This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Retrieving Semantics from the Deep: an RAG Solution for Gesture Synthesis

M. Hamza Mughal[1] Rishabh Dabral[1] Merel C.J. Scholman[3]
Vera Demberg[1,2] Christian Theobalt[1,2]
[1]Max Planck Institute for Informatics, SIC [2]Saarland University [3]Utrecht University

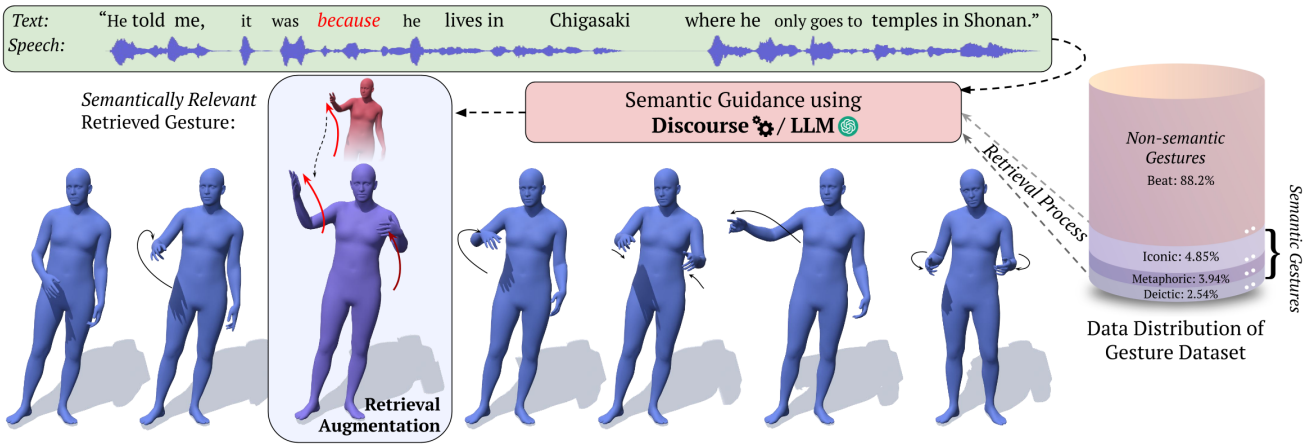


Figure 1. Our RAG-GESTURE approach produces **semantically meaningful gestures** by leveraging explicit knowledge to retrieve exemplar gestures from the sparse semantic data [34] and **guiding the diffusion-based generation process** through **Retrieval Augmentation**.

Abstract

Non-verbal communication often comprises of semantically rich gestures that help convey the meaning of an utterance. Producing such semantic co-speech gestures has been a major challenge for the existing neural systems that can generate rhythmic beat gestures, but struggle to produce semantically meaningful gestures. Therefore, we present RAG-GESTURE*, a diffusion-based gesture generation approach that leverages Retrieval Augmented Generation (RAG) to produce natural-looking and semantically rich gestures. Our neuro-explicit gesture generation approach is designed to produce semantic gestures grounded in interpretable linguistic knowledge. We achieve this by using explicit domain knowledge to retrieve exemplar motions from a database of co-speech gestures. Once retrieved, we then inject these semantic exemplar gestures into our diffusion-based gesture generation pipeline using DDIM inversion and retrieval guidance at the inference time without any need of training. Further, we propose a control paradigm for guidance, that allows the users to modulate the amount of influence each retrieval insertion has over the generated sequence. Our comparative evaluations demonstrate the validity of our approach against recent gesture generation approaches. The reader is urged to explore the results on our project page.*

1. Introduction

Human communication is a complex and multifaceted process that involves both verbal and non-verbal elements. Non-verbal communication comprises *co-speech gestures*, which are defined as body and hand movements that are temporally aligned and semantically integrated with the speech [38]. Gestures convey information in tandem with speech and language, and can carry additional meaning that enhances the semantic construct of the message. Generating meaningful full-body gestures is therefore important for the communicative efficacy of virtual humans in telepresence and content-creation domains.

McNeill [38, 39] categorizes gestures into *beat gestures*, which are rhythmic movements driven by prosody, and *semantics-driven gestures*, which carry the communicative intent. Semantic gestures are context-driven and bear a specific meaning that complements the utterance [6, 38]. Rule-based approaches retrieve these semantic gestures and combine them for generation, while neural methods leverage the data to learn the synthesis process. However, the former tends to result in an unnatural outcome and the latter struggles to generate semantic gestures because these gestures, while being a part of the existing datasets, occur rather sporadically compared to the rhythmic beat gestures [47].

In this work, we address this issue by framing the problem as a Retrieval Augmented Generation (RAG) task. We

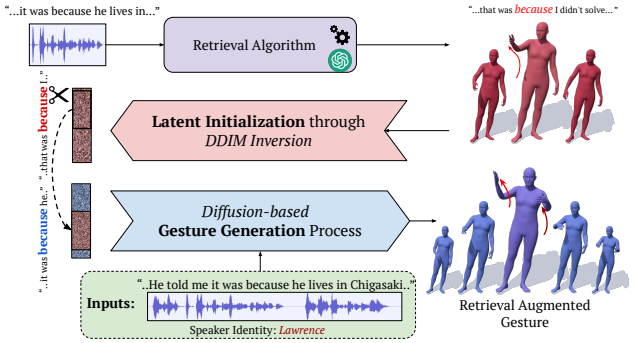


Figure 2. **Overview.** Our approach retrieves example gestures on semantically important words in speech and inserts those examples into the generated gesture by using them to guide the generation.

propose RAG-GESTURE – a diffusion-based gesture synthesis approach that generates natural and semantically rich gestures by retrieving context-appropriate exemplars from a database and injecting them into the gesture generation process. Motivated by the insights from Neff [44], we decompose the generation problem into two tasks: *specification*, which ascertains "what gesture" to produce for "which word" in the speech, and *animation*, which determines "how to generate" that specified gesture.

To solve the specification problem, we explicitly retrieve the relevant exemplar motions from a gesture database using either of the two retrieval algorithms proposed in Sec. 3.4. More specifically, we show that the chain-of-thought reasoning abilities of Large Language Models [48, 67] can be exploited to extract which parts of the utterance would likely be semantically gestured, and what 'type' of semantic gesture (*iconic, deictic, metaphoric, etc.*) the phrase would invoke. We also demonstrate that one could, alternatively, use specific linguistic elements to base the retrieval on. In particular, we focus on discourse connectives such as *because*, *while*, and *on one hand*, which have been shown to affect the gestural patterns [9, 40].

