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

Supplementary Material

This supplementary document provides a glossary of notations used for method explanation in Sec. 1 and discusses dataset statistics in Sec. 2. It also provides further details and analyses of word-excitation guidance in Sec. 3, user study in Sec. 4 and implementation details in Sec. 5. Moreover, we discuss evaluation metrics in Sec. 6 and training details for baseline methods in Sec. 7.

1. Glossary for Notations

In Tab. 1, we provide a list of variables used in our the method and implementation details (Method Sec. and Sec. 5) for ease of reference.

Variable	Description
$\mathbf{x}$	Gesture Sequence
$\mathbf{x}_b$, $\mathbf{x}_h$	Body and Hand Motions
$\mathbf{C}$	Conditioning Set
$\mathbf{z}$	Latent representation
$\mathbf{z}_b$, $\mathbf{z}_h$	Latent Representation for body and hands
ξ_b, ξ_h	Encoder for body and hands
$\mathcal{D}_b$, $\mathcal{D}_h$	Decoder for body and hands
$\mathbf{x'}_b$, $\mathbf{x'}_h$	Reconstructed motion for body and hands
$\hat{\mathbf{z}}$	Time-aware Latent Representation
ϵ_θ	Predicted noise
f_θ	Denoiser neural network
$\mathbf{a}$	Audio Signal
$\boldsymbol{\tau}$	Text Embedding
$\boldsymbol{\tau}'$	Text Embedding for co-participant
$\mathbf{s}$	Speaker Identity Token
$\mathbf{m}$	Active/Passive Bits for Latent Chunks
w_c	Modality guidance scale for condition $\mathbf{c}$
S	Number of tokens selected for WEG
G_{exc}	Word Excitation Guidance objective
$\tilde{\mathbf{z}}^{(\mathbf{t})}$	Updated latent after WEG

Table 1. **List of variables and their corresponding explanation**

2. Dataset Statistics & Discussion

The proposed DND GROUP GESTURE consists of 6 hours of mocap data comprising of 5 persons in the scene. In total, we have 2.7M poses along with synchronized, per-person audio tracks and text transcripts (see Fig. 1). The proposed dataset addresses a different aspect of human gestures, *i.e.* group conversations, which is a sparsely researched setting. This makes our dataset complementary to the existing monadic gesture datasets like BEAT. We discuss how BEAT can be used with our framework along with our dataset in Sec. 7.2.

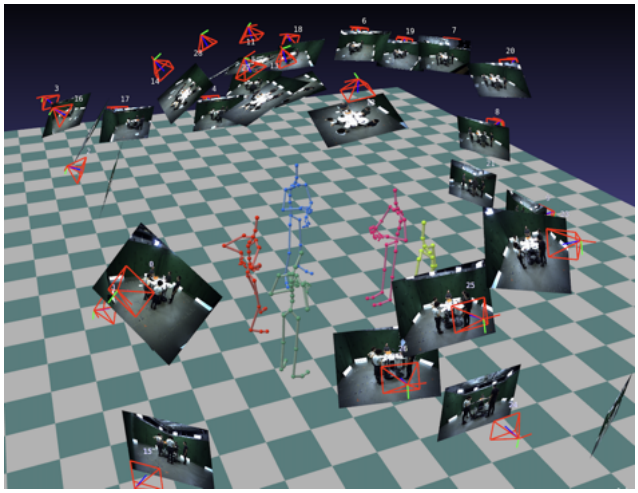


Figure 1. Here, we show an over-arching view of our data-recording setup, where we have five people interacting with each other, while their motion tracking is recorded via a state-of-the-art marker-less motion capture system. Each person also has individual microphones which feed into our audio setup.

Finding a capture setting that elicits high density of meaningful, semantic gestures is indeed a challenging task. These considerations lead us to capture the participants in a role-playing setting as they need to describe an imaginary world to each other, thereby leading to a high density of semantic gestures. The setting also offers a clear intrinsic reward to the participants (of winning the game). As we show in the video and website, the gestures in our dataset are similar to the ones appearing in daily conversation because participants are simply discussing a game plan or their next steps in certain situations using language that is colloquially used in conversations. Interestingly, the most relevant gestures to the game setting are pointing gestures (participants usually point to objects on table) which are considered deictic gestures, which happen frequently in normal conversations. We highlight that the proposed dataset is recorded in a markerless motion capture setup which it keeps the group conversation natural without restrictions of a capture suite or markers, thereby reducing the Observer's Paradox. Finally, the subjects are familiar with each other (esp. in the DnD setting) which further helps in more natural conversations.

