2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Text-Conditioned Beat Gesture Generation for a Social Robot via a Conditional Variational Autoencoder

Alejandro Climent Peñalver[1], Enrique Fernández-Rodicio[1] and Álvaro Castro-González[1]

Abstract— **Conversation can benefit from small rhythmic gestures that track prosody, reinforce structure, and help to keep attention. However, many robots used in human–robot interaction still rely on fixed templates or clip libraries that scale poorly to open-domain interactions; moreover, embedded platforms impose tight limits on motion range, speeds, and timing. Consequently, gesture generation methods must be lightweight, stable, and easy to integrate. To address this need, this work presents a lightweight gesture-generation model that generates in real time beat gestures based on the transcription of the robot's speech. First, a Conditional Variational Autoencoder (CVAE) conditioned on sentence-level BERT embeddings is trained on 2D pose–text pairs to produce upper-body pose sequences. Next, a geometry-based retargeting algorithm deterministically maps those poses to the robot's joints while enforcing kinematic limits. Finally, the joint sequence is converted into a pseudo-state machine and triggered in lockstep with the utterance. The results obtained show that the system achieves smooth, text-conditioned beat gestures with solid fidelity and temporal diversity, and demonstrates real-time performance when integrated on a social robot.**

I. INTRODUCTION

Human social interaction depends on more than words. Nonverbal cues such as posture, gaze, facial expressions, and especially beat gestures aligned with prosody work together with speech to structure discourse, highlight salient information, and sustain attention [1]. As a result, robots that speak fluently yet remain physically static are often perceived as less expressive and less natural [2].

Bringing this capability to social robots is challenging. In interactive, open-ended settings, a gesture generation module must run continuously during dialogue under tight timing constraints. Another challenge is that many robotic platforms operate with few degrees of freedom and modest onboard resources [3], [4]. Libraries of handcrafted behaviours therefore persist; while effective for fixed routines such as greetings, they scale poorly to beat gesticulation, which consists of movements driven by prosody and not iconic; in this setting, variety and responsiveness are essential [5].

Recent neural generative models are promising, yet two practical gaps hinder integration on robots. First, there is a trade-off between computational power and latency. Pipelines that couple text and motion with small language models are feasible but often reduce pose fidelity and alignment, whereas variants relying on larger models or multi-stage reasoning introduce latencies incompatible with on-board execution [6]. Many recent gesture and motion generation methods are trained and evaluated on workstations with high-end GPUs[7], [8]. This can make deploying these solutions on social robots difficult. Second, data remains a bottleneck. Public HRI corpora are modest in size, and results from the annual GENEA Challenge, the community benchmark for speech-aligned gesture generation, show that current systems still fall short of human naturalness and interlocutor alignment [9].

Beyond computational power limitations, robots impose control and kinematic constraints. Coordinating whole-body movement in real time can be demanding if humanlike poses have to be mapped to a different number of joints while enforcing limits on position, velocity, and acceleration [10], [11].

Within this context, this paper addresses the generation of beat gestures under strict execution constraints on the social robot Mini 3. The pipeline uses a lightweight conditional variational model to produce sequences of upper-body poses conditioned on the transcription of the robot's speech. Those poses are mapped deterministically to a platform with a few degrees of freedom while respecting kinematic and timing limits. The design prioritises training stability, low inference cost, and clean integration into a standard HRI stack so that gestures are scheduled and executed in sync with speech.

The rest of the manuscript is organised as follows. Section II surveys the most influential work on gesture generation, tracing the field from early rule-based systems to recent neural approaches. Section III introduces the generation model used in this study, together with the dataset employed for training and analysis. Section IV describes the social robot used in our experiments and explains the end-to-end pipeline from utterance to execution. Section V presents the evaluation protocol and the results obtained. Finally, section VI closes with conclusions.

II. RELATED WORK

Research on automatic gesture generation has progressed from early hand-crafted approaches to contemporary deep-learning methods. This section provides an analysis of different approaches for gesture generation.

Early systems for nonverbal behaviour generation in robots and embodied agents relied on explicit rules and scripted behaviours. Cassell *et al.*'s BEAT [5] mapped textual analyses (e.g., topics or action verbs) to prototypical gestures through if–then rules. In prior work, Cassell *et al.*'s *Animated Conversation* [12] planned synchronised prototype gestures from dialogue via predefined mappings. Around the same period, Thórisson (1996) [13] introduced a real-time

[1]Robotics Lab, Universidad Carlos III de Madrid, Av. de la Universidad 30, Leganés, Madrid 28911, Spain. climentdeibi@gmail.com, {enrifern,acgonzal}@ing.uc3m.es

multimodal architecture that coordinated voice, gaze, and gesture. These foundations supported clear communicative intentions but were limited by finite libraries and hand-coded rules, especially for low semantic-content beat gestures that accompany prosody rather than lexical meaning [14].

With broader motion capture becoming available, learning-based selection over libraries emerged. Levine *et al.* [15] trained Hidden Markov Models to select gesture segments conditioned on prosodic features. Follow-up work by Levine *et al.* [16] combined a Conditional Random Field to predict coarse kinematic patterns with a Markov Decision Process to choose specific clips. These methods, which generalised beyond rules but still reused snippets, were prone to overfitting with limited data and struggled outside their design domain.

