

HoleGest: Decoupled Diffusion and Motion Priors for Generating Holisticly Expressive Co-speech Gestures

Supplementary Material

1. Supplementary Information on Quantitative Experiments

Our model is trained in two ways. For the video, we used all 25 English users from the BEATX dataset. For quantitative experiments, we retrained the model only on Speaker2 to achieve a fair comparison and maintain consistency with the EMAGE setup. We also added metrics for all BEATX sequences: DSG (FGD=11.742, BA=7.3368, DIV=11.121), EMAGE (FGD=7.305, BA=7.709, DIV=10.948), and HoloGest (FGD=6.457, BA=8.0281, DIV=13.525). Although the FGD slightly decreased on the complete BEATX set, HoloGest scored the highest in all major metrics, proving its effectiveness.

2. Some Explanations in the Video

In the video, we showcase the dynamic styles of two walking speakers, Speaker2 and Speaker4, reflecting their true label motions with continuous pacing. This highlights the superiority of our method over contemporary approaches, such as EMAGE and DSG, which only produce refined in-place actions. Moreover, our method remains stable in static styles, as demonstrated in the NPC dialogue scenario in the video’s demo section. *Upon paper acceptance, we will open-source the entire code repository and project homepage, which will include generated results for various styles.*

3. Technical Details and Effectiveness Analysis

In this section, we conduct an effectiveness analysis of the decoupling mechanism, SIDD module, and semantic alignment module to substantiate the rationale and importance of these three modules in the main text.

3.1. Decoupling Mechanism Analysis

We retrained the baseline models DSG, EMAGE, HoloGest– with only hand and body decoupling, and HoloGest with upper and lower body and finger decoupling. As shown in Table 1, we evaluated the FGD metrics of the decoupled body and fingers, finding that fingers are more challenging to learn compared to the body. We speculate that this is due to the stronger correlation of fingers with semantics and upper limb movements, and their weaker relation to audio melody features. If learned as a whole, the finger motion distribution would be averaged into the body distribution. Based on this, our system

Method	BEATX				
	FGD↓	FGD_body↓	FGD_hands↓	FGD_upper↓	FGD_down↓
DSG [10]	8.811	4.81	6.82	5.79	5.11
EMAGE [4]	5.512	3.6	5.08	3.27	4.87
HoloGest–(Ours)	6.203	4.02	4.93	4.02	4.37
HoloGest(Ours)	5.3407	3.86	4.41	3.73	3.96

Table 1. Evaluation of FGD Objective Metrics for Decoupled Components.

decouples the body and fingers, learning their independent distributions to improve the generation quality of each part. The semi-implicit denoising process significantly reduces the inefficiency caused by this decoupling method and the original DDPM denoising method. Furthermore, there is a significant difference between the data distributions of the upper and lower body, and the lower body movements are often weakly correlated with the audio, posing challenges for uniform learning in the diffusion model. To address this issue, we decouple these parts and learn them individually. We validate this by comparing the FGD component metrics of DSG, HoloGest–, and HoloGest. Decoupling improves the FGD of each part, with the most significant improvement in the lower body, close to one point.

In addition, although the decoupling approach of EMAGE is consistent with HoloGest, the high-fidelity generation results of our semi-implicit diffusion model lead to superior performance compared to EMAGE and almost all VAE-based methods.

3.2. Ablation Study on the SIDD Module

In this section, we investigate the impact of diffusion step size, reconstruction loss weight, and auxiliary forward loss on SIDD Module performance. All ablation studies were conducted on the BEATX dataset, with results shown in Table 2.

Ablation on Sampling Step Size. In this experiment, we investigate the impact of varying sampling step sizes on model performance. Specifically, we train models with the same structure using 1, 5, 10, 20, 30, and 50 steps respectively. The final results, as shown in Table 2, indicate that the FGD value tends to stabilize after 50 steps, but an increase in step size also results in slower speed. When the step size is 1, our structure reverts to a traditional GAN model, leading to a drastic drop in the quality of the generated gestures.

