Leveraging Large Language Models to Annotate Activities of Daily Living Captured with Egocentric Vision

1st Sloke Shrestha
University of Texas at Austin
Austin TX 78712, USA
sloke@utexas.edu

2nd Edison Thomaz
University of Texas at Austin
Austin TX 78712, USA
ethomaz@utexas.edu

***Abstract*—Developing a system that automatically and passively recognizes activities of daily living (ADLs) would be transformative for numerous health applications. However, engineering approaches for building such a classifier today requires the availability of large and rich annotated datasets representing ADLs in a generalizable way. In this work, we evaluated state of the art large language models (LLMs) to perform fully-automated and assisted manual annotations of first-person images with ADLs. We performed automatic evaluations on four different vision language pipelines (VLPs): concept detector, concept detector + GPT-3.5, BLIP2, and GPT-4. Three of them were tested on 31,849 first person images and one of them, GPT-4, was tested on 3,446 images first person images. Among the four VLPs, BLIP2 scored the highest cosine similarity of 0.86. Furthermore, we evaluated assisted manual annotation with 20 participants who annotated 100 ADL images with recommended labels from three different VLPs. We show that annotation with BLIP2 assistance has highest pick rate of 0.698 and a subjective workload (NASA Task Load Index) score of 39.41 in a scale of 100. Despite limitations, our work demonstrates how large language model can be leveraged to optimize the difficult task of data annotation for building ADL classifiers.**

***Index Terms*—Large Language Models, Vison Language Models, Human Activity Recognition, Data Annotation, Activities of Daily Living**

I. Introduction

Recognizing activities of daily living (ADL) plays a crucial role in health applications, particularly in the fields of elderly care [1], [2], rehabilitation [3], and chronic disease management [4]. ADLs refer to the basic tasks that individuals perform every day, such as eating, bathing, dressing, toileting, and moving around. Monitoring these activities can provide valuable insights into a person's health status and functional abilities [5].

Researchers have explored numerous methods over the years aimed at building systems that can automatically and passively recognize ADLs. This task centers on building and evaluating machine learning models that are trained to discriminate ADLs using sensor data [6]–[10]. Engineering recognition models that perform well in real-world conditions is a major undertaking. A significant challenge is the acquisition of ground truth data to train supervised models. Direct observation, self-report-based diaries and experience sampling are some of the data collection methods that have been traditionally used. Unfortunately, these approaches suffer from several practical and methodological flaws such as biases and subjectivity.

More recently, the use of wearable egocentric imaging has emerged as an objective way of recording everyday activities outside the lab [6], [11]. However, reviewing and annotating thousands of first-person photos or video clips is a tedious, time-consuming, and error-prone process.

In this paper, we leveraged LLMs and vision models to develop and evaluate novel annotation approaches aimed at reducing the time and effort required in reviewing and labeling egocentric photos representing ADLs. In our proposed pipeline, as shown in figure 1, a person wears a wearable camera and captures images in regular interval while they go about their daily life. A VLP then infers the human activity based on the contents of the image. We explored four distinct vision language pipelines (VLPs): (1) Concept Detector (2) Concept Detector + GPT-3.5, (3) BLIP2 [12] and (4) GPT-4 [13]. These VLPs are evaluated under two scenarios: *fully-automated annotation* and *assisted manual annotation*. In the fully-automatic annotation case, the inferred activities from the VLPs were directly compared against the ground truth. For the assisted manual annotation scenario, the labels were selected manually using an interface in which annotation suggestions were provided by the VLPs.

The specific contribution of our work are:

- A novel *fully-automated annotation* strategy to annotate first-person images representing ADLs using LLMs. We annotated 31,849 first person images automatically from the ADL video dataset [14] using four VLPs: concept detector, concept-detector + GPT-3.5, BLIP2, and GPT-4 (3,446 images). Compared to ground truth, BLIP2 outperformed the other methods with a cosine similarity score of 0.86 on average.
- A comprehensive analysis of an approach where labels produced by the VLPs were offered as suggestions for assisted manual annotation. We ran a human study with 20 participants and evaluate the approach in terms of task completion time, annotation accuracy, workload and pick rate. We observe that the suggestions from BLIP2 has

the highest pick rate of 0.698 and a subjective workload score of 39.41 (scored out of 100).

