

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.
- Reuben M. Aronson, Thiago Santini, Thomas C. Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. Eye-hand behavior in human-robot shared manipulation. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 4–13, 2018.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023.
- Abdishakour Awale and Duygu Sarikaya. Human gaze guided attention for surgical activity recognition, 2022.
- Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network, 2025. URL <https://arxiv.org/abs/2504.13181>.
- Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *IEEE Intelligent Transportation Systems Magazine*, 9(4):10–22, 2017. doi: 10.1109/MITS.2017.2743165.
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtava Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023. URL <https://arxiv.org/abs/2209.06794>.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL <https://doi.org/10.1007/s11263-021-01531-2>.
- Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4):694–724, July 2007. ISSN 0033-2909. doi: 10.1037/0033-2909.133.4.694. URL <http://dx.doi.org/10.1037/0033-2909.133.4.694>.
- Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention, 2019.
- Tanya Goyal and Greg Durrett. Annotating and modeling fine-grained factuality in summarization, 2021. URL <https://arxiv.org/abs/2104.04302>.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray,

385 Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Car-
386 tillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano
387 Fragomeni, Qichen Fu, Abrahm Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang,
388 Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico
389 Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan
390 Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanov,
391 Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo,
392 Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall,
393 Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna
394 Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva,
395 Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba,
396 Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of
397 egocentric video, 2022.

398 Michelle R. Greene, Benjamin J. Balas, Mark D. Lescroart, Paul R. MacNeilage, Jennifer A. Hart,
399 Kamran Binaee, Peter A. Hausamann, Ronald Mezile, Bharath Shankar, Christian B. Sinnott,
400 Kaylie Capurro, Savannah Halow, Hunter Howe, Mariam Josyula, Annie Li, Abraham Mieses,
401 Amina Mohamed, Ilya Nudnou, Ezra Parkhill, Peter Riley, Brett Schmidt, Matthew W. Shinkle,
402 Wentao Si, Brian Szekely, Joaquin M. Torres, and Eliana Weissmann. The visual experience
403 dataset: Over 200 recorded hours of integrated eye movement, odometry, and egocentric video,
404 2024. URL <https://arxiv.org/abs/2404.18934>.

405 Mary M Hayhoe, Anurag Shrivastava, Ryan Mruczek, and Jeff B Pelz. Visual memory and motor
406 planning in a natural task. *J. Vis.*, 3(1):49–63, 2003.

407 Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly
408 estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806,
409 2020. ISSN 1941-0042. doi: 10.1109/tip.2020.3007841. URL [http://dx.doi.org/10.1109/](http://dx.doi.org/10.1109/TIP.2020.3007841)
410 [TIP.2020.3007841](http://dx.doi.org/10.1109/TIP.2020.3007841).

411 Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong,
412 Yali Wang, Limin Wang, and Yu Qiao. Egoxolearn: A dataset for bridging asynchronous ego- and
413 exo-centric view of procedural activities in real world, 2025. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2403.16182)
414 [2403.16182](https://arxiv.org/abs/2403.16182).

415 Junhwa Hur and Stefan Roth. Mirrorflow: Exploiting symmetries in joint optical flow and occlusion
416 estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct
417 2017.

418 Bruno Laeng, Ilona M. Bloem, Stefania D’Ascenzo, and Luca Tommasi. Scrutinizing visual images:
419 The role of gaze in mental imagery and memory. *Cognition*, 131(2):263–283, 2014. ISSN 0010-
420 0277. doi: <https://doi.org/10.1016/j.cognition.2014.01.003>. URL [https://www.sciencedirect](https://www.sciencedirect.com/science/article/pii/S0010027714000043)
421 [com/science/article/pii/S0010027714000043](https://www.sciencedirect.com/science/article/pii/S0010027714000043).

422 Bolin Lai, Miao Liu, Fiona Ryan, and James M. Rehg. In the eye of transformer: Global-local
423 correlation for egocentric gaze estimation, 2024. URL <https://arxiv.org/abs/2208.04464>.

