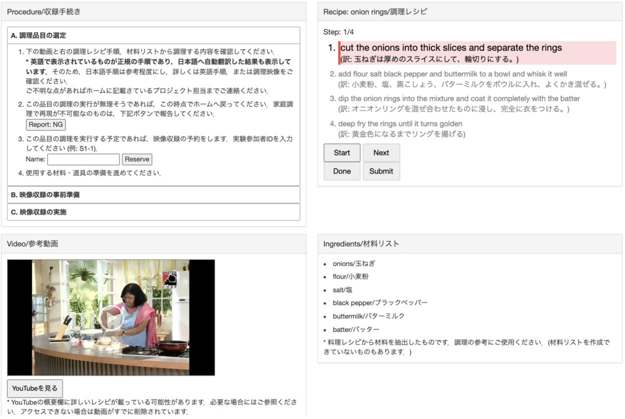
Supplementary Material
Exo2EgoDVC: Dense Video Captioning of Egocentric Procedural Activities Using Web Instructional Videos



Procedure/収録手続き

A. 調理品目の選定

1. 下の動画と右の調理レシピ手順，材料リストから調理する内容を確認してください．
＊英語で表示されているものが正規の手順であり，日本語へ自動翻訳した結果も表示しています． そのため，日本語手順は参考程度にし，詳しくは英語手順，または調理映像をご確認ください．
ご不明な点があればホームに記載されているプロジェクト担当までご連絡ください．
2. この品目の調理の実行が無理そうであれば，この時点でホームへ戻ってください．家庭調理で再現が不可能なのものは，下記ボタンで報告してください．
Report: NG
3. この品目の調理を実行する予定であれば，映像収録の予約をします．実験参加者IDを入力してください (例: S1-1).
Name: Reserve
4. 使用する材料・道具の準備を進めてください．

B. 映像収録の事前準備

C. 映像収録の実施

Recipe: onion rings/調理レシピ

Step: 1/4

1. cut the onions into thick slices and separate the rings
(訳: 玉ねぎは厚めのスライスにして、輪切りにする。)
2. add flour salt black pepper and buttermilk to a bowl and whisk it well
(訳: 小麦粉、塩、黒こしょう、バターミルクをボウルに入れ、よくかき混ぜる。)
3. dip the onion rings into the mixture and coat it completely with the batter
(訳: オニオンリングを混ぜ合わせたものに浸し、完全に衣をつける。)
4. deep fry the rings until it turns golden
(訳: 黄金色になるまでリングを揚げる)

Start Next
Done Submit

Video/参考動画

YouTubeを見る

* YouTubeの概要欄に詳しいレシピが載っている可能性があります．必要な場合にはご参照ください．アクセスできない場合は動画がすでに削除されています．

Ingredients/材料リスト

- onions/玉ねぎ
- flour/小麦粉
- salt/塩
- black pepper/ブラックペッパー
- buttermilk/バターミルク
- batter/バッター

* 料理レシピから材料を抽出したものです．調理の参考にご使用ください．(材料リストを作成できていないものもあります．)

Figure 1. **Web user interface for our recording.** Top left: Instruction of recording, Top right: Step description with the focus on the current step, Bottom left: Reference video from YouCook2 [12], Bottom right: Necessary ingredients extracted from captions.

1. Dataset Details

Video recording: We ask 44 participants to record cooking activities in their own home kitchens using a head-mounted GoPro camera. The cooking recipes are adopted from YouCook2 (YC2) [12] captions with 2,000 recipes consisting of 82 classes of recipes (*e.g.*, "BLT" is a class and multiple recipes belong to the class). Each participant chooses five recipes at will so that selected classes do not overlap and then prepares the meal by following the step descriptions written in the recipe. In total, we collect 226 videos totaling 43 hours. We also received approval for this activity data collection from an Institutional Review Board and obtained consent from participants who joined this recording.

Fig. 1 indicates our web application used for our video recording, displaying the instruction of video collection, step descriptions, reference videos from YouCook2, and necessary ingredients extracted from annotated captions. This Web interface helps participants prepare ingredients and check how to cook from the reference videos prior to recording. During recording, the highlighted step description is shown to indicate their current step and changes by manipulating the button below. The AR markers are displayed on the screen in the transition of steps, which are used to annotate temporal segments.

