Understanding Video Transformers via Universal Concept Discovery

Appendix

In this appendix, we report additional results, visualizations and implementation details. We first conduct validation experiments for VTCD to show that the discovered concepts align with human-interpretable groundtruth labels in Section 7.1. We then provide statistics of concepts importance distribution between layers in Section 7.2. Next, in Section 7.3, we provide further discussion and qualitative results showing how different concepts are captured in different self-attention heads from the same layer. We provide further implementation details in Section 8. Finally, we discuss limitations of VTCD in Section 9. Note that we include additional video results and corresponding discussions on the project web page.

7. Additional results

7.1. Concept validation

Directly measuring concept accuracy is not possible in an open world approach as we don't know a priori what concepts should be present in a model. There are, however, special cases where we can directly measure the accuracy of *some* of the concepts. For example, TCOW is trained to track through occlusions, so we can expect to find concepts that correspond to the target object and containers/occluders in it. We perform this evaluation for VTCD and the random crop baseline [24, 25, 72] and report the mIoU between the best found concepts and the groundtruth masks in Table 4 (top). These results validate the accuracy of VTCD, which is able to reach up to 94% of the performance of the fully-supervised TCOW by discovering concepts in its intermediate representations.

7.2. Quantitative analysis of per-layer concept importance

We now quantify the importance of each model layer for the two target models analyzed in Section 5.2 in the main paper. To this end, we calculate the average concept importance ranking per-layer and then normalize this value, which results in a $[0-1]$ score, where higher values indicate more important layers, and plot the results in Figure 8.

We immediately see similarities and differences between the two models. For example, the first two layers are less important than mid layers for both models. For VideoMAE, the middle (6) and end layer (12) are the most important. Interestingly, for TCOW, the most important layer by far is layer 3, while the final layer is the least important. This makes intuitive sense since TCOW is an object tracking model, hence it most heavily utilizes spatiotemporal positional information and object-centric representations in early-to-mid layers. In contrast, VideoMAE is trained for action classification, which requires fine-grained, spatiotemporal concepts in the last layers.

Method	Target	Occluders	Containers
Baseline	3.0	31.5	43.3
VTCD (Ours)	**19.2**	**69.7**	**73.8**
TCOW (supervised)	36.8	76.8	78.2

Table 4. Evaluating the accuracy of object tracking concepts found in the TCOW model by VTCD and the random crop baseline used in other recent methods [24, 25, 72].

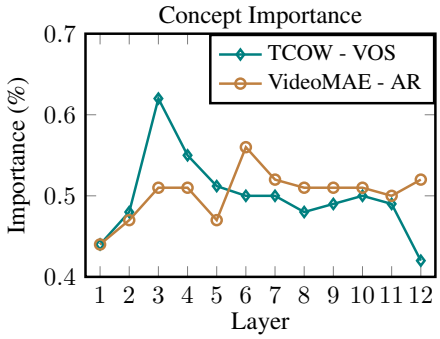

Figure 8. The average concept importance over all model layers for a VOS model (TCOW) and an action recognition model (VideoMAE). Interestingly, while VideoMAE encodes important concepts both in the middle and late in the model, TCOW encodes most important concepts at layer three and the least important in the final layer.

7.3. Uniqueness of head concepts

As discussed in Section 4 in the main paper, we qualitatively visualize the concepts from the same layer but different heads of a model to demonstrate that the heads encode diverse concepts. For example, Figure 9 shows that discovered concepts in heads one and six in layer five of the TCOW [66] model encode unrelated concepts (*e.g.* positional and falling objects). This corroborates existing work [2, 20, 36, 50, 67] that heads capture independent information and are therefore a necessary unit of study using VTCD.

8. Implementation details

Concept discovery. When generating tubelets (Section 3.1.1), we use 12 segments and set all other hyperparameters to the Scikit-Image [65] defaults, except for the compactness parameter, which is tuned on a held-out set for each model to the following values: TCOW - 0.01, VideoMAE - 0.1, SSL-VideoMAE - 0.1, InternVideo - 0.15.