The animation problem, in our context, requires integrating the retrieved gesture motions (corresponding to a subset of the sentence) into the overall motion generated for the full sentence/utterance (see Fig. 1). To this end, we design an inference-time approach to augment the gesture generation capability of diffusion models with retrieved gesture clips. Our approach allows us to surgically integrate the semantic gestures into the generation process of the base model, while not affecting the synthesis for the remaining parts of the utterance.

Through comprehensive numerical and perceptual evaluations, we demonstrate the effectiveness of our proposed neuro-explicit approach compared with purely neural state-of-the-art approaches like [34, 35, 46, 63] that typically struggle to synthesize semantically meaningful gestures. To summarize, our core contributions are as follows:

- We introduce RAG-GESTURE – a diffusion-based approach which leverages RAG to generate natural-looking and semantically rich gestures by injecting meaningful motion exemplars into the gesture generation process.
- We introduce multiple retrieval algorithms to extract semantically relevant exemplars from the sparse semantic data distribution by leveraging either the linguistic structure of utterances or LLM-based gesture type prediction.
- Further, we propose a novel inference-time RAG approach for our framework and demonstrate how gesture exemplars can be transferred onto the generated motion locally through the combination of Latent Initialization and Retrieval Guidance technique, which provides a way to control the influence of the retrieval insertion.

2. Related Work

We first provide an overview of gesture synthesis approaches, followed by a discussion of relevant approaches from RAG literature. Lastly, we discuss recurrent patterns of gestures used to express discourse structure, which motivates the design of the retrieval-based method.

2.1. Co-Speech Gesture Synthesis

Co-speech gestures consist of hand and arm movements that occur in sync with speech and help convey meaning and structure discourse. Earlier methods in co-speech gesture generation [10, 56] perform well on semantics and appropriateness of gestures [47]. These systems retrieve the most appropriate gesticulation units for given speech, resulting in high semantic alignment. However, since they insert predefined units into animation, the synthesis quality of these systems tends to be unnatural. To mitigate the lack of naturalness, recent learning-based methods [2, 16, 20, 30] employ deep networks to convert speech input to gesture output and aim to ensure smooth motions. These methods are mostly data-driven and rely on learning speech-to-gesture matching through large scale gesture datasets [15, 18, 34, 60]. Most of these recent methods [1, 30, 59, 61] suffer from reduced communicative efficacy and low semantic alignment, due to bad generalization to semantic co-speech gestures in the dataset. Furthermore, generation and reproduction of semantically appropriate gestures corresponding to speech still remains a challenge for these systems [62].

A few recent approaches directly addressed this problem. Kucherenko *et al.* [31] use speech to predict the gesture properties, which inform the model on what type of gesture to generate. Instead, our framework proposes to use RAG to follow the gesture type information instead of taking the properties as conditioning input. Mughal *et al.* [42] generate semantic gestures by providing word-level semantic control and use it to sample semantically meaningful gestures. Zhi *et al.* [68] initialize diffusion-based generation with text-aligned embeddings by adding noise to them (sim-

ilar to inpainting [36]). In contrast, our approach inverts semantic exemplars into the diffusion latent space and uses them to initiate the denoising process. SemanticGesticulator [65] is a generative framework which retrieves semantic gestures from a curated motion library and aligns them with rhythmic gestures through a training-based approach. In comparison, our framework performs inference-time RAG using a diffusion model rather than fine-tuning the gesture generator to follow the retrieval. Secondly, our approach also utilizes prosody information for every retrieval instead of only using textual transcription. Thirdly, instead of relying on a curated library, our approach introduces explicit algorithms to retrieve semantics from the existing data [34]. Lastly, our framework is flexible enough to leverage any retrieval algorithm at inference time.

2.2. Retrieval Paradigms for Generative Methods

Employing database retrieval for improved animation performance is a prevalent paradigm in motion and gesture synthesis. Classical [8, 11] and learning-based motion matching approaches [24, 25] have been applied for better character control in motion synthesis [53]. Diffusion-based text-to-motion synthesis frameworks also deal with the problem of semantic control at a global level [63] or at the local keyframe level [19, 26]. Zhang *et al*. [63] apply RAG to diffusion-based text-to-motion synthesis using global text similarity for retrieval. Unlike text-to-motion, gesture generation is affected by language content and prosody at varying levels of semantic granularity. That means one can neither perform global semantic matching between the retrieved and generated motion [63] nor simply copy and paste retrieved motion parts onto the keyframe locations [19] without considering the differences in context.

Gesture synthesis approaches have also utilized database retrieval both in classical and learning based paradigms. Early approaches [5, 29, 45] utilize statistics to retrieve gestures from a database of gesticulation units. A few learning-based approaches combine rule-based retrieval and deep learning-based synthesis to improve upon the semantic quality of gesture generation. ExpressGesture [17] introduces a database-driven framework to ensure that generated gestures retain the expressive and defined gesture form. Habibie *et al*. [21] employ nearest neighbor search to retrieve most appropriate gestures and provide a style-control mechanism over the generative framework. In comparison, our retrieval approach is not limited to style or select keywords but is grounded in explicit rules for semantic retrieval that are driven from linguistic and gestural structure.

2.3. The effect of Linguistic Dynamics on Gestures

The goal of our retrieval approach is to extract semantic gestures from the database. Our proposed design is therefore motivated by the relation between language dynamics and co-speech gestures. Laparle [32] demonstrates how gesture features like hand shape or orientation are important to distinguish between topics in discourse. Gestures can also indicate the relations between different parts of the message. For example, a contrast relation can be expressed by the lexical marker *on the one hand...on the other hand*, framing the two contrasting arguments as items placed on separate hands [22]. The raised index finger draws the attention of the interlocutor to new and important topics, and is commonly used to express exceptions or concessions (which can be expressed linguistically using *however*) [7, 27, 28].

In sum, gestures can be used to convey the semantic structure of a message, but existing generation works fail to reproduce these context-driven patterns [62], which results in repetitive gestures. We attempt to leverage the linguistic information in our retrieval framework. This framework can, in turn, improve semantic grounding of generated gestures through RAG based insertion.

