

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

Supplementary Material

We first discuss limitations of our method and provide details on the perceptual evaluation. Further, we elaborate on evaluation metrics and provide additional experiments and analyses. Lastly, we discuss our implementation and provide analysis and runtime details for it.

1. Limitations

Our method relies on the sparse semantic data which is extracted from the BEAT2 dataset [9]. This creates a scarcity of good exemplars which can be used during the database matching steps. As a result, the LLM-based Gesture Type algorithm sometimes struggles to find good contextual matches for each gesture type and word identified by the LLM.

Secondly, our method combines explicit rule-based algorithms with a neural generation framework. In rare cases, these algorithms can fail in edge cases and result in an incorrect retrieved example. However, our learned framework on top of the retrieval algorithms mitigates this and ignores out-of-distribution motion exemplars since it has been trained to produce only those gestures which match the speech. This also is affected by the extent of retrieval augmentation.

2. Details on User Study

For the perceptual evaluation of gesture generation capability, 56 participants were shown a randomly sampled set of 16 forced-choice questions. Our study consists of four sections corresponding to four different user studies. First section focuses on the comparison with the state-of-the-art approaches. Each question comprises of a side-by-side animation of our method along with one of EMAGE [9], Audio2Photoreal [11], RemoDiffuse [13] or the ground-truth. Second section is a short section where we compare LLM-driven gesture generation with discourse-based synthesis. Third, we perform pair-wise comparisons between results from one of the RAG baselines (derived from ReMoDiffuse [13]) and our method. Lastly, we evaluate different approaches to perform RAG by controlling its extent.

In first three sections, we try to evaluate naturalness and appropriateness. Specifically, we ask two questions (a) “Which of the two gestures look natural?” and (b) “Which of them looks appropriately aligned to what the person is saying?”. In the last section, we add an additional question which focuses on gauging the semantic appropriateness in the retrieval window, with a goal of evaluating the RAG capability. We highlight the identified words from the retrieval algorithms and add an additional question: “Which of the two have better gestures in the highlighted section,

especially at the capitalized word in the prompt?”

3. Evaluation Metrics

FID. We employ the Frechet Inception Distance (FID) metric inspired by Yoon *et al.* [12], which is also known as FGD. We use the autoencoder network provided by BEAT2 [9] to get the gesture encodings for FID evaluation and do not retrain our own network.

Beat Alignment Score. Originally introduced to measure alignment of music beats to dance motion, Beat Alignment Score [7] has been adapted for the gesture synthesis task where it aims to measure the correlation between gesture beats and audio beats.

L1 Divergence. This metric (also called L1 variance) measures the distance of all frames in a single generated sample from their mean. It is helpful in identifying synthesized gestures that are static and unexpressive.

Diversity. It computes the average pairwise Euclidean distance between the generated gestures from the test set.

Multi-modality. This metric requires sampling different gesture motions for a single speech input from the generative model [1]. Then, it computes Euclidean distance between the diverse generated gestures. It probes the diverse sampling capabilities of a generative model.

4. Additional Experiments

4.1. Per-speaker Quantitative Comparison with FID & Multi-modality

To perform a robust evaluation on speaker generalizability of our framework, we provide per-speaker FID and Multi-modality metrics for the all-speaker model in Tab. 1. We observe that our framework achieves best FID for large number of the speakers, which shows that our method generalizes well to the speaker specific patterns and idiosyncrasies despite taking no seed gestures at input. Overall, CaMN [8], achieves lower FID because it uses seed input from the ground truth data which results in lower scores. Due to the same reason, EMAGE [9] also gets lower FID.

