ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesis

Muhammad Hamza Mughal[1,2] Rishabh Dabral[1] Ikhsanul Habibie[1] Lucia Donatelli[3]
Marc Habermann[1] Christian Theobalt[1,2]

[1]Max Planck Institute for Informatics, SIC [2]Saarland University [3]Vrije Universiteit Amsterdam

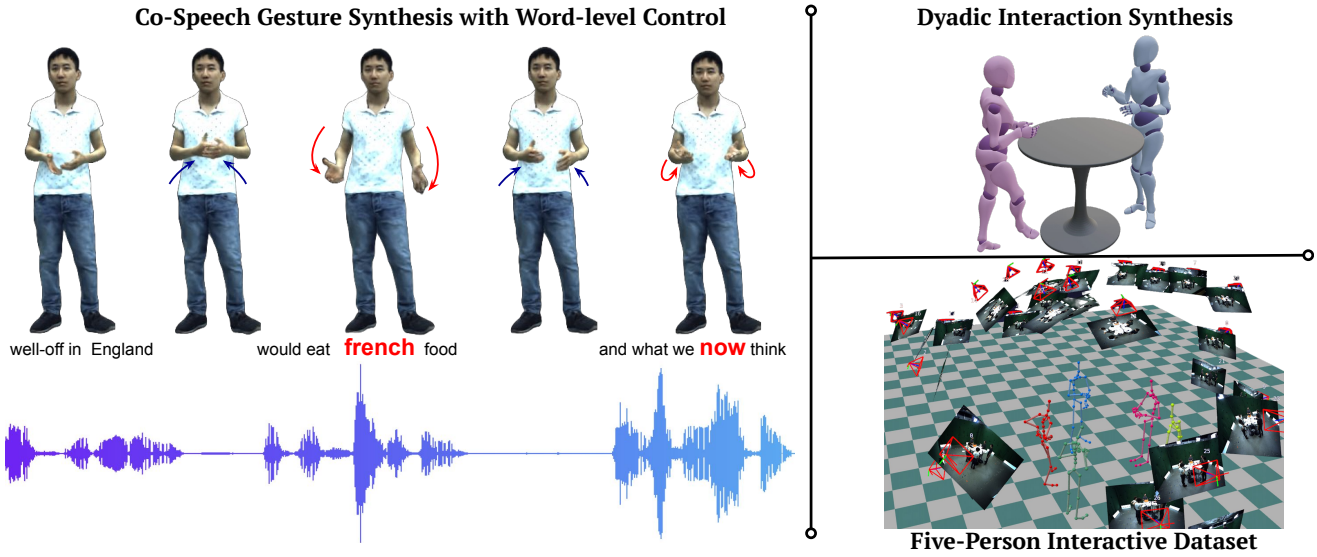


Figure 1. **Our CONVOFUSION approach generates body and hand gestures in monadic and dyadic settings, while also offering advanced control over textual and auditory modalities in speech. Lastly, we introduce the DND GROUP GESTURE dataset, showcasing rich interactions with co-speech gestures between five participants.** Motions rendered using ASH [51].

Abstract

Gestures play a key role in human communication. Recent methods for co-speech gesture generation, while managing to generate beat-aligned motions, struggle generating gestures that are semantically aligned with the utterance. Compared to beat gestures that align naturally to the audio signal, semantically coherent gestures require modeling the complex interactions between the language and human motion, and can be controlled by focusing on certain words. Therefore, we present CONVOFUSION *, a diffusion-based approach for multi-modal gesture synthesis, which can not only generate gestures based on multi-modal speech inputs, but can also facilitate controllability in gesture synthesis. Our method proposes two guidance objectives that allow the users to modulate the impact of different conditioning modalities (e.g. audio vs text) as well as to choose certain words to be emphasized during gesturing. Our method is versatile in that it can be trained either for generating monologue gestures or even the conversational gestures. To further advance the research on multi-party interactive gestures, the* DND GROUP GESTURE *dataset is released, which contains 6 hours of gesture data showing 5 people interacting with one another. We compare our method with several recent works and demonstrate effectiveness of our method on a variety of tasks. We urge the reader to watch our supplementary video at our webpage.*

1. Introduction

Gestures are one of the fundamental ways of expression and can significantly enhance the interpretation of the verbally communicated utterance [29]. As our society integrates multi-billion parameter large-language-model (LLMs) [62, 74] into our workflows and daily lives, it is only natural to consider ways to augment the LLM based on spoken language alone with *non-verbal* information essential to in-

terpreting such language. Towards this goal, speech and text-based gesture generation approaches have come a long way from symbolically representing gestures [8, 9] in a rule-based generation framework [32] to the state-of-the-art methods trained on human motion capture data [4, 73, 75].

Yet, while the majority of methods successfully capture *beat gestures* that are prosodically aligned with speech, they lack language-based control over the gesture generation and therefore, struggle to generate precise *semantic gestures* that contribute to the overall meaning of an utterance. This can be attributed to the fact that the motion of beat gestures is temporally well-aligned with the speech signals and generally follows a similar spatial pattern for all speakers and content, therefore, it is easier to model using learning techniques. On the other hand, semantic coherence has a more complex temporal interplay with the words, their meaning and who the individual speaker is.

In this work, we propose CONVOFUSION – a novel controllable gesture synthesis method to generate not only co-speech gestures, but also reactive (and passive) gestures. We follow a latent diffusion approach [13, 55], which has the benefit of learning a jitter-free motion representation. Unlike existing latent diffusion methods [13], we design our motion latents to be time-aware, thus allowing us to learn temporal correlations between motion and speech along with the ability to perform perpetual gesture synthesis.

Our synthesis model supports a variety of input signals (text and audio of the speakers in the conversation) and provides a framework to control them. To enable controllable multi-modal inference of our model, we introduce a novel classifier-free guidance training strategy. More specifically, instead of dropping the entire multi-modal conditioning signal, we show that selectively replacing the modalities with null-vectors facilitates test-time control over each modality. Finally, CONVOFUSION also allows us to enhance the micro-gestures associated with a particular word, thanks to the fine-grained textual guidance. Having the test-time modality control and word-level textual guidance provides us the unique ability to have coarse and fine control of the generated motions; a feature missing in existing gesture synthesis works [4, 24, 67].

