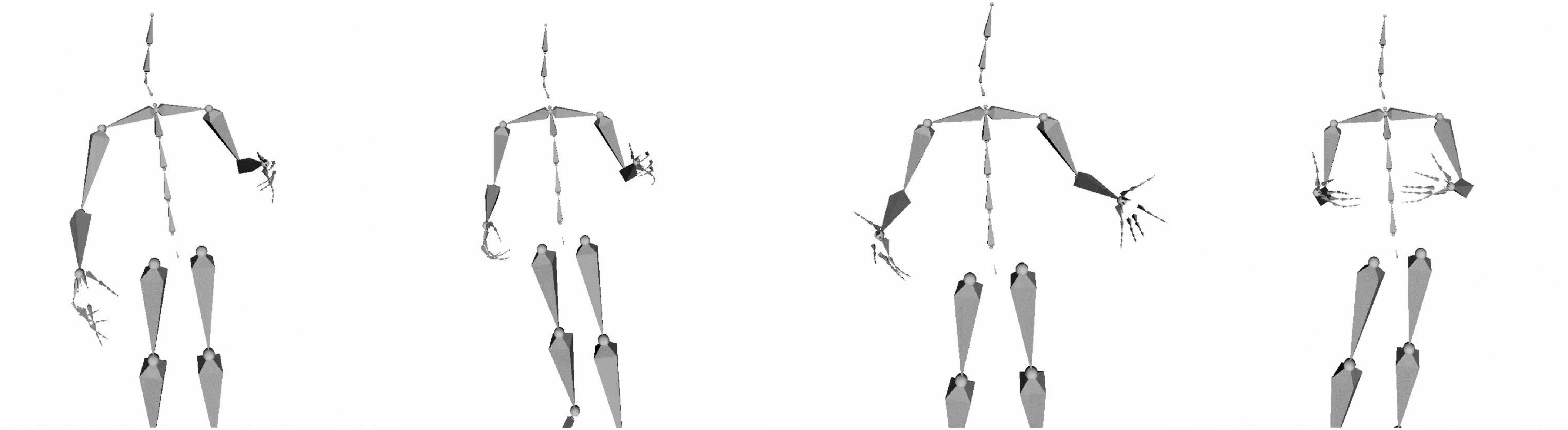
QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speech-Driven Gesture Generation

Sicheng Yang[1], Zhiyong Wu[1,4], Minglei Li[2], Zhensong Zhang[3], Lei Hao[3], Weihong Bao[1], Haolin Zhuang[1]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2] Huawei Cloud Computing Technologies Co., Ltd, China
[3] Huawei Noah's Ark Lab, China
[4] The Chinese University of Hong Kong, Hong Kong SAR, China