However, CaMN and EMAGE always generate same gestures for a given speech input, so they perform worse in terms of Multi-modality, which makes them less ideal for diverse gesture generation. In contrast, our approach gets

	FID↓						Multimodality↑					
	CaMN	EMAGE	Audio2Phoreal	ReMoDiffuse	Ours (w/ Discourse)	Ours (w/ LLM & Gesture Type)	CaMN	EMAGE	Audio2Phoreal	ReMoDiffuse	Ours (w/ Discourse)	Ours (w/ LLM & Gesture Type)
wayne	1.23	2.06	2.32	3.58	1.49	1.59	n/a	n/a	1.1	3.4	3.1	3.7
scott	0.83	1.17	1.02	1.76	0.78	0.83	n/a	n/a	1.5	7.7	5.3	5.4
solomon	1.22	1.42	1.93	2.45	0.92	0.86	n/a	n/a	0.5	3.5	4.3	4.5
lawrence	0.98	1.39	1.13	3.16	0.69	0.66	n/a	n/a	1.9	6.9	5.6	6.3
stewart	0.65	1.26	1.62	1.76	1.49	1.49	n/a	n/a	0.3	0.7	2.4	3.0
carla	0.81	1.40	1.33	2.95	1.63	1.49	n/a	n/a	0.4	1.3	1.4	1.5
sophie	0.92	1.67	1.85	2.76	1.76	1.74	n/a	n/a	0.69	3.5	2.7	3.2
miranda	0.58	0.87	1.10	1.86	1.09	1.34	n/a	n/a	0.4	0.9	1.7	1.9
kieks	1.30	2.62	1.90	7.65	1.74	1.63	n/a	n/a	1.0	2.1	3.6	4.1
nidal	0.40	0.65	0.72	1.74	0.67	0.64	n/a	n/a	0.7	1.9	2.6	3.1
zhao	1.66	2.70	1.96	3.37	1.37	1.32	n/a	n/a	1.3	3.2	3.3	3.4
lu	1.40	2.73	1.92	2.23	1.27	1.16	n/a	n/a	0.7	1.7	2.4	2.6
carlos	0.78	1.47	1.71	2.47	1.95	2.56	n/a	n/a	0.2	2.5	2.9	3.1
jorge	1.49	2.57	1.97	3.55	1.89	1.93	n/a	n/a	0.3	1.7	1.9	2.1
itoi	0.93	1.61	1.34	2.28	1.07	1.32	n/a	n/a	0.8	1.8	3.0	3.1
daiki	0.78	1.78	1.66	3.04	0.91	1.19	n/a	n/a	0.3	2.3	2.3	2.7
li	1.10	1.74	1.17	2.06	0.71	0.79	n/a	n/a	0.6	3.9	2.7	3.0
ayana	1.19	2.03	1.96	4.35	2.09	2.13	n/a	n/a	0.4	1.8	1.9	2.1
luqi	1.25	2.22	1.67	5.21	1.38	1.86	n/a	n/a	0.7	1.7	2.2	2.5
hailing	0.53	1.20	7.53	5.72	2.35	2.79	n/a	n/a	0.3	1.0	2.5	2.6
kevin	1.07	1.70	1.19	1.87	0.92	0.95	n/a	n/a	0.3	1.8	1.6	1.9
goto	0.84	1.32	2.01	2.51	1.45	2.09	n/a	n/a	0.3	2.3	2.0	2.4
yingqing	1.67	2.50	2.00	4.34	1.82	1.74	n/a	n/a	1.1	3.5	3.0	3.2
tiffnay	0.81	1.35	1.09	2.67	0.92	1.11	n/a	n/a	0.3	1.3	1.2	1.6
katya	1.10	2.09	1.57	2.65	1.15	1.23	n/a	n/a	0.6	2.9	2.3	2.6

Table 1. Per Speaker FID/Multimodality

Multi-modality ↑					
CaMN	EMAGE	Audio2Phoreal	RemoDiffuse	Ours (w/ Discourse)	Ours (w/ LLM)
n/a	n/a	16.9	66.5	69.1	76.7

Table 2. Overall Multi-modality for All-Speaker Model

the best results showing diverse gesture generation capabilities of our model (Tab. 2).