Deep learning shifted the field toward data-driven sequence models. Long Short-Term Memory networks by Hochreiter and Schmidhuber [17] enabled frame-wise synthesis with temporal coherence. Hasegawa *et al.* [18] mapped MFCC audio features to 3D motion using a bidirectional LSTM, yielding smooth, speech-synchronised sequences often refined with temporal post-filters. Yoon *et al.* [19] proposed a GRU-based encoder-decoder architecture with a soft attention mechanism that allows the decoder to focus on keywords in the sentence. Kucherenko *et al.* [20] later predicted latent gesture codes via a denoising autoencoder (*SpeechE*), reducing noise and removing the need for external smoothing in user studies. Despite these gains, recurrent models commonly face limited diversity tied to the corpus, weaker capture of semantic content under audio-only conditioning, and accumulating autoregressive error over long horizons [21].

Variational methods model variability explicitly through continuous latent variables. Doersch [22] provided a tutorial overview of Variational Autoencoders (VAEs), and Dai and Wipf [23] analyse failure modes and remedies. On the other hand, Li et al.'s *Audio2Gestures* [24] uses a conditional VAE with regularisation that encourages smooth yet diverse motion, including diversity-promoting and bidirectional consistency losses. Ahuja et al.'s *Mix-StAGE* [25] incorporates speaker-style embeddings for controllable synthesis. Ghorbani et al.'s *ZeroEGGS* [26] learns a latent style space from examples, enabling zero-shot style control and diverse outputs. Taken together, these works indicate that regularised latent-variable models provide a principled way to couple conditioning signals (speech or text) with motion variability in a single framework, yielding stable training and adjustable diversity—qualities that are attractive for real-time systems with limited compute.

Adversarial learning offers a complementary route. Goodfellow *et al.* [27] introduced Generative Adversarial Networks; in gesture generation, Yu and Tapus [11] proposed SRG3, a GAN-based system, with an auxiliary L_1 loss to encourage global structure, and Ferstl *et al.* [28] formulated a multi-objective adversarial approach that also penalises pose error to improve robustness and perceived quality. While effective, adversarial training can be unstable and susceptible to mode collapse without additional guidance.

Attention-based models capture longer-range dependencies efficiently. Bhattacharya *et al.*'s *Text2Gestures* [29] encoded emotion-annotated text and decoded skeleton poses, aligning utterance segments with motion. Xie *et al.*'s *ReCoM* [30] embedded a transformer in a recurrent loop with iterative reconstruction and temporal smoothing to reduce abrupt transitions while preserving variety. These approaches are powerful but often demand substantial data and computation; without strong temporal inductive biases, they may exhibit subtle coherence issues [14].

Diffusion models have recently gained traction: training adds noise progressively and learns to invert that process. Mughal *et al.*'s *ConvoFusion* [31] supports text, main-speaker and interlocutor audio, and speaker style, with word-excitation guidance to emphasise micro-gestures at salient words. Favali *et al.*'s *TAG2G* [32] reports stable, accurate synthesis and addresses diversity collapse seen in some adversarial setups, at the cost of multi-step inference, mitigated by accelerated sampling.

Recent work explores large language models to inject semantics into gesture generation. Pang *et al.*'s *LLM-Gesticulator* [33] combines audio and text, leveraging internal LLM representations for synchronised full-body animation and prompt-based style control. To reduce computation, Galatolo and Winkle [6] used small language models to produce high-level symbolic annotations (e.g., intent, emphasis) that downstream motion modules realise. Editing frameworks further stress the value of gestures with clear communicative function, especially for interactive robots [34].

Gesture generation has advanced in recent years, yet several gaps still block deployment on social robots. Systems must run in real time under tight compute, learn robustly without collapsing while preserving nonrepetitive variation, and execute in sync with speech within kinematic limits. Within these constraints, a text-conditioned CVAE fits well because explicit latent regularisation supports stable learning [22], [23], single-pass decoding keeps CPU latency low compared with multi-step diffusion [32], and sampling in the latent space enables controllable variation [24]. In contrast, adversarial approaches often require delicate tuning and can suffer from mode collapse [27], [28], [11], which complicates robust on-board deployment.

III. Gesture Generator Module

This section describes the basis of our gesture generation approach, including the model selected, the dataset used to train it and the preprocessing applied to this dataset.

A. Model

We have used a Conditional Variational Autoencoder for our gesture generation module. The architecture of this model can be seen in Fig. 1. Given the transcript of the robot's speech, the model generates a sequence of 2D upper-body poses represented as time-ordered joint coordinates.

Building on this setup, the model follows a simple, efficient pipeline. First, each utterance of the speech is tokenised

and encoded with a pretrained BERT model to obtain a fixed-length sentence embedding e that summarises semantics. That embedding conditions both the CVAE encoder and the decoder, and it is concatenated with a latent vector z so that linguistic information guides the internal representation and the generated motion. The architecture employs single-layer LSTMs because gesture poses form time series, and this choice captures local temporal dynamics with modest compute, keeps decoding latency low on the target hardware, and limits overfitting by avoiding unnecessary depth. Moreover, the latent space has dimension d=128.

During training, the encoder reads the ground-truth pose sequence together with e and outputs the parameters (μ, σ) of a Gaussian approximate posterior. A latent sample is then obtained through the standard reparameterization in Eq. (1).

$$z = \mu + \sigma \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{1}$$

The decoder generates poses autoregressively. At each time step, it receives the concatenated vector $[e \,\|\, z]$ and the previous pose, while the first step is initialised with a learned start token.