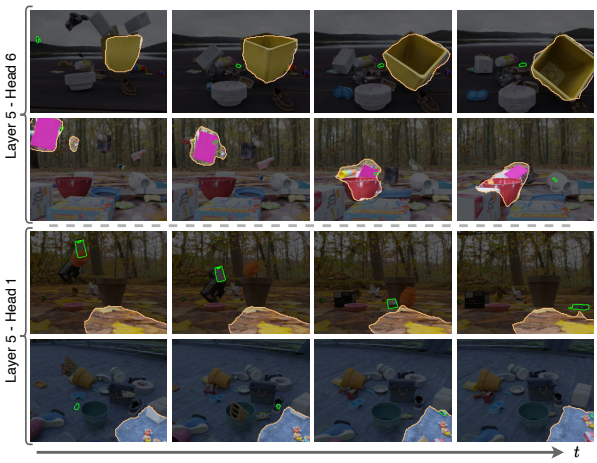

Figure 9. Different heads in the same layer encode concepts capturing different phenomena. In layer 5 of TCOW [66], head 6 (top two rows) highlights falling objects, while head 1 (bottom two rows) captures spatial position.

When clustering concepts using CNMF (Section 3.1.2) we follow the same protocol as [2] and use the Elbow method, with the Silhouette metric [57] as the distance, to select the number of clusters with a threshold of 0.9.

Concept importance. For all importance rankings using CRIS, we use the original loss the models were trained with. For InternVideo, we use logit value for the target class by encoding the class name with the text encoder, and then taking the dot product between the text and video features. We use 4,000 masking iterations for all models, except for TCOW [66], where we empirically observe longer convergence times and use 8,000 masks.

Concept pruning with VTCD. The six classes targeted in the concept pruning application (Table 2 in the main paper) are as follows:

1. Pouring something into something until it overflows
2. Spilling something behind something
3. Spilling something next to something
4. Spilling something onto something
5. Tipping something with something in it over, so something in it falls out
6. Trying to pour something into something, but missing so it spills next to it

Semi-Supervised VOS with VTCD. To evaluate VOS performance on the DAVIS'16 benchmark [52] with any pretrained video transformer, we first identify self-attention heads that encode object-centric concepts on the training set. Specifically, we calculate the mIoU between every concept found by VTCD (*i.e.* the set of tubelets belonging to that concept) and the groundtruth labels for each training video. We then record the heads in which the best performing concepts came from. Next, we run VTCD on the validation set using only the heads containing object-centric concepts. To select the final concept for evaluation, we choose the one with the highest mIoU with the first frame label (*i.e.* the query mask).

To generate tubelets for an entire video, we simply use non-overlapping sliding windows and run SLIC on the temporally concatenated features. We also leverage SAM [38] for post-processing the tubelets generated by VTCD. Note that SAM can take as input a set of points, a bounding box, or both. For each frame and corresponding mask from the VTCD tubelet, we generate the smallest bounding box surrounding the mask. We also calculate the centroid of the mask and then sample two points from a Gaussian centered at the centroid, with covariance $(L/10, W/10)$ where L and W are the length and width of the mask, respectively. We then pass both the box and points to SAM ViT-h to produce the post-processed mask for that frame.

9. Limitations

One limitation of our method is the need to manually set the SLIC compactness hyper-parameter. Additionally, the compactness property of SLIC makes it challenging to capture concepts that are not spatially localized. Finally, there is a high computational requirement for computing the Rosetta score in Equations 7 and 8 as D grows.

Acknowledgements. We acknowledge financial support from the Canadian NSERC Discovery Grants. K.G.D. contributed to this work in their personal capacity as Associate Professor at York University. Thanks to Greg Shakhnarovich for feedback on the paper.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34 (11):2274–2282, 2012. 2, 3, 6

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep VIT features as dense visual descriptors. In *ECCV Workshops*, 2022. 2, 4, 5, 7, 8, 1

[3] Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discovering objects that can move. In *CVPR*, 2022. 8

[4] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. 8

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. 1, 2

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, and Adam Letts. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 8

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 8

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM FAccT*, 2018. 1

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 5

[10] Michael Chang, Tomer Ullman, Antonio Torralba, and Joshua Tenenbaum. A compositional object-based approach to learning physical dynamics. In *ICLR*, 2016. 8

[11] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, 2021. 2

[12] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 2

[13] Jinwoo Choi, Chen Gao, C. E. Joseph Messou, and Jia-Bin Huang. Why can't I dance in the mall? Learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019. 3