3. Approach

Our primary goal is to generate semantically meaningful co-speech gesture sequences that are consistent with the content of the corresponding speech. This goal can be partly achieved by our latent-diffusion based gesture generation model, which is trained to synthesize gestures while being conditioned on the corresponding modalities (refer to Sec. 3.1). This data-driven gesture generation approach is capable of generating natural looking and beat-aligned gesture sequences from the input speech.

However, this basic generation framework is not sufficient to exhibit rich semantics in the generated gestures. Therefore, we propose additional steps that ensure the solution to two key problems – *animation* (refer to Sec. 3.2 and Sec. 3.3) and *specification* (refer to Sec. 3.4) of semantically rich gestures [44]. Our approach is illustrated in Fig. 3.

3.1. Gesture Generation with Latent-Diffusion

As the first step, we build a robust generation framework based on latent-diffusion modeling [50], which serves as our base model to generate plausible co-speech gesture sequences. The gesture sequence $\mathbf{x} \in \mathbb{R}^{N \times d_{\mathbf{x}}}$ consists of N frames of human motion with $d_{\mathbf{x}}$ representing a combination of J full body joints represented using the 6D rotation representation of [69], FLAME model parameters for face motion, root translation and foot contact labels [35]. Instead of training our diffusion framework on raw motion $\mathbf{x}$, we encode motion using VAE encoders, which provide separate encodings for different body parts [42]. These part-wise VAE encodings are then used to train our conditional latent-diffusion model that generates co-speech gestures.

Decoupled Gesture Encoding. Based on the observations that each body region has different relations with

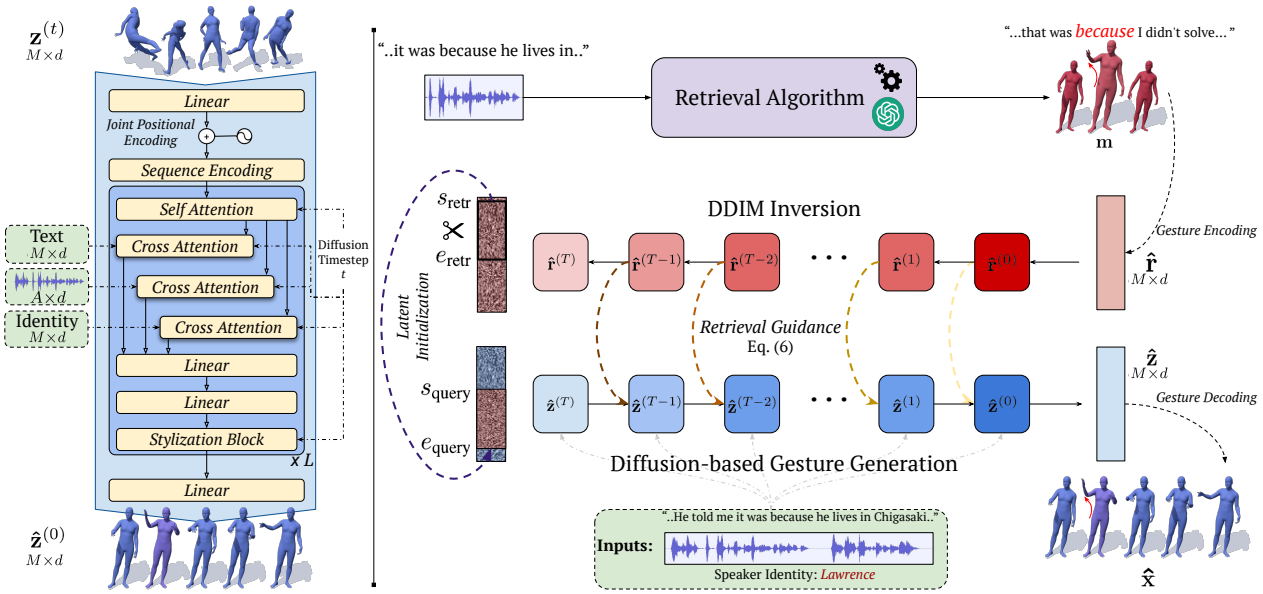


Figure 3. **RAG-GESTURE Framework.** Our approach leverages a diffusion model which predicts clean sample $\hat{\mathbf{z}}^{(0)}$ from noisy gesture sample $\mathbf{z}^{(t)}$. We then utilize retrieval algorithms (Sec. 3.4) to modify the gesture sampling at inference time by inserting the retrieved motion through Latent Initialization (Sec. 3.2) and further controlling the sampling process through Retrieval Guidance (Sec. 3.3). This results in a sampled motion which follows the semantic retrieval.

speech [35] and scale differences between them affect the generation quality [42], we decouple the gesture sequence $\mathbf{x}$ into four different body regions: $\mathbf{x_u} \in \mathbb{R}^{N\times 6J_u}$, $\mathbf{x_h} \in \mathbb{R}^{N\times 6J_h}$, $\mathbf{x_f} \in \mathbb{R}^{N\times 100}$ and $\mathbf{x_l} \in \mathbb{R}^{N\times(6J_l+3+4)}$ for upper body, hands, face and a combined representation for lower body and translation respectively. We train separate time-aware VAEs [42] for each body region, which can encode and decode the motion for each region: $\mathbf{z}_i = \xi_i(\mathbf{x}_i)$, $\mathbf{x}'_i = \mathcal{D}_i(\mathbf{z}_i)$. These VAEs are trained using a combination of standard reconstruction/geometric losses for motion and the KL-Divergence loss for the VAE latent space (see supplemental). The resulting gesture representation is a set of decoupled encodings: $\mathbf{z} = \{\mathbf{z_u}, \mathbf{z_h}, \mathbf{z_f}, \mathbf{z_l}\}$. This set can be concatenated to form a gesture representation $\mathbf{z} \in \mathbb{R}^{M\times d_\mathbf{z}}$, where $M < N$ is the length of time-compressed encoding corresponding to each $\mathbf{x}_i$ and $d_\mathbf{z}$ is embedding length of the representation. Refer to the supplemental for more details.

