Master internship subject

3D Human Motion Diffusion Model

Hosting institute

ICube Laboratory (The Engineering science, computer science and imaging laboratory) at the University of Strasbourg is a leading research center in Computer Science, with more than 300 permanent researchers, with the recently opened AI graduate school supported by the French government.

Work place and salary

The thesis work will take place in the MLMS (Machine Learning, Modeling & Simulation) research team of the ICube laboratory (The Engineering science, computer science and imaging laboratory) of the University of Strasbourg, a leading research center with more than 300 permanent researchers. The workplace is located on the hospital site of the laboratory, a 10-minute walk from the heart of downtown Strasbourg, listed as a UNESCO World Heritage Site.

650 euros net monthly

Supervisors

– director: Hyewon Seo (ICube, Univ. Strasbourg)

– co-supervisors: Cédric Bobenrieth (ICAM, Strasbourg)

Staring date

February – April 2025.

Work description

Human motion generation is a key task in computer graphics, crucial for applications involving virtual characters, such as film production or virtual reality experiences. Recent deep learning methods, particularly generative models, started to make significant contributions in this domain. While early neural methods focused on the unconditional generation of vivid and realistic human motion sequences, more recent methods guide the motion generation using various conditioning signals, including action class, text, and audio. Among them, the diffusion-based model has shown significant success, dominating research frontiers [TRG*23, KKC23, ZCP*24, DMGT23].

Motivated by these recent successes, we will develop action-conditioned human motion generator based on a diffusion model. In particular, we will aim at the generation of daily actions in residential settings, in the view of augmenting training data for the action recognition models. To achieve this goal, we will deploy a diffusion-based motion generation, based on our previous works [ZFY*24, XS24]. To condition the generation using an action class or a text description, we will adopt CLIP [RKH*21] as a text encoder to embed the text prompt and use a trainable tensor as embeddings for different action classes.

The work can be structured as the following tasks:

1. **Data rearrangements:** We will rearrange/select the datasets that are at our disposal in such a way that they can be seamlessly used as training data.
2. **Unconditional motion generation:** The first step is to train a diffusion model
3. **Action-conditioned motion generation**: The aforementioned model will be extended towards conditional generation task. To achieve highly precise conditioned sampling without the need for training auxiliary models, we will take a classifier-free guidance approach.
4. **Experiments**: The developed model will be parameter-tuned, tested, and compared with the state-of-the-art models. Additionally, a number of ablation studies will be conducted to assess the impact of various components on performance.

Candidate profile

− Solid programming skills in Python

− Working skills in Blender for 3D modeling and animation

− Experience in Deep Learning (Diffusion model)

− Good communication skills

Application

Send your CV and academic records (Bachelor and Master) to cedric.bobenrieth@icam.fr and seo@unistra.fr, for (a) possible interview(s).

Bibliography

[DMGT23] DABRAL R., MUGHAL M. H., GOLYANIK V., THEOBALT C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023), pp. 9760–9770.

[KKC23] KIM J., KIM J., CHOI S.: Flame: Free-form language-based motion synthesis & editing. In Proceedings of the AAAI Conference on Artificial Intelligence (2023), vol. 37, pp. 8255–8263.

[RKH∗21] G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., et al.: Learning transferable visual models from natural language supervision. In International conference on machine learning (2021), PMLR, pp. 8748–8763.

[TRG∗23] TEVET G., RAAB S., GORDON B., SHAFIR Y., COHEN-OR D., BERMANO A. H.: Human motion diffusion model. In The Eleventh International Conference on Learning Representations (2023).

[XS24] Xue K. and Seo H., Shape Conditioned Human Motion Generation with Diffusion Model, arXiv preprint https://arxiv.org/abs/2405.06778.

[ZCP∗24] ZHANG M., CAI Z., PAN L., HONG F., GUO X., YANG L., LIU Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[ZFY*24] Zou K., Faisan S., Yu B., Valette S., Seo H., 4D Facial Expression Diffusion Model, ACM Transactions on Multimedia Computing, Communications and Applications, https://dl.acm.org/doi/10.1145/3653455.

Master internship subject

Action Recognition by Knowledge Augmentation in Vision Language Model

Hosting institute

ICube Laboratory (The Engineering science, computer science and imaging laboratory) at the University of Strasbourg is a leading research center in Computer Science, with more than 300 permanent researchers, with the recently opened AI graduate school supported by the French government.

Work place and salary

The thesis work will take place in the MLMS (Machine Learning, Modeling & Simulation) research team of the ICube laboratory (The Engineering science, computer science and imaging laboratory) of the University of Strasbourg, a leading research center with more than 300 permanent researchers. The workplace is located on the hospital site of the laboratory, a 10-minute walk from the heart of downtown Strasbourg, listed as a UNESCO World Heritage Site.

650 euros net monthly

Supervisors

– director: Hyewon Seo (ICube, Univ. Strasbourg)

– co-supervisor: Diwei Wang (ICube, Strasbourg)

Staring date

February – April 2025.

Work description

Action recognition from video is highly important for assistive care robots, as it enables them to understand and respond appropriately to the needs and activities of the people they assist. Recent DL models for action recognition are moving toward more data-efficient, interpretable, and computationally optimized frameworks: The combination of transformer architectures, spatio-temporal attention, multimodal fusion, and self-supervised learning, just to mention a few. Meanwhile, the recent emergence of large-scale pre-trained vision-language models (VLMs) has demonstrated remarkable performance and transferability to different types of visual recognition tasks, thanks to their generalizable visual and textual representations. It has been confirmed by our recent study[1] [2], where our developed model learns and improves visual, textual,

[1]Wang D., Yuan K., Muller C., Blanc F., Padoy N., Seo H., "Enhancing Gait Video Analysis in Neurogenerative Diseases by Knowledge Augmentation in Vision Language Model", Lecture Notes in Computer Science (Proc. Medical Image Computing and Computer-Assisted Intervention), vol. 15005, pp 251–261, Springer, 2024.