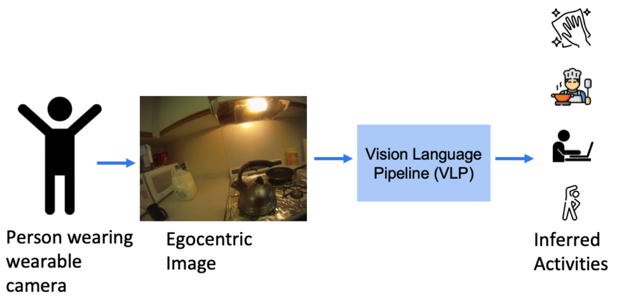

Person wearing wearable camera
Egocentric Image
Vision Language Pipeline (VLP)
Inferred Activities

Fig. 1. We present a methodology for data annotation wherein an individual utilizes a wearable camera to capture images at predetermined intervals during their everyday activities. Subsequently, a vision-language pipeline can be employed to infer human activities from the egocentric images.

II. Models

In this work, we used a combination of vision and large language models to generate and support the assignment of annotations. We compared the performance of these models against a concept detection baseline.

A. Concept Detector

We used Clarifai's general-image-recognition-v2 to list different concepts detected in the egocentric images. Concepts, as defined by clarifai, are adjectives, verbs, and objects that are associated with the image. These concepts have been used in the past works by Thomaz [11] to aid annotation of first-person images and by Fast [15] to derive language signals from images.

B. Concept Detector + GPT-3.5

This VLP adds an LLM on top of the concept detector. We used the list of concepts obtained from clarifai's image recognition model and fed it to OpenAI's GPT-3.5-turbo using a prompt with few shot examples [16].

C. Vision Language Model (VLM)

VLMs are multi-modal models which accept vision and language inputs. In this work, we chose GPT-4 [13] and BLIP2 [12] as our VLMs. BLIP2 is different from concept detector + LLMs because BLIP2 bridges the modality between vision and language through pre-training, whereas, concept detector + GPT-3.5 bridges the modality gap by using vision models' outputs to prompt the LLM. BLIP2 bootstraps language-image pre-training from off-the-shelf frozen pre-trained image encoders and frozen LLMs. BLIP2 was trained on various dataset like COCO [17], Visual Genome [18], CC3M [19], CC12M [20], SBU [21], LAION400M dataset [22]. The authors of BLIP2 also created synthetic captions for web images using CapFilt method [12], [23].

We used BLIP2 with the ViT-g/14 encoder for the image encoder and pre-trained OPT-2.7b [24] for the frozen LLM. The implementation of BLIP2 is provided by HuggingFace. The prompt used for the BLIP2 is *"What is the person doing?"*. In this case, we empirically decided to use the shorter and simpler prompt since it provided decent inferences.

We used GPT-4 as the other VLM. GPT-4 is a multi-modal chat-bot service provided by OpenAI. It can accept image and text inputs, and can have conversations with users. In this work, we used the gpt-4-vision-preview API provided by OpenAI with few-shot prompting to retrieve GPT-4 inferences.

III. Evaluation

A. Dataset

We used an open source dataset, activity of daily living (ADL) video dataset [14], to evaluate the different VLPs. The ADL dataset consists of egocentric videos collected from a chest-mounted camera from n = 20 participants. The videos were annotated with 18 different activities including combing hair, putting on makeup, and brushing teeth. We sampled frames from the videos at 1 frame per second to get egocentric images. Only windows of videos that consisted of ADL activity annotations were sampled for frames. We sampled a total of 31,849 frames. Every frame we sampled consisted of a ground truth made available by the ADL video dataset.

