Look Ma, No Hands!
Agent-Environment Factorization of Egocentric Videos

Matthew Chang
UIUC
mc48@illinois.edu

Aditya Prakash
UIUC
adityap9@illinois.edu

Saurabh Gupta
UIUC
saurabhg@illinois.edu

Abstract

The analysis and use of egocentric videos for robotic tasks is made challenging by occlusion due to the hand and the visual mismatch between the human hand and a robot end-effector. In this sense, the human hand presents a nuisance. However, often hands also provide a valuable signal, e.g. *the hand pose may suggest what kind of object is being held. In this work, we propose to extract a factored representation of the scene that separates the agent (human hand) and the environment. This alleviates both occlusion and mismatch while preserving the signal, thereby easing the design of models for downstream robotics tasks. At the heart of this factorization is our proposed Video Inpainting via Diffusion Model (VIDM) that leverages both a prior on real-world images (through a large-scale pre-trained diffusion model) and the appearance of the object in earlier frames of the video (through attention). Our experiments demonstrate the effectiveness of VIDM at improving inpainting quality on egocentric videos and the power of our factored representation for numerous tasks: from object detection to learning of reward functions, policies, and affordances from videos.*

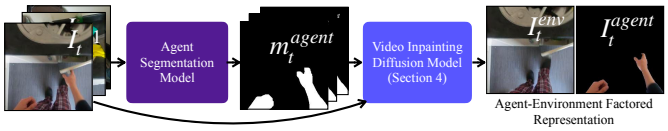

Figure 1. An agent representation I_t^{agent} is obtained using a model to segment out the agent. The environment representation I_t^{env} is obtained by inpainting out the agent from the original image using VIDM, a novel Video Inpainting Diffusion Model (Sec. 2).

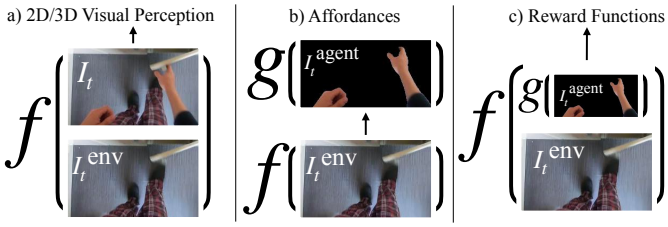

Figure 2. **Agent-Environment factored representations enable many applications.** (a) For perception tasks, I^{env} can be used in addition to the original image. (b) For affordance learning tasks, I^{env} can be used to predict relevant desirable aspects of the agent. (c) For reward learning tasks I^{agent} can be transformed into agent-agnostic formats for more effective transfer across embodiments.

1. Introduction

Observations of humans interacting with their environments, as in egocentric video datasets [2, 5], hold the potential to scale up robotic policy learning. However, a key bottleneck in these applications is the mismatch in the visual appearance of the robot and human hand, and the occlusion caused by the hand. Human hands can be a nuisance. They occlude objects and induce a domain gap between the data available for learning (egocentric videos) and the data seen by the robot at test time. However, hands also provide a valuable signal for learning. The hand pose may reveal object affordances, and the approach of the hand toward objects can define dense reward functions for learning policies.

In this work, we propose the use of a *factored agent and environment representation*. The agent representation is obtained by segmenting out the hand, while the environment representation is obtained by inpainting the hand out of the image (Fig. 1). But how do we obtain such a factored representation from raw egocentric videos? Rather than just relying on a generic prior over images, we observe that the past frames may already have revealed the true appearance of the scene occluded by the hand in the current time-step. We develop a video inpainting model that leverages both these cues. We use a large-scale pre-trained diffusion model for the former and an attention-based lookup of information from the past frames for the latter. Our method outperforms prior inpainting baselines, and improves performance on many downstream tasks.

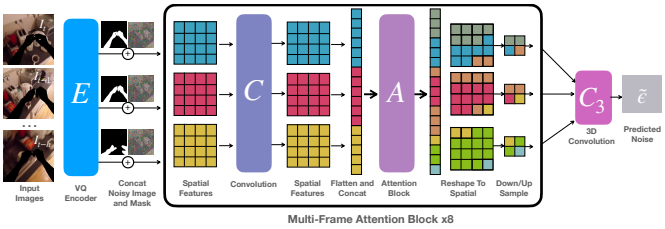

Figure 3. **Video Inpainting Diffusion Models (VIDM).** We extend pre-trained single-frame inpainting diffusion models [8] to videos. Features from context frames $(I_{t-h}, \ldots, I_{t-1})$ are introduced as additional inputs into the Attention Block A. We repeat the multi-frame attention block 8 times (4 to encode and 4 to decode) to construct the U-Net [9] that conducts 1 step of denoising. The U-Net operates in the VQ encoder latent space [8].

