Research Statement
Foundation Models for Pixel-level Understanding and Generalist Embodied AI

Mennatullah Siam, PhD
University of British Columbia
Vancouver, Canada
mennatullah.siam@ubc.ca

Foundation models have gained significant traction and success, particularly in natural language processing, exemplified by the GPT series. Specifically, multi-modal large language models (MLLMs) have recently emerged directed towards building general purpose agents with full understanding of image and text input, capable of generating images and/or language output. It has been recently equipped with visual grounding and reasoning abilities especially with the recent emergence of models such as Qwen-VL and Intern-VL variants. Nonetheless, the aforementioned models are still limited in their understanding of the world in motion on the pixel-level, which is crucial in various robotics applications. Pixel-level image and video understanding goal is to make inferences about the surrounding world from corresponding image or video data on the pixel-level, e.g., segmentation, depth estimation, flow estimation and pixel-level tracking. Such pixel-level understanding capabilities is still limited in general purpose agents built using multi-modal large language models. While there are foundation models specialized in segmentation tasks or depth estimation separately, the main goal is to go towards general purpose agents and away from such specialized models. Some of the recent MLLMs that serve as general purpose agents are capable of grounding, yet they are mainly designed for box/region-level output. On the other hand, MLLMs designed for pixel-level understanding are limited to segmentation and have degraded capabilities in simple tasks such as visual question answering. Studying the intersection of pixel-level understanding and multi-modal large language models is a major goal in my research program which paves the road towards general purpose AI agents that are both interpretable and language guided. Furthermore, I aim to study the aforementioned topics within a responsible AI framework that gives emphasis to interpretability, benchmarking MLLMs performance and their ability to operate in low data regime. Another limitation of the current foundation models generally is that in certain tasks they under-perform specialized models. Thus, another major goal of my research program is studying test-time adaptation and prompting techniques with limited labelled data. Such adaptation with limited labelled data can help in decolonizing AI and empowering low resourced communities, where my research focuses on African/African origin communities. In summary, my research program is focused on studying **foundation models** purposed for **pixel-level scene and video understanding** within a **responsible AI** framework that puts emphasis on data-efficient learning and interpretable techniques.

During my tenure-track assistant professor position in Ontario Tech University I secured in my first year of appointment a quarter million grant funds for my research. I was able to acquire the NSERC Discovery Grant and launch supplements in addition to my startup funds. I was additionally able to acquire the Digital Research Alliance resources allocation for research groups and NSERC Alliance international, catalyst grant. I am also affiliated with University of British Columbia as affiliate faculty. During my PhD and Postdoc I was able to acquire grants for my research including: (i) the Alberta Innovates Futures Technology scholarship, (ii) Verna Tate scholarship, (iii) Alberta Graduate Excellence scholarship and (iv) Vista Postdoc fellowship. Additionally, I was able to collaborate with international collaborators in RIKEN institute, Japan during my current position and different companies throughout my PhD some of which are focused on robotics including Wayve, Nvidia autonomous driving team, Valeo Vision Systems, Huawei HiSilicon Research and ElementAI. I have participated as well in the University of Alberta team in 2018 that was awarded a KUKA Innovation Award Finalist[1], for our work on online tool and task learning using human-robot interaction. Finally, because of my motivation towards fewshot learning and pixel-level understanding research I was an organizer of the Learning with Limited Labelled Data for Image and Video Understanding (L3D-IVU) Workshop in CVPR'22, 23, 24[2] and I am the primary organizer of PixFoundation workshop in CVPR'25[3].

Research philosophy. My research philosophy relies on three notions: (i) questioning the norms that are not motivated theoretically or empirically but were rather inherited historically from previous research, (ii) going beyond benchmarks and understanding the needs of deploying in the wild, (iii) using interpretability of black box models to gain insights on how to improve their generalization capabilities in low-data regime. My previous, ongoing and future work is discussed in light of these three notions in the following sections.

