A. The design choice of our method

A gesture motion is composed of a sequence of gesture units, such as swiping hands from left to right and holding hands at a position [7]. We design our algorithm according to this observation, paying special attention to the construction and selection of gesture units. Specifically, due to the good performance of VQ-VAE in quantization, we trained a gesture VQ-VAE for 200 epochs to mine these gesture units from the dataset, similar to existing works [7, 16, 21, 34] [A56]. In our settings, each code corresponds to one gesture unit which is 8 (d) frames of gesture motions. Unlike Bailando [38], Gensture2Vec [A58] and VQ-Text2Sign [A57] using position as input features, our VQ-VAE is trained with rotation instead, which can represent the motion better. To find gesture candidates that match a given piece of audio and corresponding text, we quantize the audio first, because our ablation studies in Table 2 illustrate that the Levenshtein distance based on discrete audio alleviates the inherent asynchrony problem of gesture and audio, and achieves better results than the non-discrete counterpart. We don't need to quantize the text, since it is discrete already.

In terms of audio quantization, the audio is represented by two groups containing 320 tokens, for a total of 320^2 results, or 102.4K tokens. $\mathbf{G}$ refers to the motion sequence (position, velocity, acceleration, rotation, Euler angles, quaternions, etc.). We use rotation for VQ-VAE, and rotation velocity for periodic autoencoder. For motion matching, we first calculate $\hat{\mathbf{C}}_a$ and $\hat{\mathbf{C}}_t$ based on the audio and text. We also calculate the distance between all gesture codes in codebook and the previous pose code $\mathbf{g}_{-1}$ to obtain $\hat{\mathbf{C}}_g$ for motion coherence. Then $\hat{\mathbf{C}}_g$ and $\hat{\mathbf{C}}_a$ determine audio-based candidate and $\hat{\mathbf{C}}_g$ and $\hat{\mathbf{C}}_t$ determine text-based candidate. The final gesture is selected according to the continuity of the phase of the previous gesture and the phase of the two candidate gestures.

B. Proposed Algorithm

A more detailed and procedural description of our proposed QPGesture approach is shown in Algorithm 1.