One of the goals of our framework is to model the gestures exhibited in a conversational setting. Unfortunately, most existing datasets only contain monologues, as in the TED [69] and SHOW [67] datasets. Even the datasets recorded in conversational setting [43] provide annotations only for one person. To address this, we introduce the DND GROUP GESTURE dataset. It involves five participants playing multiple sessions of Dungeons and Dragons – a popular role-playing game. The dataset comes with high quality full-body motion capture of all the participants, along with multi-channel audio recordings and text transcriptions. Thanks to around 6 hours of capture, the DND GROUP GESTURE dataset allows us to propose a novel approach to generate gestures in a dyadic setting.

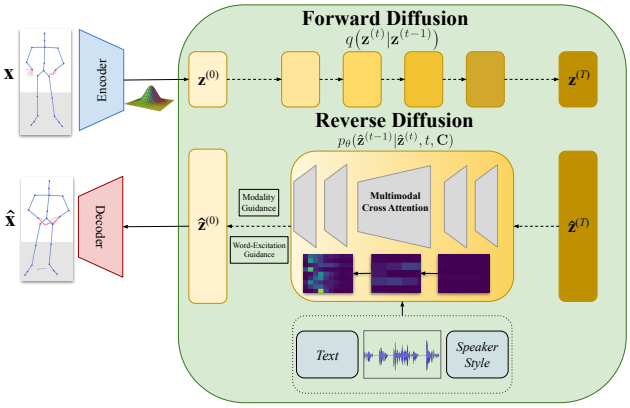


Figure 2. **Overview of the proposed approach.** We generate gestures conditioned on multiple conditioning signals such as text, audio, speaker style, etc. using a latent diffusion approach. During inference, we introduce modality guidance and word-excitation guidance to control the properties of the generated gestures.

In summary, our technical contributions are as follows:

- We propose CONVOFUSION – a diffusion-based approach for monadic and dyadic gesture synthesis. We do so not only in the co-speech setting but can also generate passive/reactive gestures.
- Thanks to the proposed coarse and fine-grained guidance, our work investigates ways to incorporate a variety of multi-modal signals and provides a framework to control their influence in the generated gestures.
- We demonstrate how generating gestures in the proposed latent mitigates the jittering artifacts prevalent in the hand-articulations of existing datasets. Unlike existing motion latent diffusion works [13], the proposed time-aware latent representation allows us to perform perpetual gesture synthesis with high synthesis quality.
- This work also introduces the DND GROUP GESTURE dataset, thereby facilitating future research on dyadic and group gesture synthesis.

2. Related Works

As our work draws inspiration from the extensive literature on gesture synthesis and recent works on diffusion-based generative models, we discuss relevant literature from these two perspectives in this section.

2.1. Co-Speech Gesture Synthesis

Co-speech gestures are a unique form of gesture, in which hand and arm movements used to communicate information are temporally synchronized and semantically integrated with speech [47]. While such gestures are thought to contribute to meaning and discourse in the same way as lexical

items and intonation patterns, their multi-functional nature makes automatic generation challenging. Non-referential *beat* gestures align with prosodically stressed words and contribute less to overall semantic meaning [29, 48]; such gestures have proved easier to generate [41]. Semantic gestures categorized as *iconic*, *metaphoric*, or *deictic* visually illustrate some aspect of the spoken utterance yet are less patterned between speakers and content; these gestures are more challenging to effectively reproduce [34].

Early works in the field of co-speech gesture synthesis can be divided into rule-based and data-driven techniques. Rule-based methods [10, 11], which usually utilize heuristics, generate gesture combinations with high semantic alignment to speech. [65] provides a comprehensive overview of these methods. However, they produce unnatural and less diverse gesture outputs. To mitigate this problem, early statistical approaches [38, 39] try to model the underlying gesture distribution using data and then predict gestures that are most appropriate for given speech input. However, both rule-based systems and early statistical approaches predict gesture sequence in terms of known gesticulation units, which makes the final output look unnatural and choppy. Therefore, recent data-driven learning-based methods [4–6, 24, 33] employ neural networks to map speech input to a gesture sequence, which allows for per-frame gesture prediction, providing an end-to-end solution for speech-to-gesture synthesis. [50] provides an in-depth overview of classical and recent data-driven methods.

Earlier deep-learning-based methods which used CNN [24], RNNs [45, 68, 70] and transformers [6] employed deterministic approaches to predict gestures for the speech input. On the other hand, generative methods offer a better alternative since they can introduce stochasticity in the generation process which leads to diverse outputs. Generative modeling approaches [2, 19, 25, 40, 44, 67] have been used for synthesis resulting in human-like gestures. But, they also suffer from low semantic relation with the speech input because there exists many-to-many relations between speech and gestures and it becomes hard for the generative approaches to realize which gesture is more semantically accurate corresponding to the speech. Therefore, recent approaches [3, 4, 35, 41, 42] try to improve intent's alignment with gesture prediction. Gesture styles are also incorporated in the gesture generation pipeline for personalized gesture synthesis [19, 66].

2.2. Speech Gesture Datasets

As the performance of learning-based methods relies on the quality of its training data, a number of gesture synthesis datasets have been proposed by the community. However, high-quality speech-driven gesture synthesis datasets are typically expensive and tedious to collect as they require hours of speech gesture motion capture (mocap) recordings in a studio setting. Because of these limitations, early works typically involve a single speaker [17, 18]. To collect a large number of training samples, several works have proposed to leverage monocular 3D estimation approaches to obtain the 3D body, face, and hand keypoints [1, 20, 23, 24, 67, 69]. Unfortunately, such monocular estimation results are subpar compared to the standard multi-view mocap approaches and are unsuitable for multi-speaker settings.

To address the lack of large-high-quality data, [43] proposed BEAT, a 76-hour mocap-based speech gesture dataset recorded from 30 different subjects. Unlike BEAT which focused on a single speaker, [37] introduced a high-quality speech gesture dataset that involved multiple speakers, but was limited to two-person conversations. In contrast to previous works, we propose a high-quality speech-gesture dataset involving 5 subjects within a conversation. In addition, different from most mocap-based datasets that use marker-based mocap technologies, we employ a state-of-the-art markerless mocap system to accurately capture the 3D body and hands of multiple speakers without being restricted by body mocap suits. Tab. 1 provides a brief overview of some notable datasets and their qualities. Moreover, we also compare them with the DnD Group Gesture dataset we present in this work.