Few-shot image and video semantic segmentation. Earlier approaches in few-shot learning followed a meta-learning scheme that emulates the inference stage during training through sampling tasks of few labelled training data and target examples. I started with questioning whether simple techniques applied during the few-shot inference [21, 14] can go beyond sophisticated meta-learning approaches. I initially investigated ***multi-resolution masked imprinting*** for few-shot object segmentation [21] that surpassed the state-of-the-art at that time without bells and whistles and with simple mechanisms during the few-shot inference. It relied on the relation between softmax classifiers and metric learning, where the final novel

[1] https://www.youtube.com/watch?v=aLcw73dt_Oo
[2] https://sites.google.com/view/l3d-ivu/
[3] https://sites.google.com/view/pixfoundation/

classifier weights are imprinted (i.e. set to) the corresponding globally average pooled features of the few training examples. It was adapted to the segmentation task with multiresolution processing and incorporating the adaptation of the background classifier. I have also extended the work towards video segmentation with utilizing the unlabelled query examples during the few-shot inference. It is considered the first attempt for non-metalearning ***temporal transductive inference*** in fewshot video segmentation [14]. In transductive inference schemes the unlabelled test set is used during inference to enhance the prediction. I proposed spatiotemporal regularizers to learn the final classifiers for the novel class while ensuring temporal consistency of the predictions globally per-video. Moreover, in collaboration with Prof. Leonid Sigal and Prof. Jim Little, I mentored their PhD student on a generalized few-shot segmentation approach [7] that proposes a multiscale visual prompting of pixel-level image understanding models. Part of the work we proposed transductive prompt-tuning that utilizes the unlabelled query images. Finally, in collaboration with RIKEN institute we extended the few-shot segmentation setup and studied MLLMs in remote sensing focused on Africa [3, 1].

Then I was motivated to go beyond single image benchmarks for few-shot learning that were widely adopted, and design tasks to utilize the temporal structure in videos when learning with limited labels. I started with the simpler techniques, which was considerably successful, but I found out that it was limited in the temporal modelling to the classifier only. Therefore, I started working on meta-learning attention based models that can utilize the temporal structure. I started with ***co-attention mechanisms guided by the semantic labels*** on single images and its extension to videos [16]. The co-attention mechanism computes attention among both the target and the few training examples modulated by the class semantic features, in order to help extract the relevant features in the target image for segmentation. In another work, I proposed ***meta-learning multiscale transformer comparators for few-shot video semantic segmentation***. During the design of this work, I questioned the norm on how multiscale transformers in the state of the art of fully supervised semantic segmentation methods operated (e.g., Mask2Former) [15]. Since their mode of operation in the multiscale processing scheme was focused on learning better compact representation to identify the different objects/instances, which lead to loss of detailed information. Thus, I proposed the first attempt to maintain the spatiotemporal dimension during multiscale processing in transformers in what we referred to as multiscale memory learning. My design for multiscale memory learning lends itself to interpretability as you can investigate what the memory entries across different scales attend to, and due to its design it learned a temporally consistent attention maps. Finally, I have extended such work towards segmenting actors/actions in videos from few labelled examples to show the versatility of my approach.

Fully supervised video understanding applied to robotics settings. During my PhD, I have explored fully supervised methods for video object segmentation and understanding before exploring its few-shot counterpart. In my earliest works, I wanted to go beyond the available video object/motion segmentation benchmarks that were not built for autonomous driving. Thus, I proposed a ***motion segmentation benchmark automatically generated from KITTI and Cityscapes datasets focused on autonomous driving*** in real-world scenarios [20, 19]. I have proposed the first attempt to learn a ***multi-task learning framework that combines motion and appearance [20]*** to enable object detection, road segmentation and motion segmentation, which was used to apply for a patent with Valeo vision systems [22]. I have also focused on deploying ***real-time motion segmentation methods using geometric priors [17]***. My design lead to an efficient motion segmentation model running at nine frames per second on an embedded device. In another method, I presented [18] ***motion adaptation in video object segmentation*** which enables an efficient adaptation for the model. Inspired by semi-supervised video object segmentation methods that use a ground-truth initialization mask to segment the object of interest, I leveraged pseudo-labelled predictions instead. My method was published as the best method on DAVIS benchmark at the time [4]. I have also collaborated and mentored undergraduate students during the project to collect a dataset for teaching robots new objects through motion, which I evaluated on. Additionally, I have collaborated and equally contributed on the project for ***online tool and task learning in robot manipulation using human robot interaction [4]*** which was awarded a finalist team in the KUKA innovation award.