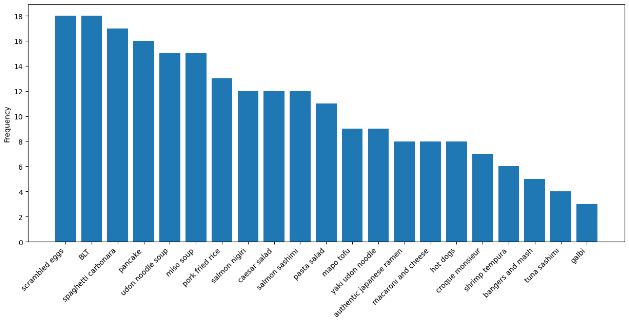


Figure 2. **Recipe distribution in EgoYC2**

To maintain the coherency of captured activities, we instruct the participants to remember the recipes beforehand, which allows them to move to the next step smoothly in the actual recording. Even though they halted midway through the recording to remember the step procedure, we treat it as acceptable behavior as it is likely to refer to the recipe on their tablets in real-life cooking.

Transfer learning setup: We use YC2 and EgoYC2 as the source and target data, respectively. We split the EgoYC2 dataset into train and evaluation sets with 151 (964) and 75 (511) videos (step descriptions), respectively. To align both datasets, we re-split the YC2 dataset according to the EgoYC2's split, where train and evaluation sets have 1,716 (13,324) and 75 (511) videos (step descriptions), respectively. The evaluation sets correspond to each other, and all the YC2 data that are not re-recorded in this work are included in the training set.

Recipe class distribution: Figs. 2 and 3 show the distribution of recipe classes for EgoYC2 and YC2. We collect 21 recipe classes out of 89 classes in YouCook2. The collected recipe list of EgoYC2 is as follows: *BLT*, *authentic*

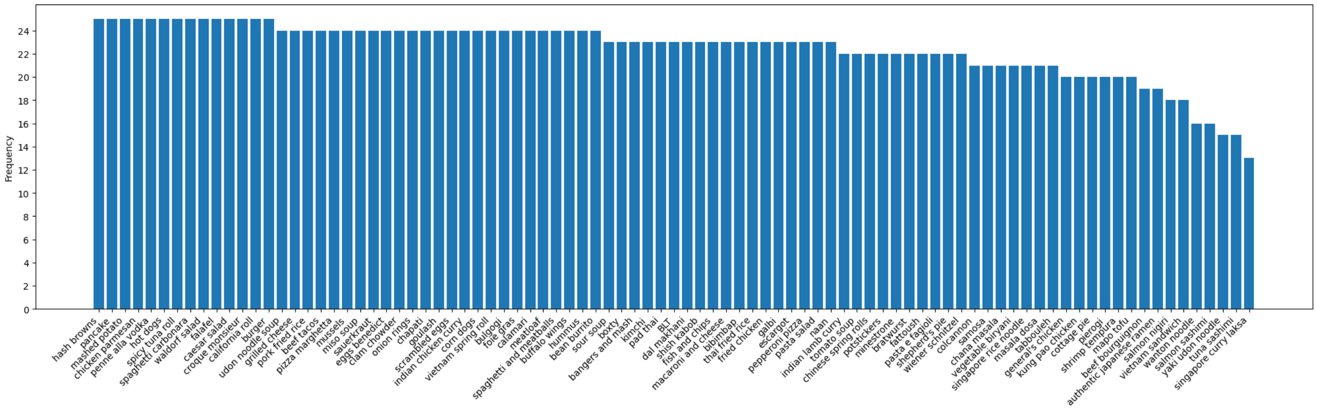


Figure 3. **Recipe distribution in YouCook2 [12]**

japanese ramen, *bangers and mash*, *caesar salad*, *croque monsieur*, *galbi*, *hot dogs*, *macaroni and cheese*, *mapo tofu*, *miso soup*, *pancake*, *pasta salad*, *pork fried rice*, *salmon nigiri*, *salmon sashimi*, *scrambled eggs*, *shrimp tempura*, *spaghetti carbonara*, *tuna sashimi*, *udon noodle soup*, *yaki udon noodle*.

Table 1. **Quantitative results in scratch training on EgoYC2.** We train models from scratch in EgoYC2 with various input feature types: raw videos (V), cropped videos (VC), and those with features of an object in hand (VC + HO).

Input	dvc_eval			SODA		
	B4	M	C	M	C	tIoU
V	0.01	3.11	12.3	3.60	5.9	30.7
VC	**0.10**	5.60	22.2	5.62	12.6	41.5
VC+HO	**0.10**	**7.34**	**29.6**	**7.04**	**17.9**	**51.4**

2. Additional Implementation Details

The architectures of the feature converter F and the view classifier C follow a two-layer one-dimensional CNN and a three-layer MLP, respectively. The video features are represented as 2,048-dimensional feature vectors for an input image. We use PDVC [8] as a baseline for dense video captioning. The PDVC uses a two-layer deformable transformer with a hidden size of 512 in the attention layers and 2,048 in the feed-forward layers. The number of event queries is set to 100 and the mini-batch size is set to 1. We use the Adam [5] optimizer with an initial learning rate of 1e-5 for the feature converter and PDVC, and 1e-4 for the view classifier. While we validate various input types for the target egocentric videos, we use the original video features generated by TSN [9] on YouCook2.