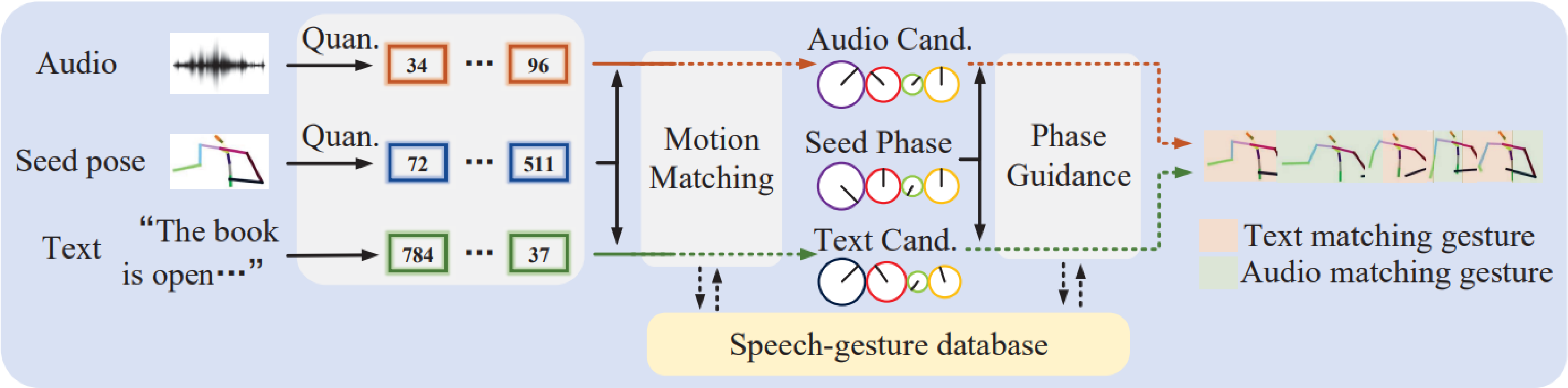
Random jitters



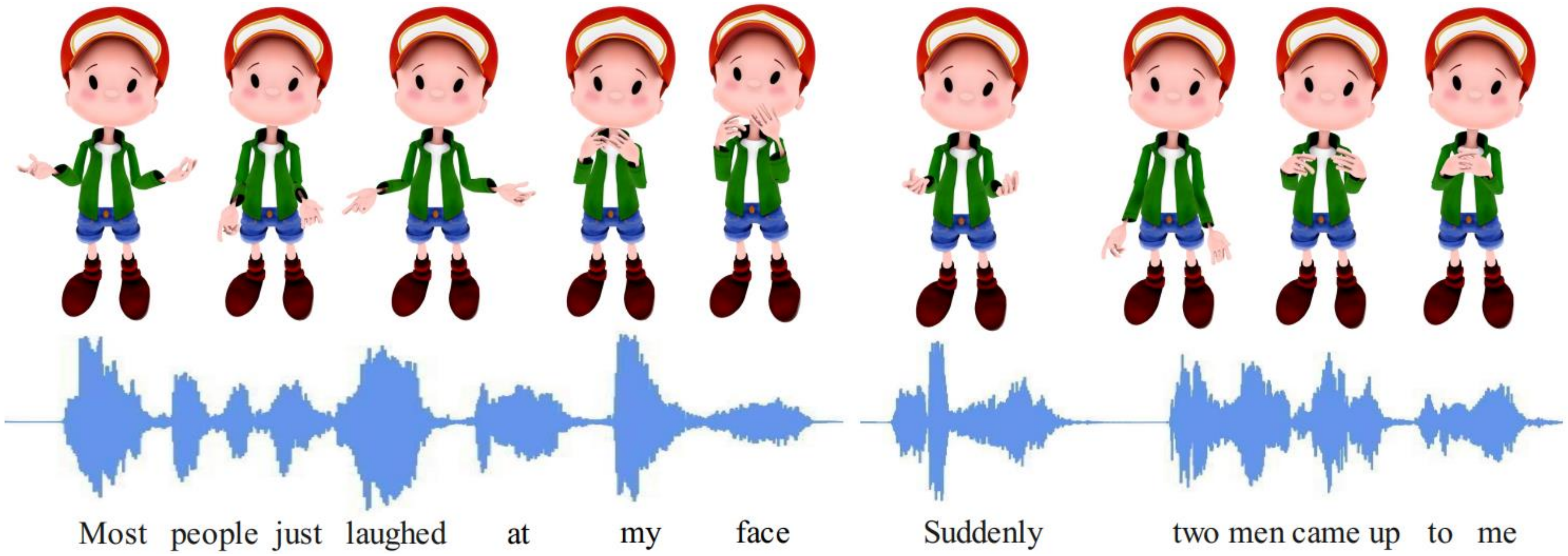
Inherent asynchronous relationship

[1] *BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis.*
[2] *https://www.huaweicloud.com/product/cbs/digitalhuman.html*