[2] Wang D., Yuan K., Seo H., "GaVA-CLIP: Refining Multimodal Representations with Clinical Knowledge and Numerical Parameters for Gait Video Analysis in Neurodegenerative Diseases", under revision, 2024.

and numerical representations of patient gait videos based on a large-scale pre-trained Vision Language Model (VLM), for several classification tasks.

Motivated by these recent successes, we will extend our previous developed model and the multimodal representation for a new classification task – action recognition from video. Similarly to our previous method, we will adopt the prompt learning strategy, keeping the pre-trained VLM frozen to preserve its general representation and leverage the pre-aligned multi-modal latent space the prompt's context with learnable vectors, which is initialized with domain-specific knowledge.

We will proceed with the following steps:

1. **Data organization:** The datasets that are at our disposal will be rearranged and selected to ensure seamless use as training data.
2. **Knowledge distillation:** Per-class description will be collected and refined in a semi-automatic manner, which we will use to initialize learnable prompts.
3. **Adaptation of the model**: Based on the above knowledge, we will adapt our previously developed model to perform the new classification task. Whenever applicable, we will design specialized loss functions tailored to the specific nature of the new task. A number of ablation studies will help improve the performance and assess the impact of various components. This may involve testing a number of VLMs as backbone.
4. **Experiments**: The developed model will be parameter-tuned, tested and compared with the state-of-the-art models.

Candidate profile

− Solid programming skills in Python/C++

− Experience in Deep Learning (Transformer, CLIP, etc.)

− Good communication skills

Application

Send your CV and academic records (Bachelor and Master) to seo@unistra.fr, for (a) possible interview(s).

Master internship subject

DeBIAN: Deep representation of the Brain Image for the Analysis of Neurodegenerative diseases

Hosting institute

ICube Laboratory (The Engineering science, computer science and imaging laboratory) at the University of Strasbourg is a leading research center in Computer Science, with more than 300 permanent researchers, with the recently opened AI graduate school supported by the French government.

Work place and salary

The thesis work will take place in the MLMS (Machine Learning, Modeling & Simulation) research team of the ICube laboratory (The Engineering science, computer science and imaging laboratory) of the University of Strasbourg, a leading research center with more than 300 permanent researchers. The workplace is located on the hospital site of the laboratory, a 10-minute walk from the heart of downtown Strasbourg, listed as a UNESCO World Heritage Site.

650 euros net monthly

Supervisors

– director: Hyewon Seo (ICube, Univ. Strasbourg)

– co-supervisors: Stephane Kremer (University hospital), Diwei Wang (ICube, Univ. Strasbourg)

Staring date

February – April 2025.

Context

Dementia with Lewy Bodies (DLB) and Alzheimer's Disease (AD) are two common neurodegenerative diseases among elderly people. Both associated with abnormal deposits of proteins in the brain, the diagnosis of these diseases can be challenging, particularly in distinguishing between them, as they exhibit similar symptoms in their early stages. Brain MRI provides detailed images of brain structures, allowing for the identification of structural changes associated with neurodegenerative diseases. Deep learning has shown great promise in analysing these images, enabling accurate predictions and interpretations. At the center of it are the recent emerging large-scale pre-trained vision-language models (VLMs), which have demonstrated remarkable performance thanks to their generalizable visual and textual representations.

Work description

We will deploy a VLM to improve the accuracy and efficiency of brain image analysis, with a specific focus on classification and associated reasoning presented in text form. Our specific focus will be on the analysis and understanding of neurodegenerative diseases, Dementia with Lewy Bodies (DLB), Alzheimer's Disease

(AD), and/or Parkinson disease. We will base our study on our recent work[1] [2], where the model we developed learns and refines visual, textual, and numerical representations of patient gait videos using a large-scale pre-trained Vision-Language Model (VLM) for several classification tasks.

We will proceed with the following tasks:

1. **Collection and organization of public dataset:** A first step is to collect the publicly available dataset and organize it to ensure seamless use as training data. Datasets from ADNI (Alzheimer's Disease Neuroimaging Initiative) are considered.
2. **Knowledge distillation:** Per-class textual description will be collected and refined in a semi-automatic manner, which we will use to initialize learnable prompts.
3. **Knowledge augmented strategy to classification of image-only data**: We will adapt our knowledge augmented prompt-tuning strategy combined with a VLM-based classifier model, to address the new task.
4. **Knowledge augmented strategy to classification of multimodal data**: We will develop methods to exploit paired demographic data, alongside images. Interestingly, this is analogous to the way neuro-radiologists interpret images.
5. **Experiments**: The developed models will be tested and compared both with each other and with state-of-the-art models.

Candidate profile

− Solid programming skills: Python/C++

− Experience in Deep Learning (Transformer, CLIP, etc.)

− Good communication skills

Application

Send your CV and academic records (Bachelor and Master) to seo@unistra.fr

[1] Wang D., Yuan K., Muller C., Blanc F., Padoy N., Seo H., "Enhancing Gait Video Analysis in Neurogenerative Diseases by Knowledge Augmentation in Vision Language Model", Lecture Notes in Computer Science (Proc. Medical Image Computing and Computer-Assisted Intervention), vol. 15005, pp 251–261, Springer, 2024.

[2] Wang D., Yuan K., Seo H., "GaVA-CLIP: Refining Multimodal Representations with Clinical Knowledge and Numerical Parameters for Gait Video Analysis in Neurodegenerative Diseases", under revision, 2024.