Training uses loss terms with complementary roles that balance fidelity, stability, and variety. Let the dataset contain N sequences, with the i-th sequence of length T_i. Denote by $x_{i,t} \in \mathbb{R}^{2J}$ the ground-truth 2D pose with J=8 keypoints at time t, and by $\hat{x}_{i,t}$ the prediction.

The reconstruction loss preserves framewise accuracy as in Eq. (2).

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \left\| x_{i,t} - \hat{x}_{i,t} \right\|_2^2. \tag{2}$$

A KL divergence regularises the latent space by nudging the approximate posterior toward a standard normal [23], given by Eq. (3).

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_{j=1}^{d} \left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2 \right). \tag{3}$$

Because public talks often include stretches with little motion, a small-movement penalty discourages near-static outputs and helps preserve rhythmicity [24], as in Eq. (4).

$$\mathcal{L}_{\text{temp}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} \max\Big(0,\ \delta - \left\| \hat{x}_{i,t+1} - \hat{x}_{i,t} \right\|_2 \Big), \tag{4}$$

with a small threshold $\delta > 0$. In addition, a diversity term encourages reasonable temporal variation when the content allows it, following prior work on gesture synthesis and VAE regularisation [35], defined in Eq. (5).

$$\mathcal{L}_{\text{div}} = -\frac{1}{N} \sum_{i=1}^{N} \text{Var}\big(\hat{x}_{i,1:T_i}\big), \tag{5}$$

where $\text{Var}(\cdot)$ is the element-wise temporal variance averaged over the sequence.

The overall objective combines the above terms with scalar weights β, λ_{temp}, and λ_{div} as Eq. (6).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \beta\, \mathcal{L}_{\text{KL}} + \lambda_{\text{temp}}\, \mathcal{L}_{\text{temp}} + \lambda_{\text{div}}\, \mathcal{L}_{\text{div}}. \tag{6}$$

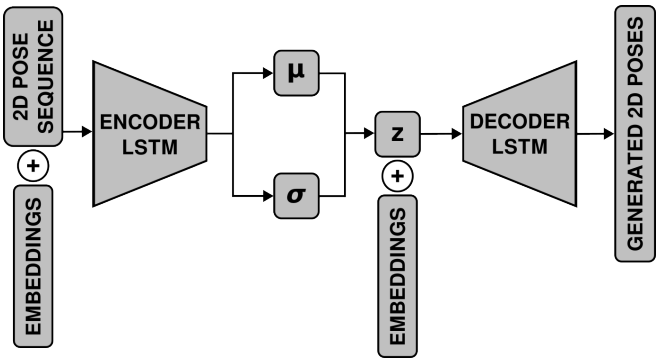

Fig. 1. Architecture of the conditional variational autoencoder with BERT conditioning.

To limit posterior collapse, the KL weight is increased gradually during the first epochs with an annealing schedule [36].

At inference time, the encoder is not used. A latent vector $z \sim \mathcal{N}(0, I)$ is sampled, and the pair $[e \,\|\, z]$ conditions the decoder at every step. The decoder then produces a sequence of 2D upper-body poses whose length follows the input sentence. Each pose contains the image-plane coordinates of the eight keypoints used in training, namely head, neck, left and right shoulders, elbows, and wrists.

B. Dataset

To train the model, this work uses the YouTube TED Gesture dataset [37]. The dataset comprises recordings of TED talks with time-aligned transcripts, and the videos are processed to extract framewise 2D poses simplified to a small set of upper-body joints. This corpus fits our case because public talks contain frequent beat gestures, typically offer long shots with a clear view of the torso and arms, and span many speakers and topics, which improves coverage and reduces speaker bias.

The dataset contains 1,766 videos segmented into clips whose number depends on the talk duration (see Fig. 2). Each clip is paired with its time-aligned transcript and the corresponding sequence of 2D poses. Poses are extracted with OpenPose and reduced to eight upper-body keypoints (head, neck, left/right shoulders, elbows, wrists), yielding 16

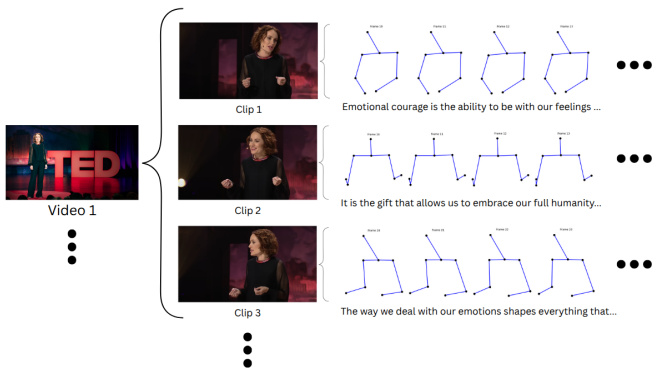

Fig. 2. Structure of the YouTube TED Gesture Dataset.

image-plane values per frame. This compact representation simplifies training. This split provides 28,945 training clips (83.92%), 2,852 validation clips (8.27%), and 2,694 test clips (7.81%) [37]. In the training set, clips average roughly 295 frames and 30 words.