Conditional Diffusion for Gesture Generation. Once we have obtained $\mathbf{z}$ as our gesture representation, we can frame the gesture synthesis task as that of conditional diffusion [23, 50], where the conditioning set $\mathbf{C}$ comprises of audio, text and speaker embeddings. In this framework, the *forward diffusion* process consists of Markovian chain of successive noising steps, where Gaussian noise $\boldsymbol{\epsilon}$ is added to the encoded gesture $\mathbf{z}^{(0)}$ for T timesteps until $\mathbf{z}^{(T)} \sim \mathcal{N}(0, \mathbf{I})$. For gesture generation, the *reverse process* is employed which iteratively denoises $\mathbf{z}^{(T)}$ to generate the gesture representation $\hat{\mathbf{z}}^{(0)}$. The process utilizes a neural network $f_\theta(\mathbf{z}^{(t)}, t, \mathbf{C})$ for denoising, that is trained to predict clean state $\hat{\mathbf{z}}^{(0)}$ with the following objective:

$$\min_\theta \mathbb{E}_{t\sim[1,T],\mathbf{z}^{(0)},\boldsymbol{\epsilon}} ||\mathbf{z}^{(0)} - f_\theta(\mathbf{z}^{(t)}, t, \mathbf{C})||_2^2 \quad (1)$$

During inference, we utilize DDIM [52] sampling for efficient generation of the gesture representation i.e. we generate a sample $\hat{\mathbf{z}}^{(t-1)}$ from $\hat{\mathbf{z}}^{(t)}$ using:

$$\hat{\mathbf{z}}^{(t-1)} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{z}}^{(0)} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\boldsymbol{\epsilon}_\theta(\hat{\mathbf{z}}^{(t)}) + \sigma_t\boldsymbol{\epsilon}^{(t)} \quad (2)$$

Here $\bar{\alpha}_t$ controls rate of diffusion and σ_t controls stochasticity such that $\sigma_t = 0$ results in deterministic sampling.

Our denoising network consists of a transformer decoder network with L decoder layers containing a separate cross-attention head for each modality from the conditioning set $\mathbf{C}$. The resulting activations from each head are combined using a linear layer in each decoder layer. Diffusion timestep t is passed to the decoder layers through Stylization block [64] after every self-attention, cross-attention and linear layer. We encode speech signal using wav2vec2 [4] and construct frame-aligned representation for transcription using BERT embeddings for each word [12]. Additionally, we create identity-specific speaker embeddings for input.

3.2. Latent Initialization through DDIM Inversion

So far, our base neural generation framework is purely data-driven and lies on one end of neural-to-explicit spectrum. It resembles the majority of state-of-the-art gesture generation frameworks [3, 46, 66], which rely on learned pat-

terns to sample relevant gestures for corresponding speech and struggle to sample the semantic ones. On the other hand of the spectrum, pure retrieval based approaches can naïvely "paste" raw exemplar motions during the generation process which hurts the naturalness of motion. For instance, one could simply add noise to the retrieval and generate the rest through diffusion-based sampling [36], which forces the generation to exactly follow the retrieval, achieving subpar results (see Sec. 4.3 for analysis). Therefore, we carefully infuse the retrieved gestures into the inference process by performing the retrieval transfer in the latent space of the trained diffusion model and provide a better sampling path for its generation through retrieval guidance (Sec. 3.3). This preserves the quality of the base model, adapts it to generate semantically rich gestures and provides further control over the influence of retrieval insertion.

In our framework, all retrieval algorithms (discussed later in Sec. 3.4) must output an exemplar motion $\mathbf{m}$, which can be encoded as the *retrieval* $\mathbf{r} \in \mathbb{R}^{M \times d}$ through our gesture encoder (Sec. 3.1). We also have access to the text and audio corresponding to the retrieved motion. Our goal is to extract the motion chunk for the retrieved semantic gesture around a word of interest (e.g. 'because' in Fig. 3) and insert it into the motion around the corresponding words in the input text. Specifically, let the retrieved motion contain the semantically relevant gesture chunk in the frame window $(s_{\text{retr}}, e_{\text{retr}})$ for the marked word. The gesture in this window must be transferred onto the corresponding word location $(s_{\text{query}}, e_{\text{query}})$ in the to-be-generated sequence, also referred to as *query* $\hat{\mathbf{z}}$ (see Fig. 3).

Here, the challenge is *how* to insert the retrieved gesture chunk such that it is seamlessly integrated into the final inference output in terms of temporal and semantic alignment. We perform this insertion through a combination of DDIM [13, 41, 52] inversion based latent initialization and a retrieval guidance objective. DDIM inversion lets us represent the retrieval $\mathbf{r}$ into the latent space of our base diffusion model, which we then use to initialize our generation.

By reversing Eq. (2) and keeping $\sigma_t = 0$ for deterministic sampling path, we can invert $\mathbf{r}^{(0)}$ as follows:

$$\hat{\mathbf{r}}^{(t+1)} = \sqrt{\bar{\alpha}_{t+1}}\hat{\mathbf{r}}^{(0)} + \sqrt{1-\bar{\alpha}_{t+1}}\boldsymbol{\epsilon}_\theta\left(\hat{\mathbf{r}}^{(t)}\right) \tag{3}$$

This iterative process inverts the clean $\mathbf{r}^{(0)}$ to full noise $\hat{\mathbf{r}}^{(T)}$, which represents retrieved gesture in the diffusion latent space and provides the initial starting point for DDIM sampling such that it ends up at $\hat{\mathbf{r}}^{(0)} \approx \mathbf{r}^{(0)}$.

To transfer the retrieved gesture onto the query gesture generation, we slice the time axis of the latent $\hat{\mathbf{r}}^{(T)}$ to extract the relevant parts using the retrieval window $[s_{\text{retr}} : e_{\text{retr}}]$ and then, place these parts onto the query window in $\hat{\mathbf{z}}^{(T)}$.

$$\hat{\mathbf{z}}^{(T)}[s_{\text{query}} : e_{\text{query}}] \leftarrow \hat{\mathbf{r}}^{(T)}[s_{\text{retr}} : e_{\text{retr}}] \tag{4}$$

The retrieval augmented $\hat{\mathbf{z}}^{(T)}$ can be represented as $\hat{\mathbf{z}}^{(T)}_{\text{retr}}$.