2.3. Diffusion-based Generative Modelling

Diffusion models [27, 59] have demonstrated remarkable potential in the field of generative modeling, consistently delivering impressive results in various synthesis applications [14, 31, 53, 56, 60, 63, 71]. New paradigms like guidance mechanisms [15, 26] and latent diffusion models [54] have been introduced to enhance quality and alignment of diffusion-based synthesis w.r.t given conditionings.

This approach has been extensively applied for conditional human motion synthesis [13, 14, 61, 63]. Similarly, co-speech gesture generation has also greatly benefited from this generative modeling technique. DiffGesture [75] uses a transformer-based diffusion pipeline with an annealed noise sampling strategy for temporally consistent gesture generation. GestureDiffuCLIP [4] employs latent-diffusion models [54] and CLIP [52] based conditioning to improve control over co-speech gesture generation. [60] presents a model to predict the movement of multiple speakers in a social setting. However, contrary to other diffusion-based gesture synthesis approaches, their model focuses on predicting the correctness of the 3D body keypoint trajectory for a few seconds in the future instead of improving the speech-gesture alignment. Instead of simply predicting the motion trajectory, our method proposes a multi-person speech-driven 3D gesture synthesis approach that can be used to predict the 3D reactive body and hand motion between various speakers and listeners within a conversation.

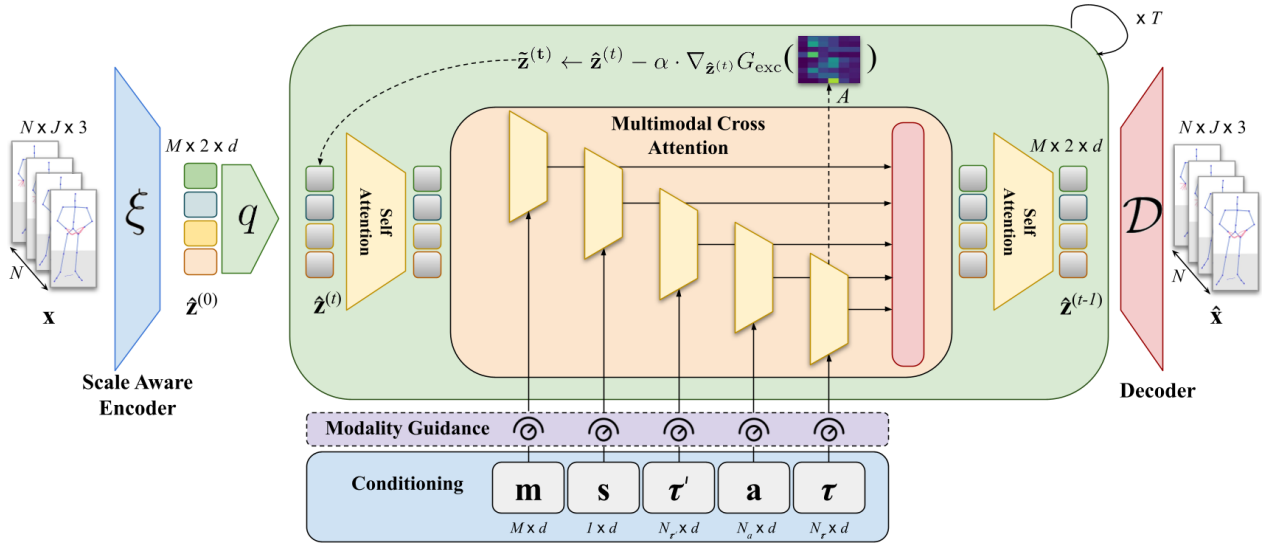


Figure 3. **The model schema.** Given a training motion $\mathbf{x} \in \mathbb{R}^{N \times J \times 3}$, we first extract its latent encoding $\hat{\mathbf{z}}^{(0)}$ (Sec. 3.1), which is then denoised by a network that incorporates the various modalities in the denoising process. At inference time, the denoised latents are decoded to produce the final generation, $\hat{\mathbf{x}}$ (Sec. 3.2). During this process, our method allows to control the generation through coarse-grained modality guidance or fine-grained word-excitation guidance (Sec. 3.3). Dotted lines represent components used only during inference.

3. Approach

The goal of our method is to generate co-speech gesture sequences for monadic and dyadic settings in correspondence with input speech. A gesture sequence $\mathbf{x} \in \mathbb{R}^{N \times J \times 3}$ consists of N frames of human motion with J articulating 3D joints. The generated gesture motion ought to be consistent with the multi-modal conditioning signal, $\mathbf{C}$, representing the speech and identity-related attributes of the persons in conversation (discussed later in Sec. 3.2).

We design our gesture synthesis method around a latent denoising diffusion probabilistic model (DDPM) framework [55]. The proposed diffusion model is trained to denoise the latent representation of the gesture motions (refer to Sec. 3.1). The generated motion latents can later be decoded using a motion decoder. Unlike existing motion latent diffusion methods [13], we design our latent space in a time-decomposable manner, thereby allowing us to learn fine-grained interplay between motion and speech. Crucially, our method also allows the end-user to *control* the attributes of the generated gestures at inference time (see Sec. 3.3). We now discuss each component in detail. Refer to the supplemental document for a glossary of major notations used in the method explanation.

3.1. Scale-aware Temporal Latent Representation

Instead of directly denoising the raw motion $\mathbf{x}$, our diffusion model operates in the latent space of human motion. Thus, we propose to learn such a latent space with two characteristics: 1) We disentangle the finger motions from the rest of body motions by encoding them into a latent space through separate encoders. 2) Instead of projecting the entire motion into one single latent vector, we encode motion into chunked latents that can be decoded jointly by a decoder.