Before training the model, we have conducted a preprocessing stage to address the many near-duplicate frames in the raw TED clips. These talks often include long holds and micro-movements, which overweights stillness and biases learning. To mitigate this, a variation-aware subsampling scans each clip in steps of ten frames and accepts a frame only if its pose differs from the last accepted one by more than a fixed threshold; otherwise, the scan advances frame by frame until the threshold is met. Clips with fewer than five valid poses after filtering are discarded. Next, each pose is centred by translating the neck to the origin, and all coordinates are scaled to $[-1, 1]$ to normalise across speakers and camera setups. Finally, anomalous frames are removed, including collapses at the origin and stretched skeletons likely caused by extreme camera orientation. These steps reduce duplication, improve numerical stability during training, and keep the data consistent with the retargeting stage.

IV. System Pipeline

This section presents the end-to-end flow of the algorithm from input to execution. It first introduces the Mini social robot and its software architecture to set the operational context. Next, it describes the process used to integrate the gesture generator into the robot. Finally, it details how the generator's output was converted into joint positions that Mini can interpret.

A. Mini and Software Architecture

Mini, shown in Fig. 3, is a tabletop robot developed for assisting older adults with mild cases of cognitive impairment [4]. It offers five degrees of freedom: waist yaw, left and right shoulder elevation, and two head axes. Motion runs within conservative envelopes to ensure safety and repeatability. The waist operates around $\pm 30^\circ$, the shoulders reach up to roughly 100° of elevation. The hardware includes an Intel i7 1260P processor, 16 GB of RAM, an Arduino Mega 2560 for low–level I/O, AX–12A servos in the upper body, an Intel RealSense SR300 RGB–D camera, a noise–cancelling microphone, capacitive touch sensors, BLE beacons, dual OLED eye displays, heart and cheek LEDs for non-verbal communication, another LED in the mouth as a visual indicator of the robot's speech, a stereo speaker, and a front touchscreen.

On the other hand, the HRI architecture follows a hierarchical and modular layout that keeps responsibilities clear and coupling low. At the core of the architecture are the skills, which represent each of the tasks that the robot can perform (reading the news, playing a game, showing photos or videos to the user...). A Decision Making System (DMS) chooses at any given time which of the skills has to be active. The Perception Manager receives the information captured by the robot's sensors and packages it into structured messages that downstream modules can consume. The HRI Manager receives requests from the skills and the DMS for conducting short interactions (ask a question, give information), manages priority conflicts among requests and uses the information received from the Perception Manager to advance these interactions. Liveliness produces random actions to endow the robot with an animate appearance.

Fig. 3. Mini, the social robot.

At the core of expressive output is the *Expression Manager*, the module that orchestrates the expressiveness capabilities of the robot. It receives requests from the rest of the architecture to use Mini's actuators, plans the execution of the requested actions, and ensures that there are no conflicts among them. The Expression Manager is divided into three levels. First, the *Expression Scheduler* acts as a gatekeeper and a planner. It validates each request, checks the availability of interfaces, locks the required resources, pauses Liveliness on the affected devices, groups actions that must start together, and orders execution with prioritised queues while supporting preemption when required. Then, the *Expression Executor* enacts the plan. It instantiates the state machine, builds or loads the motion trajectory, applies kinematic limits and smoothing, aligns speech and joints to a shared time base, and monitors execution to handle deviations, cancellations, and faults. Finally, the *Interface Players* send commands to the device drivers for joints, voice, on–screen eyes, lighting, and touchscreen. They report progress and completion to the Scheduler and the Executor so that resources are released promptly, and the next expression can start without losing multimodal synchrony.

B. System Integration on Mini

Our gesture generation approach has been integrated into Mini's software architecture as a module that interacts with the Expression Scheduler (see Fig. 4). When the robot has to utter a sentence without any associated nonverbal components, the Expression Scheduler sends the utterance to the beat gesture generator, a CVAE that returns a time sequence of 2D upper-body poses for the utterance.

Then, the Scheduler performs a deterministic retargeting

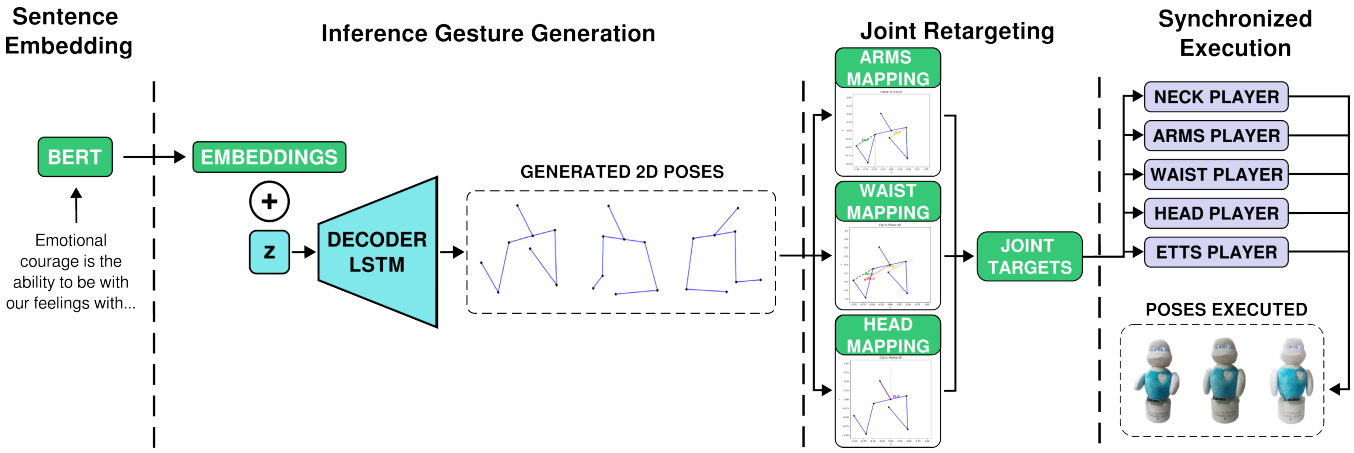

Fig. 4. Text-to-gesture pipeline used at inference. (1) BERT encodes the sentence into an embedding e; (2) the CVAE decoder samples z and autoregressively generates 2D upper-body poses; (3) deterministic 2D to joint retargeting (arms, waist, head) with limits and resampling produces joint targets; (4) FlexBE packages and synchronises the trajectory and executes it via device players.

