Object Retrieval-Guided Vision Language Modeling for Embodied Interaction

Constantin Patsch, Yuankai Wu, Marsil Zakour, Eckehard Steinbach
Technical University of Munich
{constantin.patsch,yuankai.wu,marsil.zakour,eckehard.steinbach}@tum.de

Abstract

Vision-language model (VLM)-based agents often struggle to name specific or unseen objects in hand-object interactions. We propose a zero-shot, real-time method that enhances VLM outputs by retrieving object features from a custom database and injecting prior knowledge into the captioning process during hand-object interactions. Our proposed approach enables users to guide an agent towards object-aware descriptions with task or job-specific objects, which are returned as speech output running in real time, as shown on GTEA and a smartphone-based user study with our collected dataset. The code is available on GitHub.

1. Introduction and Related Work

Recently, various vision-language models have emerged for image captioning [3, 6, 7, 12]. Focusing an agent on the egocentric perspective offers key advantages by capturing tasks from the user's viewpoint and reducing occlusions seen in static cameras. Advances in augmented reality highlight the rising importance of such systems. [1, 4]. Although VLMs are pretrained on large datasets, they often miss or confuse task-specific objects, leading to hallucinations—caption content not grounded in the visual input [2, 8, 14]. In order to improve the VLM performance, retrieval-augmented approaches leverage additional knowledge sources for tasks like visual question answering or image captioning [5, 13, 15].

Using object priors during inference, our approach improves object-specific consistency between visual input and captions. Unlike methods requiring retraining, our zero-shot retrieval integrates with existing VLM-based agents without fine-tuning, saving time and computational cost.

2. Methodology

Figure 1 illustrates the efficient construction of our object database and the overall architecture of the real-time, object-guided captioning agent introduced in this section. The agent generates captions upon detecting human-object interactions.

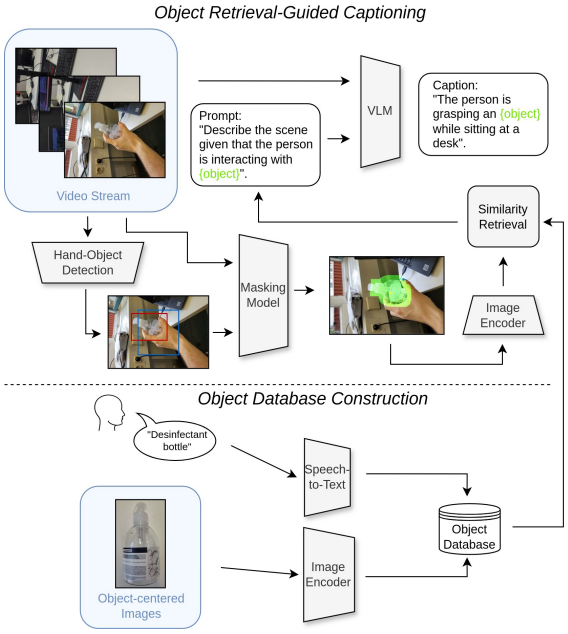

Figure 1. The user first captures object-centric images while verbally providing labels, which are transcribed via speech-to-text to build a database. Hands and objects are detected during real-time inference to identify interactions. Object mask cutouts are used for feature-based retrieval from the database.

2.1. Object-Centered Database

To enable similarity calculations with masked cutouts during inference, we build an object database focused on setup-specific, detailed labels. Objects are centered in the frame during database image capture and encoded with an image encoder. These features construct the object database $\mathcal{D}$, which is defined as $\mathcal{D} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\} \subset \mathbb{R}^d$, with $\mathbf{e}_i \in \mathbb{R}^d$ for $i = 1, \ldots, N$, where d is the feature dimension, and N denotes the number of overall objects.

During image recording, the operator verbally provides the object label, which is transcribed into text using a speech-to-text (STT) model. The resulting label is then associated with the corresponding feature vector. We use the Google Web Speech API[1] for speech transcription.

[1] https://cloud.google.com/speech-to-text

Method	B-3	B-4	R-L	M	C	S
IVL [3]	3.1	0.4	23.2	18.2	13.9	25.5
IVL [3] + O	5.2	2.1	26.8	20.4	24.7	31.1
mGPT-4 [16]	3.3	1.4	20.8	17.3	13.7	22.4
mGPT-4 [16] + O	5.2	2.6	28.1	20.0	30.9	27.6
LV [6]	0.6	0.0	27.5	16.0	16.3	18.7
LV [6] + O	**10.9**	**5.5**	**38.2**	**21.9**	**48.2**	**32.4**

Table 1. Zeroshot performance on GTEA of captioning models using BLEU-3 (B-3), BLEU-4 (B-4), ROUGE-L (R-L), METEOR (M), CIDEr(C), and SPICE(S) metrics. Method + O denotes the captioning results with our object-prior retrieval mechanism. Bold and underlined numbers indicate the best and second-best method.

