UDE: A Unified Driving Engine for Human Motion Generation

–Supplementary Materials

Zixiang Zhou
zhouzixiang@xiaobing.ai

Baoyuan Wang
wangbaoyuan@xiaobing.ai

A. Unified *vs*. Modality-Specific

We first provide more justification for unifying training of two tasks, text-to-motion, and audio-to-motion, as opposed to training modality-specific models.

A.1. Unification Brings Smooth Transition of Generated Motion between Modalities

We justify why we choose to unify the two tasks in one shared model by showing qualitative examples of driving motion sequences with both text descriptions and audio sequences as input and providing a brief analysis based on the experimental results. To evaluate, we first feed text descriptions to our model to generate a motion sequence, then feed both an audio clip and the last 8 tokens of the generated motion sequence as primitive to our model to generate subsequent motion sequences. By feeding both the audio sequence and the last tokens of the motion sequence, we are hoping that the generated motion will be conditioned on both audio input and motion primitives, simultaneously, and then a smooth transition can be obtained from one to another naturally.

Qualitative Examples Fig. 1 shows results of much more complex motion sequences driven by text descriptions and audio clips sequentially. There are 3 regions shown in each row of the figure. For *Text-to-Motion* region, the motion is controlled by text description dominantly, while for *Audio-to-Motion* region, the motion is driven by audio clip mainly. The *Transition* region, in the middle of each row, shows how the text-driven motion sequence smoothly transits to an audio-driven sequence without introducing additional motion in-between modules.

The qualitative results suggest that complex scenarios correlated with multimodal inputs could be generated by our UDE model without introducing additional motion in-between modules. We show that by feeding the multimodal input to our model sequentially, and by conditioning the current task on previously predicted tokens, we can generate more complex motion sequences smoothly transit from one scenario(text) to another(audio).

Please refer to our supplementary video for better visualization of the results of such mixed input driving tasks.

Analysis Sec.A.1 shows why we propose to unify these tasks visually, we give a brief analysis of this problem here. As stated in sec. A.1, we feed text to generate motion sequence at first, which we denote as $\tilde{x}^t = \mathcal{D}_{DMD}(\mathcal{E}_{UTT}(e^t))$, here $\mathcal{D}_{DMD}(\cdot)$ is the Diffusion Motion Decoder $\tilde{x}^t = \mathcal{D}_{DMD}(z^q)$, z^q is the motion token sequence, and $\mathcal{E}_{UTT}(\cdot)$ is the Unified Token Transformer which maps embedding to motion token sequence as $z^q = \mathcal{E}_{UTT}(e^t)$. Let's denote the last n tokens of z^q as $z^{q,T-n:T}$, where T is the length of the token sequence. Then we feed an audio clip and the last n tokens $z^{q,T-n:T}$ to generate motion sequence as $\tilde{x}^a = \mathcal{D}_{DMD}(\mathcal{E}_{UTT}(e^a, z^{q,T-n:T}))$. In this step, we notice that the input to UTT $\mathcal{E}_{UTT}(\cdot)$ has two items, 1) the first item is the embedding of audio clip e^a, and 2) the second term is the last n tokens $z^{q,T-n:T}$ which corresponds to text description. If we adopt a modality-specific paradigm, $\mathcal{E}_{UTT}(e^a, z^{q,T-n:T})$ will give unexpected results because $\mathcal{E}_{UTT}(\cdot)$ is trained either on text modality only or audio modality only. However, in our setting, the input to $\mathcal{E}_{UTT}(\cdot)$ covers two modalities, the embedding e^a corresponds to audio modality, and $z^{q,T-n:T}$ corresponds to text modality because it is obtained by text description, and vise versa. To conclude, the codebook corresponding to different scenario will not be shared in Motion Quantization and Unified Token Transformer modules, hindering the token prediction conditioned on cross modality scenario. As a consequence, a model trained on a modality-specific paradigm will not perform well in generating smoothly transited motion driven by one modality to another.