4.2. Comparison of RAG with Motion Blending

Linear motion blending can be considered as an alternative for exemplar insertion in the generated motions. Therefore, we explore this alternative by pasting semantic examples onto the retrieval windows and blending motion at the window boundaries. We observe smoothing artifacts with motion blending where motion looks unnatural and gesture beats are smoothened around window boundaries due to interpolation. Compared to this, motion naturalness and speech-to-gesture alignment are preserved by using the proposed retrieval insertion method, which “blends” motion in the diffusion latent space. *Video results* can be seen in the supplementary video (@12:54).

4.3. Ablation on single speaker training

We observe high FID for the single-speaker setting. Therefore, we analyze this further in Tab. 3 by comparing single-speaker model with Non-RAG version (to check underfitting), and also with an RAG version which uses larger all-speaker database but same model trained on a single speaker (to check if smaller database causes worse performance). We observe the same trend as the ablative analysis, that RAG versions of the model perform better. Interestingly, usage of larger DB performs slightly better due to better example matching. However, FID remains higher than CaMN/EMAGE and metrics show little difference between

	FID↓	BeatAlign→	L1Div→	Diversity→
GT		0.703	11.97	127
No RAG	0.911	0.727	12.78	130
RAG with all-speaker DB (w/ Discourse)	0.872	0.727	12.53	127
RAG with 1-speaker DB (w/ Discourse)	0.879	0.730	12.62	129

Table 3. Quantitative Ablation with 1-speaker training.

1-speaker model and “No RAG” model, which shows underfitting due to small data size. Our experiments concur that diffusion models are more data-hungry and train better with larger data. Moreover, we observe that deterministic models (CaMN/EMAGE) that use seed motion, perform better with the smaller data by overfitting and predicting samples closer to GT (Sec. 4.1).

5. Implementation Details & Analysis

5.1. Input Representations

Representing speech and its transcription is highly important aspect of diffusion-based gesture modelling process. In our experiments, we found that changing the structure of text embeddings affects gesture understanding during the learning process, which consequently is reflected during the synthesis phase as well. To construct our text representation, we build a per-frame embedding with corresponding word embeddings residing on each frame. To extract discourse connectives, we pass text transcriptions of speech samples through discopy [6] and store resulting outputs along with dataset samples. We compute the word embeddings by aggregating sub-word token activations from last 4 layers of BERT model [2].

5.2. Decoupled gesture encoding

We utilize time-aware VAE architecture by Mughal *et al.* [10]. This architecture utilizes separate encoders for

frame window chunks of original motion to encode each chunk into an encoding. Then, it jointly decodes all of the chunks together to reconstruct the original motion. We use $N = 150$ representing 10 seconds of motion at 15 frames per second. Moreover, the frame chunk length of our time-aware VAE is 15, making each chunk encoding correspond to 1 second of motion. This results in a chunked gesture encoding of length 10 for each body part. Finally, we concatenate all 4 body part encodings along the time axis with separators in between them, resulting in $M = 40 + 3 = 43$.

We train the VAE on the reconstruction task by utilizing a set of losses to optimize the model. We apply Geodesic Loss on rotation matrices and standard MSE losses on 6D, axis-angle and joint position representation of the motion. Moreover, we also apply additional MSE losses to optimize velocity/acceleration of motion [4]. Lastly, we apply loss on foot contact predictions during VAE training to reduce foot sliding [5, 14].

5.3. RAG-driven Gesture Diffusion model

To optimize our diffusion model, we utilize Adam [3] with a learning rate of $1e - 4$. We utilize “scaled linear” as our β_t schedule and use 1000 steps while training. For inference, we use spaced 50 steps with DDIM scheduler. The transformer network contains 16 attention heads and 8 decoder layers. To better disambiguate body parts in our gesture encoding, we also add a separate sinusoidal positional encoding for body parts.

Retrieved Motion Insertion. In order to insert the retrieved gestures into the query latents, we only consider latents for upper body and hands. The encodings of these body parts are transferred from retrieval to query sample because speech has the most amount of semantic significance on these two body regions in terms of co-verbal gestures.

