Semantic information shapes gaze patterns during naturalistic viewing of movies

Sophie (Xing) Su, Aditya Upadhyayula, Jeffrey M. Zacks

Department of Psychological & Brain Sciences, Washington University in St. Louis

Dynamic Cognition Laboratory

Background

- Contrastive Language–Image Pretraining (CLIP)[1]: aligns image and text embeddings in a shared space.
- CLIP embeddings capture semantic information that can be predictive of human gaze pattern.[2]
- Scene inversion disrupts processing of semantic information in a scene, but preserves low-level visual salience inversion disrupts viewing patterns.[3]
- However, no studies have examined the role of semantic information on gaze patterns in dynamic movies.
- This study investigates how semantic information relates to gaze patterns during passive movie viewing by mapping CLIP-extracted features to eye movements.

Hypotheses

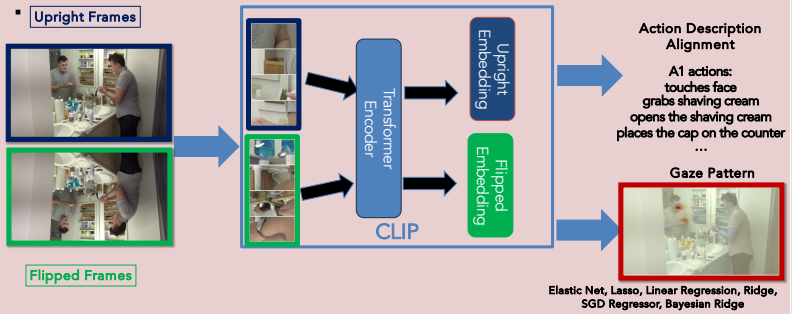
Upright images have more accurate semantic action information, making their embeddings more likely to accurately predict human gaze patterns than those derived from flipped images.

Eye-tracking Study

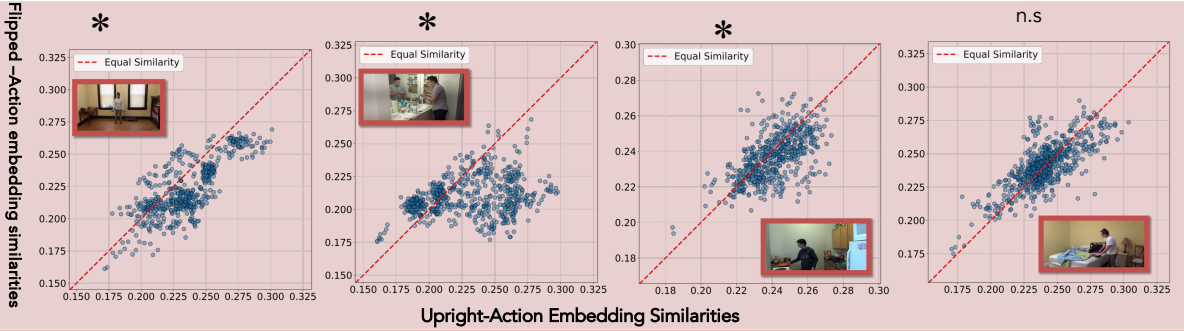
Participants:
- N = 100, Female = 67, Male = 33, Mean Age: 20, SD Age: 2.39

Task:
- Participants watched four movies (exercise, grooming, cleaning, breakfast) of daily activities. The order of the movie was counterbalanced.
- While watching movies, participants' gaze locations were recorded using EyeLink 1000.

Mapping CLIP Embeddings to Gaze Patterns and Action Descriptions: Upright vs. Flipped Frames

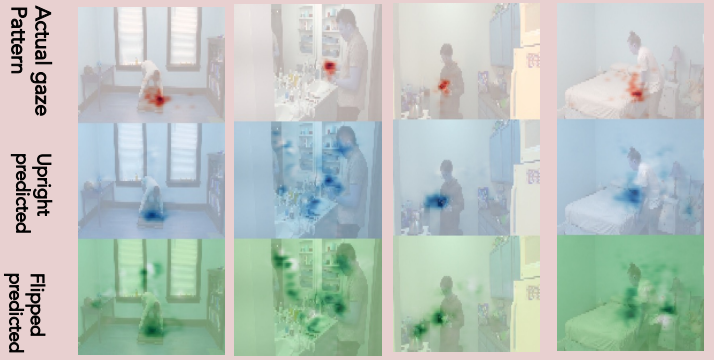


Upright Frames Align More Closely with Action Text than Flipped Frames



Upright CLIP Embeddings Better Predict Passive Gaze Patterns than Flipped Embeddings

Train of the 80% frames in the beginning, test on 20% frames in the end for each movie.



We fit a linear mixed-effects model predicting Jensen-Shannon divergence from embedding type (upright vs. flip), with a random intercept for movie and an AR(1) structure to account for frame-to-frame autocorrelation within each movie and embedding type.

Model shows a significant effect of embedding type, $b = -0.016$, $SE = 0.004$, $t(995) = -3.49$, $p < .001$

Conclusions & Future Directions

- Upright embeddings better predicts actual gaze patterns than flipped embeddings, likely because they preserve action-related semantic information that guides visual attention
- Other transformations that preserve low-level alignment information—such as diffeomorphic—could be applied to test whether our findings generalize beyond simple flip
- Comparing model variants to naturalistic gaze patterns provides a testbed for theory evaluation.

References

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763). PMLR.
2. Hayes, T. R., &Henderson, J. M. (2023). Transformers bridge vision and language to estimate and understand scene meaning. Research Square, rs-3.
3. Hayes, T. R., & Henderson, J. M. (2022). Scene inversion reveals distinct patterns of attention to semantically interpreted and uninterpreted features. Cognition, 229, 105231.

Acknowledgements

We thank the Dynamic Cognition Lab for their support and helpful discussion.
We thank Dr. Tan Nguyen for his help with poster production

Email: s.sophie@wustl.edu | Tweet at me: @sophiejoesu