3.3. Retrieval Guidance

As a result of latent initialization step, $\hat{\mathbf{z}}^{(T)}_{\text{retr}}$ comes out to be the starting point for gesture sampling process, which can be used to generate a sequence that follows the retrieved semantic gesture around the marked word. However, there is still no control over how strongly the retrieval is followed and how the diffusion sampling process goes from $\hat{\mathbf{z}}^{(T)}_{\text{retr}}$ to $\hat{\mathbf{z}}^{(0)}_{\text{retr}}$. Therefore, we present *Retrieval Guidance* mechanism, that provides modulation control over the retrieval-augmented diffusion generation process from $\hat{\mathbf{z}}^{(T)}_{\text{retr}}$ to $\hat{\mathbf{z}}^{(0)}_{\text{retr}}$. We utilize the sequence of inverted diffusion latents $(\hat{\mathbf{r}}^{(0)}, \hat{\mathbf{r}}^{(1)}, \ldots, \hat{\mathbf{r}}^{(T-1)}, \hat{\mathbf{r}}^{(T)})$ and form our guidance objective. At a given diffusion timestep t, guidance objective and its corresponding latent update are as follows:

$$G_{retrieval} = \left\| \hat{\mathbf{z}}^{(t)}_{\text{retr}}[s_{\text{query}} : e_{\text{query}}] - \hat{\mathbf{r}}^{(t)}[s_{\text{retr}} : e_{\text{retr}}] \right\|_2^2 \tag{5}$$

$$\tilde{\mathbf{z}}^{(t)}_{\text{retr}} \leftarrow \hat{\mathbf{z}}^{(t)}_{\text{retr}} - \lambda \nabla_{\hat{\mathbf{z}}^{(t)}_{\text{retr}}} G_{retrieval} \tag{6}$$

The updated $\tilde{\mathbf{z}}^{(t)}_{\text{retr}}$ can be used further in the diffusion sampling by acting as input to the denoising network $f_\theta(\tilde{\mathbf{z}}^{(t)}, t, \mathbf{C})$. This guidance paradigm ensures that latent $\hat{\mathbf{z}}^{(t)}_{\text{retr}}$ follows $\hat{\mathbf{r}}^{(t)}$ at the generation timestep t.

To control the amount of guidance at each step, one can control the number of iterations for latent update at each diffusion timestep t. As an example, the generation from $\hat{\mathbf{z}}^{(T)}_{\text{retr}}$ without any guidance will be the weakest form of retrieval augmentation where gestures are constrained to match only at the starting timestep $t = T$. On the other hand, if one were to make sure retrieval is followed to its strongest extent, $\tilde{\mathbf{z}}^{(t)}_{\text{retr}}$ can be updated at each diffusion timestep t such that $G_{retrieval} \to 0$. Similarly, one can perform more latent updates when t is closer to T and lesser or no updates when $t \to 0$, in order to only perform guidance at the start of the sampling process. Refer to Sec. 4.3 for analysis.

3.4. Retrieval for Semantic Gesture Extraction

Having discussed *how* to insert a retrieved gesture chunk into the gesture generation process, we now discuss retrieval algorithms that utilize linguistic context to determine *which* words in the query utterance are likely to invoke semantic gestures and *what* are those gestures. These algorithms retrieve semantically relevant gestures using domain knowledge of language and gesture type information. However, our approach is not limited to these algorithms and it can be extended using any retrieval paradigm for gestures.

At the core, each retrieval algorithm ranks examples in a database (the BEAT2 dataset [35] in our case) according to a specific focus. During ranking, some of the steps are common across algorithms and are driven by overarching themes in gesture understanding. For example, motion exemplars from the same speaker typically have less

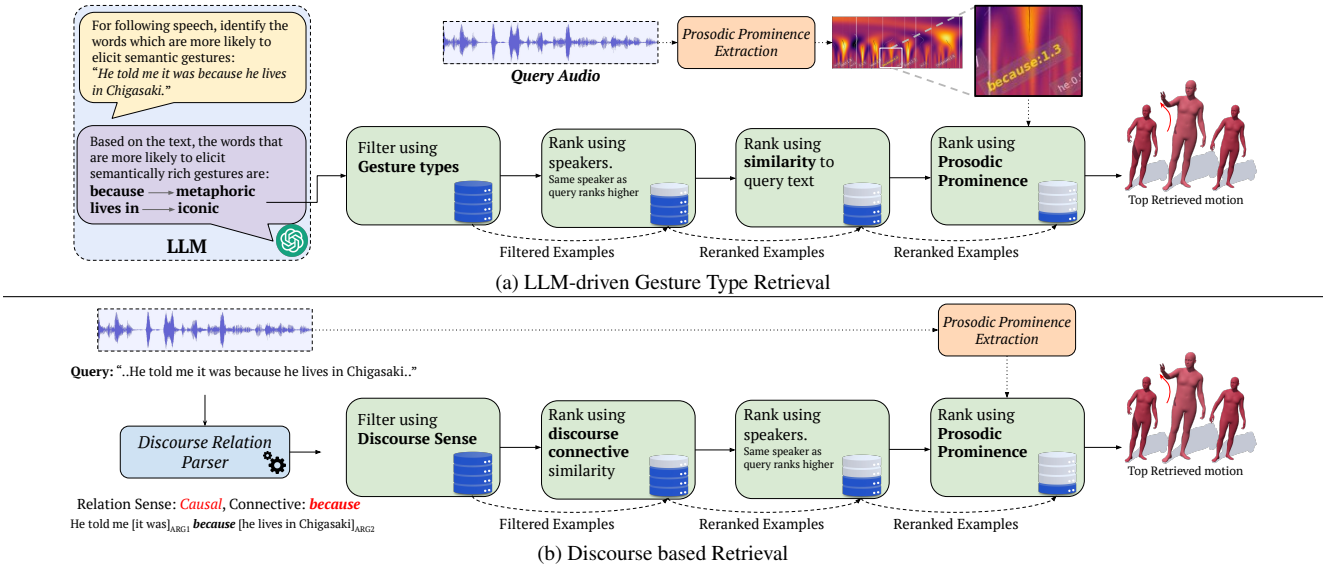


Figure 4. **Retrieval Algorithms.** Each algorithm parses the relevant semantic information (gesture types from LLM or discourse relations) and extracts gestures from a database by filtering examples using that information. Moreover, it also considers textual and prosodic context.

speaker-dependent variation and should be ranked higher. Another common step across algorithms is ranking based on prosodic prominence extraction [14]. Certain elements of speech typically stand out due to variations in pitch, loudness, or duration, referred to as *prosodic prominence* [55, 57]. Prosodic prominence and gestures are closely connected: co-speech gestures often align with prosodically stressed parts of speech [49]. Therefore, we identify utterances in the database that match the prominence values of the marked word in the input speech [54]. An overview of individual steps is provided in Fig. 4.