Pipeline



Gesture examples



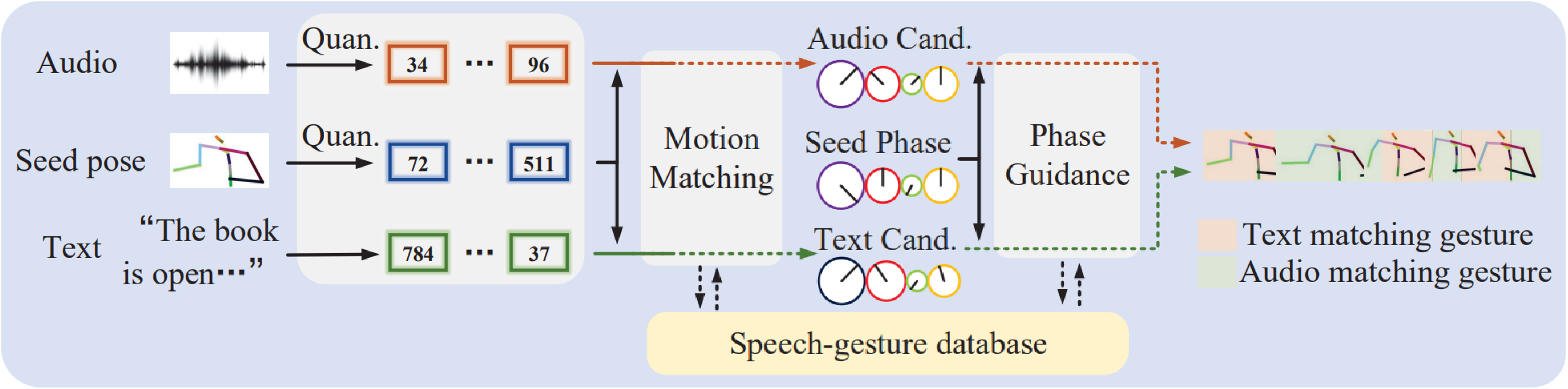
Motivation

- Problems :
  - Random jittering
  - Inherent asynchronicity with speech
- Goal:
  - Solve jittering problems, such as grabbing hands or pushing glasses
  - Better alignment of speech and gestures
  - Further improve the quality of gesture generation

Contribution

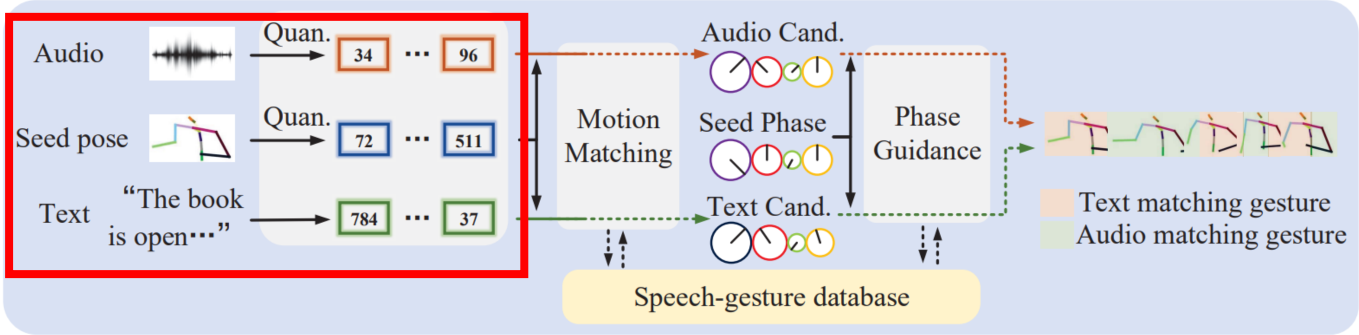
- A novel quantization-based motion matching framework for speech-driven gesture generation.
- Align diverse gestures with different speech using Levenshtein distance.
- A phase guidance strategy to select optimal audio and text candidates.