Algorithm 1: QPGesture search

```
Data: database contains quantized audio, quantized
      gesture, context, and phase
Input: a discrete text sequence
       $\mathbf{t} = [\mathbf{t}_0, \mathbf{t}_1, \ldots, \mathbf{t}_{T'-1}]$, a discrete audio
       sequence $\mathbf{a_q} = [\mathbf{a}_{\mathbf{q},0}, \mathbf{a}_{\mathbf{q},1}, \ldots, \mathbf{a}_{\mathbf{q},T'-1}]$,
       initial pose code $\mathbf{g}_{-1}$, initial phase $\mathcal{P}_{-1}$,
       $k \in \mathbb{Z}$, the desired k-best candidates, control
       masks $\mathbf{M} = [\mathbf{m}_0, \mathbf{m}_1, \ldots, \mathbf{m}_{T'-1}]$
       (optional)
Output: $\hat{\mathbf{G}}_o = [\hat{\mathbf{G}}_{o,0}, \hat{\mathbf{G}}_{o,1} \ldots, \hat{\mathbf{G}}_{o,T'-1}]$
t = 0, codebook size $C_b$
initialize $\hat{\mathbf{G}}_o = [\mathbf{g}_{-1}]$, $\hat{\mathcal{P}}_o = [\mathcal{P}_{-1}]$,
while $t$ in len(testing dataset) do
    c_dist = []×$C_b$, c_a = []×$C_b$, c_t = []×$C_b$
    a_dist = [∞]×$C_b$, t_dist = [∞]×$C_b$
    for $code = 0; code < C_b$ do
        $c\_dist[code] =$
          $d(D_g\left(\hat{\mathbf{G}}_o[-1]\right), D_g\,(code))$
    for $s = 0; s < len(database)$ do
        for $code$ in $database[s]$ do
            if $\mathbf{m}_s$ is not masked then
                if $d(quantized\ audio[s][code]) <$
                  $a\_dist[code]$ then
                    $a\_dist[code] =$
                      $d(quantized\ audio[s][code])$
                    $c\_a[code] =$
                      $quantized\ audio[s][code : code + stepsize - 1]$
                if
                  $d(context[s][code]) < t\_dist[code]$
                  then
                    $t\_dist[code] =$
                      $d(context[s][code])$
                    $c\_t[code] = context[s][code :$
                      $code + stepsize - 1]$
    $R_c = relrank(c\_dist)$, $R_a = relrank(a\_dist)$,
      $R_t = relrank(t\_dist)$
    $R_{c,a} = R_c + R_a$ (elem. wise)
    $R_{c,t} = R_c + R_t$ (elem. wise)
    sort $R_{c,a}$, sort its indices into $I_{c,a}$
    sort $R_{c,t}$, sort its indices into $I_{c,t}$
    $\hat{\mathbf{C}}_{a,t} = I_{c,a}[k]$, $\hat{\mathbf{C}}_{t,t} = I_{c,t}[k]$
    if $d(concat[\hat{\mathcal{P}}_o[-1]^{[(N_{strid}-N_{phase}):]}, \mathcal{P}_{a,t}^{[N_{strid}:]}],$
      $concat[\hat{\mathcal{P}}_o[-1]^{[-N_{strid}:]}, \mathcal{P}_{a,t}^{[(N_{phase}-N_{strid}):]}]) <$
      $d(concat[\hat{\mathcal{P}}_o[-1]^{[(N_{strid}-N_{phase}):]}, \mathcal{P}_{t,t}^{[N_{strid}:]}],$
      $concat[\hat{\mathcal{P}}_o[-1]^{[-N_{strid}:]}, \mathcal{P}_{t,t}^{[(N_{phase}-N_{strid}):]}])$
      then
        append($\hat{\mathbf{G}}_o, \hat{\mathbf{C}}_{a,t}$), append($\hat{\mathcal{P}}_o, \hat{\mathcal{P}}_{a,t}$)
    else
        append($\hat{\mathbf{G}}_o, \hat{\mathbf{C}}_{t,t}$), append($\hat{\mathcal{P}}_o, \hat{\mathcal{P}}_{t,t}$)
return $\hat{\mathbf{G}}_o[1 :]$
```

C. Dataset and processing.

We chose BEAT dataset because to our knowledge it is the largest publicly available motion capture dataset. And we will add more results of the baseline model for comparison later.

Since 2D datasets converted to 3D coordinates (pseudo GT) are low quality that are difficult to use, we plan to add more experiments on other motion capture datasets. Even based on motion capture, the hand quality of most datasets is still low [A52]. Datasets claimed with high-quality hand motion capture were still reported to have poor hand motion, e.g., ZEGGS Dataset in [A54] and Talking With Hands in [48]. We found the hand quality of BEAT is not good enough, especially when retargeted to an avatar, so we ignore hand motion currently, and leave it to future work.

D. Details of Baseline Implementation

We used the 15 joints of the upper body(spine, spine1, spine2, spine3, head, neck, neck1, L/R shoulders, L/R arms, and L/R forearms, L/R hands). The gestures for all models were at 60 frames per second (fps). Because we found that using a pre-trained model to extract features was better than using 1D convolution, for Trimodal [46], we used WavLM features instead of the original 1D convolution, while aligning the temporal dimensions using linear interpolation. For KNN [17], we found that changing the step size from 2 frames at the original 15 fps to 30 frames at 60 fps had comparable results. However, we found that generating fake gestures for training the GAN in the second stage without overlapping frames and with 5 frames as the step size takes several months, which is intolerable. This could be due to 1) a large amount of data in the BEAT dataset itself, 2) the significant increase in the number of frames at 60 fps, and 3) the time-consuming KNN search itself (the time complexity of KNN search is O(n^4) compared to time complexity of O(n^2) of our method using audio quantization and gesture quantization). So we used mismatched gestures instead of KNN-matched gestures with 50% likelihood from top2-top15 in the original KNN method as the gestures used for training the GAN in the second stage. For CaMN [31], at the time we used the BEAT dataset, facial modality was not yet available[1], so we used text, speech, speaker identity, and emotion as inputs to the CaMN network.