from 2D pose space to Mini's joint space, a process described in greater detail in the next subsection. Shoulder elevation is recovered on both sides, waist yaw is estimated, and head pitch is computed from the neck to head vector. During this conversion, the kinematic envelope is enforced: the waist is limited to about $\pm 30^\circ$, the shoulders to roughly 100° of elevation, and head yaw is kept neutral to stabilise gaze. The resulting joint references are smoothed and resampled to the control rate to ensure safe and repeatable motion.

After that, the generated motions and the utterance are packaged into a FlexBE state machine. Each step of the sequence is represented as a state that launches, in parallel, the sub-states for arms, waist, head and voice, and the template advances through the poses in the intended order.

Finally, the Expression Executor runs the behaviour. It instantiates the state machine, sends the joint targets to the corresponding Players, and starts audio playback. Hence, the gesture co-occurs with the utterance without stretching or compressing its duration. A demonstration of Mini using the proposed model can be seen in the following YouTube video[1]

C. Retargeting from 2D Poses to Mini's Joints

Because of the limited number of degrees of freedom that Mini offers, we convert the 2D pose sequences generated by the CVAE into a compact set of joint commands that the robot can execute reliably. The goal is a deterministic and lightweight mapping that preserves the visual rhythm of the gesture while enforcing kinematic limits. We operate on five keypoints per frame—shoulder, elbow, and wrist for each arm, plus neck and head. Coordinates are neck-centred and normalised on the image plane. Let the image vertical be $\mathbf{u} = [0, 1]^\top$; denote shoulder, elbow, and wrist as P_h, P_e, and $P_w \in \mathbb{R}^2$, and neck and head as P_n and P_c.

1) Arms (shoulder elevation, left and right): Shoulder elevation is estimated with three cues that act as fallbacks when needed. First, for each arm we form $\mathbf{v}_b = P_w - P_h$ and measure its angle to the image vertical using Eq. (7).

$$\theta_{\text{arm}} = \cos^{-1}\left(\frac{(\mathbf{v}_b)_y}{\|\mathbf{v}_b\|_2}\right). \tag{7}$$

This angle captures how high the wrist is relative to the shoulder in the image. After computing θ_{arm}, a linear map sends the practical data range to the robot's shoulder range, followed by saturation at the mechanical limit (about 100°).

When the wrist lies roughly above or below the shoulder in the image, the primary cue in Eq. (7) becomes insensitive. If $\theta_{\text{arm}} < \theta_{\min}$ with $\theta_{\min} = 20^\circ$, we switch to a flexion surrogate based on segment straightness. This is calculated using Eqs. 8, 9, and 10.

$$L_{\text{arm}} = \|P_h - P_e\|_2 + \|P_e - P_w\|_2, \tag{8}$$

$$L_{\text{vec}} = \|P_h - P_w\|_2, \tag{9}$$

$$\Delta = L_{\text{arm}} - L_{\text{vec}}. \tag{10}$$

The scalar Δ increases as the wrist approaches the shoulder and serves as a robust cue for arm raising even when depth cannot be inferred. We rescale Δ to the joint range and apply saturation.

If the wrist height exceeds the shoulder height, we clamp the command to the shoulder's maximum so the posture corresponds to a fully raised arm without exceeding limits.

2) Waist yaw: Mini's waist operates within approximately $[-30^\circ, +30^\circ]$. We estimate torso orientation from the average horizontal directions of both arms. Let θ_R and θ_L be the shoulder–wrist angles with respect to the horizontal for right and left arms, computed from their (P_h, P_w) pairs. The circular mean is then calculated using Eq. 11.

$$\bar{\theta} = \operatorname{atan2}\big(\sin\theta_R + \sin\theta_L,\ \cos\theta_R + \cos\theta_L\big) \tag{11}$$

provides a single direction that respects angle periodicity. We re-centre $\bar{\theta}$ so that a symmetric, front-facing pose yields zero, then clip to $[-30^\circ, +30^\circ]$ and express in radians for actuation. If both arms are nearly vertical and provide

[1]https://youtube.com/shorts/NT7CtDvGWSI?feature=share

little information, we fall back to the previous valid waist command or to zero.

3) Head pitch: Head inclination is computed from the neck–to–head vector. Let $\mathbf{v}_h = P_c - P_n$. The angle to the image vertical is calculated with Eq. 12.

$$\theta_{\text{head}} = \cos^{-1}\left(\frac{(\mathbf{v}_h)_y}{\|\mathbf{v}_h\|_2}\right), \tag{12}$$

which we map linearly to the servo range and clamp to a conservative range for comfort and display legibility. Head yaw remains neutral to stabilise gaze.