In the current setup, the insertion of retrieved gestures happens at $t = T$ where the latents are fully noised. However, our implementation also allows to arbitrarily choose a timestamp $t = K$ for retrieval insertion. Consequently, one can then perform Retrieval Guidance for steps $t < K$.

5.4. Details on LLM Prompting

We utilize OpenAI’s gpt-4o-mini model for semantic gesture type prediction. We provide a system prompt containing a brief explanation of gesture types and a user prompt which contains text from the test dataset and the question.

System Prompt. “You are an expert in human gestures. You need to identify words that may elicit semantically meaningful gestures(deictic, iconic, metaphoric) and their types: (a) Metaphoric Gesture: Represents abstract ideas or concepts physically, creating a vivid mental image. (b) Iconic Gesture: Mimics the shape or action of the object

	FID↓	BeatAlign→	L1Div→	Diversity→
GT		0.477	7.29	110
1-word	0.483	0.486	9.38	115
2-word	0.487	0.514	9.94	118
3-word	0.510	0.536	10.26	120

Table 4. Quantitative Comparison using different quantities of identified words in LLM prompt.

or concept being described. (c) Deictic Gesture: Points to or indicates a person, object, or location. Format your response as a python list of python tuples of (word, type). For example: [(‘hello’, ‘beat’), (‘world’, ‘iconic’)]”.

User Prompt. Identify at most 2 important words which are more likely to elicit semantically meaningful gestures and what are types of those gestures in following text: “SAMPLE TEXT”.

Effect of Word Number in Prompt. As shown above, the user prompt contains a maximum number of words for which LLM needs to predict gesture types. As this number of words is a hyperparameter, we perform quantitative comparison for different quantities in Tab. 4. Results show slight variation in metrics and even smaller difference in terms of perceptual quality. Therefore, we conclude that retrieval algorithm is flexible enough to be used with any configuration.

5.5. Baseline Retraining Details

To be consistent with single speaker evaluation on BEAT2 dataset, we utilize released model weights by EMAGE [9]. For other approaches (including ours), we retrain the method on single speaker data belonging to the speaker “Scott”. Since there are no available models for the chosen baselines which have been trained on all speakers in BEAT2 dataset, we train all the methods on complete dataset through their provided codebases. Methods which do not contain speaker specific generalizations like Audio2Photoreal [11], are modified to include a speaker embedding along with text and speech embeddings. Moreover, Audio2Photoreal is adapted to support the skeletal format of BEAT2. Lastly, ReMoDiffuse, originally released for text-to-motion task, is modified for gesture synthesis and their retrieval process is implemented using text feature similarity method.

For the comparison of our approach with training-based RAG (Sec. 4.3), we further modify the ReMoDiffuse architecture and train it using our gesture encodings instead of raw motion. Importantly, we implement retrieval merging strategy of SemanticGesticulator [15] to incorporate the output of our proposed retrieval algorithms into this

training-based approach. We perform this experiment by training it on all speakers and utilizing Discourse-based retrieval algorithm.

5.6. Runtime Information:

Retrieval algorithms involve multiple ranking and filtering steps, each contributing to a certain computation time. Specifically, Discourse-based algorithm takes **0.03s** to run on 1 data sample. LLM-based algorithm also includes an API call along with ranking steps and therefore, total time taken is increased to **1.52s** which includes **0.95s** for API call. RAG-driven inference for the diffusion model takes **26.93s** on NVIDIA RTX3090 for a batch size of 32.

References

- [1] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, 2023. 1
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2
- [3] P Kingma Diederik. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3
- [4] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023. 3
- [5] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *ECCV*, 2024. 3
- [6] René Knaebel. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, 2021. 2
- [7] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 1
- [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. In *ECCV*, 2022. 1
- [9] 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. 1, 3
- [10] 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
- [11] 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. 1, 3
- [12] 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. 1
- [13] 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. 1
- [14] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. In *International Conference on 3D Vision (3DV)*, 2024. 3
- [15] 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