LLM-based Gesture Type Retrieval. One of the commonly used classification for semantic gestures divides them into *iconic*, *metaphoric* and *deictic* gestures [38]. Kucherenko *et al*. [31] demonstrate the benefit of utilizing these gesture properties in the synthesis framework, as gesture type information can decouple the specification problem from the animation. Therefore, we leverage the reasoning abilities of an LLM [48] and prompt it to identify words which may elicit a semantically meaningful gesture. We also use it to predict the gesture type on each predicted word. This results in a word-to-gesture type mapping, which we can use to retrieve examples from the database that contains type labels. Details on prompts and LLM are given in the supplemental material.

Given the word-to-gesture type mapping, we take an identified word and its corresponding gesture type and filter examples in the database based on gesture type. We leverage the semantic gesture type annotations in the BEAT2 dataset [35] for this task. We then perform ranking based on speaker similarity, followed by ranking according to the text feature similarity between the query and database examples in order to match the semantic context. Lastly, we re-rank the top semantically similar examples based on prosodic prominence values. Fig. 4a illustrates the steps involved.

Discourse-based Retrieval. A text has a structure that organizes its information into a coherent flow. This structure is achieved through *discourse relations*, semantic-pragmatic links such as CAUSE-CONSEQUENCE and CONTRAST, which hold between clauses and sentences [37, 51]. These relations can be signaled by lexical cues referred to as connectives, such as *on the other hand* or *because*. For example, in an utterance from BEAT2 dataset [35]: *"I'll go shopping **if** I'm not that tired"*, the word *"if"* marks the conditional relation between two sentence segments. Prior works have demonstrated that discourse connectives have an effect on the gestural patterns [9, 40].

We leverage connectives to retrieve the relevant semantic gestures that co-occur with specific discourse relations. However, connectives can be ambiguous. For example, the connective *since* can signal a causal relation or a temporal one. We therefore retrieve the gesture co-occurring with a connective with the same meaning, but not necessarily the same word. For example, we can retrieve a gesture co-occurring with *because* to inform the generation of a gesture co-occurring with a causal *since*. Additionally, it allows us to retrieve more samples for less frequent connectives.

In sum, we find examples in the database that carry the same discourse relational sense as the query and rank again based on the word similarity of the connective. We conclude the algorithm with speaker-based and prosodic similarity ranking. The steps are shown in Fig. 4b.

	1 Speaker				All Speakers			
	FID↓	BeatAlign→	L1Div→	Diversity→	FID↓	BeatAlign→	L1Div→	Diversity→
GT		0.703	11.97	127		0.477	7.29	110
CaMN [34]	0.604	0.711	9.97	107	0.512	0.200	5.58	98
EMAGE [35]	0.570	0.793	11.41	124	0.692	0.284	6.06	88
Audio2Photoreal [46]	1.02	0.550	12.47	145	0.849	0.326	6.24	99
ReMoDiffuse [63]	0.702	0.824	12.46	123	1.120	0.218	5.06	116
Ours (w/ No RAG)	0.911	0.727	12.78	130	0.519	0.447	8.64	112
Ours (w/ Discourse)	0.879	0.730	12.62	129	0.447	0.471	9.03	114
Ours (w/ LLM & Gesture Type)	0.808	0.734	11.97	122	0.487	0.514	9.94	118

Table 1. **Comparison with state-of-the-art methods trained on BEAT2.** We demonstrate superior performance, especially when generalizing across multiple speaker identities.

4. Experiments

We first compare our approach with neural state-of-the-art baselines through quantitative (Sec. 4.1) and perceptual evaluation (Sec. 4.2). Further, we demonstrate how our RAG approach improves upon the training-based RAG baselines. Lastly, we validate our method design through extensive ablations and user studies (Sec. 4.3).

Specifically, we compare with recent data-driven gesture generation approaches that are CaMN [34] (LSTM-based), EMAGE [35] (transformer-based) and Audio2Photoreal [46] (diffusion-based). Furthermore, to validate RAG performance for gesture synthesis, we not only compare our method with ReMoDiffuse [63] (re-trained for gesture synthesis), but we also analyze it further by tailoring their approach and training it with our retrieval algorithms. We provide seed motions for baselines [34, 35], but do not use seed motions to generate our results.

Evaluation Dataset. We evaluate our performance on train/val/test split from BEAT2 dataset [35], which contains 25 speakers. Unlike baseline approaches for BEAT2 which report results only for a set of speakers, we perform evaluation on the test set of one speaker (*Scott*) and all speakers in order to evaluate performance on large-scale multi-speaker data. Our test set contains 265 utterances chunked into 10-sec sequences. We retrain methods on BEAT2 if needed.

4.1. Quantitative Evaluation

We evaluate our method on established metrics: Beat-Alignment [33], FID [60], L1 Divergence and Diversity, which measure different aspects of the motion quality. We tabulate the quantitative results in Tab. 1, where the results are divided on the basis of number of speakers used for training and evaluation. We observe that our approach achieves state-of-the-art metric performance when trained with all speakers. This also entails that our RAG-based diffusion approach generalizes and scales well across large-scale data with multiple speaker identities. For the case of single speaker training, we observe best performance in terms of diversity and L1 Divergence and second-best in Beat Alignment score. However, our diffusion-based method has higher FID compared to transformer and LSTM based methods [34, 35] that utilize ground-truth seed and can be over-fitted to a small-scale data of single speaker, while not being able to generalize to all speakers.