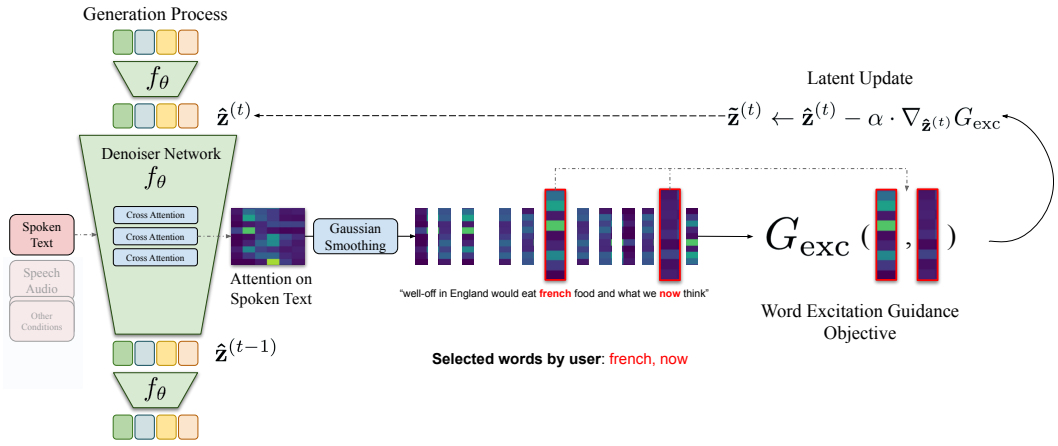


Figure 2. **Algorithm overview of Word Excitation Guidance.**

3. On Word Excitation Guidance

In the following sections, we provide details of algorithm for word-excitation guidance and then perform additional in-depth analysis on its results.

3.1. Algorithm Details

Algorithm 1 Word-Excitation Guidance

Input: Set of tokens $\{\boldsymbol{\tau}_i\}_{i=1}^{S}$
Trained Diffusion Model f_θ
Text Prompt $\boldsymbol{\tau} \in \mathbf{C}$,
Diffusion Timesteps T
Step size α
Output: Denoised Latent $\hat{\mathbf{z}}^{(0)}$
Initialize $\hat{\mathbf{z}}^{(T)} \sim \mathcal{N}(0, \mathbf{I})$
for $t = T$ *to* 0 **do**
  $\_, A = f_\theta(\hat{\mathbf{z}}^{(t)}, t, \{\boldsymbol{\tau}\})$ ▷ get attention for text
  $A \leftarrow \text{Softmax}(A_{\text{start:end}})$ ▷ remove start/end of text
  $A \leftarrow \text{Gaussian}(A)$ ▷ smooth out attentions
  $G_{\text{exc}} = \frac{1}{S}\sum_{i=1}^{S}(1 - \max(A_i))$ ▷ calculate loss i
  $\tilde{\mathbf{z}}^{(t)} \leftarrow \hat{\mathbf{z}}^{(t)} - \alpha \cdot \nabla_{\hat{\mathbf{z}}^{(t)}} G_{\text{exc}}$ ▷ update latent
  Perform Iterative refinement [1]
  $\epsilon_\theta^{(t)}, \_ = f_\theta(\tilde{\mathbf{z}}^{(t)}, t, \mathbf{C})$ ▷ estimate noise
  Perform Modality Guidance
  $\hat{\mathbf{z}}^{(t-1)} \leftarrow \text{SchedulerStep}(\tilde{\mathbf{z}}^{(t)}, \epsilon_\theta^{(t)})$
end for
return $\hat{\mathbf{z}}^{(0)}$

The process of word-excitation guidance involves modifying the usual denoising loop by updating the latents at each timestep. Before updating the latents, we normalize the attention maps by removing attention on the start and end tokens because each training batch contains text prompts of different lengths. Moreover, we observe that our latent diffusion framework assigns high attention to the start token in text (shown in Figure 4), therefore, we mitigate this issue by considering the attention on the actual text tokens. Then we apply Gaussian smoothing over the remaining attention map for stable generation results without any jerks in motion. This ensures flexibility to focus on a neighbourhood of words instead of one word by avoiding gradient updates at only the chosen tokens ignoring its neighbourhood. Next, we calculate the average for the loss over all the *focus* tokens to equally transfer gradients for all the focused words. Note that, this is different from image-based semantic guidance [1], where Chefer *et al*. apply smoothing on attention for only the chosen words/tokens which ignores the neighbourhood tokens. Moreover, their loss aggregation only enables gradient transfer for tokens with the lowest attention instead of all focused tokens by using a $\max$ function instead of mean like us. The complete process is presented in Algorithm 1.

3.2. Additional Analysis

Joints Affected Per Word. Recall that word-level excitation guidance steers the gesture generation process through the denoising network to have pronounced gestures at certain words in the text. It gives a fine mechanism for semantic control over gesture generation. In Fig. 3, we present an analysis of how this mechanism affects each joint in the generation. The figure encodes as heatmap the velocity of each joint in response to the text tokens; the assumption being that high velocity implies heavier gesturing. We see that the hand and the arm joints are affected the most at the focused words. Interestingly, minimal attention is focused on the lower body; this is expected as most gestures are predominantly upper-body motions.