A.2. Unification Brings Strong Results & Engineering Efficiency

Here we provide more quantitative analysis. To compare, we also train modality-specific models on text-driven and audio-driven tasks, respectively, and separately. We keep the model architecture fixed for both Modality-Agnostic

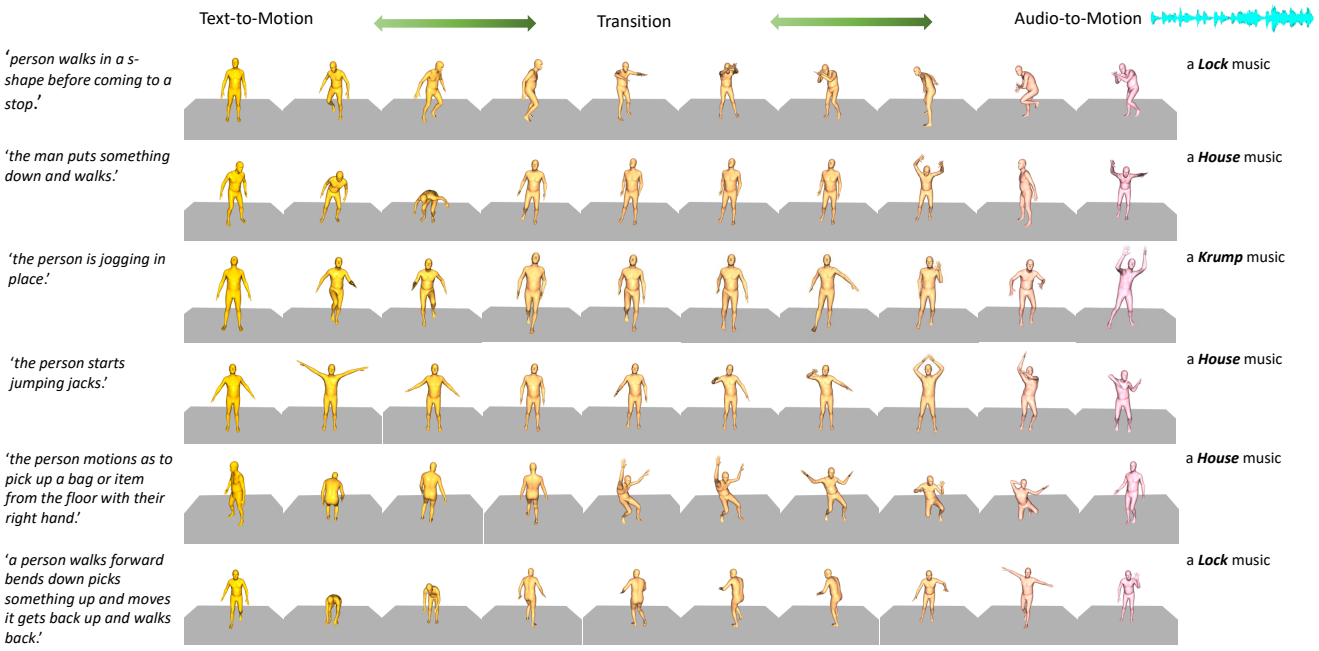

Figure 1. **Examples of unified driven samples.**Each row shows a motion sequence driven by a text description and an audio clip sequentially. The text descriptions are fed to UDE first to generate a text-conditioned motion sequence. Then the audio clip and the last 8 tokens of text-conditioned motion are fed to UDE to generate an audio-conditioned motion sequence. For each row, we extracted 10 frames sequentially to show the transition between text-driven and audio-driven sequences. Demo videos could be found in our supplementary materials.

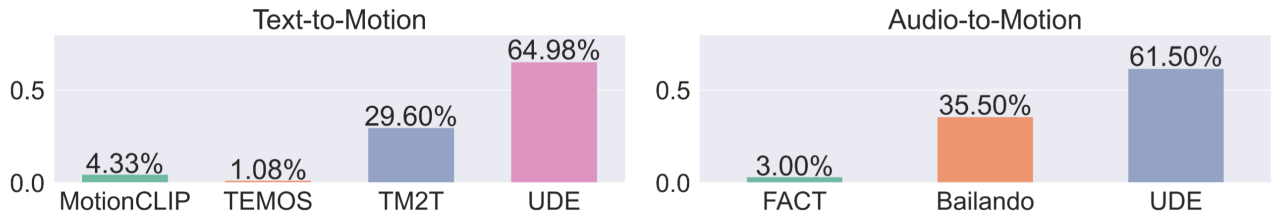

Figure 2. **User study**: We generate samples using different models from same text prompt and audio clip, respectively. And the users are asked to select the most corresponding ones.