E. Objective evaluation

E.1. Evaluation Metrics

Average jerk and Acceleration. The third and second time derivatives of the joint positions are called jerk and acceleration [A55], respectively. The average of these two metrics is usually used to evaluate the smoothness of the motion. A natural system should have the average jerk and acceleration similar to natural motion.

Canonical Correlation Analysis. The purpose of Canonical correlation analysis (CCA) [A56] is to project two sets of vectors into a joint subspace and then find a sequence of linear transformations of each set of variables that maximizes the relationship between the transformed variables. CCA values can be used to measure the similarity between the generated gestures and the real ones. The closer the CCA to 1, the better.

Diversity and Beat Align Score. We use the method in [29] to calculate the beats of audio, and follow [38] to calculate the beats and diversity of gesture. The greater these metrics are, the better.

E.2. Objective Evaluation Results

We used Trinity dataset to calculate FGD because both Trinity and BEAT are captured with Vicon, having the same names and number of joints, as in [46]. The results of our additional objective evaluation compared to the existing model are shown in Table 3. From the results, we can observe that KNN performs better than our proposed framework on three metrics: average jerk, average acceleration and global CCA. StyleGestures performs best on Average acceleration. And Trimodal has the best performance on CCA for each sequence. We can see that our model is the best match to the beats of the audio, but not as good as StyleGesture in terms of diversity. The video results show that StyleGesture has a lot of cluttered movements, increasing diversity while decreasing human-likeness and appropriateness.

The results of additional objective evaluations of our ablation studies are shown in Table 4. When we do not use vq-wav2vec or Levenshtein distance to measure the similarity of corresponding speech of gestures, but use WavLM and cosine similarity instead, the average jerk and average acceleration are worst. When the framework is inferenced without text, the average jerk, average acceleration and CCA for each sequence are better, but the global CCA is decreased. When the model is trained using deep gated recurrent unit (GRU) to learn pose code instead of motion matching, the best CCA for each sequence is obtained. For diversity, more diverse may indicate a more clutter-free gesture; and for scores, a better match with rhythm does not indicate a better semantic match. These objective measures are not consistent with subjective scoring.

However, this is consistent with current human subjective perception [26, 48] that speech-driven gestures lack proper objective metrics, even for FGD [A53]. Current research on speech-driven gestures prefers to conduct only subjective evaluation [A54]. In conclusion, we would like to emphasize that objective evaluation is currently not particularly relevant for assessing gesture generation [26]. Subjective evaluation remains the gold standard for comparing gesture generation models [26].

F. User Study

Segments should be more or less complete phrases, starting at the start of a word and ending at the end of a word. We made sure there were no spoken phrases that ended on a “cliffhanger” in the evaluation. The user study was conducted by subjects with good English proficiency. The reward is about 7.5 USD each person, which is about the average wage level [48]. More detailed demographic data of the

[1] https://pantomatrix.github.io/BEAT-Dataset/

Table 3. Quantitative results on test set. Bold indicates the best metric, i.e. the one closest to the ground truth.

Name	Average jerk	Average acceleration	Global CCA	CCA for each sequence	Diversity on feature space ↑	Diversity on raw data space ↑	Beat Align Score ↑
Ground Truth	996.32 ± 235.86	31.89 ± 6.80	1.000	1.00 ± 0.00	2.81	50.87	0.2064
End2End [47]	143.68 ± 10.45	7.09 ± 0.34	0.429	0.72 ± 0.14	1.45	20.82	0.2370
Trimodal [46]	157.87 ± 12.08	7.98 ± 0.53	0.807	**0.74** ± **0.19**	1.91	17.21	0.1221
StyleGestures [5]	280.44 ± 21.43	**23.58** ± **7.21**	0.953	0.71 ± 0.12	**5.80**	**29.88**	0.1871
KNN [17]	**423.83** ± **100.10**	40.77 ± 8.12	**0.998**	0.63 ± 0.21	3.23	19.42	0.2009
CaMN [31]	159.54 ± 13.99	8.96 ± 0.55	0.626	0.70 ± 0.17	2.26	18.60	0.1489
Ours	182.11 ± 18.15	9.87 ± 0.66	0.985	0.69 ± 0.14	4.05	23.13	**0.2557**