B. Fully-automated annotation

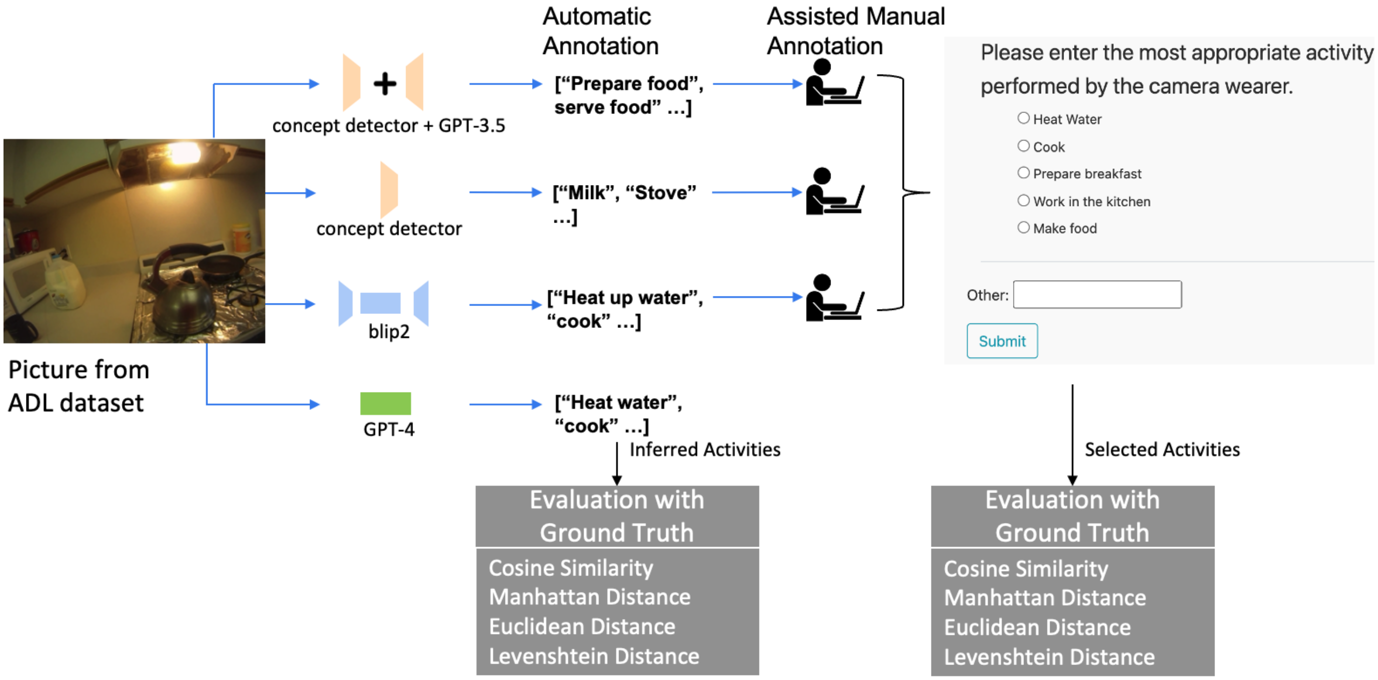
For the automatic annotation scheme, we took the 31,849 egocentric images extracted from the ADL dataset and passed them through the the three VLPs: BLIP2, concept detector, and concept detector + GPT-3.5. We passed 3,446 egocentric images from the dataset to GPT-4. The smaller dataset for GPT-4 was obtained by sampling the frames at 1/10 frames per second. This was done because OpenAI severely rate limits the GPT-4 API and running inference on the full dataset was infeasible. We later calculated similarity metrics between the inferred activity and ground truth pairs. These metrics are described in section III-D.

C. Assisted manual annotation

To understand if the VLPs can help the annotators perform the annotation faster with better accuracy, we performed a user study. We hypothesized that with the VLP suggestions provided to a human annotator, the annotation performance would increase in terms of speed, subjective workload, and accuracy.

We enrolled 20 participants in a study involving the presentation of a user interface (UI) featuring first-person images for annotation by the participants. Along with the picture, the UI showed suggestions for possible annotations for the first-person image which were obtained from the different VLPs. The user either selected one of the suggestions or typed a more appropriate annotation for the image. The study was performed in a laboratory.

We used four different sessions for the study, namely; 1) Unassisted 2) Concept Detector 3) Concept Detector + GPT-3.5 and 4) BLIP2. Unassisted provided no options for the user to choose from. So, the users needed to type annotations for



Please enter the most appropriate activity performed by the camera wearer.

