We thank the reviewers for the constructive comments and suggestions. Below we address the main issues. We will fix all other minor issues in the revised paper.

R1, R3: Dance unit. We will elaborate on the dance units, which are not a simple division of a dance sequence. We first extract kinematic beats and then align the extracted beat times to the pre-defined time steps (i.e., 8, 16, 24 and 32 in a dance unit of length 32). Finally, we interpolate and extrapolate the original pose sequences to obtain the temporally normalized dance units, which are further encoded into the disentangled initial pose space and movement space.

R2: Title. Thanks for the suggestion. We will change the title in the revision.

R2: Term usage. (1) **Abstract:** We agree with the comment and will remove the terms "top-down" and "bottom-up" in the abstract. (2) **Multimodality:** We defined the term "multimodality" in L272 and in the caption of Table 1. However, we agree that the usage of this term might be ambiguous. We will replace it with more explicit expressions (e.g., "overall diversity" and "instance diversity"). (3) **Style-consistency:** We will replace it with a more explicit term such as "music-dance consistency"or "music-dance matching".

R2: Unclear observations. The observations in L21 refer to "dance to music is a creative process that is both innate and acquired". Similarly, we leverage our prior knowledge (innate) to design the framework and then use a data-driven learning process, which corresponds to the acquired properties.

R2: What is assumed on the model. To facilitate the training process, we assume that the latent factors of the music-to-dance generation process can be disentangled to two components: beat and style. In the decomposition, we learn how to dance according to the beat. In the composition, we learn how to dance to the beat and style at the same time.

R2: More details. Our code and trained model will be released, where all implementation details can be found. (1) **Number of poses:** We use 32 frames for each dance unit in our experiments (i.e., two seconds at 16 fps). (2) **Music style classifier:** We have evaluated several methods to extract features: features from a pre-trained network SoundNet, from a music autoencoder, and from a music classifier. We adopt the music classifier as our music style extractor due to the capability to better separate different types of music. This classifier consists of multiple fully-connected layers and takes as inputs the MFCC features extracted from music of various types. The training data are the audios from videos used for training our dance model, totaling 360K clips and three categories. (3) **Baselines:** LSTM baseline is basically the same as [28] (see Figure 5 in [28]). It takes a sequence of extracted audio features as inputs and directly predicts a sequences of poses. Aud-MoCoGAN is an extention of MoCoGAN [31]. The original MoCoGAN takes a sequence of random variables as inputs. We use audio features (MFCC) and beats (a sequence of binary variables) as additional inputs. (4) **Action classifier:** We utilize an RNN to take as inputs the pose sequences of arbitrary length, and append multiple fully-connected layers on the last hidden state to perform classification. We use the features from the last fully-connected layer to compute the FID. (5) **Subjective test:** We conduct the subjective test online. The professional backgrounds (high-school students, doctors, professors) and ages (from 15 to 61) of the subjects are diverse.

R2: Quantitative and qualitative results. We will reorganize the section order according to the suggestion. Since the trained model will be released, it is easy to qualitatively evaluate the results. For the quantitative results, only motion realism and style consistency are from the subjective test. Other evaluation metrics including FID, beat hit-rate, beat coverage, diversity, and multimodality are all measured quantitatively.

R2, R3: More references. The LSTM-based methods are similar to our first baseline and suffer from diversity and multimodality. We will cite and discuss these papers in the revised manuscript.

R3: Long sequence generation. In our framework, we can generate up to five consecutive dance units (about 10 seconds) in each step, and a sequence of arbitrary length can be seamlessly generated in an iterative manner. In this work, we use the last pose of the previous step as the initial pose of the next step, as shown in the bottom right of Figure 1 and Eq. (9). We will design methods to better capture the long-term relationship among sub-sequences (e.g., through a hierarchical recurrent network) in our future work.

R3: Complicated dance. In term of complicated poses, the hip-hop dance in our selected dance styles is quite complicated. We will continue to collect and incorporate more dance styles including pop dance and partner dance.