Transformer Encoder(**MATE**), and Unified Token Transformer(**UTT**), and we don't train them with Diffusion Motion Decoder(**DMD**) because we don't want to introduce diversity at this time for a fair comparison. Therefore, we just adopt the pretrained VQ-Decoder in Motion Quantization(**MQ**) stage. For both text-to-motion and audio-to-motion tasks, we follow the same optimization strategy described above and trained 300 epochs for each task. During the evaluation, we follow the deterministic token prediction strategy described above, where we don't inject $z \sim \mathcal{N}(0, I)$ to **UTT** because diversity is not desired at this stage. We evaluate the performance of our **UDE** model over the same metrics as above: 1) For text-to-motion, we evaluate our method on *Text Retrieval Acc.*, *FID* scores, and *Diversity*. 2) For audio-to-motion, we evaluate our method on *Beat Align Score*, *FID*, and *Diversity*, respectively.

Tab. 1 summarizes the quantitative results. As we can observe from the results, for the text-to-motion task, training our model on text-to-motion dataset only does not bring obvious performance gain. On the contrary, training a text-only model brings even worse Top-1 Acc. and **FID**s noticeably. For Top-1 Acc. text-only training brings around 5% accuracy drop against unified training (8.81 to 7.77). If we take a look at the **FID**s, text-only training also brings worse results. A similar conclusion could be drawn from the audio-to-motion task. If we train an audio-only model, it does not improve performance. For *Beat Align*, unified training shows a slightly better beat synchronization property. For feature-wise quality and diversity, audio-only and unified models draw a tie. For audio only model, it has better performance on FID_m and Div_m, while for the unified model, better kinetic quality FID_k and kinetic diversity

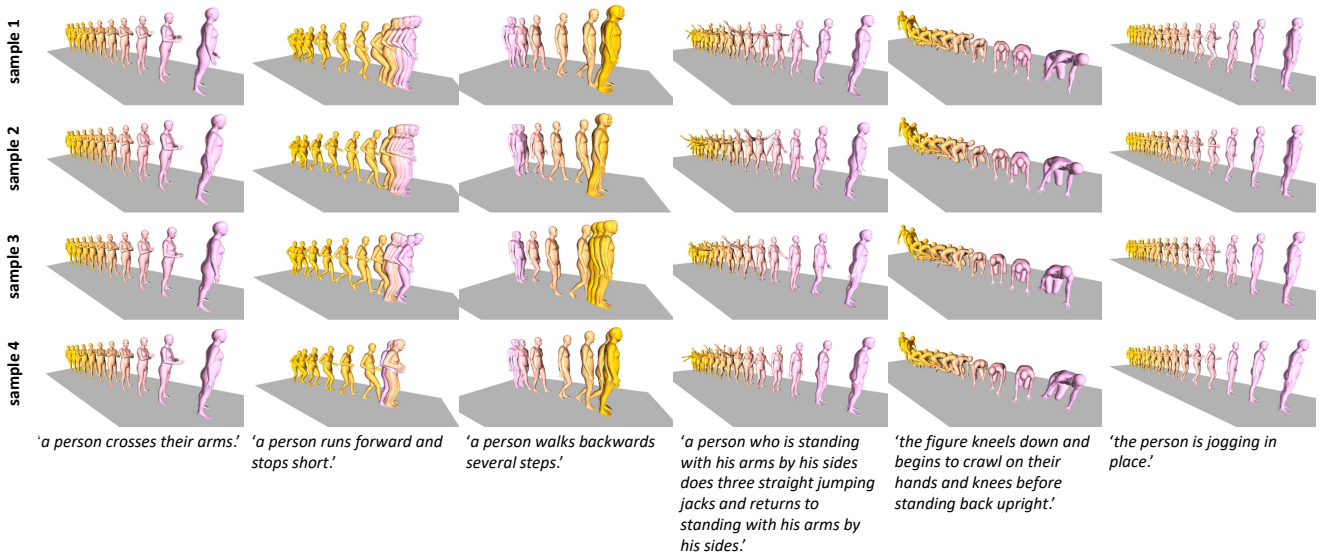

Figure 3. **Diversity of Text-to-Motion.** We show more qualitative results on text-to-motion tasks. For each column, we show 4 samples driven by the same text description with high diversity. We adjust the trajectory of some motion sequences for better visualization. Demo videos could be found in our supplementary materials.