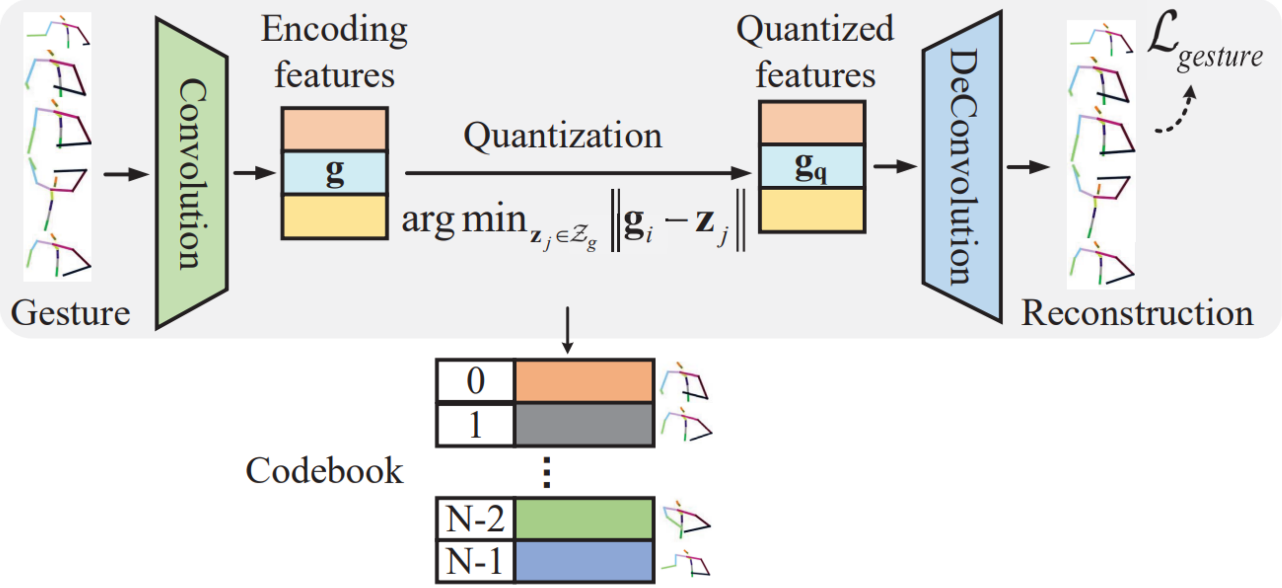
Approach



Overall

Approach

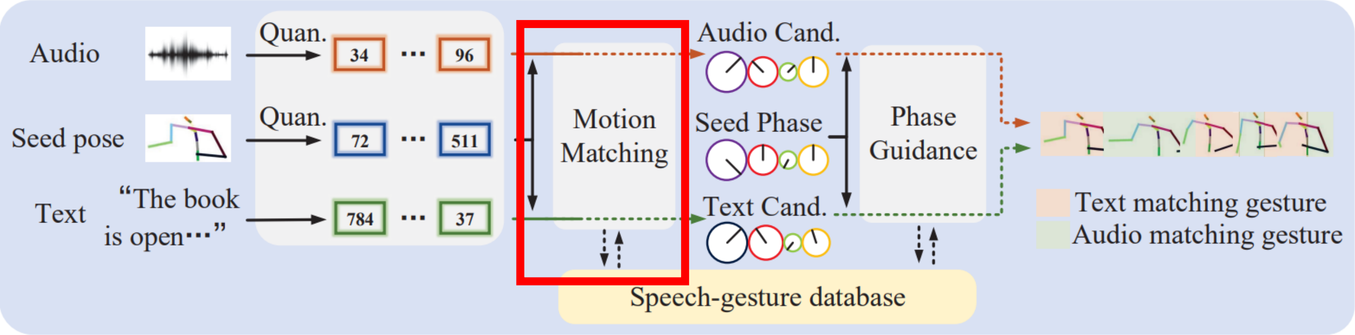




Gesture VQ-VAE

- Gesture VQ-VAE
  - Encode the joint sequence
    $$\mathbf{g} = E_g(\mathbf{G})$$
  - Decoder
    $$\mathbf{G}_1 = D_g\left(\mathbf{g}_q\right) = D_g(\mathbf{q}(E_g(\mathbf{G})))$$
  - Loss function
    $$\mathcal{L}_{gesture(E_g, D_g, \mathcal{Z}_g)} = \left\| \mathbf{G}_1 - \mathbf{G}_1 \right\|_1 + \alpha_1 \left\| \mathbf{G}_1' - \mathbf{G}_1' \right\|_1 + \alpha_2 \left\| \mathbf{G}_1'' - \mathbf{G}_1'' \right\|_1 + \left\| \mathrm{sg}[\mathbf{g}] - \mathbf{g_q} \right\| + \beta \left\| \mathbf{g} - \mathrm{sg}\left[\mathbf{g_q}\right] \right\|$$
- vq-wav2vec
- Sentence-BERT

Approach



Current segment audio →
Quantized audio database →
Sort using Levenshtein distance → $\mathbf{C}_a$ { 34, 50, …, 96 }