Decoupled Latent Representations. The articulation of the finger joints is critical to the quality of gesture synthesis. However, the fingers articulate in a significantly different space and scale compared to the rest of the body and naïvely encoding the full-body gestures results in inaccurate reconstruction of hands. We therefore follow prior works that decouple the two sets of joints [21, 22] and represent the motion $\mathbf{x}$ as a latent vector $\mathbf{z} = \{\mathbf{z}_b, \mathbf{z}_h\}$, where $\mathbf{z}_b \in \mathbb{R}^d$ and $\mathbf{z}_h \in \mathbb{R}^d$ are separate encodings of the body and hand motion.

The latent vectors are learned using a VAE framework. The hand and body motions, $\mathbf{x}_h$ and $\mathbf{x}_b$, are encoded using transformer encoders: $\mathbf{z}_b = \xi_b(\mathbf{x}_b)$, $\mathbf{z}_h = \xi_h(\mathbf{x}_h)$. The latent vectors represent the mean of the distribution, which can be sampled using the reparameterization trick [30] and fed into a decoder to reconstruct the motion $\mathbf{x}'_b = \mathcal{D}_b(\mathbf{z}_b), \mathbf{x}'_h = \mathcal{D}_h(\mathbf{z}_h)$. We train the VAE with the standard reconstruction loss, $\mathcal{L}_2$, Bone-length regularization loss $\mathcal{L}_{bone}$ [14] and the KL-Divergence of the latents, $\mathcal{L}_{KL}$. Additionally, we reduce the jitter in reconstruction proposing a Laplacian regularization term:

$$\mathcal{L}_{lap} = \|\mathscr{L}\{\hat{\mathbf{x}}\} - \mathscr{L}\{\mathbf{x}\}\|_2 \quad (1)$$

where $\mathscr{L}\{\cdot\}$ is the Laplace transform operator along N frames. Refer to Sec. 5.4 and supplemental for analysis.

Time-Aware Latent Representation. The motion latents learned by the VAE represent a large motion sequence (>100 frames) with a single d-dimensional vector. This,

rather coarse, granularity prohibits applications such as perpetual rollout where the motion can be autoregressively decoded with an overlapping window. To enable such applications, we propose to encode shorter motion chunks in the latent $\mathbf{z}$ but decode multiple such chunked latents, $\{\hat{\mathbf{z}}_i\}_{i=1}^{M}$ together with a single decoder, as shown in Fig. 4.

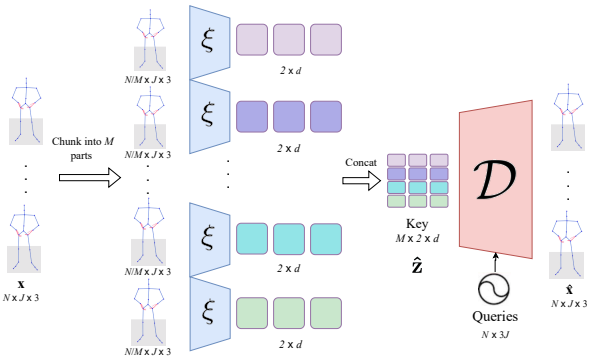


Figure 4. **Chunked latent encoding-decoding.** We encode a motion of N frames into a sequence of M latent vectors, which are jointly decoded by the decoder $\mathcal{D}$. Encoding into chunked latents allows for perpetual rollout and decoding jointly induces temporal consistency while converting the latents back into motion.

Given a gesture sequence $\mathbf{x}$, we first split the sequence into M equally sized chunks $\{\mathbf{x}'_i\}_{i=1}^{M}$, where $\mathbf{x}'_i \in \mathbb{R}^{N/M \times J \times 3}$. Next, each of the chunks $\mathbf{x}'_i$ is encoded in isolation using $\hat{\mathbf{z}}_i = \xi_b(\mathbf{x}'_i)$. However, while decoding, the decoder collectively decodes a sequence of chunked latents, $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}_i\}_{i=1}^{M}$, following: $\hat{\mathbf{x}}' = \mathcal{D}(\hat{\mathbf{z}})$. In summary, our latent encodings transform a motion sequence $\mathbf{x} \in \mathbb{R}^{N \times J \times 3}$ into latent representations $\hat{\mathbf{z}} \in \mathbb{R}^{M \times 2 \times d}$. This can enable perpetual gesture generation using diffusion inpainting technique [46] as we discuss and analyze in Sec. 5.4.

3.2. Modality-Conditional Gesture Generation

Having obtained $\hat{\mathbf{z}}$ as the time-aware latent representation of gesture motions, we formulate the gesture synthesis task as that of conditional latent diffusion [55]. The forward diffusion process, successively corrupts the latent sequence $\hat{\mathbf{z}}^{(0)}$ by adding Gaussian noise ϵ for T timesteps with the assumption that $\hat{\mathbf{z}}^{(T)} \sim \mathcal{N}(0, I)$. For generation, the *reverse diffusion* process is performed on $\hat{\mathbf{z}}^{(T)}$ by iteratively denoising $\hat{\mathbf{z}}^{(T)} \sim \mathcal{N}(0, \mathbf{I})$ to generate a latent sequence $\hat{\mathbf{z}}^{(0)}$, and can be formulated as

$$p_\theta\big(\hat{\mathbf{z}}^{(0:T)}\big) = p\big(\hat{\mathbf{z}}^{(T)}\big) \prod_{t=1}^{T} p_\theta\big(\hat{\mathbf{z}}^{(t-1)} | \hat{\mathbf{z}}^{(t)}\big), \qquad (2)$$

where $p_\theta(\hat{\mathbf{z}}^{(t-1)}|\hat{\mathbf{z}}^{(t)})$ is approximated using a neural network parameterized by weights θ. This neural network f_θ is trained to predict noise $\epsilon_\theta(\hat{\mathbf{z}}^{(t)}, t)$ [27], which can be used in the training objective $\mathcal{L}_d = ||\epsilon - \epsilon_\theta(\hat{\mathbf{z}}^{(t)}, t)||^2$.