Choice Of Words. Specific types of gestures tend to correlate with certain linguistic structures and parts of speech. Thus, for analysis, we conduct experiments with attention

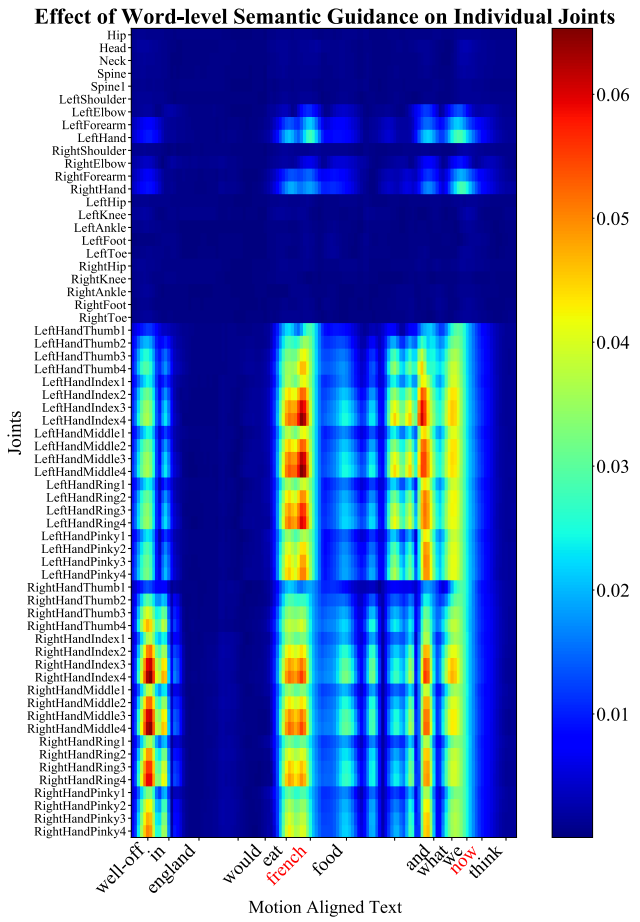


Figure 3. Heatmap showing the distribution of velocity of all joints compared with motion-aligned text (focused words are highlighted in red). We see high joint velocity for hand and arm joints around the words "french" and "now".

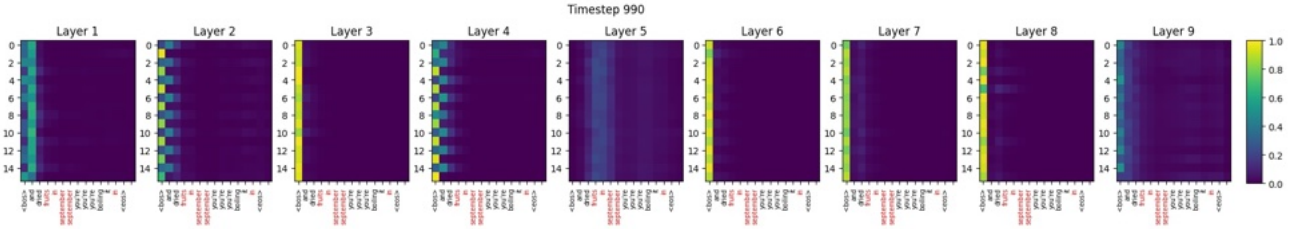
focused on these elements. To extract phrases that may map onto a semantic gesture, we select *random* three-word phrases in the text. To experiment with individual words, we can focus on nouns and verbs as they have a higher chance of mapping onto *iconic* gestures. Adverbs and adjectives can also be chosen since they can convey spatiotemporal properties of events and entities. This choice mechanism, which is motivated by the mapping of gestures to linguistic structures, is also flexible enough for the users to choose different linguistic features to focus on. We also consider optimal stress word discovery as a future endeavour. Lastly, the success of word-level semantic guidance is also affected by the amount of stress certain word has in the audio. We show attention results for phrases and words in Fig. 4 and gesture generation results in supplementary video.

Interpreting Word-Excitation through Attention Maps. Since we perform word-level guidance on text attention maps, we include example results in Fig. 4. The attention map A is dependent on text conditioning and diffusion timestep and shows the relation between chunks of latent representation and text tokens. Therefore, performing word-level guidance at each diffusion timestep yields slightly different attention distribution over words. We observe that as we move from $t = T$ to 0, focused word tokens (highlighted in red) start to get high attention, especially after $t = T/2$. We also see the effect of Gaussian smoothing as the attention on focus words is not sharply focused on only those words. Rather it is spread over its neighbouring tokens as well. Lastly, notice the striped pattern of the attention weights. This arrangement is a manifestation of the separated body and hand latents that have been stacked alternatively. It shows that the network learns to perform attention in a separate manner for both types of latents and guidance affects them differently across different layers as well.

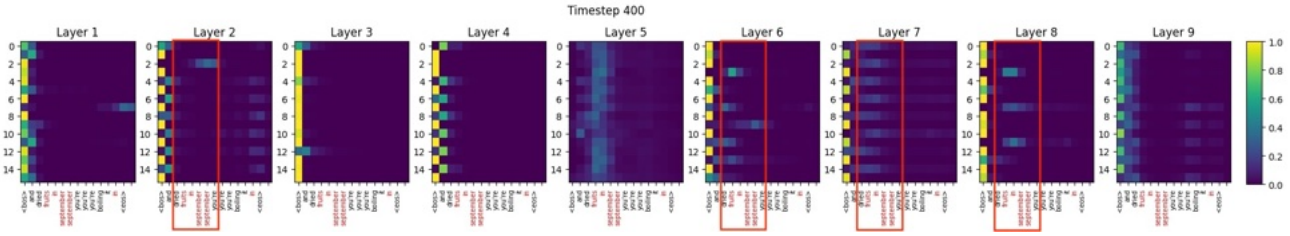
Limitations. Our method is, after all, a data-driven method. It depends on the learned conditional gesture distribution of text and other modalities, which can lead to it generating the most common gesture type (*beat gestures*) seen for some words. Consequently, performing word-level guidance does not always guarantee the specific motion of accurate semantic sub-gesture type (*iconic, deictic, metaphoric etc.*) at the focused word or phrase. However, as we analyzed, the usage of word-excitation guidance (WEG) mostly results in a semantically meaningful gesture as compared to the base prediction without WEG. For future works, a more explicit representation of gesture types and their mapping to words can be provided as a conditioning, which might help in predicting semantically accurate gestures. Secondly, the amount of focus each word/phrase attains in terms of gesture movements is dependent on the fact that speech also contains certain prosodic stress for that word. Similarly, if the gestures around focused words are already stressed adequately in motion or those words already have high attention on them, then the change introduced by guidance will only be subtle. Lastly, the choice of words affects the type of stress in gesture movements predicted and this can be highly subjective.