Method	B-3	B-4	R-L	MET	CIDEr	SPICE
LV	30.6	23.4	53.4	22.9	34.7	19.1
LV + O	**50.4**	**44.7**	**72.3**	**44.2**	**80.2**	**61.8**

Table 2. Zeroshot performance on our user-study of Llava-Vicuna7B [6] (LV) using BLEU-3 (B-3), BLEU-4 (B-4), ROUGE-L (R-L), METEOR (MET), CIDEr(C), and SPICE(S) metrics. LV + O denotes the captioning performance with our object-prior retrieval mechanism.

2.2. Object Retrieval-Guided Captioning

The continuous video stream is processed by a hand-object detection, a masking, an image encoder (Dinov2 [9]), and a vision language model. The Faster-RCNN [10] based hand-object detection model [11] extracts the interacted object and outputs bounding box coordinates $\mathbf{h} \in \mathbb{R}^{z\times 4}$, where z indicates the number of detected interacted objects, and the second dimension represents the coordinates of two bounding box corners. The segmentation network can then be prompted based on the interacted object's bounding box coordinates. The mask cutout of the object is subsequently passed to the video encoder to output the feature query $\mathbf{q} \in \mathbb{R}^d$. Based on the Euclidean distance between the query and the database embeddings, we retrieve the closest embedding $\mathbf{e}^*$, which determines the corresponding label, resulting in the following

$$\mathbf{e}^* = \arg\min_{\mathbf{e}_i \in \mathcal{D}} \|\mathbf{q} - \mathbf{e}_i\|_2. \quad (1)$$

The successfully retrieved label is incorporated into the prompt as prior knowledge for the VLM to guide the captioning process towards the correct object. The text prompt, along with the frame where the object interaction has been detected, is passed to the VLM for inference, which outputs the resulting caption. Compared to a regular VLM agent, which might misinterpret objects in the scene, our agent can integrate these 'visually unidentified' objects in the form of retrieved objects into its output due to its semantic reasoning capabilities.

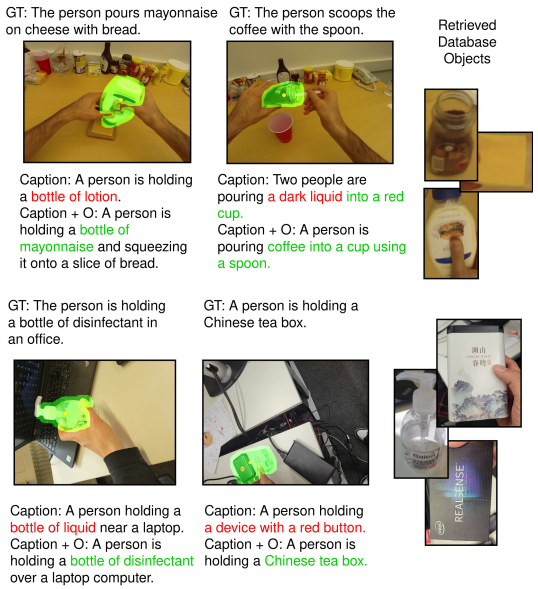

Figure 2. Qualitative examples from both datasets.

3. Evaluation

To evaluate the performance gain of our approach, we use the image captioning models MiniGPT-4 [16] (mGPT-4), InternVL [3] (IVL), and Llava [6] (LV). On both datasets, we consider the zero-shot performance without retraining any of the involved model components.

For the GTEA dataset in Table 1, we can observe an overall performance improvement regardless of the image captioning model being used when utilizing our approach compared to solely using the captioning model. In particular, the large increase in the CIDEr metric across all models indicates the improved semantic correctness. To investigate the real-world applicability, we evaluate our approach in a real-time user study, which consists of 107 samples from 7 participants captured in varying environments. We use the 4-bit quantized version of Llava-Vicuna7B [6], where the results averaged over all participants in varying environments are displayed in Table 2. In particular, the CIDEr and SPICE scores improved significantly, suggesting that the semantic object consistency improved.

4. Conclusion

In this work, we present a real-time agent for object retrieval VLM-based captions, explaining hand object interactions by a user derived from egocentric interactions, without requiring model retraining. By leveraging hand-centric object segmentation, speech-driven labeling, and an object-retrieval database, our method improves the semantic consistency of the agent's output, especially for unseen or task-specific objects. Experiments on both datasets indicated a superior performance of an object retrieval-guided agent.

5. Acknowledgement

We gratefully acknowledge the funding of the Lighthouse Initiative Geriatronics by StMWi Bayern (Project X, grant no. 5140951) and LongLeif GaPa GmbH (Project Y, grant no. 5140953).

References

[1] Fabio Arena, Mario Collotta, Giovanni Pau, and Francesco Termine. An overview of augmented reality. *Computers*, 11 (2):28, 2022. 1

[2] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1

[3] 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. 1, 2

[4] Shaveta Dargan, Shally Bansal, Munish Kumar, Ajay Mittal, and Krishan Kumar. Augmented reality: A comprehensive review. *Archives of Computational Methods in Engineering*, 30(2):1057–1080, 2023. 1

[5] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379, 2023. 1

[6] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[8] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1

[9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[11] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2

[12] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[13] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11844–11857, 2023. 1

[14] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, pages 232–248. Springer, 2024. 1

[15] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. 1

[16] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2