○ Heat Water
○ Cook
○ Prepare breakfast
○ Work in the kitchen
○ Make food

Other:

Fig. 2. Egocentric image frames were extracted from ADL video dataset. The images were passed through the four VLPs: BLIP2, Concept Detector, Concept Detector + GPT 3.5, and GPT-4. The outputs of the VLPs were compared with the ground truth to get similarity scores between inferred activity and ground truth. For evaluation of manual annotation with VLP suggestions, we took a subset (100 images) of egocentric images from ADL video dataset. The inferred activities from the VLP outputs were presented to the annotators as options to choose from. The selected activity and the ground truth were compared to calculate a similarity metric.

all of the images in the session. Rest of the sessions provided suggestions using their respective VLPs.

All participants went through all four sessions. Each session consisted of 25 first-person images which were randomly sampled from the ADL video dataset frames. Given that there were four sessions, each participant annotated 100 first-person images in total.

We recorded four metrics during the user study: 1) NASA TLX score 2) Annotation time 3) Pick rate and 4) Cosine similarity of picked annotation. For each image annotated, we recorded the *time* it took to annotate each image. After each session, the participant was presented with Hart and Staveland's *NASA Task Load Index (TLX)* [25].

NASA TLX is a multi-dimensional rating procedure that assesses the overall workload score based on six sub-scales: *Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration.* It is widely used for measuring subjective workload across a wide range of industries. The TLX score ranges from 0 to 100.

After all the sessions were completed, the users were asked to provide open-ended comments about their experience with the annotation process. We provided some interesting quotes from the users which are reported in section IV-B. For any VLP, we define the *pick rate* as ratio of the number of times the user selected an option from the VLP to the total number of annotations performed by the user. We also calculated the cosine similarity between the selected annotation and the ground truth to get a notion of accuracy of the selected annotation.

D. *Calculation of Similarity*

For both the fully-automated annotation and assisted-manual annotation with VLP suggestions schemes, we obtained inferred/selected activity and ground truth pairs. For each ground truth and activity pair, we pre-processed the text to have a lemmatized verb + object form by using Spacy's part of speech (POS) tagging tool. We calculated distance metrics between the inferred activity and the ground truth using the distance metrics: 1) Cosine Similarity 2) Euclidean Distance 3) Manhattan Distance, and 4) Levenshtein distance

Levenshtein distance is calculated between two strings. It is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. On the other hand, cosine similarity, euclidean distance, and the Manhattan distance are computed on embedding vectors. We operated on OpenAI's embedding vectors, called text-embedding-ada-002.

IV. Results

We calculated various metrics to compute similarity as described in section III-D. This section presents the results in two sections: fully-automated annotation and assisted-manual annotation.

A. *Fully-automated annotation*

Figure 3 shows the probability of the scores falling in different bins for each of three VLPs: BLIP2, concept detector, and concept detector + GPT-3.5, and Figure 4 shows the probabilities for GPT-4.

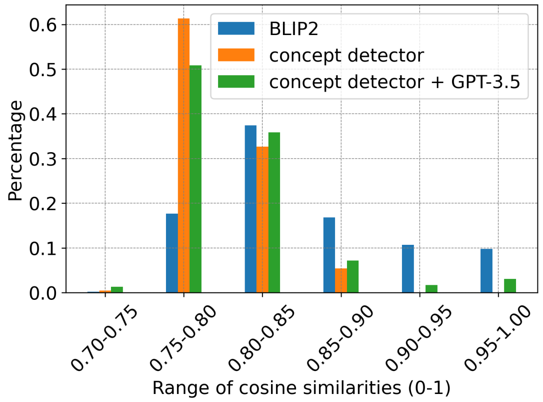


Fig. 3. Histogram of cosine similarity scores for differnt VLMs.

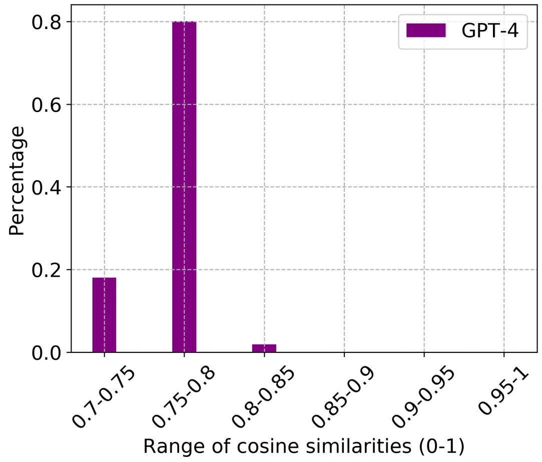


Fig. 4. Histogram of cosine similarity scores for GPT-4.

