As a multimodal AI researcher, I collect, model, and evaluate *language* and *visual* information to create computer systems that are **inclusive, responsible** and have a **positive social impact**. I am passionate about helping others and addressing the main challenges in today's society through technology. My research contributions consist of proposing innovative tasks [3, 9–13], building automatic unimodal and multimodal (textual and visual) models to solve them [1, 3, 6, 9–13, 18], creating challenging datasets [3, 4, 9–13, 16, 17, 19], and evaluating models across various dimensions of human identity, such as race and gender [19], language [4, 14, 16, 17], income [20], and geographical location [8].

1 Towards Human Action Understanding

Understanding human actions is one of a computer system's most impactful and challenging multimodal tasks. This is due to the *complexity of our environment*, such as lighting conditions and background clutter, as well as the *complexity of our minds*, including *why* and *how* we choose to perform actions. However, the benefits outweigh the complexity costs, as once a computer system learns how to interact with humans, it can assist us in our everyday activities and significantly improve our quality of life.

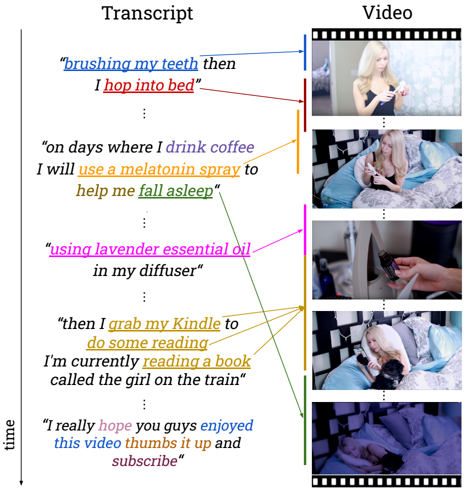
Transcript
Video
"brushing my teeth then
I hop into bed"
"on days where I drink coffee
I will use a melatonin spray to
help me fall asleep"
"using lavender essential oil
in my diffuser"
"then I grab my Kindle to
do some reading
I'm currently reading a book
called the girl on the train"
"I really hope you guys enjoyed
this video thumbs it up and
subscribe"
time

Figure 1: Temporal action localization in social media videos. The main data challenges: (1) vocabulary complexity, as people name actions in various ways (e.g., "reading a book", "grab my Kindle"), (2) not all actions can be localized (e.g., "drink coffee"), (3) misalignment between when the action is mentioned in the transcript and shown in the video (e.g., "fall asleep").

In my dissertation [7], I researched how to use multimodal information from what people *say* and *do* in social media videos to enable automatic models to learn about everyday human actions. Using machine learning techniques applied to social media videos, **my thesis was among the first to provide empirical evidence that computers can learn about human actions from social media multimodal data**. Specifically, I proposed new tasks and built annotated datasets for human action understanding, and developed multimodal models to solve them [10–13].

As a *novel and challenging data source about human actions, I proposed using vlogs, or video blogs on YouTube and demonstrated that they are suitable for learning about human actions and behaviors.* Lifestyle vlogs, in particular, are ideal for this purpose, as they showcase a person's everyday routines. The vlogger visually records their activities during a typical day and verbally expresses their intentions and feelings about activities. Additionally, vlogs typically include transcripts with complex natural language expressions, which serve as an alternative to the costly process of manual annotations.

I categorized the proposed tasks into *physical* and *commonsense* tasks, based on the type of information that the models need to solve them. Physical tasks primarily depend on visual information, such as detecting if an action appears in the video [10] or identifying the temporal location of an action in the video [13] (Figure 1). Commonsense tasks, on the other hand, require the use of both visual and contextual information. For instance, these tasks involve identifying the reason behind a particular action [12], or determining if two actions are likely to occur in the same temporal context [11].

To solve these tasks, I built multimodal models and *performed extensive ablations on how each modality contributes to solving the tasks and when the multimodal information is beneficial.* Learning the connections between vision and language signals is essential to human action understanding, and effectively using multimodal information is a challenging, open problem. Through my research, I tackle some of these challenges and pave the way for exciting opportunities for future work, including my current post-doctoral work on inclusive representations for multimodal models.

2 Inclusive Representations of AI Multimodal Models

While multimodal models have shown impressive results in various benchmarks, little is known about their limitations, as the high dimensional space learned by these models makes it difficult to identify semantic errors. Recent work has addressed this problem by designing highly controlled probing task benchmarks. I propose a more scalable framework [2] that relies on already annotated benchmarks. The method consists of extracting a large set of diverse features from a vision-language benchmark and measuring their correlation with the output of the target model. Through this method, I *confirm previous findings and uncover novel insights on what semantic concepts are challenging for current state-of-the-art vision-language models.*

My research also takes a more comprehensive approach by examining model performance across diverse demographics. Many models aim to achieve a "general understanding" by utilizing English data from Western countries. However, my research shows this approach results in a considerable performance gap across demographics. This gap is important in that demographic factors shape our identities and directly impact the model's effectiveness in the real world. Neglecting these factors could exacerbate discrimination and poverty. My research aims to bridge this gap and pave the way for more inclusive and reliable models [8, 20].

Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models. AI models have been known to overlook their impact on specific groups, especially those in low-income households. Current state-of-the-art multimodal models cannot deal with this diversity of representations. To address this issue, I engaged in a very successful collaboration with my mentee, Joan, a Ph.D. student passionate about social justice and technology. We address this limitation by comprehensively evaluating a state-of-the-art multimodal model on a geo-diverse dataset containing household images associated with different income values [20].

We find that the diversity in topic appearance is more predominant in lower-income households, probably due to a need to improvise. To illustrate, in Figure 2, in low-income households, the topic "toilet paper" appears as "grass", "newspaper" or "water". Our findings indicate that *model performance for the poorer groups is consistently lower than the wealthier groups across various topics and countries.* We highlight insights that can help mitigate these issues and *propose actionable steps for economic-level inclusive AI development, such as investing in crowd-sourced geo-diverse datasets, and developing flexible labels that consider the data provider.*

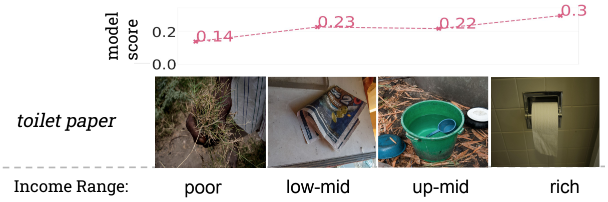

Figure 2: The model performance on the same topic is influenced by the diverse appearance of entities from the *same* topic, which often correlates with income. Our analysis draws attention to how diverse objects and actions appear in our everyday lives and calls for future work to consider this when building models and datasets.

Annotations on a Budget: Leveraging Geo-Data Similarity to Balance Model Performance and Annotation Cost. Current multimodal state-of-the-art models do not work equally well for everyone due to the imbalanced geographical and economic representation of the data used in the training process. More data must be collected from underrepresented countries to address this issue, but the cost of annotating this data is a significant bottleneck ($\sim 1\$/image$). As a complementary solution, that balances model performance and annotation costs, I propose identifying the data that is most effective to annotate [8]. My approach, depicted in Figure 3, first involves finding the countries and corresponding topics that are less represented in the training data of vision-language models and propose focusing future annotation efforts on this data. Second, across 52 countries and 94 topics, I identify the groups of countries visually similar in their representation of a given topic. This is particularly useful when there is insufficient data for one of the countries in the group, and there is no annotation budget. We can effectively supplement the data from this country by using data from the other countries in the group.

Furthermore, I find that geographical distance does not correlate with visual similarity between countries, and therefore, collecting globally diverse annotations requires considering additional information such as income, cultural heritage, and history. Importantly, my research findings *create opportunities for affordable and geo-diverse data collection, encouraging contributions to creating datasets and models that work for everyone.*

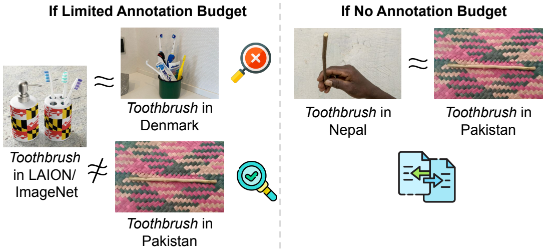

Figure 3: To budget annotations and balance model performance: (1) annotate the images visually different from the ones in high-resource data (e.g., ImageNet); (2) supplement data from low-resource countries with visually similar countries.

3 AI for Positive Social Impact

I prioritize positive social impact when selecting research directions: multimodal evaluation across income [20] and country [8], automatic detection of inspiring content across cultures [9, 14], mental health data generation across race and gender [19] and machine translation for low-resource languages [16]. Ultimately, research is for people and can enable culture and demographic-aware technologies that make a lasting positive impact on society [15].