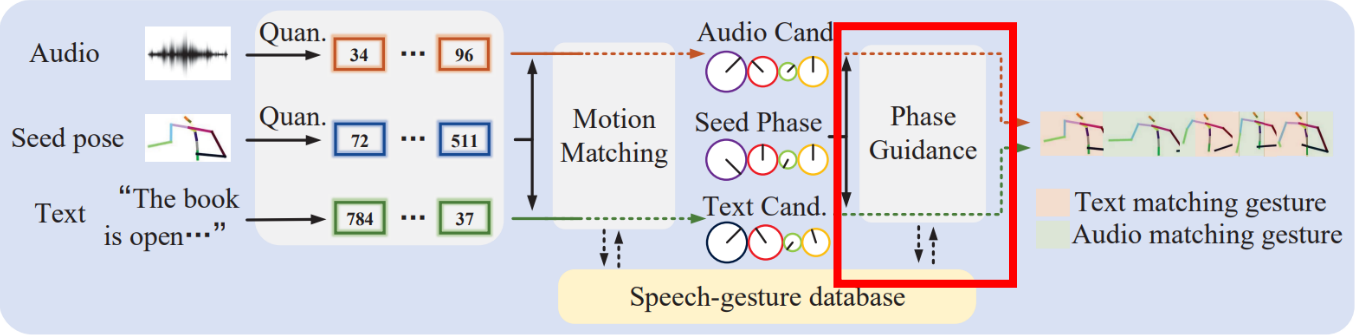
Initial/Previously searched pose →
Quantized pose database →
Sort using Euler distance → $\mathbf{C}_g$ { 72, 320, …, 511 }

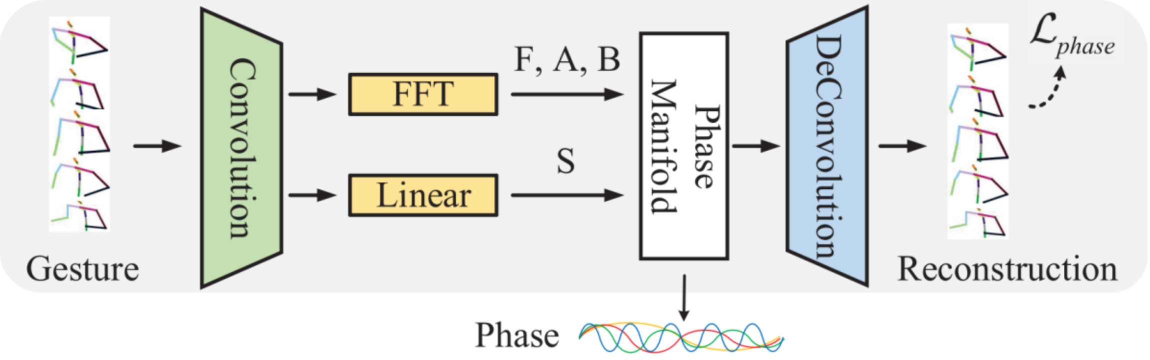
Current segment audio →
Text database →
Sort using Cosine distance → $\mathbf{C}_t$ { 784, 425, …, 37 }

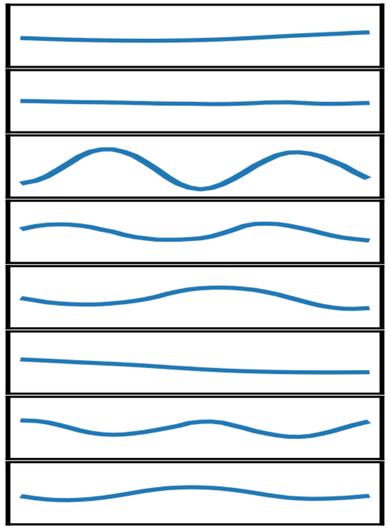
$\mathbf{C}_a + \mathbf{C}_g \xrightarrow[\text{weighting}]{\text{Ranking}}$ Audio candidate $\mathbf{C}_a$

$\mathbf{C}_t + \mathbf{C}_g \xrightarrow[\text{weighting}]{\text{Ranking}}$ Text candidate $\mathbf{C}_t$

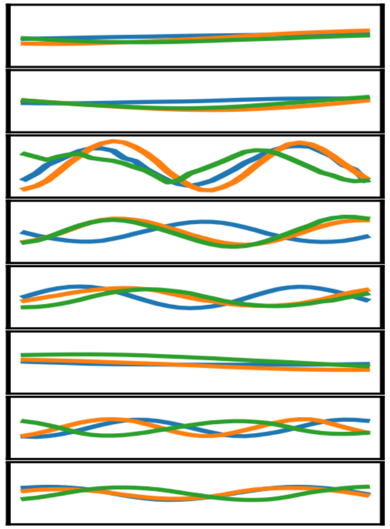
Approach







Phase manifold of the seed code



Phase manifold of the candidate

- Encode the joint sequence

$$\mathbf{L} = E_p(\mathbf{G})$$

- Periodic parameters

$$\mathbf{A}_i = \sqrt{\frac{2}{T}\sum_{j=1}^{K}\mathbf{p}_{i,j}}, \quad \mathbf{F}_i = \frac{\sum_{j=1}^{K}\left(\mathbf{f}_j \cdot \mathbf{p}_{i,j}\right)}{\sum_{j=1}^{K}\mathbf{p}_{i,j}}, \quad \mathbf{B}_i = \frac{\mathbf{c}_{i,0}}{T},$$