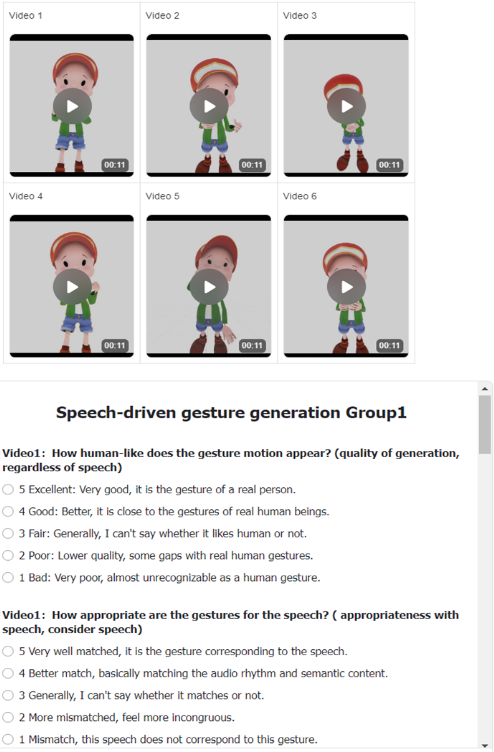
Table 4. Ablation studies results. 'w/o' is short for 'without'. Bold indicates the best metric, i.e. the one closest to the ground truth.

Name	Average jerk	Average acceleration	Global CCA	CCA for each sequence	Diversity on feature space ↑	Diversity on raw data space ↑	Beat Align Score ↑
Ground Truth (GT)	996.32 ± 235.86	31.89 ± 6.80	1.000	1.00 ± 0.00	2.81	50.87	0.2064
w/o wavvq + WavLM	168.09 ± 22.44	9.18 ± 0.81	**0.993**	0.69 ± 0.13	8.49	18.82	0.2098
w/o audio	176.84 ± 14.61	9.60 ± 0.50	**0.993**	0.68 ± 0.13	8.42	**25.83**	0.2001
w/o text	**196.61** ± **29.34**	**10.68** ± **1.22**	0.961	0.71 ± 0.15	7.53	15.78	0.1699
w/o phase	176.94 ± 21.41	9.60 ± 0.80	0.986	0.72 ± 0.13	4.83	15.30	**0.3076**
w/o motion matching (GRU + codebook)	141.52 ± 9.65	7.56 ± 0.56	0.694	**0.75** ± **0.14**	**10.98**	12.51	0.2303
Ours	182.11 ± 18.15	9.87 ± 0.66	0.985	0.69 ± 0.14	4.05	23.13	0.2557

subjects who participated in the subjective evaluation are as follows.

- Gender: Participants were approximately 90% were male and 10% were female.
- Region: They were overwhelmingly residents of mainland China, and one international student from Malaysia. They are all students from our lab[2].
- Age: All participants were between the ages of 20-28.

The questions for user study follow GENEA 2022 [48]. If there is no overlap in the 95% confidence intervals of the ratings between the different models, then the difference is considered to be statistically significant.



Speech-driven gesture generation Group1

* **Video1: How human-like does the gesture motion appear? (quality of generation, regardless of speech)**

○ 5 Excellent: Very good, it is the gesture of a real person.
○ 4 Good: Better, it is close to the gestures of real human beings.
○ 3 Fair: Generally, I can't say whether it likes human or not.
○ 2 Poor: Lower quality, some gaps with real human gestures.
○ 1 Bad: Very poor, almost unrecognizable as a human gesture.