The motion generation framework discussed above is so far *unconditional*. Our gesture synthesis approach can be conditioned in primarily two settings: monadic and dyadic. The *monadic setting* refers to the co-speech gesture generation based solely on the speaker's own utterance and typically occurs in monologue scenarios. For this, we represent the conditioning signal as $\mathbf{C} = \{\mathbf{a}, \boldsymbol{\tau}, \mathbf{s}\}$, consisting of the audio signal $\mathbf{a} \in \mathbb{R}^{N_a \times d}$ and the text tokens $\boldsymbol{\tau} \in \mathbb{R}^{N_{\boldsymbol{\tau}} \times d}$, as well as $\mathbf{s} \in \mathbb{R}^{1 \times d}$ representing the speaker identity token. Generally, N_a corresponds to the number of audio frames, $N_{\boldsymbol{\tau}}$ corresponds to the number of text tokens in the utterance. Speaker identity $\mathbf{s}$ can enable applications like stylized gesture synthesis which can generalize to different gesture styles. For the *dyadic setting*—which takes place in conversation scenarios—the generated gestures must be in accordance with the co-participant's utterance as well. In this case, we have $\mathbf{C} = \{\mathbf{a}, \boldsymbol{\tau}, \boldsymbol{\tau}', \mathbf{s}, \mathbf{m}\}$, where $\boldsymbol{\tau}'$ refers to the co-participant's speech content i.e. their text. Here, we can also choose their audio instead of their text as well. Finally, $\mathbf{m} \in \{0, 1\}^M$ indicates whether the speaker is actively responding with speech, or passively back-channeling e.g. by laughing or nodding (see also supplemental video).

We use a transformer decoder network [64] with multi-head attention to approximate the denoising function producing $\epsilon_\theta(\hat{\mathbf{z}}^{(t)}, t, \mathbf{C})$. This allows us to elegantly integrate multiple modalities in $\mathbf{C}$ with separate cross-attention heads, as shown in Fig. 2. Let us consider the case of the audio signal, $\mathbf{a}$. The cross-attention features, ϕ_a, are computed using the attention matrix $\text{Attn}(\hat{\mathbf{z}}, \mathbf{a}) \in \mathbb{R}^{N_a \times M}$ as:

$$\text{Attn}(\hat{\mathbf{z}}, \mathbf{a}) = \sigma\left(\frac{Q_z K_a}{\sqrt{d}}\right), \phi_a = V_a \cdot \text{Attn}(\hat{\mathbf{z}}, \mathbf{a}) \qquad (3)$$

where σ is the softmax operator, Q_z, K_a, V_a are the query, key and value vectors recovered form the motion latent features $\hat{\mathbf{z}}$ and the audio features $\mathbf{a}$. We similarly recover text features, $\phi_\tau = \text{Attn}(\hat{\mathbf{z}}, \boldsymbol{\tau})$, also for the text.

3.3. Towards Controllable Gesture Generation

In addition to multi-modal gesture synthesis, our method is designed to allow coarse and fine-grained control. For coarse control, one can adjust the impact of a specific modality on the generated motion by utilizing our *modality-level guidance strategy*. For fine control, the user can choose specific words to enhance the gestures for the words using the proposed *word-excitation guidance* (WEG) objective.

Modality-Guidance. Classifier-free guidance [26] has been used to improve the generation quality of various diffusion-based motion and gesture generation methods [4, 13, 36, 61]. Typically, this is done by randomly replacing the conditioning vectors with a null-embedding $\mathbf{C} \leftarrow \emptyset$. At inference, the noise predictions are blended at each diffusion timestamp t to get the noise prediction $\epsilon_\theta^{(t)}$:

$$\epsilon_\theta^{(t)} = \epsilon_\theta(\hat{\mathbf{z}}^{(t)}, t, \emptyset) + \lambda(\epsilon_\theta(\hat{\mathbf{z}}^{(t)}, t, \mathbf{C}) - \epsilon_\theta(\hat{\mathbf{z}}^{(t)}, t, \emptyset)) \quad (4)$$

where, λ represents the guidance scale. Once estimated, $\epsilon_\theta^{(t)}$ can be used to sample $\hat{\mathbf{z}}^{(t-1)}$ for the next iteration using Eq. 11 of [27]. However, recall that our conditioning set $\mathbf{C} = \{\mathbf{a}, \boldsymbol{\tau}, \boldsymbol{\tau}', \mathbf{s}, \mathbf{m}\}$ consists of several modality-specific conditions. Naïvely setting all the elements to $\emptyset$ for random iterations prohibits separately learning the effect of each individual modality within $\mathbf{C}$ on the conditional distribution. Instead, we train our model with random modality dropouts (with null-embedding replacement) with 10% drop probability. This encourages the model to learn several combinations of marginalized conditional probability distributions.

At inference, we sample with modality-guidance:

$$\epsilon_\theta^{(t)} = \epsilon_\theta^{\emptyset} + \lambda_m \sum_{\mathbf{c} \in \mathbf{C}} w_c \big(\epsilon_\theta^{c}(\hat{\mathbf{z}}^{(t)}, t, \mathbf{c}) - \epsilon_\theta(\hat{\mathbf{z}}^{(t)}, t, \emptyset) \big) \quad (5)$$

where the scale parameters, $w_c \geq 0$, determine the contribution of each modality towards the generated gesture and λ_m is the global guidance scale. Adjusting the modality scale, w_c allows us to coarsely control the gesture quality and also analyze the sensitivity of the generation process to specific modalities. Note, that this is an optional sampling strategy required only for modality-level control.

Word-Excitation Guidance. Inspired by the controllable image generation methods [12, 16], we propose a word-level guidance mechanism that allows us to finely control the gesture generation based on a user-defined set of words during the sampling process.