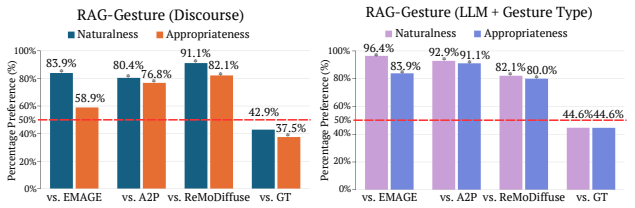


Figure 5. **Results of Perceptual Evaluation.** A2P: Audio2Photoreal [46] GT: Ground Truth. * denotes p-value < 0.05

4.2. Perceptual Evaluation

Evaluating gesture synthesis performance is a highly challenging task due to the stochasticity in gesture production and subjective nature of gesture perception [47, 58]. Therefore, we perform four different user studies for comparison and ablative analysis (Sec. 4.3). First, we evaluate the synthesis quality compared to the baseline methods. We investigate the quality of gestures based on well-established measures of *naturalness* and *appropriateness* in relation to speech [43]. We ask the participants to perform pairwise comparisons between generations from baselines and our method. Results compare RAG-GESTURE with discourse and LLM-based retrievals against state-of-the-art methods (Fig. 5). We observe that our method achieves better perceptual quality for both naturalness and appropriateness. For LLM based approach, we are marginally below the ground-truth preference, underscoring that our method produces not only natural, but also highly appropriate gestures.

4.3. Ablative Analysis

Improvement over RAG Baselines. ReMoDiffuse [63] uses global text similarity for retrieval and trains the diffusion network to follow it. In contrast, we design our retrieval to match local context around the semantically important word and do not require training/fine-tuning for RAG unlike existing approaches [63, 65].

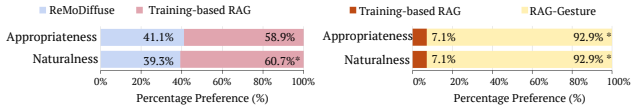


Figure 6. **Perceptual Comparison** between training-based baselines and our inference-time approach. *: p-value < 0.05

	FID↓	BeatAlign→	L1Div→	MPJPE (mm)↓
GT		0.477	7.29	
ReMoDiffuse [63]	1.120	0.218	5.06	200.5
Training-based RAG	0.525	0.414	6.91	34.6
RAG-GESTURE	0.447	0.471	9.03	35.0

Table 2. Quantitative Comparison between RAG baselines.

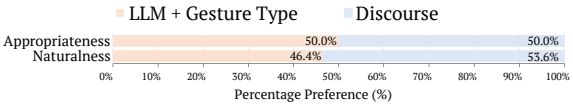


Figure 7. User Study comparing **LLM and Discourse** algorithms.

Besides comparing these differences with ReMoDiffuse, we also evaluate RAG-GESTURE (Discourse) against a training-based approach where we incorporate our discourse-based retrievals with ReMoDiffuse's generator through the local retrieval merging strategy of SemanticGesticulator [65] (see supplementary for details). Note that this architecture is trained with encoded gestures instead of raw motion and is similar to SemanticGesticulator [65] at a higher level, which also performs retrieval merging with encoded gestures during fine-tuning.

We evaluate this quantitatively in Tab. 2 and perform an additional user study (Fig. 6). Moreover, we also report MPJPE between the generated gesture and retrieved motion during the retrieval insertion window to measure the retrieval following performance. Firstly, we observe that the training-based RAG which uses local semantic retrieval during training performs better than ReMoDiffuse, both in terms of perceptual quality and metrics. Moreover, we also see higher faithfulness to the retrievals through low MPJPE which entails that properly specifying the retrieved motion on the exact word affects the gesture quality. Secondly, our inference-time RAG-GESTURE is better than the training-based model, even after utilizing our retrieval algorithm. This underscores our method's superiority and the importance of having a good gesture animation performance.

LLM-based Gesture Type vs. Discourse Retrieval. For evaluating perceptual differences between LLM-based gesture type retrieval and discourse-based retrieval, we conduct an additional user study where we ask the participants to evaluate differences between generated gestures from the two algorithms in terms of naturalness and appropriateness (Fig. 7). Interestingly, we find that evaluators do not show any strong preference between the two algorithms.

	FID↓	BeatAlign→	L1Div→	MPJPE (mm)↓
GT		0.477	7.29	
No RAG	0.519	0.447	8.64	59.6
RAG by Inpainting [36]	0.446	0.450	8.46	28.1
RAG by LI only	0.589	0.486	9.29	54.6
RAG-GESTURE (LI+RG)	0.447	0.471	9.04	35.0

Table 3. **Extent of Semantic Retrieval Insertion.** LI: Latent Initialization; RG: Retrieval Guidance.

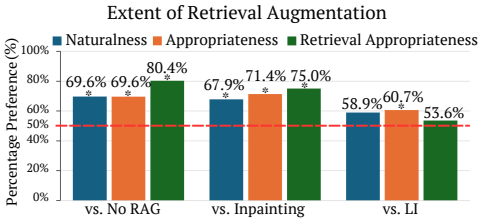


Figure 8. **Ablative Study** which illustrates preference of RAG-GESTURE over varying degrees of insertions. *: p-value < 0.05

Extent of Retrieval Augmentation Since our approach can control the influence of the retrieved gesture onto the generated one, we analyze how much influence is optimal to synthesize semantically accurate yet natural looking gestures. We compare RAG-GESTURE's prediction against varying degrees of retrieval insertion (Tab. 3) using Discourse-based retrieval. "Inpainting" means adding noise to the retrieval and synthesizing the remaining motion and "No RAG" represents no retrieval insertion. We conduct another user study with a third "Retrieval Appropriateness" question to gauge semantic appropriateness of the generated gesture during the insertion window (Fig. 8). Interestingly, we observe increasing preference of our retrieval-guided model as we move towards both extremes of the insertion. In terms of gesture perception, evaluators neither prefer generation without RAG (purely data-driven) nor a strong insertion through Inpainting (purely retrieval-based). Lastly, our model achieves good balance of retrieval following in terms of MPJPE, prosodic alignment and diversity.

5. Conclusion

In this work, we present RAG-GESTURE – a novel approach that leverages interpretable retrieval algorithms to introduce inference-time RAG to a diffusion-based gesture generator. Therefore, it combines advantages of the pure retrieval-based and fully data-driven approaches to synthesize natural yet semantically meaningful gestures. Perceptual evaluation demonstrates that this type of neuro-explicit approach is better than either of the two approaches and existing state-of-the-art methods. Moreover, we demonstrate that LLM-driven gesture type prediction and discourse relations can enhance semantic understanding in current gesture generation architectures. Lastly, our framework can be extended to synthesize gestures that exhibit task-specific gestural patterns like referential or emotion-specific gestures.