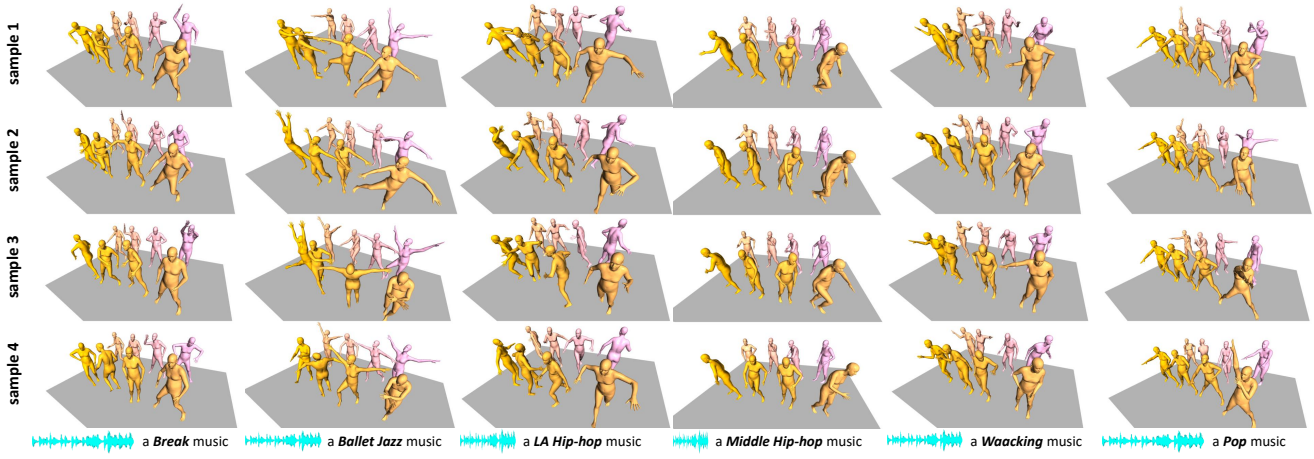

Figure 4. **Diversity of Audio-to-Motion.** We show more qualitative results on audio-to-motion tasks. For each column, we show 4 samples driven by the same audio clip with high diversity. We show samples driven by 6 audio clips with different genres. Demo videos could be found in our supplementary materials.

Div_k are obtained.

This study suggests that our model achieves competitive performance on both text-driven and audio-driven scenarios when trained in a unified paradigm, compared with training on the uni-task paradigm. With the performance maintained, our method successfully puts these two driven tasks to one unified solution. Through unification, engineering efficiency is improved because only one model needs to be maintained and improved for possible future applications. This suggests there is potential for unification on multi-modal human motion generation.

B. Detail of Model Architecture

We describe the detail architecture of Unified Transformer Encoder(**UTT**) and Diffusion Motion Decoder(**DMD**) here. Fig. 5 describes the architecture of **UTT**. We describe the Unified Token Transformer module and the conditional discriminator in an end-to-end manner, and the transformer encoder layer with causal self-attention is demonstrated at the bottom panel of Fig. 5. The detailed architecture of **DMD** is shown in Fig. 6, where the left panel illustrates the token transformer module, and the right panel shows the diffusion transformer decoder module.

Method	Text-to-Motion						Audio-to-Motion				
	Text Retrieval Acc.		FID		Diversity		Beat Align ↑	FID		Diversity	
	Top-1 Acc. ↑	Top-5 Acc. ↑	FID_k ↓	FID_m ↓	Div_k ↑	Div_m ↑		FID_k ↓	FID_m ↓	Div_k ↑	Div_m ↑
text only	7.77	**26.01**	31.73	5.69	**4.42**	**6.96**	-	-	-	-	-
audio only	-	-	-	-	-	-	0.2231	39.27	**11.65**	5.68	**8.03**
unified	**8.11**	25.01	**27.66**	**4.92**	4.28	6.77	**0.2268**	**28.44**	15.70	**6.13**	4.07

Table 1. **Ablation on Unification v.s. Modality-Specific.** We explore our method trained in a unified paradigm against that trained in a modality-specific paradigm. All three models, use exactly the same architecture, and we don't inject any random term to eliminate the influence of diversity. For **text only** and **audio only**, they are trained on HumanML3D and AIST++ datasets solely.