Detecting Inspiring Content on Social Media across Cultures. Inspiration is a promising target of study because it is richly associated with a range of positive outcomes, enabling access to more creativity, productivity, and happiness. Inspiration moves a person to see new possibilities and transforms the way they perceive their own potential. Despite the attractive promise of inspiration, there has been little work on automatically detecting content that is specifically inspiring, rather than merely engaging or positive. My work [9] is the first to study inspiration through Machine Learning methods by automatically detecting inspiring content from social media data. To this end, I *provide a detailed analysis of data labeled as inspiring to gain insights on which topics are inspiring, and how they influence the readers.* One important finding is that inspiring posts fall into two main categories: posts that make people *feel* good by changing their mood or their perspective (e.g., feel gratitude, admiration), and posts that make people *act* on their thoughts and wishes (e.g., start a new hobby or a new job). I released a dataset of inspiring and non-inspiring English-language public posts from Reddit. Furthermore, in [14], together with my mentee, Gayathri, we extend this research to Romanian and Indian cultures.

OCR Improves Machine Translation for Low-Resource Languages. Although Machine Translation (MT) has achieved many recent successes, it still lacks support for most low-resource languages. A solution to this issue is to use Optical Character Recognition (OCR) tools to extract the text from large data "locked away" in digitized format. However, very little is known about OCR performance on low-resource languages and how OCR-ed data impacts MT performance. Therefore, I created and made available OCR4MT [16], an OCR benchmark for 60 low-resource languages in low-resource scripts. The languages are grouped into 8 groups according to their location and script: Latin, Cyrillic, Perso Arabic, North Indic, South Indic, Southeast Asian (SEA), China-Japan-Korea (CJK), and Other/Unique (Armenian, Amharic, Georgian, Greek, Hebrew). I used this benchmark to evaluate state-of-the-art OCR models and determine the minimum quality needed for OCR-extracted text to benefit MT. The most important takeaway from my work is that *OCR-ed data improves MT for low-resource languages and scripts.*

Mental Health across Race and Gender in Synthetic vs. Human Data. The emergence of Large Language Models (LLMs) poses many exciting use cases in various applications for synthetic data generation. At the same time, we need to develop procedures to understand the behaviors that LLMs exhibit to prevent misrepresentations of minority voices or, specifically in mental health, a misdiagnosis. To address this need, we measure how accurately GPT-3 represents depression stressors across race and gender. Specifically, we created HeadRoom [19], a synthetic dataset of depression-triggering stressors across demographics using GPT-3 while controlling for race (African American, Asian, Hispanic, White), gender (female and male), and time phase (before and after COVID-19). We use this dataset to identify the most predominant depression stressors for each demographic group and to compare our findings to human data through semantic and syntactic analyses. Our findings show that *GPT-3 generated data mimics real-life data distributions for the most prevalent depression stressors among diverse demographics*. Finally, I would like to highlight that this project was led by Shinka, one of my mentees, a master's student passionate about the intersection of AI and mental health. Throughout the project, I provided guidance and support, assisting her with experiments and paper writing.

4 My Research Agenda

My long-term vision is to see multimodal AI models being safely and responsibly used for the benefit of all people worldwide. To achieve this goal, I will continue collaborating with interdisciplinary teams to integrate diverse perspectives into the systems we build. I believe this collaboration is a two-way street, and through my efforts in Discover CS and AI in Education research and teaching, I aim to contribute to making CS accessible across disciplines. I have three main broad research areas that I am excited to pursue in the future:

Ensuring People's Identity is Represented in AI Models. To describe a lack of diversity in research, in 2010, Henrich et al. [5] coined the term WEIRD, which stands for Western, Educated, Industrialized, Rich, and Democratic. This construct highlights how most of the databases in the experimental branches of cognitive science and economics are Western-centric and primarily focused on undergraduates. This trend has continued in the AI research community, which has tended to overlook the impact of AI models on various demographics, despite their impressive performance across various tasks and datasets. This neglect exacerbates the "digital divide", which excludes non-WEIRD individuals from accessing AI benefits. Therefore, I plan to continue democratizing AI globally by creating inclusive datasets and models that represent people's diverse identities and needs.

AI for Encouraging Creativity and Inspiration across Cultures. Creativity is an essential skill in our current era of uncertainty. Technology is advancing rapidly, making it impossible to predict the world's future in five years. Despite major technological advances, education systems have not yet adapted to this reality, and companies are in dire need of innovators who think outside the box. At the same time, due to social media, we live in times of immense opportunity for information and fast worldwide impact. Social media can be a powerful tool for researching and promoting creativity and inspiration. Specifically, in [9], I am the first to show that AI models can successfully detect inspirational content on social media platforms like Reddit or Facebook. I plan to expand this line of work across more cultures and modalities. In addition, through my Launch your Confidence workshop, I aim to explore the impact of improvisation on creativity, how to integrate it into our education system, and how AI can assist in this process.