V. Evaluation and Results

The system runs in pure-generation mode, receiving only text and a latent sample and producing a new pose sequence. Model selection also considers a reconstruction regime to probe the encoder–decoder pathway and the influence of auxiliary losses (mitigating posterior collapse and over-regularisation [38], [35]), while pure generation measures sampling behaviour that matches deployment.

A. Protocol and Metrics

Two regimes are used in the evaluation. In *reconstruction*, the model receives the utterance together with the ground-truth pose sequence and must reproduce it, exposing issues such as posterior collapse [38] and excessive regularisation [35]. In *pure generation*, the model receives the utterance and a latent sample z and synthesises a new sequence, matching the deployed setting. This separation decouples the faithfulness of the learned representation from sampling-time behaviour.

The analysis combines three complementary criteria so that fidelity, variety, and kinematics are considered together. Fréchet Gesture Distance (FGD) compares the distributions of real and generated poses through their means and covariances [39], [40]. Formally, it is calculated using Eq. 13.

$$\text{FGD} \;=\; \left\|\mu_r - \mu_g\right\|_2^2 \;+\; \text{Tr}\Big(\Sigma_r + \Sigma_g - 2\,(\Sigma_r\Sigma_g)^{1/2}\Big) \tag{13}$$

where μ_r, Σ_r are the mean and covariance of real poses and μ_g, Σ_g are those of the generated poses; lower values indicate better coverage of the real distribution and therefore higher dataset-level fidelity. In parallel, the *diversity metric* measures the per-component temporal variance within each generated sequence [41], which reflects how rich the motion is over time; high values are desirable provided that fidelity remains acceptable. Finally, the *smoothness metric* is the mean jerk (the discrete third derivative of motion) [40]; lower jerk penalises abrupt changes and favours visually pleasant trajectories.

B. Model Selection with Reconstruction

To balance faithful reconstruction against expressive variability, several CVAE variants were trained by adjusting the auxiliary-loss weights defined in Eq. (6): λ_T scales the small-movement/temporal penalty, and λ_D scales the diversity term. Table I summarizes four configurations (Model 1–4). As can be seen, **Model 1** corresponds to a *collapse* case: the diversity metric is extremely low, and motion becomes near static, which is consistent with the decoder ignoring z [38]. At the other extreme, **Model 4** illustrates *over-regularisation*: the diversity metric inflates, but FGD worsens substantially, signalling weaker text–motion linkage [35]. Between these limits, **Model 2** attains the lowest FGD yet occasionally shows limited variation, whereas **Model 3** accepts a slight increase in FGD and delivers stronger diversity and shoulder activity, which improves legibility on Mini. For these reasons, Model 3 is selected for the deployment.

Model	λ_T	λ_D	FGD	Diversity	Smoothness
Model 1	0.30	0.30	0.7750	0.0009	0.1932
Model 2	0.40	0.55	**0.6303**	0.1129	1.4935
Model 3	0.45	0.60	0.7456	**0.1329**	1.9173
Model 4	0.70	0.70	5.7876	0.9056	3.0363

TABLE I

Reconstruction results for CVAE variants.

C. Pure Generation and Temperature

Having selected **Model 3**, the evaluation then focuses on the deployment regime. Pure generation is analysed by sampling z and controlling dispersion with a fixed temperature τ that scales the latent sample [42], [43]. In practice, this parameter balances coverage against stability: as τ increases, the model explores a wider latent region and variety grows, up to a point where fidelity degrades. In our case, as shown by the results presented in Table II, FGD improves and reaches a minimum around $\tau \approx 1.6$, whereas larger values deteriorate quality [44]. This pattern also explains why, despite using the same data, reconstruction can yield higher FGD than well-tempered sampling: reconstruction tends to compress variance, whereas moderate sampling explores the data space more faithfully [23].

Temp τ	FGD	Diversity	Smoothness
0.7	1.1704	0.0142	1.1795
1.0	0.8688	0.0241	1.2155
1.3	0.6031	0.0481	1.2951
1.6	**0.4944**	0.0797	1.4258
1.8	0.5337	0.1149	1.5144
2.0	0.6241	0.1437	1.6182
2.5	1.1596	0.2221	1.9256

TABLE II

Results for Model 3 at different temperatures τ.

D. Discussion

Taken together, the sequences generated by Model 3 are human-like and follow consistent kinematic patterns rather than random motion. Nevertheless, biases from the TED domain are visible, for example, frequent downward head pitch and relatively few high arm raises. To keep motion perceptible on Mini, very small arm flexions in the range of 10–20 degrees were slightly amplified during retargeting, and the temporal density was reduced to roughly one pose per

five words. With these adjustments, motion remains readable on hardware with limited ranges and speeds. In part, this reflects the use of TED 2D poses for training and evaluation, which introduce domain and representation bias. Without depth, out-of-plane motion is ambiguous, the monologue style encourages downward head pitch and few high arm raises, and the variation-aware subsampling removes long holds. In addition, duration is estimated from text length rather than prosody, so timing follows a coarse word rate instead of intonation and pauses. These factors can shift both objective metrics and the qualitative appearance of the gestures.

At execution time, shoulder elevation, waist yaw, and head pitch stay within safe limits thanks to clipping and mild smoothing, which avoids jitter while preserving the rhythmic character. When speech is played, the system keeps the generator's rhythm without rescaling duration, favouring voice–gesture alignment.