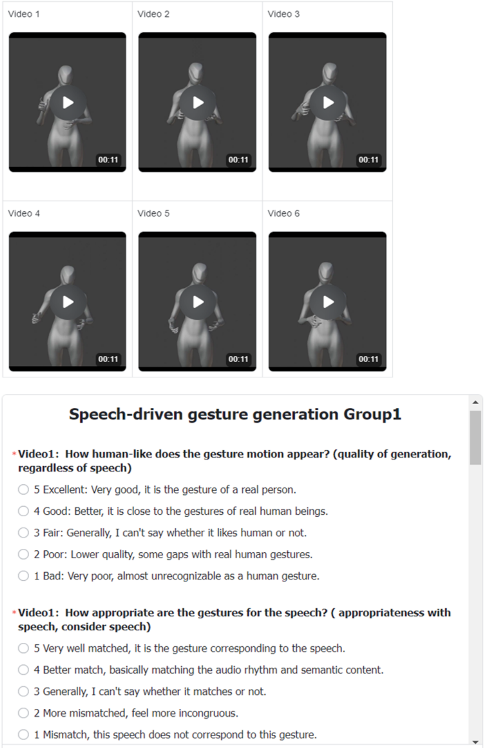
* **Video1: How appropriate are the gestures for the speech? (appropriateness with speech, consider speech)**

○ 5 Very well matched, it is the gesture corresponding to the speech.
○ 4 Better match, basically matching the audio rhythm and semantic content.
○ 3 Generally, I can't say whether it matches or not.
○ 2 More mismatched, feel more incongruous.
○ 1 Mismatch, this speech does not correspond to this gesture.

Figure 9. Screenshot of the parallel rating interface from the user study for comparison with existing methods.

The experiment is conducted with 23 participants with good English proficiency to evaluate the human-likeness and appropriateness. We use two avatar characters to test the robustness of the results, both of them are publicly accessible. During the evaluation, we prompted the participants to ignore the finger movements and lower body movements, as well as to ignore the problems in skeletal rigging and to pay attention to the upper body gestures. For human-likeness, it is mainly to evaluate whether the motion of the avatar looks like the motion of a real human. In terms of appropriateness, it is the evaluation of whether the motion of the avatar is appropriate for the given speech. A screenshot of the evaluation interface used for comparison with existing methods is presented in Figure 9. An example of

[2] https://thuhcsi.github.io/labintro.html



Speech-driven gesture generation Group1

* **Video1: How human-like does the gesture motion appear? (quality of generation, regardless of speech)**

○ 5 Excellent: Very good, it is the gesture of a real person.

○ 4 Good: Better, it is close to the gestures of real human beings.

○ 3 Fair: Generally, I can't say whether it likes human or not.

○ 2 Poor: Lower quality, some gaps with real human gestures.

○ 1 Bad: Very poor, almost unrecognizable as a human gesture.

* **Video1: How appropriate are the gestures for the speech? (appropriateness with speech, consider speech)**

○ 5 Very well matched, it is the gesture corresponding to the speech.

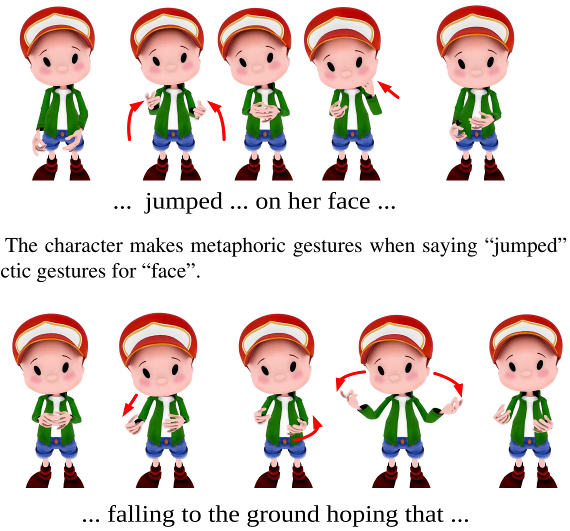
○ 4 Better match, basically matching the audio rhythm and semantic content.