4. User Study

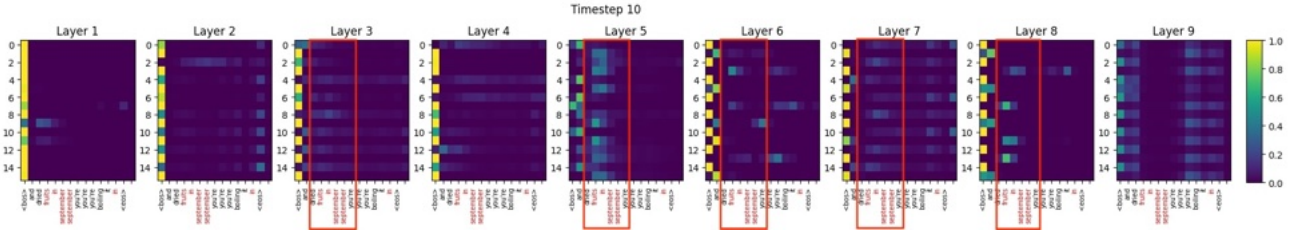
For evaluation of monadic synthesis, the user were shown a randomly sampled set of 10 forced-choice questions. Each question included a side-by-side animation of our method along with one of MLD [2], CaMN [8], or the ground-truth. The participants had to answer two question, (a) *"Which of the two gesture motions appears more natural?* and (b) *"Which of the two gesture motions corresponds better with the spoken utterance?"*. These questions try to gauge plausibility of the motions and alignment of the generated gestures with the utterance. For the task of dyadic synthe-



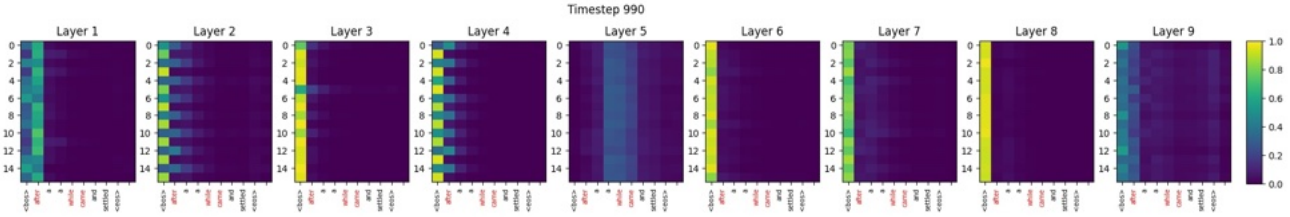
(a) Text: "and dried fruits in september you're boiling it in"



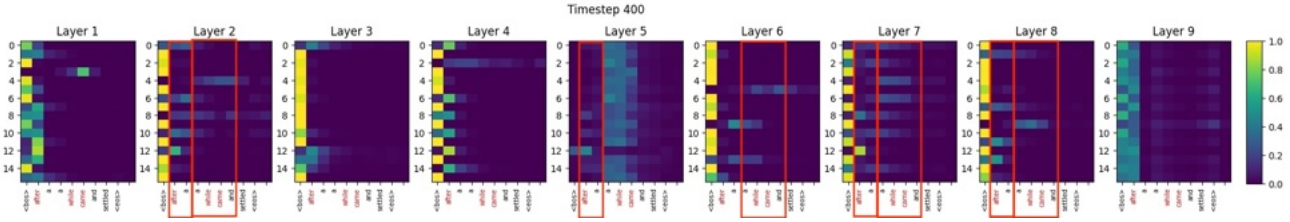
(b) Text: "and dried fruits in september you're boiling it in"



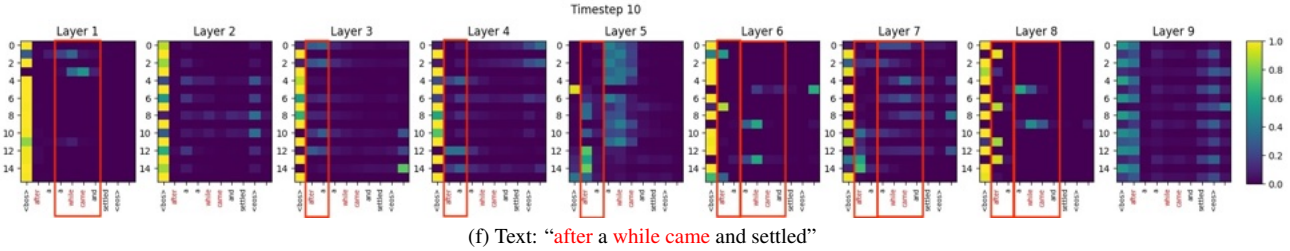
(c) Text: "and dried fruits in september you're boiling it in"