Let $\mathrm{Attn}(\hat{\mathbf{z}}^{(t)}, t, \boldsymbol{\tau}) \in \mathbb{R}^{N_{\boldsymbol{\tau}} \times M}$ be the text attention matrix at the t^{th} iteration of the denoising process. For a set of text tokens $\{\boldsymbol{\tau}_i\}_{i=1}^{S}$, selected by a *user* with the intention of gesture enhancement, we focus on the corresponding column, $A_i \in \mathbb{R}^M$ in the text attention matrix. Now, with the assumption that the element with maximum attention in A_i aligns with the motion chunk associated with the text, we introduce a guidance objective to further enhance (or, excite) the same attention:

$$G_{\mathrm{exc}} = \frac{1}{S} \sum_{i=1}^{S} (1 - \max(A_i)) \quad (6)$$

Next, we use the gradient of G_{exc} w.r.t the latent $\hat{\mathbf{z}}^{(t)}$ to perform the word-excitation guidance:

$$\tilde{\mathbf{z}}^{(t)} \leftarrow \hat{\mathbf{z}}^{(t)} - \alpha \cdot \nabla_{\hat{\mathbf{z}}^{(t)}} G_{\mathrm{exc}}, \epsilon_\theta^{(t)} = f_\theta(\tilde{\mathbf{z}}^{(t)}, t, \mathbf{C}) \quad (7)$$

where α is the guidance scale for the word excitation guidance, which also serves as a step size for latent update.

4. Dataset

To enable a high-quality, speech-driven gesture synthesis method involving multiple speakers, we introduce the DND GROUP GESTURE dataset. Our dataset is designed to also invoke a wide range of non-verbal gestures during the speaker interactions. We based our dataset recordings on D&D tabletop roleplaying game, where five different players are standing in a circle around a game map. Each participant is equipped with a dedicated wireless microphone to ensure a clean audio recording and audio source separation. The setup of the gameplay involves various types of interaction between the actors that often require semantically meaningful gestures such as pointing to a certain location on the map. In total, the dataset consists of 4 separate recording sessions with a total duration of 6 hours.

Name	# Identities	Size	Body Parts	Multi-party Interaction	# Interacting Speakers
IEMOCAP [7]	10	12h	Face	✓	2
Creative-IT [49]	16	2h	Body †	✓	2
CMU Haggling Dataset [28]	122	3h	Face, Body, Hands	✓	3
TED Dataset [69]	1295	52.7h	Upper Body		
Speech Gesture 3D [24]	10	144h	Upper Body, Hands, Face		
Talking with Hands [37]	50	50h	Body, Hands	✓	2
PATS [1]	25	250h	Upper Body, Hands		
SaGA++ [34]	25	4h	Body hands		
ZeroEGGS Dataset [20]	1	2h	Body, Hands		
BEAT [43]	30	76h	Body, Hands, Face		
DND GROUP GESTURE	5	6h	Body, Hands	✓	5

Table 1. Comparison of currently available datasets to our DND GROUP GESTURE dataset. *Body parts* refer to the parts where the 2D or 3D pose tracking is available. † indicates that the body tracking is only available for one of one interacting actors.

Our proposed dataset is recorded using a state-of-the-art multi-view markerless mocap to obtain accurate 3D body and hand pose estimates of multiple subjects at a given time. This allows our participants to move freely without being obstructed by the tight mocap suit or gloves. In addition to audio and the 3D pose annotations, we also provide text and gesture annotations for each individual speaker that distinguishes different types of observable gestures, including beats, iconic, deictic, and metaphoric. Our dataset will be made publicly available to the community.

5. Experiments

Our method, in its vanilla form, is designed to generate human gestures from speech, yet it goes several steps beyond this task. For instance, we adapt our method to perform dyadic conversations. More importantly, we show how different modalities contribute to the generation and perform fine-grained text-based control. Naturally, it is difficult to find suitable baselines to compare with. To perform fair evaluations, we, therefore, compare with methods that can be trivially adapted to our setting. Specifically, we compare with MLD [13] (a generic latent diffusion method), CaMN [43], Multi-Context [70], DiffGesture [75] (specifically monadic gesture baselines) and DiffuGesture [72] (two-person motion synthesis works). Notably, CaMN [43], DiffGesture [75] and DiffuGesture [72] require a seed motion sequence to build the gesture generation on. This is different from our setting and provides vital clues about the gesture style. We provide the seed motions for the two

	FID ↓	BeatAlign →	Diversity →	L1 Div →	SRGR ↑
GT	-	0.89	13.21	13.12	-
Multi-Context [70]	$\geq 10^3$	0.8	26.71	43.31	0.140
DiffGesture [75]	$\geq 10^3$	0.96	176	17.8	0.003
CaMN [43]	142	0.74	9.66	5.85	0.443
MLD [13]	475	0.76	16.98	5.42	0.214
Ours	271	0.82	9.82	6.24	0.365

Table 2. **Comparison on the BEAT [43].** Two methods [70, 75] produce extremely jittery motions. We demonstrate superior beat alignment and diversity scores among the remaining methods.

methods, but do not use the seed motions to generate our results. The methods are compared using the established motion synthesis metrics as well as a user-study.

Evaluation Datasets. We evaluate our performance in monadic gesture generation on the recently introduced BEAT dataset [43]. The test set includes 2492 5-sec motion sequences and includes a set of 5 unseen speakers. For evaluating the motion in dyadic setting, we use the test set of the proposed DND GROUP GESTURE dataset. The test set contains 3932 sequences of 5-second conversations.

Metrics. Evaluating synthesized motions is challenging due to the subjective nature of perceiving good gestures. Yet, we evaluate our method on the established metrics like Beat-Alignment [58], FID, Semantic Relevance Gesture Recall (SRGR) [43] that evaluate different aspects of the motion. We also use Diversity and L1 Divergence to evaluate the ability of models to span the space of gesture motions with enough coverage.

5.1. Monadic Co-speech Gesture Synthesis

We tabulate our results on the BEAT test set for monadic co-speech gesture synthesis in Tab. 2. We observe that DiffGesture [75] and Multi-ContextNet [70] struggle with the FID which, upon visualization, can be attributed to the extremely jittery nature of the generated motions. Interestingly, this also leads to Multi-ContextNet [70] to perform the best in the Beat Alignment metric as for every beat in the audio, there is always a jittery motion to align with. Among other methods, we observe better performance in terms of diversity and beat alignment. It is interesting to note that MLD, which is trained on a non-temporal latent representation, achieves a reasonable beat alignment but worse semantic recall. We hypothesize that the semantic alignment benefits from a finely discretized motion representation. Our method lies in the middle of the discretization spectrum, where CaMN operates on raw motion frames while MLD collapses the temporal axis within a single latent.

5.2. Dyadic Co-speech Gesture Synthesis

We adapted two baselines to the dyadic setting for comparison. MLD's architecture was extended by adding additional conditioning blocks of the co-participant's speech.