Cross-disciplinary AI for Positive Social Impact. AI technologies are starting to mature enough to have a broad impact and to help address significant global issues such as poverty, healthcare, education, and climate change. Hence, it is an ideal time to bring together experts from diverse backgrounds to explore how to use AI responsibly and effectively for the betterment of society. With this in mind, I am helping to organize the NLP for Positive Impact EMNLP 2024 workshop, which addresses various social-good themes and aims to bring together and foster conversations among people with diverse backgrounds. I plan to continue organizing AI for positive impact workshops and encourage my future students to work on projects that benefit such initiatives. This commitment is directly related to my aim of collaborating with individuals and organizations with social-good initiatives that seek responsible and inclusive AI solutions. I am particularly interested in how AI can personalize, attract, and retain more minorities in CS.

In summary, as a multimodal AI researcher, I specialize in building and evaluating *language-vision* tasks, datasets, and models, focusing on *inclusive* and *responsible* usage. I aim to work with interdisciplinary teams to identify real-world problems, create intelligent systems to address them, and then help deploy these systems to collect data for further improvements. My broad background in Computer Science, Natural Language Processing, and Computer Vision allows me to collaborate effectively with professionals from various fields.

References

[1] Laura Burdick, **Oana Ignat**, Yiming Zhang, Rada Mihalcea, Mingzhe Wang, Steve Wilson, Yumou Wei, and Jia Deng. Building a flexible knowledge graph to capture real-world events. In *TAC*, 2019.

[2] Santiago Castro*, **Oana Ignat***, and Rada Mihalcea. Scalable performance analysis for vision-language models. In *Proceedings of 12th Joint Conference on Lexical and Computational Semantics*, July 2023.

[3] Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, **Oana Ignat**, Nan Liu, Jonathan Stroud, and Rada Mihalcea. Fill-in-the-blanks as a challenging video understanding framework. In *Proceedings of 60th Annual Meeting of the Association for Computational Linguistics*, May 2022.

[4] Aylin Gunal, Verónica Pérez-Rosas, **Oana Ignat**, and Rada Mihalcea. Cross language motivational interviewing. *Preprint*.

[5] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 2010.

[6] **Oana Ignat**. Disparity image segmentation for free-space detection. In *IEEE 12th International Conference on Intelligent Computer Communication and Processing*, 2016.

[7] **Oana Ignat**. *Towards Human Action Understanding in Social Media Videos Using Multimodal Models*. PhD thesis, 2022.

[8] **Oana Ignat**, Longju Bai, Joan Nwatu, and Rada Mihalcea. Annotations on a budget: Leveraging geo-data similarity to balance model performance and annotation cost. *Under Review*, 2023.

[9] **Oana Ignat**, Y-Lan Boureau, Jane A. Yu, and Alon Y. Halevy. Detecting inspiring content on social media. *9th International Conference on Affective Computing and Intelligent Interaction*, 2021.

[10] **Oana Ignat**, Laura Burdick, Jia Deng, and Rada Mihalcea. Identifying visible actions in lifestyle vlogs. In *Proceedings 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[11] **Oana Ignat**, Santiago Castro, Weiji Li, and Rada Mihalcea. Human action co-occurrence in lifestyle vlogs using graph link prediction. *Under Review*, abs/2309.06219, 2023.

[12] **Oana Ignat**, Santiago Castro, Hanwen Miao, Weiji Li, and Rada Mihalcea. WhyAct: Identifying action reasons in lifestyle vlogs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, November 2021.

[13] **Oana Ignat**, Santiago Castro, Yuhang Zhou, Jiajun Bao, Dandan Shan, and Rada Mihalcea. When did it happen? duration-informed temporal localization of narrated actions in vlogs. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2022.

[14] **Oana Ignat*** and Gayathri GI*. Detecting inspiration on social media across cultures. *Preprint*.

[15] **Oana Ignat***, Zhijing Jin*, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Nam Ho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan Nwatu, Verónica Pérez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea. A phd student's perspective on research in nlp in the era of very large language models. *ArXiv*, abs/2305.12544, 2023.

[16] **Oana Ignat**, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. OCR improves mt for low-resource languages. In *Findings of the Association for Computational Linguistics*, May 2022.

[17] **Oana Ignat**, Phoebe Xu, Verónica Pérez-Rosas, Artem Abzaliev, and Rada Mihalcea. Multilingual deception detection of gpt-generated hotel reviews. *Preprint*.

[18] **Oana Ignat**, Jingru Yi, Burak Uzkent, and Linda Liu. Augment the pairs: Semantics-preserving image-caption pair augmentation for grounding-based vision and language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, January 2024.

[19] Shinka Mori, **Oana Ignat**, Andrew Lee, and Rada Mihalcea. Towards algorithmic fidelity: Mental health representation across demographics in synthetic vs. human-generated data. *Under Review*.

[20] Joan Nwatu*, **Oana Ignat***, and Rada Mihalcea. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.