(d) Text: "after a while came and settled"



(e) Text: "after a while came and settled"



(f) Text: "after a while came and settled"

Figure 4. **Text attention example with focus on a phrase: (a)** $t = 990$ **(b)** $t = 400$ **(c)** $t = 10$ **and example with multiple individual words: (d)** $t = 990$ **(e)** $t = 400$ **(f)** $t = 10$**.** Vertical axes show $M \times 2$ latent chunks where even and odd indices stand for body and hand joints respectively. Horizontal axes show word tokens where focused words are highlighted in red. Attention changes are highlighted in red boxes where neighboring tokens are also included to show the effect of Gaussian smoothing. Lastly, "<bos>" & "<eos>" tokens represent start and end of the text (refer to Method Sec.

sis, we showed 5 randomly sampled forced choice questions to each participant, comparing our method with the adapted MLD and the ground-truth. The participants had to judge the naturalness of the motions similar to previous task and also answer the question: *"In which of the two interactions, the motion of interacting character fits well with both speech of the the main agent as well as their own speech, if any."* We report percentage preference for both tasks. In the third section, we asked the users to evaluate the word-excitation guidance. Each question included three motions—corresponding to the ground-truth motion, non-guided motion, and word-excitation guided motion — as well as the words that need to be excited during synthesis. We compare ground-truth, non-guided gesture, and word-excitation guided gesture. The users were asked to rate each motion on a Likert scale of 1-5, with 5 indicating the most semantically aligned gesture.

5. Implementation Details

Motion Representation. The motion $\mathbf{x}$ corresponds to the root-relative 3D coordinates for all $J-1$ joints and camera-relative translation of the root joint [4, 14, 19, 20]. The hand joints are also made relative to their corresponding wrist joint. We also pre-process the joint positions following [4] by normalizing the motion sequence to start the root trajectory from the origin while facing the positive z-axis.

VAE. We implement decoupled scale-aware VAE using two transformer encoders in order to make two halves of the latent representation focus separately on body and hands. Each encoder is based on transformer architecture with long skip-connections utilized by Chen *et al.* [2] as they prove this method to be effective in retaining high information density in latent representation. The output of each encoder is combined into two quantities to represent Gaussian distribution parameters $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\Sigma}_\phi$ of the combined scale-aware latent space $\mathcal{Z}$, where ϕ represent learnable weights of encoders. We can sample $\mathbf{z}^{2\times d}$ using reparameterization trick [6].

We train the VAE until convergence with a combination of losses to achieve the desired reconstruction quality. MSE-based reconstuction loss is applied on the reconstructed motion $\hat{\mathbf{x}}$:

$$\mathcal{L}_2 = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \tag{1}$$

Moreover, Kullback-Liebler divergence $\mathcal{L}_{KL}$ is used for regularizing the latent space:

$$\mathcal{L}_{KL} = D_{KL}(\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)||\mathcal{N}(\mathbf{z}; 0, \mathbf{I})) \tag{2}$$

We also apply Bone Length Consistency Loss [3], which ensures that bone lengths do not vary across frames in a gesture sequence by minimizing the variance of bone lengths

Hyperparameter	Value
Latent dimension d	128
Motion Length N	128
Number of Joints J	63
Motion chunks M	8
λ_{KL}	0.05
λ_{lap}	1
λ_{bone}	1
Transformer Layers	5
Attention Heads	2
Learning Rate	1×10^{-4}
Optimizer	AdamW [9]
FPS	25

Table 2. List of values used for training VAE for our method

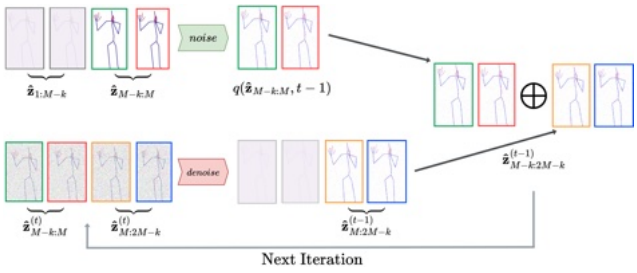


Figure 5. **Iterative process for the perpetual rollout of arbitrary-length generation.** This is based on the diffusion inpainting technique.

l_n.

$$\mathcal{L}_{bone} = \frac{\sum_{n=1}^{N}(l_n - \bar{l})^2}{n-1}, \tag{3}$$

Lastly, the VAE loss also contains a Laplacian regularization term $\mathcal{L}_{lap}$ as described earlier, to better reconstruct subtle jerks in gestures and reduce jitter.

$$\mathcal{L}_{VAE} = \mathcal{L}_2 + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{lap}\mathcal{L}_{lap} + \lambda_{bone}\mathcal{L}_{bone} \tag{4}$$

In order to achieve time-aware latent representation, we encode time-aligned M chunks of motion $\{\mathbf{x}'_i\}_{i=1}^{M}$ using encoders by passing each $\mathbf{x}'_i$ from ξ_b and ξ_h to get $\hat{\mathbf{z}}_i$. This sequence $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}_i\}_{i=1}^{M}$ is applied with a positional encoding [15] along M to represent time-alignment. Along with positional encoding as queries, $\hat{\mathbf{z}}$ passed onto decoder $\mathcal{D}$ as a memory to obtain $\hat{\mathbf{x}}$. This unique structure of $\hat{\mathbf{z}}$ allows us to perform arbitrary length generation with latent diffusion models, which generally are constrained to due to fixed-length generation.