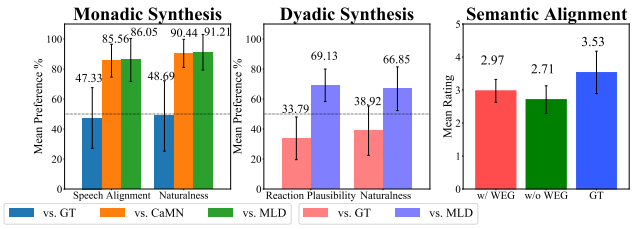


Figure 5. **Results of the user study.** We compare with CaMN [43] and MLD [13], and achieve an overall favourable preference scores for monadic and dyadic settings. We also evaluate the effectiveness of the word-excitation guidance (WEG).

	BeatAlign →	Diversity →	L1 Div →
GT	0.90	17.7	5.12
MLD	0.96	20	0.31
DiffGesture	0.97	2176	1308
Ours	0.90	6.38	1.19

Table 3. **Qualitative comparison of dyadic motion synthesis on the DND GROUP GESTURE dataset.**

Likewise, DiffuGesture was adapted to our setting as detailed in [72]. We observe similar patterns of jittery motion with DiffuGesture, whereas MLD produced suboptimal results in terms of beat alignment. In contrast, we achieve similar beat alignment as the ground-truth while also producing higher L1 Diversity, thus indicating non-static motions.

5.3. User Study

As noted above, evaluating motion synthesis models on a set of numerical metrics hides several aspects of the gesture synthesis. Prior works [14, 63] report mismatch between metrics and the subjective evaluations by the users. Hence, we perform a perceptual user study to evaluate the quality of our synthesis results w.r.t state-of-the-art methods. For evaluating the monadic results, we aim to evaluate the general plausibility of the motions and probe the coherence of the gestures with the utterance. Likewise for dyadic synthesis, the goal is to measure if participant's generated gestures align well with their speech as well as co-participant's speech content. To evaluate the word-excitation guidance, we ask the users to evaluate if the generated gestures have distinct gesticulation at the focus words.

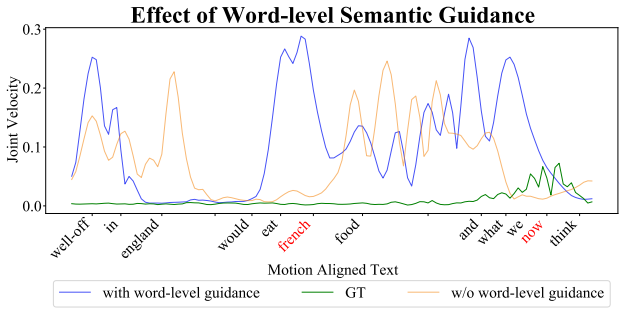
Results. We plot the results of our user study in Fig. 5. For the monadic setting, the participants preferred our motions over those of CaMN and MLD for both questions. At the same time, we were marginally below the ground-truth preference. The inference remains similar for the dyadic evaluations as well, although with significantly lower margins. Finally, the user study demonstrates better semantic alignment with the generated motions with the use of WEG.

5.4. Ablative Analysis

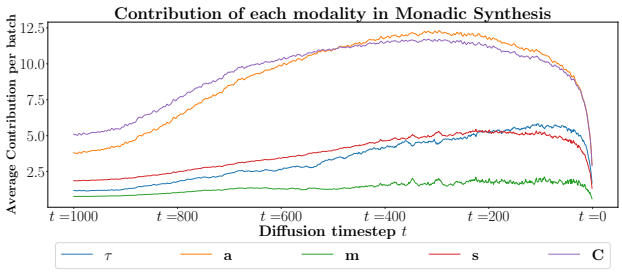
Latent Representation. Our chunked, scale-aware latent representation is motivated by various factors, such as per-

	Reconstruction Loss ↓	Smoothness Error [57] ↓
MLD [13]	10×10^{-3}	4.4×10^{-3}
Our VAE	5×10^{-3}	3.5×10^{-3}
w/o $\mathcal{L}_{lap}$	3×10^{-4}	3.7×10^{-3}
w/o Time Aware	9×10^{-3}	4×10^{-3}
w/o Scale Aware	5.5×10^{-3}	3.7×10^{-3}

Table 4. **Ablation study on the VAE design.** $\mathcal{L}_{lap}$ ensures the motions retain the velocity of ground truth, even though removing it leads to lower reconstruction loss. While training without time-aware representation gives slight increase in reconstruction loss, it cannot support unbounded generation.



(a) Fine-grained alignment



(b) The effect of modalities.

Figure 6. (a) Given a text prompt with focus words, "french" and "now", we observe that WEG significantly increases the joint velocities for the two words compared to the non-guided case. In (b) we show the contributions of each modality as diffusion denoising progresses. Audio tends to dominate the generation process.

petual motion synthesis, better temporal alignment with the conditioning modalities, and the scale difference between the hands and the fingers. We tabulate the influence of the three main design choices in Tab. 4. We also show that on the VAE reconstruction task alone, our latent representation outperforms MLD's latent representation.

Influence of Modalities. With a variety of conditioning modalities within our framework, it is natural to question which modalities bear a greater effect on the final generation. We analyze this by plotting the norm of the contributions of each modality in Eq. (5) (computed before scaling with w_c). As Fig. 6b demonstrates, the audio modality bears the largest influence on the gesture generation process. Interestingly, we notice an overall trend of increasing contributions until they drop down significantly towards the final stages of denoising, indicating that the diffusion process makes smaller edits in the final stages and takes heavier updates during the middle phases of denoising. In Fig. 6a, one can observe a significant bump in the joint velocities (indicating more animated behaviour) at the precise moment of the excitation word. These observations highlight the overall effectiveness of our two-level guidance objectives. We refer the reader to the supplemental for more results.

On Semantic Consistency: Thanks to the proposed Word Excitation Guidance (WEG), our method samples gestures that produce more pronounced attention features for the user-selected words We demonstrate this by training a gesture type classifier to recognize beat and semantic gestures. For synthesized gestures without WEG, we observe that the recall for semantic labels is **0.34**. However, this recall increased to **0.40** when WEG was employed, *indicating that the use of WEG enhances semantic coherence in generated gestures*. Refer to supplemental for implementation details.