During my Postdoc I have mentored a PhD student in York university to learn multiscale video transformers with temporally consistent predictions. Our method presented the first attempt for ***bidirectional label propagation with masked attention*** which was deployed across different video understanding benchmarks [9, 10]. Label propagation encompasses the learning of a similarity graph with the nodes as the pixels of the spatiotemporal volume of features and using the normalized graph laplacian to propagate the labels across the input clip. Our models' predictions are temporally consistent with minimal design efforts and is learned end-to-end unlike postprocessing with conditional random fields. Finally, it is not only sufficient to have a working model with the state of the art performance but the ability to interpret what the representation learned is equally crucial. Towards this end, I collaborated with a PhD student on the creation of tools to ***interpret black box models for video segmentation and understanding***. Our tool can quantify on the intermediate representations whether the model relies on dynamics conveyed from multiple frames or static information in a single frame [11]. It further developed into a mechanism to improve the dynamic bias which resulted in better performance on rare classes in action recognition [12]. During that project I guided the student on the action recognition study, while I worked on the video object segmentation one. Currently, I am supervising a visiting PhD student from German University in Cairo on studying video understanding models from a neuroscience perspective, in modelling the human brain responses when perceiving short videos of specific actions and objects [5]. We published the short version of our work in NeuroAI and WiML workshop, NeurIPS, 2024, while the full version is under review. In addition to collaborating with Prof. Leonid Sigal in UBC, Canada and his PhD students on video object segmentation and tracking in ego-centric videos that is mainly useful in robotics settings [6]. Finally, I have explored pixel-level vision foundation models and proposed pixel-level vision centric benchmarks [13]. In this work, I showed that simple mechanisms of mining attention maps of MLLMs not trained for pixel-level grounding surpass state-of-the-art pixel-level MLLMs when evaluating both grounding and visual question answering.

[4] `https://davischallenge.org/davis2016/soa_compare.html`

1 Future directions

Video understanding with dense predictions and its fewshot counterpart is under-explored, although it has a significant impact on a multitude of applications. Thus, it serves as an interesting direction for future contributions. There are four main aspects I am interested to explore: (i) building foundation models for pixel-level video understanding, (ii) benchmarking efforts for vision foundation models for pixel-level understanding within a general purpose agent-like framework, (iii) exploring multilingual video and language modelling, and (iv) utilizing the previous in creating general purpose embedied agents.

Building foundation models for pixel-level video understanding. I am interested to develop general purpose MLLMs with pixel-level video understanding capabilities without the use of seperate specialized decoders for referring video segmentation, image segmentation, depth or flow estimation that could hinder their generalization ability. Foundation models [2] refer to a set of models that are pre-trained on large-scale data using a task that allows powerful generalization. Building such pixel-level understanding on the output level while maintaining the general-purpose framework is crucial in robotics and embodied AI. Primarily, this research direction will focus on how to generate on the output level such pixel-level predictions without the use of dedicated decoders that can parse special type of tokens. I am also interested to develop test-time adaptation techniques to such video understanding foundation models. Specifically, I plan to explore interactive and transductive prompt tuning for such pixel-level understanding foundation models. Two main considerations that are required to adapt to the video understanding task which are, spatiotemporal modelling on the feature level and spatiotemporal consistency on the prediction level. For the former, I plan to explore gated cross attention in the image encoder to allow for, not only encoding features per frame, but fusing the information across the input clip and learning unified spatiotemporal features. As for the latter, I plan to continue on my previous work [9] with a many-to-many label propagation module to enforce consistency on the predictions. Another interesting development is the use of unlabelled test data to adapt prompting in what we refer to as transductive prompt-tuning. Certain developments have proposed prompt-tuning that allows learning soft prompts using back propagation [8]. Nonetheless, previous developments did not explore prompt-tuning within pixel-level understanding models nor using the unlabelled test data to help improve the learned prompts.