Perpetual Generation Rollout. During inference of our diffusion framework, we leverage the time-aware latent sequence $\{\hat{\mathbf{z}}_i\}_{i=1}^{M}$ to autoregressively generate latent sequences for arbitrarily long sequences. As compared to ear-

lier approaches [8, 18], we do not concatenate our model's output which may cause irregular motion at the point of joining. We also do not encode variable length sequences in our VAE framework as done by MLD [2]. Instead, we propose an autoregressive generation approach to predict the time-aware latent sequence beyond M number of chunks. The key to this approach is an iterative process (shown by Fig. 5) where a sequence of future latent chunks is predicted based on the last k current latent chunks through a denoising process. Given the sequence $\hat{\mathbf{z}} \in \mathbb{R}^{M\times 2\times d}$ represents motion $\mathbf{x}$ of first N frames, we call it $\hat{\mathbf{z}}_{1:M}$ which is known to us. We utilize the last k latent chunks from this known sub-sequence, i.e. $\hat{\mathbf{z}}_{(M-k):M}$, and generate the next $M-k$ latent chunks through the denoising process to obtain a new overlapping sequence $\hat{\mathbf{z}}_{(M-k):(2M-k)}$. Every time we need to generate the next $(M-k):(2M-k)$ sequence, we first inject noise to the previously known $(M-k):M$ sub-sequence until the $t-1$ diffusion timestep. Then, this sub-sequence is concatenated with the latent sub-sequence at $M:(2M-k)$ which contains new latents for the non-overlapping part in $(M-k):(2M-k)$ sequence. This concatenated sequence $\hat{\mathbf{z}}^{(t-1)}_{(M-k):(2M-k)}$ is then passed to the next denoising iteration where the process repeats by noising the known part for the next diffusion timestep. This technique follows the masked denoising technique used for diffusion image inpainting [10].

$$\hat{\mathbf{z}}^{(t-1)}_{M-k:2M-k} = \oplus(q(\hat{\mathbf{z}}_{M-k:M}, t-1), \hat{\mathbf{z}}^{(t-1)}_{M:2M-k}) \quad (5)$$

Here, the $\oplus$ operator concatenates along latent chunks to total length of M for each sequence. When applied iteratively to the subsequent new frames, this process enables an autoregressive rollout of fixed-length gesture sequences into infinite-length synthesis. We set the value of the hyperparameter k as $k = M/2$ for simplicity.

Details on Denoising Network. We design denoising network for the latent diffusion framework to predict $\epsilon_\theta(\hat{\mathbf{z}}^{(t)}, t)$. We implement the denoising schedule based on DDPM framework with hyperparameters presented in Tab. 3. This framework consists of a Markovian chain of successively adding Gaussian noise ϵ to $\hat{\mathbf{z}}^{(0)}$ for T timesteps i.e. *forward diffusion* process. Through this process, $\hat{\mathbf{z}}^{(0)}$, which was sampled from data distribution, becomes $\hat{\mathbf{z}}^{(T)}$, which follows noise distribution $\mathcal{N}(0, \mathbf{I})$ assuming T is sufficiently large.

$$q(\hat{\mathbf{z}}^{(1:T)}|\hat{\mathbf{z}}^{(0)}) = \prod_{t=1}^{t=T} q(\hat{\mathbf{z}}^{(t)}|\hat{\mathbf{z}}^{(t-1)}) \quad (6)$$

where $q(\hat{\mathbf{z}}^{(t)}|\hat{\mathbf{z}}^{(t-1)}) = \mathcal{N}(\hat{\mathbf{z}}^{(t)}|\sqrt{1-\beta_t}\hat{\mathbf{z}}^{(t-1)}, \beta_t\mathbf{I})$, describes evolution of latent distribution during the noising process at time step t. Here, β_t represents the rate of diffusion. The *reverse diffusion* process consists of denoising $\hat{\mathbf{z}}^{(T)} \sim \mathcal{N}(0, \mathbf{I})$ for T timesteps to generate a latent sequence $\hat{\mathbf{z}}^{(0)}$:

$$p_\theta(\hat{\mathbf{z}}^{(0:T)}) = p(\hat{\mathbf{z}}^{(T)}) \prod_{t=1}^{T} p_\theta(\hat{\mathbf{z}}^{(t-1)}|\hat{\mathbf{z}}^{(t)}), \quad (7)$$

where $p_\theta(\hat{\mathbf{z}}^{(t-1)}|\hat{\mathbf{z}}^{(t)})$ is approximated using a denoiser neural network $f_\theta(\hat{\mathbf{z}}^{(t-1)}|\hat{\mathbf{z}}^{(t)}, t, \mathbf{C})$, which is trained to predict noise We use transformer decoder network as f_θ which takes $\hat{\mathbf{z}}^{(t-1)}$ as queries along with diffusion timestep t and conditioning set $\mathbf{C}$ as memory input. We apply positional encoding to queries and individual memory inputs similar to [15]. To better distinguish between body and hand latents in $\hat{\mathbf{z}}^{(t-1)}$, we add a learned embedding that aims to differentiate between body and hand parts of the latent representation. We also add a learned embedding to each element of our conditioning set $\mathbf{C}$ separately which helps the network differentiate between different conditioning types. Each transformer layer starts with Self-Attention and LayerNorm layers, along with a time-layer based on Stylization Block [19] to incorporate diffusion timestep embedding. Multi-modal cross attention consists of the same number of heads as the number of elements in the conditioning set $\mathbf{C}$. The outputs of all heads are aggregated using a linear projection, which is followed by another linear layer with GeLU activation [5].