○ 3 Generally, I can't say whether it matches or not.

○ 2 More mismatched, feel more incongruous.

○ 1 Mismatch, this speech does not correspond to this gesture.

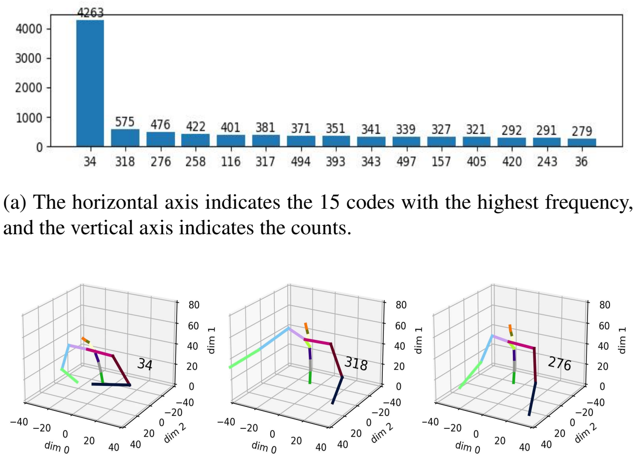
Figure 10. Screenshot of the parallel rating interface from the user study for ablation studies.


... jumped ... on her face ...
... falling to the ground hoping that ...

(a) The character makes metaphoric gestures when saying “jumped” and deictic gestures for “face”.

(b) The character makes beat gestures when saying “falling”, “ground” and “hoping”.

Figure 11. Sample results of co-speech gesture generation from our method. Motion history images for some parts are depicted along with the speech text.



(a) The horizontal axis indicates the 15 codes with the highest frequency, and the vertical axis indicates the counts.

(b) 3D joints visualization of the first three codes.

Figure 12. The histogram of the first 15 code frequencies of speaker “wayne” and 3D joints visualization results of the first three codes.

the evaluation interface for ablation studies can be seen in Figure 10. Participants reported that the gestures generated by our framework contain many semantic and rhythmically related gestures, as shown in the figure 11. Please refer to our supplementary video for comparisons with the baseline model and ablation studies.

G. Controllability

For the speaker “wayne”, the histogram of the first 15 code frequencies is shown in the Figure 12. It can be seen that the most frequent code is ‘34’, which can be considered to represent the average gesture, that is, the gesture without speech and in silence. We visualized the three most frequent codes: ‘34’, ‘318’ and ‘276’, and we can find that ‘318’ is a code with a preference for right-handedness. We chose a very typical motion clip using the right-handedness (72s to 76s of gesture “1_wayne_0_87_94”), a 4s video with a total of 30 codes at 60FPS and 8 codebook sampling rates, of which there are twelve ‘318’ codes. We use a code with a preference for left-handedness instead of ‘318’ (e.g. ‘260’), and the results are shown in our supplementary video.

Appendix References

[A52] Nyatsanga *et al.* A comprehensive review of data-driven co-speech gesture generation. *arXiv:2301.05339*, 2023. 11

[A53] Rishabh Dabral *et al.* Mofusion: A framework for denoising-diffusion-based motion synthesis. *arXiv:2212.04495*, 2022. 12

[A54] Simon Alexanderson *et al.* Listen, denoise, action! audio-driven motion synthesis with diffusion models. *arXiv:2211.09707*, 2022. 11, 12

[A55] 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*, IVA '19, page 97–104, New York, NY, USA, 2019. Association for Computing Machinery. 12

[A56] Najmeh Sadoughi and Carlos Busso. Speech-driven animation with meaningful behaviors. *Speech Commun.*, 110:90–100, 2019. 11, 12

[A57] Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu. Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141*, 2022. 11

[A58] Payam Jome Yazdian, Mo Chen, and Angelica Lim. Gesture2vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3100–3107, 2022. 11