Impact of Reconstruction Loss. When the reconstruction loss weight is set to 0, the model degenerates to a process similar to the inverse of the original diffusion model,

Steps			Recon loss		AFD loss	
num	FGD↓	Inf. time↓	weight	FGD ↓	FGD↓	
1	55.44	0.03	0	63.10	w/o	26.9
5	29.91	0.23	1	9.41	w/	5.34
10	8.87	0.29	10	5.34	-	-
20	7.802	0.4	100	5.40	-	-
30	6.110	0.64	-	-	-	-
50	5.34	0.88	-	-	-	-
80	5.48	1.26	-	-	-	-
100	5.39	1.55	-	-	-	-

Table 2. Ablation experiments for the SIDD module. We find that when the diffusion steps reach 10, the FGD metric tends to stabilize, while when simplified to a traditional GAN network with only one step, the FGD metric sharply declines. The absence of body reconstruction loss severely affects the quality of the generated gestures, but the weight of this constraint has little impact on model learning. AFD explicitly constrains the difference between the forward noise and the noise sampled from the denoising distribution at the same time step, thereby further improving the quality of the generated gesture sequences.

where the output corresponds to the sampling path of the predicted noise distribution rather than the reconstructed motion. In this case, the quality of the generated gestures significantly deteriorates. Conversely, introducing an explicit reconstruction loss substantially improves the gesture quality. Based on empirical evidence, we conducted experiments with weights set to 1, 10, and 100, observing that the magnitude of the weight does not affect the FGD metric and the quality of the generated gestures.

Impact of Forward Noise Constraint. The term "w/o" denotes that we eliminate the forward loss and train the model solely based on adversarial learning. It is worth noting that while pure adversarial training can help the model learn the marginal distribution, the complexity of the large step-size distribution still results in a significant discrepancy between the posterior sampling outcomes and the forward process noise at the same time step, leading to a substantial decline in the FGD metric.

3.3. Comparison with Other Acceleration Methods

In Table 3, we juxtapose our proposed method with other bespoke acceleration strategies specifically designed for diffusion-based generative models. Specifically, our empirical results are benchmarked against strategies that have been accelerated employing DPM-Solver, as well as the original configurations of DDGAN [8] and SIDDMs [9]. The experimental findings reveal that for the DPM-Solver strategy, its first-order Taylor series expansion is tantamount to the widely recognized DDIM sampling strategy. Merely accelerating the sampling process by diminishing the number of sampling steps often culminates in imprecise approximations of the intricate multimodal distributions, thereby precipitating a drastic deterioration in the generation qual-

	BEATX					
	FGD↓	SA↑	BA↑	DIV↑	Inf. time↓	steps
DPM-Solver-1(DDIM) [5]	21.7	0.11	6.0	-	0.27	10
DPM-Solver-2 [5]	19.92	0.09	6.24	-	0.41	10
Naive DDGAN [8]	16.4	0.17	6.16	10.94	0.28	10
Naive SIDDMs [9]	16.2	0.19	6.27	11.5	0.26	10
HoloGest(Ours)	8.87	0.61	7.47	13.62	0.29	10

Table 3. Comparative results with contemporary accelerated diffusion methods are presented. Ensuring fairness, all methods employ

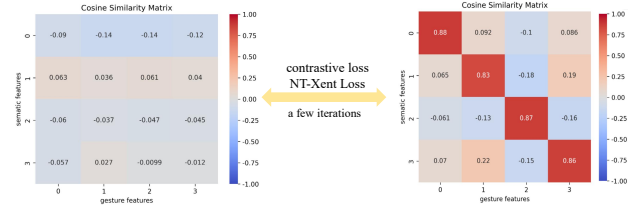


Figure 1. Illustration of training the gesture and transcription text latent space alignment module based on contrastive loss.

ity. Owing to the presence of second-order derivatives, the second-order Taylor series expansion necessitates invoking the denoising function twice at the midpoint during the sampling process, which inadvertently hampers the generation speed, while only offering marginal enhancements in the generation quality. Higher-order Taylor series expansions demand more frequent invocations of the denoising function, which is incongruous with the principal objective of acceleration.

Additionally, we conducted a comparative analysis of the acceleration strategies of the original DDGAN and SIDDm. Specifically, we trained an unconditional discriminator for DiffGes on the BEAT dataset, eschewing the explicit geometric loss (a deviation from our method), and juxtaposed it against our proposed method. The outcomes (as delineated in the table) underscore that the original configurations of the implicit and semi-implicit strategies exhibit subpar performance in terms of the overall quality of the generated gestures. This can be attributed to the fact that unlike images, human representations typically adhere to more stringent geometric constraints and necessitate more specific constraints.