Hyperparameter	Value
d	128
Range of β_t	$[8.5\times10^{-4}, 1.2\times10^{-2}]$
T	1000
β_t Schedule	Scaled Linear [13]
Self-Attention Heads	4
Decoder Layers	9
Learning Rate	7×10^{-5}
Optimizer	AdamW

Table 3. List of hyperparameters for denoising network in our method

Guidance Parameters. We modify classifier-free guidance to add modality-level control for each element in our conditioning set. The random modality dropout rate is set to 10% and global guidance scale λ_m is set to 7.5. The values of w_c is determined by the task at hand. For example, if we want to extract only the gesture styles of different speakers regardless of input text and audio, we set all w_c to 0 except $w_\mathbf{s} = 1$, which corresponds to speaker identity. This will generate *unconditional* gestures in the style of a specific speaker (see supplemental video for the example). For word-excitation guidance, step size α, goes from 100 to 70.71 as it varies w.r.t. diffusion timestep. The kernel size

for Gaussian smoothing is 3.

Semantic Consistency Evaluation Model: Our method can generate semantically meaningful gestures (as shown in Suppl. Video), thanks to the proposed Word Excitation Guidance (WEG). We conduct this ablation the following way. First, we trained a binary classifier that classifies 1s motions of the BEAT dataset into either beat gesture, or semantic gesture type (based on the GT labels). Here semantic class consists of *iconic*, *metaphoric*, and *deictic* classes. This classifier is then used as an oracle to compute the recall of our generated motions for semantic class predictions. Specifically, we extract the speech and text for the sentence in which a semantic gesture has been labeled in dataset. These are then input to CONVOFUSION to generate the corresponding gestures, with and without WEG. For the case of WEG, we focus on the exact words wherein the semantic gesture occurs in the sentence.

6. Evaluation Metrics

We report quantitative results on Beat Alignment Score [7], FID, Diversity, L1 Divergence and Semantic Relevance Gesture Recall (SRGR) [8] and here we briefly describe each one of them. Beat Alignment Score was initially introduced [7] to measure the alignment of music beats to dance motion for the task of music-to-dance synthesis. This has also been adapted for the task of gesture synthesis, where it measures the correlation between gesture beats and audio beats. It is useful in differentiating between static motions which do not align well with the audio from natural-looking gestures which have speech-aligned kinematic beats. However, it can report false high values if the motion has a large amount of jitter because it would assume beats created by jitter align well with most of the audio beats. We can see this happening for methods that show high jitter [18, 22] in our experiments. They have high Beat Alignment score while their FID is also large.

We employ the Frechet Inception Distance (FID) metric provided by Yoon *et al*. [18], also known as FGD. We trained our FID network using implementation by Liu *et al*. [8]. It is based on an autoencoder network that is trained for reconstruction task, and is calculated by comparing features of the ground truth data $\mathbf{x}$ and generated data $\hat{\mathbf{x}}$ through:

$$\mathrm{FGD}(\mathbf{x}, \hat{\mathbf{x}}) = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \quad (8)$$

Methods that generate diverse gestures like ours and do not contain pre-pose information unlike CaMN [8] and Multi-Context [18], may suffer on this metric because our gestures will not try to match ground truth motion. Diversity computes the average pairwise Euclidean distance of the gesture generations in the test set. L1 Divergence (also called L1 variance) measures the distance of all frames N in a gesture sequence from their mean μ_N. Here B is size of test set.

$$\mathrm{L1div}(\mathbf{x}) = \frac{1}{B} \sum_{i=1}^{N} |\mathbf{x}_i - \mu_N| \quad (9)$$

This metric specifically identifies if the gestures are static in movement and make less diverse movements along the generation length. As shown in supplemental video, CaMN [8] and MLD [2] suffer from this problem whereas, our method predicts different gestures according to the text and audio conditionings and does not have static motion.

Semantic Relevance Gesture Recall (SRGR) uses semantic score labelled in BEAT [8] as a weight for the Probability of Correct Keypoint (PCK) between the generated gestures and ground truth gestures. It aims to reward being close to ground truth motion at points where a semantically relevant gesture exists while also predicting diverse gestures, as mentioned by Liu *et al*. [8]. This metric has a similar issue as FID as it compares PCK between ground truth and prediction and due to many-to-many correspondence between gestures and speech, this might not be suitable. Therefore, we conclude that each metric focuses on certain aspects of gesture generation and human-annotated user study results are more conclusive to determine better generation quality and to perform a holistic analysis of gesture quality

7. Baseline Training Details

In the following, we provide details on how we process all the different modalities for our dataset (Sec. 7.1) and provide details for training each method we use for comparison (Sec. 7.2).