$$\left(s_x, s_y\right) = FC\left(\mathbf{L}_i\right), \quad \mathbf{S}_i = \operatorname{atan}2\left(s_y, s_x\right)$$

$$\mathbf{L} = f(\mathcal{T};\mathbf{A},\mathbf{F},\mathbf{B},\mathbf{S}) = \mathbf{A}\cdot\sin(2\pi\cdot(\mathbf{F}\cdot\mathcal{T} - \mathrm{S})) + \mathbf{B}$$

- Loss function $\mathcal{L}_{phase} = \mathcal{L}_{phase-recon}(\mathbf{G}, h(\mathbf{L}))$

Experiments

- Dataset and processing
  - BEAT dataset
  - 15 joints corresponding to the upper body
  - 8:1:1 by training, validation, and testing
- Experiment setup
  - VQ-VAE: ADAM optimizer (learning rate is e^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.98$). Batch size 128. 200 epochs. $\beta = 0.1$, $\alpha_1 = 1$, $\alpha_2 = 1$. Down-sampling rate is 8. T = 240.
  - Motion matching: Window lengths is 4 pose codes. d=32.
  - Phase guidance: AdamW optimizer (weight decay 10^{-4}). Batch size 128. 100 epochs. Phase channels M is 8. $N_{\text{phase}} = 8$, $N_{\text{stride}} = 3$.

Evaluation

Name	Objective evaluation			Subjective evaluation	
	Hellinger distance average ↓	FGD on feature space ↓	FGD on raw data space ↓	Human-likeness	Appropriateness
Ground Truth (GT)	0.0	0.0	0.0	3.79 ± 0.19	3.62 ± 0.21
End2End [47]	0.146	64.990	16739.978	3.64 ± 0.11	3.23 ± 0.14
Trimodal [46]	0.155	48.322	12869.98	3.31 ± 0.17	3.20 ± 0.19
StyleGestures [5]	**0.136**	35.842	9846.927	3.66 ± 0.08	3.30 ± 0.11
KNN [17]	0.364	43.030	12470.061	2.38 ± 0.10	2.35 ± 0.13
CaMN [31]	0.149	52.496	10549.455	3.65 ± 0.16	3.29 ± 0.15
Ours	**0.136**	**19.921**	**5742.281**	$\mathbf{4.00 \pm 0.14}$	$\mathbf{3.66 \pm 0.23}$