Finally, end-to-end latency is about $1.616\,\mathrm{s}$ and the inference–retargeting stage about $0.947\,\mathrm{s}$. CPU usage shows brief peaks near 85% and RAM usage is around 0.8% of the total. Overall, these figures confirm that a CVAE-based pipeline is compatible with real-time interaction on a compact social robot.

VI. CONCLUSIONS

In this work, we built and deployed an end-to-end pipeline that turns text into co-speech gestures and executes them on a social robot with limited expressiveness and computational resources. A CVAE generates 2D upper-body poses, a deterministic retargeting maps them to Mini's five controllable joints under conservative limits, and the sequence is packaged as a FlexBE behaviour that runs in sync with speech. This turns gesture generation from a standalone prototype into an operational module within an HRI stack.

However, some limitations remain. First, training and evaluation rely on TED 2D poses, which inject domain and representation bias. Without depth, out-of-plane motion is ambiguous, the monologue format encourages downward head pitch and few high arm raises, and the variation-aware subsampling removes long holds. Together, these factors can shift both objective metrics and the qualitative appearance of the gestures. Second, timing is derived from text length rather than prosody, so alignment is preserved qualitatively at run time. Overall, our findings indicate that expressive co-speech movement is viable on limited hardware when generation, retargeting, and execution are designed jointly. Future work could focus on evaluating the proposed solution in robotic platforms with different DOF configurations, conditioning the model on other features like the prosody, and performing a user study that evaluates the subjective effect of the inclusion of our model in real human-robot interactions.

ACKNOWLEDGMENT

The research leading to these results has received funding from the projects: Robots sociales para mitigar la soledad y el aislamiento en mayores (SOROLI), PID2021-123941OA-I00, funded by Agencia Estatal de Investigación (AEI), Spanish Ministerio de Ciencia e Innovación. Interacción social multiusuario y multirobot para actividades de estimulación grupales en personas mayores, PID2024-157304OB-I00, funded by Agencia Estatal de Investigación (AEI), Spanish Ministerio de Ciencia e Innovación. The results have been funded by AdEBot: mejorando la expresividad de robots sociales para entornos terapéuticos, 2024/00766/001, under the program Ayudas para la Actividad Investigadora de los Jóvenes Doctores, Programa Propio de Investigación awarded by Universidad Carlos III de Madrid.

REFERENCES

[1] J. Niu, C.-F. Wu, X. Dou, and K.-C. Lin, "Designing gestures of robots in specific fields for different perceived personality traits," *Frontiers in Psychology*, vol. 13, p. 876972, June 2022, accessed: 2025-06-07.

[2] A. Clark and I. Ahmad, "Touchless and nonverbal human–robot interfaces: An overview of the state-of-the-art," *Smart Health*, vol. 27, p. 100365, 2023.

[3] E. Fernández-Rodicio, Castro-González, J. J. Gamboa-Montero, S. Carrasco-Martínez, and M. A. Salichs, "Creating expressive social robots that convey symbolic and spontaneous communication," *Sensors*, vol. 24, no. 11, p. 3671, 2024. [Online]. Available: https://doi.org/10.3390/s24113671

[4] M. Salichs, Castro-González, E. Salichs, E. Fernández-Rodicio, M. Maroto Gómez, J. J. Gamboa, S. Marques, J. Castillo, F. Alonso-Martín, and M. Malfaz, "Mini: A new social robot for the elderly," *International Journal of Social Robotics*, vol. 12, 12 2020.

[5] J. Cassell, H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," vol. 2001, 08 2001, pp. 477–486.

[6] A. Galatolo and K. Winkle, "Simultaneous text and gesture generation for social robots with small language models," *Frontiers in Robotics and AI*, vol. 12, 05 2025.

[7] B. Liu, L. Liu, S. Zhang, S. Gu, Y. Zhi, T. Zhu, L. Yang, and L. Ye, "Mag: Multi-modal aligned autoregressive co-speech gesture generation without vector quantization," 2025. [Online]. Available: https://arxiv.org/abs/2503.14040

[8] K. Zhao, G. Li, and S. Tang, "Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control," 2025. [Online]. Available: https://arxiv.org/abs/2410.05260

[9] T. Kucherenko, R. Nagy, Y. Yoon, J. Woo, T. Nikolov, M. Tsakov, and G. E. Henter, "The genea challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 792–801.

[10] P. Olikkal, D. Pei, B. K. Karri, A. Satyanarayana, N. M. Kakoty, and R. Vinjamuri, "Biomimetic learning of hand gestures in a humanoid robot," *Frontiers in Human Neuroscience*, vol. 18, 2024. [Online]. Available: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2024.1391531

[11] C. Yu and A. Tapus, "Srg3: Speech-driven robot gesture generation with gan," in *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2020, pp. 759–766.

[12] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA: Association for Computing Machinery, 1994, p. 413–420.

[13] K. R. órisson, "Communicative humanoids: a computational model of psychosocial dialogue skills," Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996. [Online]. Available: http://hdl.handle.net/1721.1/29118

[14] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff, "A comprehensive review of data-driven co-speech gesture generation," *Computer Graphics Forum*, vol. 42, no. 2, p. 569–596, May 2023. [Online]. Available: http://dx.doi.org/10.1111/cgf.14776

[15] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," *ACM Trans. Graph.*, vol. 28, 12 2009.

