Statement on Future Research

Chi Keung Tang

Future Research: Multimodal Large Language Model for High-Quality Content Generation

Large Language Models (LLMs) have recently attracted much attention and research effort in both industry and academia. Many LLMs, such as GPT-4, LLaMA, Gemma, have shown their advanced capabilities, robustness and generalization across various downstream tasks. These progresses have motivated researchers to explore the application of LLMs in multimodal tasks, integrating them with modalities such as images, videos, audio, and more, resulting in promising outcomes in understanding these different modalities. The of LLMs in the context of multimodal generation remains relatively underexplored to date despite its high potential in various important applications. For instance, text-driven 3D human motion generation can facilitate and advance robots and humanoid applications. Multi-event long-condition text/video-to-audio generation has high potential in various content generations where high-quality audio is essential in aligning the input text and video. This project taps into the rapid development of LLMs, moving forward with multimodal LLMs with different modalities and contribute useful, convesational AI frameworks employing GPT-4 to support a set of versatile generation tasks customizable through interactive conversational exchanges with the human user. Specifically, we propose:

1. ChatCam for Empowering Camera control through Conversational AI

Cinematographers skillfully capture the essence of the 3D world by maneuvering their cameras, creating an array of compelling visual narratives. Achieving aesthetically pleasing results requires not only a deep understanding of scene elements and their interplay but also meticulous execution of techniques. Recent progress of large language models (LLMs) has marked a significant milestone in AI development, demonstrating their capability to understand and act within the 3D world. Witnessing this evolution, our work explores the feasibility of empowering camera control through conversational AI, thus enhancing the video production process across diverse domains such as documentary filmmaking, live event broadcasting, and virtual reality experiences. While techniques have been proposed and used for controlling the trajectories of objects and cameras in video generation approaches for practical usage, or predicting similar sequences through autoregressive decoding processes, generating camera trajectories has yet to be explored. This task involves multiple elements such as language, images, 3D assets, and, beyond mere accuracy, necessitates visually pleasing rendered videos as the ultimate goal.

We propose *ChatCam*, a system that allows users to control camera operations through natural language interaction. Leveraging an LLM agent to orchestrate camera operations, our method assists users in generating desired camera trajectories, which can be used to render videos on radiance field representations such as NeRF or 3DGS. At the core of our approach, we introduce *CineGPT*, a GPT-based autoregressive model that integrates language understanding with camera trajectory generation. We train this model using a paired text-trajectory dataset to equip it with the ability for text-conditioned trajectory generation. We also propose an *Anchor Determinator*, a module that identifies relevant objects within the 3D scene to serve as anchors, ensuring correct trajectory placement based on user specifications. Our LLM agent parses compositional natural language queries into semantic concepts. With these parsed sub-queries as inputs, the agent then calls our proposed *CineGPT* and *Anchor Determinator*. It composes the final trajectory with the outputs from these tools, which can ultimately be used to render a video that fulfills the user's request.

With comprehensive evaluations and comparisons to other state-of-the-art methods, our method exhibits a pronounced ability to interpret and execute complex instructions for camera operation. Our user studies further demonstrate its promising application prospects in actual production settings.

In summary, the contributions will be:

- We introduce *ChatCam*, a system that, for the first time, enables users to operate cameras through natural language interactions. It simplifies sophisticated camera movements and reduces technical hurdles for creators.

- We develop *CineGPT* for text-conditioned camera trajectory generation and an *Anchor Determinator* for precise camera trajectory placement. Our LLM agent understands users' requests and leverages our proposed tools to complete the task.

2. VP-LLM for text-driven 3D volume completion with LLMs

We propose Volume Patch Large Language Model, VP-LLM, which achieves 3D completion with precise textual control. Inspired by the recent progress in 3D multi-modality models, we believe that Large Language Models (LLMs) can underpin our approach by decoding the complex associations between 3D structures and textual descriptions. LLMs, pretrained on large-scale text datasets, have the capability to process long sequences and comprehend complex human languages, while 3D models represented by voxel grids can be straightforwardly converted into a one-dimensional format through flattening. Therefore, we investigate how to enable LLMs to understand a 3D model by decoding complex correlations between 3D structures and textual descriptions, or "translating" it into a "sentence".

For seamless incorporation of 3D data into the LLM tokenization framework, 3D models are initially segmented into smaller patches, facilitating independent encoding and decoding. Different from most previous methods that manage the 3D object as a unified, this idea of patchification is more scalable and extendable.

The 3D volumes are first patchified into individual patches and processed by a patch-wise Variational Autoencoder (VAE) to encode individually. The encoded patches are then projected and concatenated with user-specified text conditions to the LLM. Finally, the output projection layer extracts the features generated by the LLM and lets the VAE decode back each patch individually.

