to imitate social human-human interaction. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. pp. 238–242 (2023)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
53. Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P.S.: Generalizing to unseen domains: A survey on domain generalization. IEEE transactions on knowledge and data engineering **35**(8), 8052–8072 (2022)
54. Wu, J., Chen, S., Gan, S., Li, W., Yang, C., Sun, L.: Cultural self-adaptive multimodal gesture generation based on multiple culture gesture dataset. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3538–3549 (2023)
55. Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Cheng, M., Xiao, L.: Diffusestylegesture: stylized audio-driven co-speech gesture generation with diffusion models. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 5860–5868 (2023)
56. Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Zhuang, H.: Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2321–2330 (2023)
57. Yang, S., Xue, H., Zhang, Z., Li, M., Wu, Z., Wu, X., Xu, S., Dai, Z.: The diffusestylegesture+ entry to the genea challenge 2023. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 779–785 (2023)
58. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
59. Yoon, Y., Ko, W.R., Jang, M., Lee, J., Kim, J., Lee, G.: Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 4303–4309 (2019)
60. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019)
61. Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., Yu, L.: Taming diffusion models for audio-driven co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10544–10553 (2023)