3.4. Technical Details of Semantic Alignment Module

Drawing inspiration from GestureCLIP [1], we introduce a semantic gesture alignment module based on the JEPA structure. This method maps gestures and text data into a shared embedding space. Differing from GestureCLIP’s JEA, we abstract the latent representation of the audio and employ Speaker_id to bolster the prediction, seeking mutually predictable representations under supplementary conditions. Our JEPA employs a 12-layer ViT (Vision Transformer) structure and is trained using a contrastive learning-based strategy, as depicted in Figure 1. The gesture encoder bears similarity to the VAE structure in MLD [2], with mod-

Method	$PFC \downarrow$	$BA \uparrow$	$Dist_k$	$Dist_g$
GT	1.332	0.24	10.61	7.48
Bailando [6]	1.754	0.23	7.92	7.72
FACT [3]	2.2543	0.22	10.85	6.14
EDGE [7]	1.6815	0.27	9.17	7.22
HoloGest(Ours)	1.4913	0.292	9.81	7.49

Table 4. Quantitative comparison for the music and dance task. Using EDGE as the baseline, our semi-implicit decoupled structure significantly accelerates the generation speed of diffusion models while improving the generation quality. The introduction of motion priors substantially enhances the PFC value.

ifications made solely to the dimensions of the input and output. As illustrated in Table 1, our method outperforms JEA and GestureCLIP in terms of the SA (Semantic Alignment) metric, with GestureCLIP’s SA=0.58 and HoloGest’s SA=0.66. All evaluation configurations and computations are maintained consistent with those in the original GestureCLIP manuscript.

3.5. Inference Efficiency Analysis

Compared to other diffusion methods, our approach requires only 10 steps to generate output, with a 2-second motion slice taking just 0.29 seconds. For 50 steps, it takes 0.88 seconds. The transcription time from speech to text for each slice is 0.33 seconds, and when combined with the generation time, it remains under 2 seconds. By setting a buffer for two slices, we can satisfy the requirements of real-time applications. The response latency for each slice ranges from 0.62s (10 steps) to 1.21s (50 steps). After smoothing the transition action between the two buffers over 12 frames, the generated speech action sequence is output. As the maximum delay for each slice is 1.21s, which is less than its demonstration time of 2 seconds, no additional delay is needed for each slice after the two buffers.

4. Generalization Experiments

To evaluate our system’s core mechanisms, decoupled diffusion and motion priors, in the human motion generation domain, we extended the framework to the music-driven dance generation domain using the AIST++ [3] dataset. This dataset, not obtained through marker-based motion capture, contains artifacts like jitter and floating, causing physically unnatural issues like skating in methods like EDGE [7] and FACT [3].

We chose the diffusion-based EDGEE [7] as the baseline model and introduced the decoupled structure and locomotion prior module to validate our core ideas’ generalization capability. We trained a music-to-dance generation model on AIST++ and compared it with contemporary methods. Results in Table 4 show that our method signifi-

cantly improves generation quality and speed compared to non-diffusion-based methods and EDGE, while producing stable and natural motion results. PFC represents the physical feasibility of footsteps (lower values are better), BA represents beat alignment, and $Dist_k$ and $Dist_g$ represent diversity measurements of generated dances on “dynamic” and “geometric” levels, respectively. *For more qualitative comparisons, please refer to the accompanying video.*

5. Applications

Our proposed method holds immense potential to revolutionize various domains that involve real-time virtual human interaction. Below, we discuss two prominent application areas where our method could have a significant impact.

5.1. Virtual Assistants

Consider a scenario where a user is interacting with an advanced virtual assistant, such as an enhanced version of ChatGPT. By incorporating our method, the virtual assistant would not only be able to respond verbally but also utilize gestures to communicate, adding a layer of non-verbal communication that plays a crucial role in human interaction. The real-time generation of gestures based on the user’s speech input would make the interaction more engaging, immersive, and lifelike, thereby elevating the overall user experience.

5.2. Video Gaming

In the context of video gaming, our method could be employed to augment the realism of non-player characters (NPCs). By leveraging our approach, NPCs could generate gestures for their dialogues in real-time, adding depth to their character and enriching the gaming experience. This application could potentially redefine the standards of character realism in video games, adding a new dimension to the gaming experience and setting a new benchmark for future advancements in the field.

We highly recommend the readers refer to the accompanying video for the qualitative results obtained using our developed prototype system.

References

- [1] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with CLIP latents. *ACM Trans. Graph.*, 42(4):42:1–42:18, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2
- [3] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance

- generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 3
- [4] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling. *arXiv preprint arXiv:2401.00374*, 2023. 1
- [5] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 2
- [6] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 3
- [7] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. Edge: Editable dance generation from music. *arXiv preprint arXiv:2211.10658*, 2022. 3
- [8] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022. 2
- [9] Yanwu Xu, Mingming Gong, Shaoan Xie, Wei Wei, Matthias Grundmann, Tingbo Hou, et al. Semi-implicit denoising diffusion models (siddms). *arXiv preprint arXiv:2306.12511*, 2023. 2
- [10] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffus-estylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. 2023. 1