Acknowledgements. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 "Neuroexplicit Models of Language, Vision, and Action" - project number 471607914. The third author of this study (MS) was supported through the NWO-funded project "Searching for meaning: Identifying and interpreting alternative discourse relation signals in multimodal language comprehension" (VI.Veni.231C.021).

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. In *ECCV*, 2020. 2

[2] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum*, 2020. 2

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

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 4

[5] Kirsten Bergmann and Stefan Kopp. Gnetic–using bayesian decision networks for iconic gesture generation. In *Intelligent Virtual Agents: 9th International Conference, IVA*, 2009. 3

[6] Jessica A. Bernard, Zachary B. Millman, and Vijay A. Mittal. Beat and metaphoric gestures are differentially associated with regional cerebellar and cortical volumes. *Human Brain Mapping*, 36(10):4016–4030, 2015. 1

[7] Jana Bressem and Cornelia Müller. A repertoire of German recurrent gestures with pragmatic functions. In *Body–language– communication: An international handbook on multi-modality in human interaction [Handbooks of Linguistics and Communication Science 38.2]*, pages 1575–1591. Mouton de Gruyter, 2014. 3

[8] Michael Büttner and Simon Clavet. Motion matching-the road to next gen animation. *Proc. of Nucl. ai*, 1(2015):2, 2015. 3

[9] G. Calbris. *Elements of Meaning in Gesture*. John Benjamins Publishing Company, 2011. 2, 6

[10] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: The behavior expression animation toolkit. In *SIGGRAPH Conference Proceedings*, 2001. 2

[11] Simon Clavet et al. Motion matching and the road to next-gen animation. In *Proc. of GDC*, page 4, 2016. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 4

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 5

[14] Adso Fernández-Baena, Raúl Montaño, Marc Antonijoan, Arturo Roversi, David Miralles, and Francesc Alías. Gesture synthesis adapted to speech emphasis. *Speech communication*, 57:331–350, 2014. 6

[15] 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. 2

[16] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 89:117–130, 2020. 2

[17] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Expressgesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds*, 32(3-4):e2016, 2021. 3

[18] 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. 2

[19] Purvi Goel, Kuan-Chieh Wang, C Karen Liu, and Kayvon Fatahalian. Iterative motion editing with natural language. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 3

[20] 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

[21] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 conference proceedings*, 2022. 3

[22] Jennifer Hinnell. The verbal-kinesic enactment of contrast in north american english. *The American Journal of Semiotics*, 35(1-2):55–92, 2019. 3

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 4

[24] Daniel Holden. Character control with neural networks and machine learning. *Proc. of GDC 2018*, 1:2, 2018. 3

[25] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM TOG*, 39(4): 53–1, 2020. 3

[26] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *ECCV*, 2024. 3

[27] Anna Inbar. The raised index finger gesture in Hebrew multimodal interaction. *Gesture*, 21(2-3):264–295, 2022. 3

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

[29] Michael Kipp. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005. 3

[30] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019. 2

[31] 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. 2, 6

[32] Schuyler Laparle. Tracking discourse topics in co-speech gesture. In *International Conference on Human-Computer Interaction (HCII)*, 2021. 3

[33] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 7

[34] 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. In *ECCV*, 2022. 1, 2, 3, 7

[35] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *CVPR*, 2024. 2, 3, 4, 5, 6, 7

[36] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3, 5, 8

[37] William C Mann and Sandra A Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8 (3):243–281, 1988. 6

[38] David McNeill. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, 1992. 1, 6

[39] David McNeill. *Gesture and Thought*. University of Chicago Press, 2005. 1

[40] David Mcneill, Elena Levy, and Susan Duncan. Gesture in discourse. *The Handbook of Discourse Analysis*, 2014. 2, 6

[41] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 5

[42] Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *CVPR*, 2024. 2, 3, 4

[43] Rajmund Nagy, Hendric Voss, Youngwoo Yoon, Taras Kucherenko, Teodor Nikolov, Thanh Hoang-Minh, Rachel McDonnell, Stefan Kopp, Michael Neff, and Gustav Eje Henter. Towards a genea leaderboard–an extended, living benchmark for evaluating and advancing conversational motion synthesis. *arXiv preprint arXiv:2410.06327*, 2024. 7

[44] Michael Neff. Hand gesture synthesis for conversational characters. *Handbook of Human Motion*, pages 1–12, 2016. 2, 3

[45] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM TOG*, 27(1):1–24, 2008. 3

[46] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *CVPR*, 2024. 2, 4, 7

[47] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, 2023. 1, 2, 7

[48] OpenAI, Achiam J, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2024. *arXiv preprint arXiv:2303.08774*, 2024. 2, 6

[49] Patrick Louis Rohrer, Elisabeth Delais-Roussarie, and Pilar Prieto. Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in english academic discourses. *Lingua*, 293:103583, 2023. 6

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2021. 3, 4

[51] Ted J M Sanders, Wilbert P M S Spooren, and Leo G M Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35, 1992. 6

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 4, 5

[53] Sebastian Starke, Paul Starke, Nicky He, Taku Komura, and Yuting Ye. Categorical codebook matching for embodied character controllers. *ACM TOG*, 43(4):1–14, 2024. 3

[54] Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136, 2017. 6

[55] Jacques Terken and Dik Hermes. The perception of prosodic prominence. In *Prosody: Theory and experiment: Studies presented to Gösta Bruce*, 2000. 6

[56] Marcus Thiebaux, Stacy Marsella, Andrew N Marshall, and Marcelo Kallmann. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, 2008. 2

[57] Michael Wagner and Duane G Watson. Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945, 2010. 6

[58] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, 52(3):379–389, 2022. 7

[59] 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

[60] 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. 2, 7

[61] 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. 2

[62] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the International Conference on Multimodal Interaction*, 2022. 2, 3

[63] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, 2023. 2, 3, 7, 8

[64] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4

[65] Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Trans. Graph.*, 2024. 3, 7, 8

[66] 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. 4

[67] 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. 2

[68] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *ICCV*, 2023. 2

[69] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3