Attention Maps. We visualize the attention maps for analysis (see supplemental) to interpret what spatio-temporal properties are highlighted in the model training. The first property is a clear separation between the hand and body latents, shown by the striped patterns of the attention maps. Secondly, WEG boosts the attention weights for the highlighted words. Refer to supplemental for detailed analysis.

Perpetual Rollout. In addition to allowing for temporal alignment with several modalities, our chunked latent representation also benefits us by allowing perpetual rollout. To do so, one can simply follow the auto-regressive denoising process followed by the existing motion diffusion methods [14, 61, 63] with the difference that instead of inpainting the actual motion, we inpaint the latents. Refer to supplementary material for implementation details.

6. Conclusion

In this work, we proposed a novel approach towards controllable co-speech gesture synthesis. With the aim of generating long term, jitter-free gestures, we proposed a time-aware latent representation that can be denoised using a diffusion model. To control the effects of individual modalities, we proposed a variant of classifier-free guidance. We also proposed WEG to enhance the gestures for a user-selected set of words in the text, thus facilitating text level fine-grained control. Our analysis shows that word-excitation induces more animated behaviour for the selected words. Finally, with the introduction of the DnD Group Gesture dataset we hope the field will further propel the research on multi-party gesture synthesis.

Acknowledgements. This work was supported by the ERC Consolidator Grant 4DReply (770784). We also thank Andrea Boscolo Camiletto & Heming Zhu for help with visualizations and Christopher Hyek for designing the game for the dataset.

References

[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. 2020. 3, 6

[2] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. *Comput. Graph. Forum*, 39(2):487–496, 2020. 3

[3] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM TOG*, 41(6):1–19, 2022. 3

[4] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM TOG*, 42(4): 1–18, 2023. 2, 3, 5

[5] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *ACM MM*, 2021.

[6] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021. 3

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 2008. 6

[8] Justine Cassell. Embodied conversational interface agents. *Commun. ACM*, 2000. 2

[9] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew 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*, 1994. 2

[10] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *SIGGRAPH Conference Proceedings*, 1994. 3

[11] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: The behavior expression animation toolkit. In *SIGGRAPH Conference Proceedings*, 2001. 3

[12] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM TOG*, 42(4), 2023. 6

[13] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023. 2, 3, 4, 5, 6, 7, 8

[14] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, 2023. 3, 4, 7, 8

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[16] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation, 2023. 6

[17] Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018. 3

[18] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 89:117–130, 2020. 3

[19] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. *Comput. Graph. Forum*, 42(1):206–216, 2023. 3

[20] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum*, 42(1):206–216, 2023. 3, 6

[21] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)*, 2021. 4

[22] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023. 4

[23] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. 3

[24] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the International Conference on Intelligent Virtual Agents*, 2021. 2, 3, 6

[25] Ikhsanul Habibie, Mohamed Elgharib, Kripashindu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *SIGGRAPH '22 Conference Proceedings*, 2022. 3

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3, 5

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 5, 6

[28] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart

Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 6

[29] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004. 1, 3

[30] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4

[31] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2021. 3

[32] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents*, 2006. 2

[33] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020. 3

[34] Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. Speech2Properties2Gestures: Gesture-property prediction as a tool for generating representational gestures from speech. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*, 2021. 3, 6

[35] Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. Speech2properties2gestures: Gesture-property prediction as a tool for generating representational gestures from speech. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 2021. 3

[36] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis, 2023. 5

[37] G. Lee, Z. Deng, S. Ma, T. Shiratori, S. Srinivasa, and Y. Sheikh. Talking with hands 16.2m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3, 6

[38] Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. *ACM TOG*, 28(5):1–10, 2009. 3

[39] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. *ACM TOG*, 29(4):1–11, 2010. 3

[40] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *ICCV*, 2021. 3

[41] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. Seeg: Semantic energized co-speech gesture generation. In *CVPR*, 2022. 3

[42] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *ACM MM*, 2022. 3

[43] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *European Conference on Computer Vision*, 2022. 2, 3, 6, 7

[44] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. In *NeurIPS*, 2022. 3

[45] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *CVPR*, 2022. 3

[46] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 5

[47] Lars Marstaller and Hana Burianová. The multisensory perception of co-speech gestures–a review and meta-analysis of neuroimaging studies. *Journal of Neurolinguistics*, 30:69–77, 2014. 2

[48] David McNeill. *Gesture and thought*. University of Chicago press, 2008. 3

[49] Angeliki Metallinou, Zhaojun Yang, Chi-chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. The usc creativeit database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Language resources and evaluation*, 50:497–521, 2016. 6

[50] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff. A comprehensive review of data-driven co-speech gesture generation. *Comput. Graph. Forum*, 42(2):569–596, 2023. 3

[51] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *CVPR*, 2024. 1

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2022. 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2021. 3

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 4, 5

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed

Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv*, 2022. 3

[57] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 39, 2020. 8

[58] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation via actor-critic gpt with choreographic memory. In *CVPR*, 2022. 7

[59] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[60] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *ICCV*, 2023. 3

[61] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. In *ICLR*, 2023. 3, 5, 8

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 1

[63] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *CVPR*, pages 448–458, 2023. 3, 7, 8

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5

[65] Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014. 3

[66] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. In *IJCAI*, 2023. 3

[67] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3D human motion from speech. In *CVPR*, 2023. 2, 3

[68] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk 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)*, 2019. 3

[69] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proc. of The International Conference in Robotics and Automation (ICRA)*, 2019. 2, 3, 6

[70] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM TOG*, page 1–16, 2020. 3, 6, 7

[71] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 3

[72] Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. Diffugesture: Generating human gesture from two-person dialogue with diffusion models. In *International Conference on Multimodal Interaction*, 2023. 6, 7

[73] Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. Diffugesture: Generating human gesture from two-person dialogue with diffusion models. In *International Cconference on Multimodal Interaction*, pages 179–185. 2023. 2

[74] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 1

[75] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, 2023. 2, 3, 6, 7