2. Method

Motivated by the recent success of generative models, we develop our factorization directly in the pixel space. Given an image I_t from an egocentric video, our factored representation decomposes it into I_t^{agent} and I_t^{env}. Here, I_t^{env} shows the environment without the agent, while I_t^{agent} shows the agent (Fig. 1) without the environment. This factorization enables the independent use of agent/environment information, which can be tailored to the downstream task, see Fig. 2.

Video Inpainting via Diffusion Models (VIDM) The inpainting function inpaints the mask m_t^{agent} in image I_t using information in images I_{t-h} through I_t. This is realized through a neural model that uses attention [10] to extract information from the previous frames.

We finetune a latent diffusion model [8] which has been pre-trained on the Places [11] dataset for single-frame inpainting. To incorporate information from previous frames, VIDM performs convolutions on the $h+1$ sets of spatial features in parallel but allows attention across all frames at attention blocks. This way weights from the pre-trained network can be used as-is. The diffusion targets for this model are frames extracted from Epic-Kitchens [2] and a subset of Ego4D [5] (kitchen videos). We generate masks based on human hand shapes from VISOR annotations and do not propagate loss on human hand pixels in the target images.

3. Experiments

On a reconstruction quality benchmark on held-out scenes from Epic-Kitchens [2], our model is able to improve PSNR scores to 32.26, compared to the 28.27 and 26.98 of the single-frame latent diffusion model [8] (fine-tuned on our data) and SoTA video inpainter DLformer [7] respectively (Figure 4). Similarly, VIDM improved SSIM scores to 0.956 *vs*. 0.932 and 0.922 and achieves an FID score of 10.37 *vs*. 27.50 and 51.74. On an egocentric object detection benchmark, using I^{env} combined with I is able to outperform using

a) Raw b) LDM FT [8] c) DLFormer d) VIDM (Ours)

Figure 4. Our approach (VIDM) is able to correctly steal background information from past frames (top row, oranges on the bottom right) and also correctly reconstructs the wok handle using strong object appearance priors (bottom row).

Table 1. Average recall of detections from a COCO-trained Mask RCNN [6] on active objects from VISOR [3].

Image Used	$AR_{all}@1$	$AR_{all}@5$	$AR_{all}@10$
Raw Image (*i.e.* I_t)	0.137	0.263	0.272
I_t^{env} inpainted using Latent Diffusion (finetuned)	0.154	0.262	0.271
I_t^{env} inpainted using VIDM (Ours w/o factorization)	0.163	0.268	0.277
I_t and I_t^{env} inpainted using VIDM (Ours w/ factorization)	**0.170**	**0.379**	**0.411**

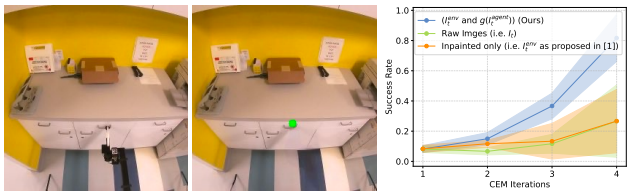

Figure 5. **Real-world experiment setup and results.** (left) Raw views from camera, (center) Agent-agnostic representation. (right) Success rate as a function of CEM iterations.

the raw image alone, or just inpainting the hand (Table 1). For predicting affordances, on the benchmark from [4], using I_{env} to predict the type of grasp mAP from 0.355 (previous SoTA masking out hand) to 0.410. In a cross-embodiment reward learning setting, we transform I_{agent} to an agent-agnostic format. We simply place a green dot at the tip of the human hands/end-effector. We learn reward functions using in-the-wild human videos from [2] for the tasks of drawer, cabinet, and refrigerator opening. Following [1] we measure the Spearman's rank correlation between the learned reward function and ground truth on pseudo-robotic trajectories (collected manually with a 2-finger gripper). Using the factored representation improves performance from 0.558 (raw videos) to 0.614. Finally, we test the efficacy of this representation in a real-world robot learning setting. We use the reward function for drawer opening above (learned entirely from human videos) and train a policy on a Stretch RE2 to open a novel drawer in the real world (Figure 5). We find that using the factored representation greatly speeds up learning compared to raw images or inpainting only.

References

[1] Matthew Chang, Arjun Gupta, and Saurabh Gupta. Learning value functions from undirected state-only experience. In *International Conference on Learning Representations*, 2022. 2

[2] 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. 1, 2

[3] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 2

[4] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[5] 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, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. 2022. 1, 2

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. 2017. 2

[7] Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, and Chen Li. Dlformer: Discrete latent transformer for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3511–3520, 2022. 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[11] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2