[16] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," *ACM Transactions on Graphics*, vol. 29, 07 2010.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[18] D. Hasegawa, N. KANEKO, S. Shirakawa, H. Sakuta, and K. Sumi, "Evaluation of speech-to-gesture generation using bi-directional lstm network," 11 2018.

[19] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4303–4309.

[20] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, ser. IVA '19. ACM, Jul. 2019, p. 97–104. [Online]. Available: http://dx.doi.org/10.1145/3308532.3329472

[21] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, ser. IVA '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 93–98. [Online]. Available: https://doi.org/10.1145/3267851.3267898

[22] C. Doersch, "Tutorial on variational autoencoders," 2021. [Online]. Available: https://arxiv.org/abs/1606.05908

[23] B. Dai and D. Wipf, "Diagnosing and enhancing vae models," 2019. [Online]. Available: https://arxiv.org/abs/1903.05789

[24] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, Z. He, and L. Bao, "Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders," 2021. [Online]. Available: https://arxiv.org/abs/2108.06720

[25] C. Ahuja, D. W. Lee, Y. I. Nakano, and L.-P. Morency, "Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach," 2020. [Online]. Available: https://arxiv.org/abs/2007.12553

[26] S. Ghorbani, Y. Ferstl, D. Holden, N. F. Troje, and M.-A. Carbonneau, "Zeroeggs: Zero-shot example-based gesture generation from speech," 2022. [Online]. Available: https://arxiv.org/abs/2209.07556

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[28] Y. Ferstl, M. Neff, and R. McDonnell, "Multi-objective adversarial gesture generation," ser. MIG '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3359566.3360053

[29] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha, "Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, Mar. 2021, p. 1–10. [Online]. Available: http://dx.doi.org/10.1109/VR50410.2021.00037

[30] Y. Xie, Y. Sun, H. Zhang, Y. Liu, and J. Tang, "Recom: Realistic co-speech motion generation with recurrent embedded transformer," 2025. [Online]. Available: https://arxiv.org/abs/2503.21847

[31] M. H. Mughal, R. Dabral, I. Habibie, L. Donatelli, M. Habermann, and C. Theobalt, "Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis," 2024. [Online]. Available: https://arxiv.org/abs/2403.17936

[32] F. Favali, V. Schmuck, V. Villani, and O. Celiktutan, "Tag2g: A diffusion-based approach to interlocutor-aware co-speech gesture generation," *Electronics*, vol. 13, no. 17, 2024.

[33] H. Pang, T. Ding, L. He, M. Tao, L. Zhang, and Q. Gan, "Llm gesticulator: Leveraging large language models for scalable and controllable co-speech gesture synthesis," 2024. [Online]. Available: https://arxiv.org/abs/2410.10851

[34] Y. Bao, D. Weng, and N. Gao, "Editable co-speech gesture synthesis enhanced with individual representative gestures," *Electronics*, vol. 13, no. 16, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/16/3315

[35] D. Shen, C. Qin, C. Wang, H. Zhu, E. Chen, and H. Xiong, "Regularizing variational autoencoder with diversity and uncertainty awareness," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, ser. IJCAI-2021. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, p. 2964–2970. [Online]. Available: http://dx.doi.org/10.24963/ijcai.2021/408

[36] Y. Ichikawa and K. Hukushima, "Learning dynamics in linear vae: Posterior collapse threshold, superfluous latent space pitfalls, and speedup with kl annealing," 2023. [Online]. Available: https://arxiv.org/abs/2310.15440

[37] Y. Yoon, "Youtube gesture dataset," 2025, gitHub repository, acceso: 28 junio 2025. [Online]. Available: https://github.com/youngwoo-yoon/youtube-gesture-dataset

[38] A. D. McCarthy, X. Li, J. Gu, and N. Dong, "Addressing posterior collapse with mutual information for improved variational neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8512–8525. [Online]. Available: https://aclanthology.org/2020.acl-main.753/

[39] T. Guichoux, L. Soulier, N. Obin, and C. Pelachaud, "2d or not 2d: How does the dimensionality of gesture representation affect 3d co-speech gesture generation?" 2024. [Online]. Available: https://arxiv.org/abs/2409.10357

[40] T. Kucherenko, P. Wolfert, Y. Yoon, C. Viegas, T. Nikolov, M. Tsakov, and G. E. Henter, "Evaluating gesture generation in a large-scale open challenge: The genea challenge 2022," *ACM Transactions on Graphics*, vol. 43, no. 3, p. 1–28, Jun. 2024. [Online]. Available: http://dx.doi.org/10.1145/3656374

[41] M. Meng, K. Mu, Y. Zhu, Z. Zhu, H. Sun, H. Yan, and Z. Fan, "Varges: Improving variation in co-speech 3d gesture generation via styleclips," 2025. [Online]. Available: https://arxiv.org/abs/2502.10729

[42] A. Shih, D. Sadigh, and S. Ermon, "Long horizon temperature scaling," 2023. [Online]. Available: https://arxiv.org/abs/2302.03686

[43] T. Song, J. Sun, X. Liu, and W. Peng, "Scale-VAE: Preventing posterior collapse in variational autoencoder," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 14 347–14 357. [Online]. Available: https://aclanthology.org/2024.lrec-main.1250/

[44] J. Prost, A. Houdard, A. Almansa, and N. Papadakis, "Diverse super-resolution with pretrained deep hiererarchical vaes," 2024. [Online]. Available: https://arxiv.org/abs/2205.10347