Note that the cosine similarity we used for the textual embeddings operates in a very small region (0.70 to 1). To get a better sense of what cosine similarity means, we show a table of different ranges of cosine similarities and some examples of the pairs of texts that the score represents in Table I.

Table II shows the mean and standard deviation of different types of scoring metrics for ground truth and inferred activity pairs for different types of VLPs.

B. Assisted manual annotation

We measured four metrics in the user study: 1) NASA TLX scores 2) Time 3) Pick rate 4) Accuracy measure. They are described in section III-C. Table III shows the different scores averaged across different users along with the standard deviation.

Range	(Ground Truth, Inferred Activity) pairs
0.7-0.75	(making cold food/snack, indoors), (drinking water/bottle, clean), (drinking coffee/tea, relax)
0.75-0.8	(watching tv, relax), (watching tv, furniture), (reading book, room)
0.8-0.85	(watching tv, room), (brushing teeth, mirror), (reading book, read a crossword puzzle), (brushing teeth, taking a selfie)
0.85-0.9	(using computer, people), (reading book, study), (using computer, surf the web)
0.9-0.95	(washing hands/face, washing their hands), (laundry, washing clothes), (eating food/snack, eating)
0.95-1.0	(watching tv, watch TV), (using computer, work), (reading book, reading a book)

TABLE I
EXAMPLES OF INFERRED ACTIVITY AND GROUND TRUTH PAIRS FOR DIFFERENT COSINE SIMILARITY BINS. THE PAIRS SHOWN ARE BEFORE LEMMATIZATION.

V. DISCUSSION

A. Fully-automated annotation

Figure 3 illustrates that BLIP2 exhibits a higher probability of yielding a high cosine similarity score compared to its counterparts. Approximately 10% of activities inferred by BLIP2 scored between 0.95 and 1, and between 0.90 and 0.95, indicating a perfect match between inferred activity and ground truth pairs for approximately 20% of the dataset. Moreover, BLIP2 demonstrates a lower probability of obtaining a score below 0.80 compared to its counterparts, implying that BLIP2's human activity annotations are more aligned with the ground truth.

Similarly, in Figure 4, the likelihood of obtaining a high score for GPT-4 is notably low. However, it's important to note that the images for GPT-4 were sampled at a frame rate of 0.1 frames per second, resulting in a smaller number of tested images compared to the other three VLPs.

B. Assisted-manual annotation

Table III shows that BLIP2 has one of the highest average time taken per image, one of the lowest NASA TLX score, and the highest pick rate. The unassisted session achieved the lowest NASA TLX score with average time taken per image of 10.84 seconds. This shows unassisted session is better than using any VLPs. However, we think this is attributed to the design of the annotation tool rather than the VLPs themselves.

Some users noted that it took more time to make annotations using BLIP2. This was because the text output form BLIP2 was longer and needed more time to scan through the options. One of the reasons was that different options suggested by BLIP2 were very similar to one another, which increased the time it took to scan through them all. Other reasons included high cognitive load required to read through the descriptive texts. Some users mentioned it was easier to type out the answer by themselves in the unassisted session rather than going through the descriptive texts.

VI. CONCLUSION

In this work we explored the use of state of the art LLMs, vision models, and VLMs to aid in the annotation of first-

[1] GPT-4 was tested on $1/10^{th}$ of data.

VLP	Levishtein Distance ↓	Cosine Similarity ↑	Euclidean Distance ↓	Manhattan Distance ↓
BLIP2	12.33 ± 8.52	**0.86 ± 0.07**	**0.50 ± 0.22**	**15.48 ± 6.20**
concept detector	11.78 ± 3.26	0.80 ± 0.03	0.65 ± 0.12	20.13 ± 2.36
concept detector + GPT-3.5	11.70 ± 3.67	0.81 ± 0.04	0.63 ± 0.14	19.35 ± 3.23
GPT-4 [1]	11.26 ± 1.20	0.76 ± 0.015	0.68 ± 0.027	21.30 ± 0.87

TABLE II

Similarity metrics for different VLMs.