Benchmarking general purpose pixel-level multi-modal large language models. I plan to continue on building benchmarks that are vision centric and focused on evaluating the capabilities of these developed pixel-level understanding MLLMs without being reduced to specialized ones. Towards this goal I plan to extend my previous work [13] towards a unified view of pixel-level understanding to include depth, flow and tracking. Additionally, I plan to build simulation environments that can help the deployment of these general purpose agents in a near real-world robotic setting. These benchmarks should be capable of evaluating various capabilities including the basic visual question answering and chat performance in addition to complex reasoning task. Furthermore, evaluating these general purpose agents abilities to generate pixel-level output as part of their scene understanding capabilities including segmentation, depth, flow and long-term tracking.

Multilingual video language modelling. Additionally, I am interested to explore vision and language and how the language structure can embed inductive biases that are necessary for parsing video data. Questioning the widely adopted norm that primarily uses the English language or the Chinese language, I am interested to explore the effect of the language structure on the language & video modelling. I am especially interested to study the Arabic language that has special properties which does not exist in others. A multitude of work has recently started investigating video and language modelling through the use of large-scale datasets with weakly annotated English captions. The Arabic language has its own advantages unlike other languages as it has a unique root system, where the root indicates a three or four consonant based words that refer to a concept. These roots can then be used to build multiple words according to a predefined set of forms to construct the corresponding adjectives, nouns, verbs and others. Such a system can enforce certain inductive biases that is necessary for handling out of distribution data and for better generalization capabilities. Weakly supervised captions from Arabic subtitles on large-scale datasets collected on movies can be used to build the corresponding Arabic language video dataset. Furthermore, interpretability studies on the learned video & language models and how the language structure affects the model can be studied across languages.

General-purpose embodied agents. Furthermore, I am interested to link the previous developments in pixel-level multi-modal large language models to building general purpose agents. Towards this I plan to explore the ability to use both supervised instruction tuning and online reinforcement learning to learn general purpose agents that can predict actions within various robotics tasks. These would include tokenizers/detokenizers for continuous actions that include navigation and manipulation tasks. Such agents would greatly benefit from the pixel-level understanding in two ways: (i) it can act as an interpretability tool to verify the scene understanding capabilities of these agents while predicting the various actions, (ii) it can improve the action prediction itself especially in complex reasoning tasks. The developed agents can be evaluated in various robotics tasks including navigation (e.g., in Habitat Nav environments), manipulation (e.g., Meta World or CALVIN) or even outside the robotics settings such as user interface control (e.g., desktop or mobile phones). This holistic view that merges pixel-level understanding with general purpose embodied agents should be the main driving force towards improved agent behaviors in complex and unseen settings. Finally, I plan to interact with researchers from mechanical engineering departments towards deploying in the real-world.

2 Summary

In summary, I plan to explore building general purpose agents with special focus on adapting pixel-level multi-modal large language models within a responsible AI framework. It provides a holistic framework to improve the generalization of general purpose agents that can impact robotics among other applications while being interpretable, which is strongly tied to a responsible approach to AI. My research agenda aligns well with TTIC and its faculty. I plan to apply for various funding opportunities from NSF and NIH in addition to industry collaborations and project grants. I also plan to apply for AWS credits for computational resources.

References

[1] Abduljaleel Adejumo, Faegheh Yeganli, Clifford Broni-bediako, Aoran Xiao, Naoto Yokoya, and Mennatullah Siam. A vision centric remote sensing benchmark. *arXiv preprint arXiv:2503.15816*, 2025.

[2] R. Bommasani, D. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] Clifford Broni-Bediako, Junshi Xia, Jian Song, Hongruixuan Chen, **Mennatullah Siam**, and Naoto Yokoya. Generalized few-shot semantic segmentation in remote sensing: Challenge and benchmark. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[4] Masood Dehghan*, Zichen Zhang*, **Mennatullah Siam***, Jun Jin, Laura Petrich, and Martin Jagersand (equally contributing). Online object and task learning via human robot interaction. In *Proceedings of the International Conference on Robotics and Automation*, pages 2132–2138, 2019.