In summary, the major contributions are

- VP-LLM is the first work leveraging *LLM* to achieve text-guided 3D completion, outperforming existing state-of-the-art text-conditioned 3D completion works.
- The *patchification* to be proposed enables a scalable integration of 3D volumes into the LLM, solving the difficulty in handling high-resolution voxel grids faced by existing works.
- Our work can perform 3D completion and denoising for multiple categories using detailed text control within one unified model.

3. Motion-Agent for 3D human motion generation with LLMs

We propose *Motion-Agent*, a multimodal framework that leverages the generalization and flexibility of pre-trained LLMs. Central to the framework is our new generative agent, *MotionLLM*, the incorporation of which eliminates the need for extensive pre-training by employing lightweight adapter-based fine-tuning of a pre-trained LLM.

In Motion-Agent, we first train a pair of motion tokenizer and detokenizer. The motion tokenizer encodes motions into motion embeddings and quantizes them into a set of discrete LLM-understandable tokens using a codebook, while the detokenizer reconstructs tokens back to their original continuous forms. This tokenizer-detokenizer pair enables the translation between continuous motion sequences and discrete tokens, facilitating interaction with the LLM while still allowing for the recovery of the original motions from the tokens. MotionLLM is trained by enriching a pre-trained LLM's vocabulary with these additional motion tokens, while keeping the original text tokens unchanged. Given that motions can be represented as temporal sequences, our tokenization process translates motions into token sequences akin to sentences in natural language. MotionLLM takes in a description of simple motions and outputs motion token sequences. On top of this, GPT-4 acts as a coordinator, decomposing user instructions to determine the number of calls to MotionLLM and how to structure those calls effectively. The resulting motion token sequences from multiple calls are concatenated and decoded by the detokenizer to produce the final output.

Our Motion-Agent framework leverages pre-trained LLMs in two key ways: (1) fine-tuning a lightweight LLM via adapters to serve as a text-motion translation agent, and (2) using an LLM for conversational interactions without training, thus facilitating multi-turn dialogue for refining generated motions and producing

extended motions by iteratively generating and concatenating sequences. Despite training only a small number of parameters, MotionLLM can achieve competitive results in motion generation (text to motion) compared to those trained-from-scratch models with specialized architectures. The other notable benefit of our MotionLLM is exhibited by its enhanced comprehension of detailed and intricate text descriptions, enabling the accommodation of diverse and customized language inputs. In motion captioning (motion to text), MotionLLM achieves state-of-the-art performances, generating semantically accurate and contextually appropriate text descriptions. By combining MotionLLM with GPT-4, Motion-Agent enables versatile dialogue-based motion generation and reasoning, without requiring specific datasets or extra training for these tasks.

To summarize, our contributions will include:

- We introduce a simple, efficient conversational framework, Motion-Agent, that utilizes pre-trained LLMs and produces state-of-the-art results in various motion-language tasks.
- We demonstrate the flexibility and versatility of our method by achieving highly customizable motion-language tasks, including long and complex motion generation, multi-turn editing, and multi-turn reasoning.

4. Audio-Agent for audio generation, editing and composition leveraging LLMs

We introduce Audio-Agent, which breaks down intricate user inputs using GPT-4 into multiple generation steps. Each step includes a description along with start and end times to effectively guide the audio generation process. Our framework integrates two key tasks: Text-to-Audio (TTA) and Video-to-Audio (VTA). We leverage a pre-trained TTA diffusion model, Auffusion with essential adaptations serving as the backbone for our generation process. In the TTA task, Auffusion focuses solely on generating simple, atomic text inputs. Our framework supports audio generation, editing, and composition. For the VTA task, we recognize that models such as GPT-4 and other large language models lack sufficient temporal understanding for video clips. To address this problem, we employ moderate fine-tuning to align the two modalities. We utilize the smaller Gemma2-2B-it model, which has 2 billion parameters, and fine-tune an adapter and a projection layer to convert visual inputs into semantic tokens. We then implement cross-attention guidance between the diffusion layers of Auffusion. This approach eliminates the need for additional training on a temporal detector, as the semantic tokens inherently contain time-aligned information.

The summary of our contributions is as follows:

- we propose Audio-Agent which utilizes a pre-trained diffusion model as a generation agent, for both TTA and VTA task;
- For TTA, Audio-Agent can handle complex text input, which is broken down into simple and atomic generation conditions for the diffusion model to make inference on;
- For VTA, we fine-tune an open-source LLM (Gemma2-2B-it) to bridge the modality gap between video and audio modalities to align the underlying semantic and temporal information. Through extensive evaluation, our work demonstrates on-par results compared to the state-of-the-art task-specific models trained from scratch. We hope our work can motivate more relevant works on multi-event long-condition TTA generation, which to our knowledge has not yet been fully explored despite its high potential in various content generations where high-quality audio is essential.