VLP	Average time taken per image (s) ↓	NASA TLX score ↓	Pick Rate ↑	Cosine Similarity ↑
BLIP2	11.15 ± 7.86	**39.41 ± 21.35**	**0.698**	0.8512 ± 0.070347
concept detector	11.35 ± 7.38	41.48 ± 20.91	0.24	0.8461 ± 0.074309
concept detector + GPT-3.5	10.56 ± 7.55	46.9 ± 20.18	0.45	0.8607 ± 0.078150
unassisted	10.84 ± 8.69	**39.16 ± 24.10**	N/A	0.8599 ± 0.078028

TABLE III

Average time taken, NASA TLX score, pick rate, and cosine similarity associated with each assisted-manual annotation sessions.

person images with ADLs. We showed that BLIP2 performed the best among concept-detector, concept-detector + GPT-3.5, and GPT-4 for the fully-automated annotation scheme. This work represents a step forward towards AI assisted data annotation in the field of human activity recognition.

References

[1] H. Medjahed, D. Istrate, J. Boudy, and B. Dorizzi, "Human activities of daily living recognition using fuzzy logic for elderly home monitoring," in *2009 IEEE International Conference on Fuzzy Systems*. IEEE, 2009, pp. 2001–2006.

[2] N. Zouba, B. Boulay, F. Bremond, and M. Thonnat, "Monitoring activities of daily living (adls) of elderly based on 3d key human postures," in *Cognitive Vision: 4th International Workshop, ICVW 2008, Santorini, Greece, May 12, 2008, Revised Selected Papers*. Springer, 2008, pp. 37–50.

[3] L. Schrader, A. Vargas Toro, S. Konietzny, S. Rüping, B. Schäpers, M. Steinböck, C. Krewer, F. Müller, J. Güttler, and T. Bock, "Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people," *Journal of Population Ageing*, vol. 13, pp. 139–165, 2020.

[4] S. Guidetti, K. T. Nielsen, C. Von Bülow, M. S. Pilegaard, L. Klokker, and E. E. Wæhrens, "Evaluation of an intervention programme addressing ability to perform activities of daily living among persons with chronic conditions: study protocol for a feasibility trial (able)," *BMJ open*, vol. 8, no. 5, 2018.

[5] P. F. Edemekong, D. Bomgaars, S. Sukumaran, and S. B. Levy, "Activities of daily living," 2019.

[6] S. Bhattacharya, R. Adaimi, and E. Thomaz, "Leveraging sound and wrist motion to detect activities of daily living with commodity smartwatches," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–28, 2022.

[7] A. Anjum and M. U. Ilyas, "Activity recognition using smartphone sensors," in *2013 ieee 10th consumer communications and networking conference (ccnc)*. IEEE, 2013, pp. 914–919.

[8] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[9] A. Madabhushi and J. Aggarwal, "A bayesian approach to human activity recognition," in *Proceedings Second IEEE Workshop on Visual Surveillance (VS'99)(Cat. No. 98-89223)*. IEEE, 1999, pp. 25–32.

[10] R. Adaimi, H. Yong, and E. Thomaz, "Ok google, what am i doing? acoustic activity recognition bounded by conversational assistant interactions," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–24, 2021.

[11] E. Thomaz, "Activiome: A system for annotating first-person photos and multimodal activity sensor data," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2020, pp. 1–6.

[12] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[13] OpenAI, "Gpt-4 technical report," 2023.

[14] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2847–2854.

[15] E. Fast, W. McGrath, P. Rajpurkar, and M. S. Bernstein, "Augur: Mining human behaviors from fiction to power interactive systems," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 237–247.

[16] E. Jiang, K. Olson, E. Toh, A. Molina, A. Donsbach, M. Terry, and C. J. Cai, "Promptmaker: Prompt-based prototyping with large language models," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–8.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

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

[20] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3558–3568.

[21] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.

[22] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[23] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[24] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[25] N. A. Stanton, P. M. Salmon, L. A. Rafferty, G. H. Walker, C. Baber, and D. P. Jenkins, *Human factors methods: a practical guide for engineering and design*. CRC Press, 2017.