7.1. Dataset

BEAT. We utilize the BEAT dataset [8] in order to augment the training data for our method so that it better generalizes to the task of monadic gesture synthesis. It consists of 60 hours of English speaking training data, spanning 30 subjects that perform gesture motions. The dataset is rich in good training examples for monologue setting which can serve as a good baseline dataset to train our method. Inherently, BEAT's motion representation is different than what we use to train our method. Therefore, we re-target their skeleton definition to our skeleton definition in order to match it with our DND GROUP GESTURE dataset. Moreover, we convert their representation from BVH-based euler angle representation to joint positions using forward kinematics. Then we resample their dataset from 120 FPS to 25 FPS in order to match it with our training configuration. Lastly, we apply the preprocessing steps mentioned in Sec. 5 to get the final motions which we separate into 5.12-sec chunks i.e. 128 frames for training our method.

DND GROUP GESTURE . To apply our method to the task of dyadic synthesis, we utilize our recorded DND GROUP GESTURE dataset (see Fig. 1) to extract interactions between people in our dataset. We record the dataset in BVH format as well, however, we extract joint positions for training our method. We standardize dataset FPS to 25. Each of the five people in our dataset has their own separate audio channel which we have post-processed to get clean and denoised audio. We align audio channels with the recorded tracking and verify it manually as well. Finally, we separate out motions for each person and assign them identities which are kept consistent across multiple recording sessions. Then, we preprocess the dataset and separate it out in chunks of 128 frames.

Training/Test Splits. We split the BEAT dataset by reserving 5 out of 30 English speakers for the testing set, while the remaining go into training and validation splits. Therefore, all the results and comparisons on monadic synthesis using BEAT dataset are provided on unseen speakers which shows a method's generalizability to unseen audio and text inputs. For dyadic synthesis, we randomly sample and take out 10 percent for testing and rest for training and validation.

Representation of Modalities. We process audio by sampling it to 16000 Hz and extracting melspectrograms using librosa toolbox [11]. We use 80 mel-bands and a hop length of 512 for melspectrogram conversion. We process text through text tokenizer and convert them to embeddings through T5 text encoder [12] implementation by Hugging Face [16].

Lastly, all methods are trained on these dataset splits to ensure fairness. There are some differences between the type of representation used for audio and text in each method, which we elaborate on in the next section.

7.2. Methods for Comparison

ConvoFusion (Ours). We train our method on 128-frame sequences by learning a latent space representation of them Then we use our diffusion framework on top of it. Interestingly, we can incorporate both monadic and dyadic gesture synthesis tasks into single training. Thanks to Modality Guidance, we can use different modalities interchangeably by dropping them out of the training batch and setting an unconditional token in their place. For example, BEAT dataset only contains single-person gesture annotations and does not contain a co-participant, hence making it non-trainable for the task of dyadic synthesis. However, we can simply provide an unconditional token for the co-participant's text which automatically turns the contribution of the corresponding modality guidance term to zero, and BEAT dataset can be trained jointly with DND GROUP GESTURE dataset. A similar approach can be taken for semantic annotation labels provided by BEAT dataset, which we do not provide for ours.

CaMN [8] & Multi-Context [18]. We train both of these methods using the official implementation of CaMN by Liu *et al*. on GitHub. The only modification that took place was the addition of our dataset pipeline which includes our version of BEAT dataset and DND GROUP GESTURE dataset, which makes the motion dimensions from 141 to 189 to match to our setting. We use the provided WavEncoder in the implementation to process audio signals instead of melspectrograms. Lastly, we use motion-aligned text instead of normal text inputs to be consistent with them. We also use the provided text-encoder to add our textual vocabulary to the text tokenizer for text preprocessing.

MLD [2]. This method by Chen *et al*. which uses latent diffusion models, was presented for the task of text-to-motion synthesis. We extend this method for the gesture synthesis task by utilizing our training procedure. To be consistent with their method, we use the text encoder which was used by MLD.

DiffGesture [22]. We use their official implementation on GitHub to train this method for the task of monadic gesture synthesis. Since DiffuGesture [21] does not provide an implementation for dyadic synthesis task and it is highly based on DiffGesture, we follow DiffuGesture's implementation details as close as possible and adapt DiffGesture to the dyadic synthesis task. As this method was originally trained on TED Dataset [17], which contains only the upper body, we double the capacity of their transformer network to cater to the increase in dimensionality in our setting to ensure fairness. The audio and text processing is kept consistent with DiffGesture's implementation. Lastly, for the task of dyadic synthesis, we provide co-participant's text as an additional conditioning input to match it with our training pipeline.

References

[1] 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. 2

[2] 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. 3, 5, 6, 7, 8

[3] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, 2023. 5

[4] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 5

[5] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 6

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5

[7] Ruilong Li, Sha Yang, David A. Ross, and Angjoo

Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 7

[8] 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. 3, 6, 7, 8

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[10] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 6

[11] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, 2015. 8

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), 2020. 8

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2021. 6

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

[15] 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, 6

[16] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020. 8

[17] 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. 8

[18] 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. 6, 7, 8

[19] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 5, 6

[20] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. *International Conference on 3D Vision (3DV)*, 2024. 5

[21] 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. 8

[22] 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. 7, 8