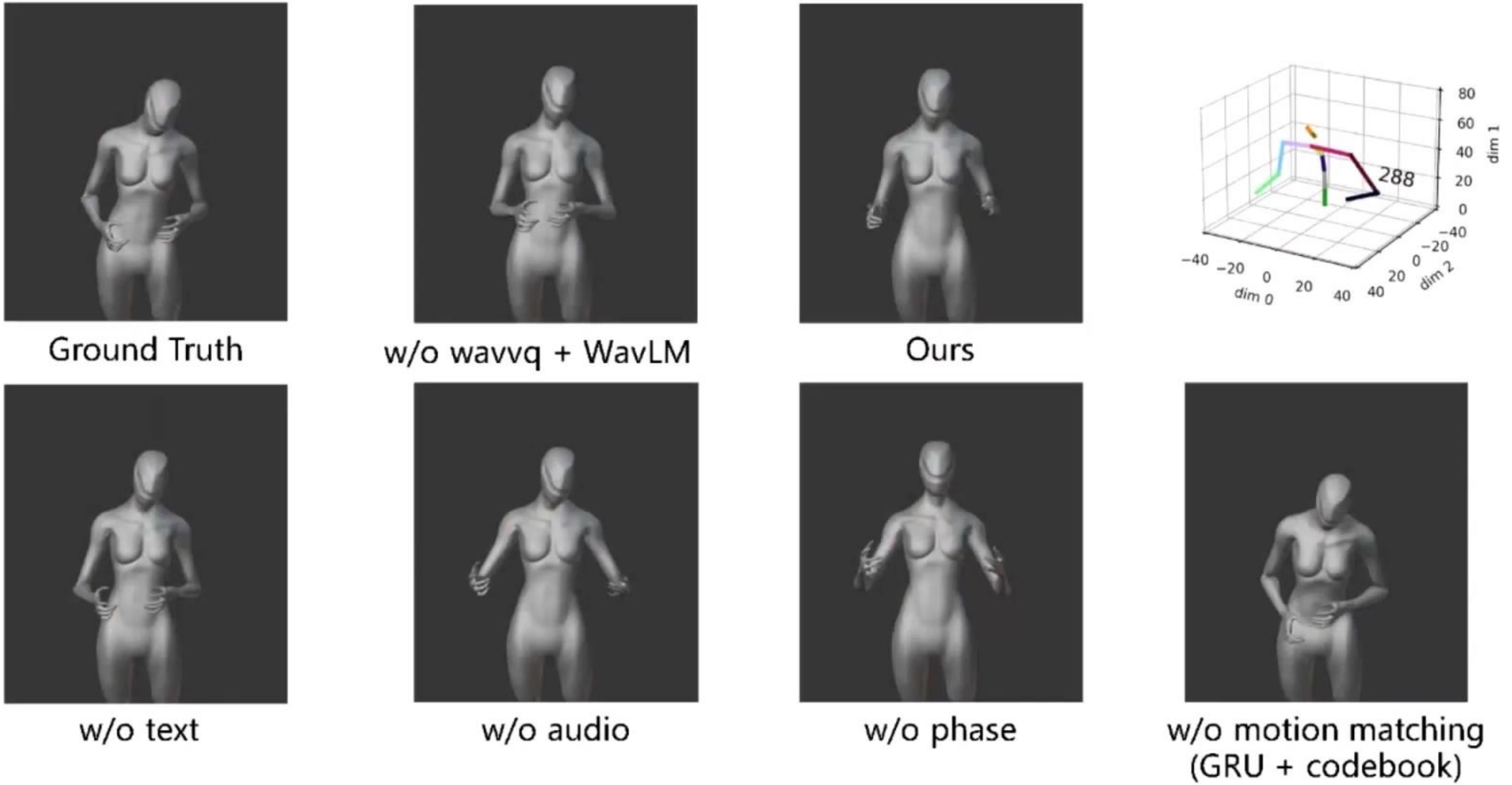
- Our method outperforms all existing methods in an objective evaluation.
- Compared to the best baseline model (StyleGestures), our method significantly improves human-likeness and appropriateness, with no significant difference in appropriateness compared to ground truth (GT).

Ablation Studies

Name	Objective evaluation			Subjective evaluation	
	Hellinger distance average ↓	FGD on feature space ↓	FGD on raw data space ↓	Human-likeness	Appropriateness
w/o wavvq + WavLM	0.151	19.943	6009.859	3.87 ± 0.21	3.64 ± 0.21
w/o audio	0.134	20.401	5871.044	3.87 ± 0.21	3.63 ± 0.20
w/o text	**0.118**	23.929	6389.866	3.57 ± 0.29	3.41 ± 0.23
w/o phase	0.138	**19.195**	5759.167	3.90 ± 0.11	3.65 ± 0.17
w/o motion matching (GRU + codebook)	0.140	30.404	11642.641	3.78 ± 0.14	3.43 ± 0.16
Ours	0.136	19.921	**5742.281**	$\mathbf{4.07 \pm 0.15}$	$\mathbf{3.77 \pm 0.21}$

- Without wavvq but use WavLM instead
  - All metrics deteriorate.
- Without audio or text
  - FGD metric increases, Hellinger distance average decreases
- Without phase guidance
  - Objective evaluations changed insignificantly, and subjective evaluations became worse
- Without motion matching but use GRU instead
  - Subjective evaluation results are the worst

Ablation Studies



Text: So now I have less money. This was an awful experience.

Conclusion

- Employing discrete gestures encoding
  - Address random jittering
- Levenshtein distance based on audio quantization
  - Solve the issue of speech and gesture asynchrony
  - Motion matching model inflexibility
- Phase-guided gesture generation
  - Switching candidates based on speech and text

Project page