[5] Mai Gamal, Mohamed Rashad, Eman Ehab, Seif Eldawlatly, and **Mennatullah Siam**. System identification of neural systems: Going beyond images to modelling dynamics. *arXiv preprint arXiv:2402.12519*, 2024.

[6] Raghav Goyal, Wan-Cyuan Fan, **Mennatullah Siam**, and Leonid Sigal. Tam-vt: Transformation-aware multi-scale video transformer for segmentation and tracking. *WACV*, 2025.

[7] Mir Rayat Imtiaz Hossain, **Mennatullah Siam**, Leonid Sigal, and James J Little. Visual prompting for generalized few-shot segmentation: A multi-scale approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23470–23480, 2024.

[8] M. Jia, L. Tang, B. Chen, C. Cardie, S. Belongie, B. Hariharan, and S. Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[9] Rezaul Karim, He Zhao, Richard P. Wildes, and **Mennatullah Siam**. MED-VT: Multiscale encoder-decoder video transformer with application to object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6323–6333, June 2023.

[10] Rezaul Karim, He Zhao, Richard P. Wildes, and **Mennatullah Siam**. A unified multiscale encoder-decoder transformer for video segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* ***(under review)***, 2024.

[11] Matthew Kowal, **Mennatullah Siam**, Md Amirul Islam, Neil D.B. Bruce, Richard P. Wildes, and Konstantinos G. Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13999–14009, 2022.

[12] Matthew Kowal, **Mennatullah Siam**, Md Amirul Islam, Neil D.B. Bruce, Richard P. Wildes, and Konstantinos G. Derpanis. Quantifying and learning static vs. dynamic information in deep spatiotemporal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (accepted)*, 2022.

[13] **Mennatullah Siam**. Pixfoundation: Are we heading in the right direction with pixel-level vision foundation models? *arXiv preprint arXiv:2502.04192*, 2025.

[14] **Mennatullah Siam**. Temporal transductive inference for few-shot video object segmentation. *International Journal of Computer Vision*. `https://doi.org/10.1007/s11263-025-02390-x`, 2025.

[15] **Mennatullah Siam**, Konstantinos G. Derpanis, and Richard P. Wildes. Multiscale memory comparator transformer for few-shot video segmentation. In *arXiv preprint arXiv:2307.07812*, 2023.

[16] **Mennatullah Siam**, Naren Doraiswamy, Boris N. Oreshkin, Hengshuai Yao, and Martin Jagersand. Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 860–867, 7 2020.

[17] **Mennatullah Siam**, Sara Eikerdawy, Mostafa Gamal, Moemen Abdel-Razek, Martin Jagersand, and Hong Zhang. Real-time segmentation with appearance, motion and geometry. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5793–5800, 2018.

[18] **Mennatullah Siam**, Chen Jiang, Steven Lu, Laura Petrich, Mahmoud Gamal, Mohamed Elhoseiny, and Martin Jagersand. Video object segmentation using teacher-student adaptation in a human robot interaction (HRI) setting. In *Proceedings of the International Conference on Robotics and Automation*, pages 50–56, 2019.

[19] **Mennatullah Siam**, Alex Kendall, and Martin Jagersand. Video class agnostic segmentation benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2825–2834, 2021.

[20] **Mennatullah Siam**, Heba Mahgoub, Mohamed Zahran, Senthil Yogamani, Martin Jagersand, and Ahmad El-Sallab. MODNet: Moving object detection network with motion and appearance for autonomous driving. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2018.

[21] **Mennatullah Siam**, Boris N. Oreshkin, and Martin Jagersand. AMP: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5249–5258, 2019.

[22] **Mennatullah Siam**, Senthil Yogamani, Ahmad ElSallab, and Heba Mahgoub. Verfahren zum bestimmen eines bewegungszustands eines objekts in abhängigkeit einer erzeugten bewegungsmaske und eines erzeugten begrenzungsrahmens, fahrerassistenzsystem sowie kraftfahrzeug. `https://worldwide.espacenet.com/publicationDetails/biblio?CC=DE&NR=102018114229&KC=&FT=E&locale=en_EP#`, 2019.