Research Statement

As a machine learning researcher, it is my strong desire to research safe, fair, and explainable approaches in multimodal learning. I want to address the current shortcomings of multimodal foundation models while building towards unbiased and transparent machine learning research for the future.

My research interests. My interests revolve around building fair and more explainable multimodal models. Nowadays, a very common method for training large multimodal models is by collecting data from web crawlers that roam the entire known internet. Therefore, it is no surprise that this data is not curated to be balanced, safe, unbiased, or even of relatively high quality; it is also not a surprise that the resulting training results in models with **downstream biases**. Given the practical simplicity of fine-tuning from such models for researchers and its wide adoption within the community (e.g., as I've seen from many works in ICML'26), I believe this is a crucial research problem affecting current ML trends and one that I am very much interested in. I am aware of many popular tools for explainability and reliability (e.g., concept bottleneck models, conformal prediction, approaches focusing on uncertainty quantification) in this direction and I believe there is an immense potential here for further research in safety-focused AI.

About me. My background is in robust and fair data-imbalanced models for computer vision and machine learning. In my previous work, we derived class-intrinsic measures of imbalance and difficulty for improving robustness and fairness in long-tailed recognition. I have prior experience in investigating representations of CLIP-like VLMs and audio foundation models and learning from images, point clouds, language, and audio.

As an example of a hypothetical research project in the intersection of these interests, I would like to present the following idea.

Investigating the Imbalances of Multimodal Foundation Models

In training of multimodal models (such as CLIP), some notion of imbalance appears at multiple stages.

First, common pretraining datasets are often long-tailed, with some concepts being overrepresented and many others having a very low number of samples. This is the natural distribution of image datasets such as iNaturalist-2018 (Van Horn et al., 2018), of autonomous driving scenarios (Xu et al., 2025), as well as the internet itself retrieved through web crawlers (Sharma et al., 2018). Notably, this has serious consequences as recent research seems to indicate that for multimodal models trained on such data, linear improvements in zero-shot generalization require exponentially more data for learned concepts (Udandarao et al., 2024), which signals a severe case of data-inefficient learning.

Second, in multimodal learning it is also the case that there are imbalances between the modalities as well, leading to modality gaps (Liang et al., 2022) and mismatches in concept alignment between modalities. Although the popular Platonic Representation Hypothesis (Huh et al., 2024) viewpoint argues that models learning from different modalities are converging to a shared view of the world, in practice there is active work on ensuring alignment of modalities (Roschmann et al., 2026; Schnaus et al., 2025).

My hypothesis. In their core, multimodal models such as CLIP aim to construct a unified representation space that connects many modalities. The effects of imbalanced pretraining and modality mismatches are directly linked in representation space, mainly caused by naive application of contrastive objectives (e.g., InfoNCE). I believe a solution towards designing better models is to manipulate representations of concepts in multimodal foundation models. Leveraging the geometric distribution of the

same concepts in different modalities within the embedding space is in my opinion is the key to improving multimodal alignment, as well as the starting point for building safer, fairer, and more interpretable multimodal models.

References

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. Y. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, *35*, 17612–17625.

Roschmann, S., Krzakala, P., Mazelet, S., Bouniot, Q., & Akata, Z. (2026). SOTAlign: Semi-supervised alignment of unimodal vision and language models via optimal transport. *Forty-third International Conference on Machine Learning*. https://openreview.net/forum?id=8PMYGcTv3L

Schnaus, D., Araslanov, N., & Cremers, D. (2025). It's a (blind) match! towards vision-language correspondence without parallel data. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24983–24992.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Udandarao, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P., Bibi, A., Albanie, S., & Bethge, M. (2024). No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *Advances in Neural Information Processing Systems*, *37*, 61735–61792.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The inaturalist species classification and detection dataset. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Xu, R., Lin, H., Jeon, W., Feng, H., Zou, Y., Sun, L., Gorman, J., Tolstaya, E., Tang, S., White, B., et al. (2025). Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*.