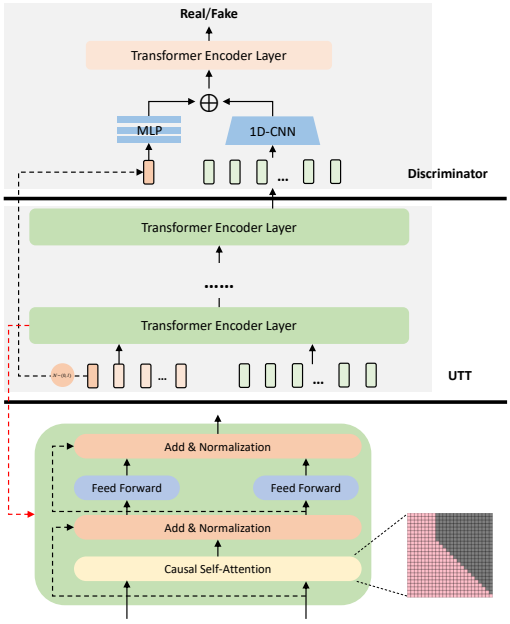

Figure 5. **Architecture of UTT. Bottom** panel shows detail of the transformer encoder layer in **middle** panel, where a causal self-attention is adopted in replacing with conventional full self-attention. The **top** panel is the detail of the conditional discriminator. The global embedding and predicted token sequences are transformed and summed together and fed to a transformer encoder to get a patch-wise validity score.

Specifically, given the predicted token sequence, the token transformer first encodes it to a sequential embedding by stacked transformer encoder layers. Then we convert the sequential embedding to a single embedding by applying a max-pooling operation along the temporal dimension. This single embedding is then adopted as condition embedding. For every step of reversed diffusion, we feed the condition embedding, as well as the timestep embedding, and the latent to the diffusion transformer decoder, and estimate the noise ϵ_t. We repeat this reversed diffusion step 1000 times to get the final denoised sample X_0.

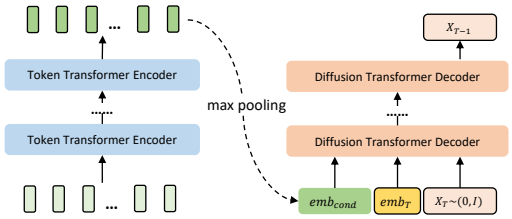

Figure 6. **Architecture of DMD. Left** panel shows the detail of Token Transformer Encoder. The predicted token sequences are fed to this module, and the sequential embedding is obtained. Then we get a single embedding by max-pooling operation along the temporal dimension. This single embedding is used as a condition at the Diffusion Transformer Decoder module. **Right** panel shows the detail of the Diffusion Transformer Decoder. We feed the condition embedding, timestep embedding, and latent to it, and the gaussian noise ϵ_t is obtained at each reversed diffusion step.

C. User Study

We also conducted a user study to evaluate the quality of our method compared with prior works. For text-to-motion task, we generate samples conditioned on same text description using different models, and users are asked to tell which sample matches best to the description. Same evaluation strategy is followed for audio-to-motion task. The results of user study are shown in Fig. 2. For text-to-motion task, among all the samples, 65% of ours matches best to the description, which is 35% higher than the second best method. For audio-to-motion task, ours also achieves 61.5% best ratio. The user study suggests that our method outperforms existing methods in terms of human perception.

D. More Qualitative Examples

We show more qualitative examples of our method on Text-to-Motion and Audio-to-Motion tasks, respectively. Specifically, we demonstrate the diversity of motion samples generated by our method. Fig. 3 shows more results on the Text-to-Motion task. In the figure, each column represents 4 samples driven by the same text description. We appropriately adjust the trajectory of some samples for better visualization, so the poses will not clutter together. Fig. 4 shows more results on the Audio-to-Motion task. Similarly,

we show 4 samples driven by the same audio clip in the same column. And we adjust the trajectory of each pose to make them in a two-row formation. As can be observed, our method achieves diversity in both text-driven and audio-driven scenarios, while maintaining semantic correlation. We also provide multiple demonstration videos in our supplementary materials on both text-to-motion and audio-to-motion tasks.