424 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-
425 training for unified vision-language understanding and generation, 2022. URL [https://arxiv](https://arxiv.org/abs/2201.12086)
426 [org/abs/2201.12086](https://arxiv.org/abs/2201.12086).

427 Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang,
428 Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models
429 as effective robot imitators, 2024. URL <https://arxiv.org/abs/2311.01378>.

430 Yin Li, Miao Liu, and James M. Rehg. In the eye of the beholder: Gaze and actions in first person
431 video, 2020a.

432 Yin Li, Miao Liu, and James M. Rehg. In the eye of the beholder: Gaze and actions in first person
433 video, 2020b. URL <https://arxiv.org/abs/2006.00626>.

434 Chen Lin and Xing Long. Open-llava-next: An open-source implementation of llava-next series
435 for facilitating the large multi-modal model community. [https://github.com/xiaoachen98/](https://github.com/xiaoachen98/Open-LLaVA-NeXT)
436 [Open-LLaVA-NeXT](https://github.com/xiaoachen98/Open-LLaVA-NeXT) 2024.

437 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL
438 <https://arxiv.org/abs/2304.08485>.

439 Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic
440 representations for vision-and-language tasks, 2019a. URL <https://arxiv.org/abs/1908.02265>.

441

442 Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action
443 recognition. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*,
444 pages 4425–4434, 2019b. doi: 10.1109/ICCVW.2019.00543.

445 Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human
446 egocentric action forecasting, 2024.

447 Kyle Min and Jason Corso. Integrating human gaze into attention for egocentric activity recognition.
448 11 2020.

449 OpenAI. Gpt-4v(ision), 2023. URL <https://cdn.openai.com/contributions/gpt-4v.pdf>.

450 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
451 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
452 Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>.

453

454 Keith Rayner. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene
455 perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506,
456 2009.

457 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks,
458 2019. URL <https://arxiv.org/abs/1908.10084>.

459 Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer
460 for egocentric action anticipation, 2024. URL <https://arxiv.org/abs/2211.14154>.

461 Ali Shafti, Pavel Orlov, and A. Aldo Faisal. Gaze-based, context-aware robotic system for assisted
462 reaching and grasping. In *2019 International Conference on Robotics and Automation (ICRA)*,
463 pages 863–869, 2019. doi: 10.1109/ICRA.2019.8793804.

464 Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric
465 attention for egocentric activity recognition, 2018.

466 Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from trans-
467 formers, 2019. URL <https://arxiv.org/abs/1908.07490>.

468 Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. URL
469 <https://arxiv.org/abs/2003.12039>.

470 Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Enhancing
471 next active object-based egocentric action anticipation with guided attention, 2023. URL <https://arxiv.org/abs/2305.12953>.

472

473 Steven P. Tipper. Eps mid-career award 2009: From observation to action simulation: The role of
474 attention, eye-gaze, emotion, and body state. *Quarterly Journal of Experimental Psychology*, 63
475 (11):2081–2105, November 2010. ISSN 1747-0226. doi: 10.1080/17470211003624002. URL
476 <http://dx.doi.org/10.1080/17470211003624002>.

477 Daniel Weber, Thiago Santini, Andreas Zell, and Enkelejda Kasneci. Distilling location proposals
478 of unknown objects through gaze information for human-robot interaction. In *2020 IEEE/RSJ*
479 *International Conference on Intelligent Robots and Systems (IROS)*, pages 11086–11093, 2020.
480 doi: 10.1109/IROS45743.2020.9340893.

481 Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu,
482 Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init
483 attention, 2023. URL <https://arxiv.org/abs/2303.16199>.

- 484 Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee,
485 and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos?,
486 2024a. URL <https://arxiv.org/abs/2307.16368>
- 487 Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. Vialm: A survey and benchmark of
488 visually impaired assistance with large models, 2024b. URL <https://arxiv.org/abs/2402.01735>.
- 490 Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from
491 large language models, 2022. URL <https://arxiv.org/abs/2212.04501>
- 492 Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and
493 Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook, 2024. URL
494 <https://arxiv.org/abs/2310.14414>.