[14] European Commision. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *European Commision*, 2021. 1

[15] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *ICLR*, 2020. 2

[16] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 1

[17] Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *TPAMI*, 32(1):45–55, 2008. 2, 4

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[19] Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *ICCV*, 2023. 2, 5

[20] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html. 2, 4, 1

[21] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html. 2

[22] Christoph Feichtenhofer, Axel Pinz, Richard P Wildes, and Andrew Zisserman. Deep insights into convolutional networks for video recognition. *IJCV*, 128:420–437, 2020. 3

[23] Christoph Feichtenhofer, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *NeurIPS*, 2022. 3

[24] Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *NeurIPS*, 2023. 1, 2, 3, 4, 5, 6

[25] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept recursive activation factorization for explainability. In *CVPR*, 2023. 1, 2, 3, 4, 5, 6

[26] Ruohan Gao, Dinesh Jayaraman, and Kristen Grauman. Object-centric representation learning from unlabeled videos. In *ACCV*, 2017. 8

[27] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022. 1, 7, 8

[28] Amir Ghodrati, Efstratios Gavves, and Cees G. M. Snoek. Video time: Properties, encoders and evaluation. In *BMVC*, 2018. 3

[29] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *NeurIPS*, 2019. 1, 2, 3, 6

[30] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 5

[31] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, and Charles Herrmann. Kubric: A scalable dataset generator. In *CVPR*, 2022. 5

[32] Isma Hadji and Richard P Wildes. A new large scale dynamic texture dataset with application to convnet understanding. In *ECCV*, 2018. 3

[33] Sven Ove Hansson, Matts-Åke Belin, and Björn Lundgren. Self-driving vehicles-An ethical overview. *Philosophy & Technology*, pages 1–26, 2021. 1

[34] The White House. President biden issues executive order on safe, secure, and trustworthy artificial intelligence. *The White House*, 2023. 1

[35] Filip Ilic, Thomas Pock, and Richard P Wildes. Is appearance free action recognition possible? In *ECCV*, 2022. 3

[36] Rezaul Karim, He Zhao, Richard P. Wildes, and Mennatullah Siam. MED-VT: Multiscale encoder-decoder video transformer with application to object segmentation. In *CVPR*, 2023. 2, 4, 1

[37] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018. 1, 2, 6

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023. 8, 2

[39] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *IJCV*, 130(5):1366–1401, 2022. 3

[40] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *CVPR*, 2022. 3

[41] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. Quantifying and learning static vs. dynamic information in deep spatiotemporal networks. *arXiv preprint arXiv:2211.01783*, 2022. 3

[42] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755):788–791, 1999. 2

[43] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013. 3

[44] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *CVPR*, 2019. 3

[45] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018. 3

[46] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. 4

[47] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 2020. 8

[48] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *NeurIPS*, 2019. 2, 4

[49] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in. 1

[50] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html. 4, 1

[51] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *ICLR*, 2023. 2

[52] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2, 3, 8

[53] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 2

[54] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *NeurIPS*, 2022. 2

[55] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *NeurIPS*, 2021. 1, 2

[56] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *CVPR*, 2022. 3

[57] Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 2

[58] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[59] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, and Thomas Brox. Bridging the gap to real-world object-centric learning. In *ICLR*, 2023. 8

[60] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[61] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *ECCV*, 2018. 8

[62] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, and Chuang Gan. Masked motion encoding for self-supervised video representation learning. In *CVPR*, 2023. 3

[63] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *CVPR*, 2023. 3

[64] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022. 3, 5, 8

[65] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. 1

[66] Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, and Carl Vondrick. Tracking through containers and occluders in the wild. In *CVPR*, 2023. 1, 2, 3, 5

[67] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL*, 2019. 2, 4, 1

[68] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *CVPR*, 2023. 1, 2

[69] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023. 3

[70] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, and Zun Wang. InternVideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 5, 8

[71] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *NeurIPS*, 2015. 8

[72] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *NeurIPS*, 2020. 5, 6, 1

[73] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware ego-centric video recognition model. In *ICCV*, 2023. 8

[74] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for CNN models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 1, 2, 4, 6

[75] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *ICML*, 2022. 2