3. Additional Results

Results with egocentric data only: Tab. 1 shows the results of scratch training on EgoYC2 only. This demonstrates consistent improvement with hand-object encoding similar to the transfer setup (Tab. 3 in the main paper). With paired videos of YC2 (Rows 2-4 in Tab. 3 in the main paper), we observe significant gains over scratch performance, which confirms the effectiveness of transfer learning in limited data regimes for egocentric video captioning.

Analysis of hyperparameter settings: We set the hyperparameter of view-invariant learning (λ_{adv}) by observing the source performance of the view-invariant pre-training (VI-PT). We use the sum of two METEOR metrics (sum_METEOR) for the model selection during the pre-training. We choose the hyperparameter with the highest sum_METEOR value and set λ_{adv} as 0.1 consistently for the fine-tuning in the target domain.

We also evaluate performance in the pre-training and fine-tuning stages, according to different hyperparameters in Tab. 2. Pre-training with $\lambda_{\text{adv}} = 0.01, 0.1$ achieves relatively high performance, while fine-tuning with $\lambda_{\text{adv}} = 0.01, 1$ worsens performance than the PT+FT baseline (top row). When adding the view-invariant technique to both the pre-training and fine-tuning, we observe an improvement of captioning ability with $\lambda_{\text{adv}} = 0.01, 0.1$, as they are adapted from the pre-training models where the view-invariant learning performs well. Based on this study, our setting of $\lambda_{\text{adv}} = 0.1$ chosen from the source pre-training performs stably in the target domain with both the pre-training and fine-tuning stages.

Hand-object segmentation results: We propose a segmentation refinement scheme based on two segmentation models: EgoHOS [11] and SAM [6]. We show the segmentation results for each method in Fig. 4. The EgoHOS inference (left) often has noisy results (*e.g.*, undersegmetation on the top row and incorrect localization of long and narrow ob-

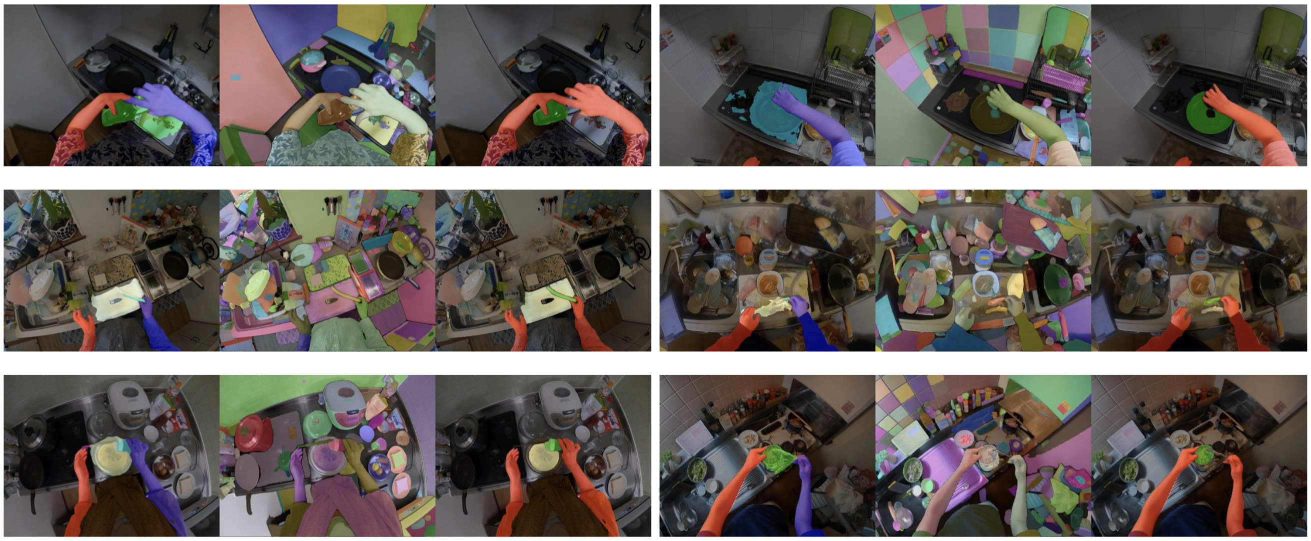


Figure 4. **Our hand-object segmentation refinement.** Each panel shows segmentation results of EgoHOS [11] (left), SAM [6] (middle), and our refined scheme (right), respectively. Since we don't use hand identity information (right/left), we show merged hand masks compared to the results of EgoHOS.

Table 2. **Analysis of hyperparameter settings.** We validate different hyperparameters for the view-invariant learning (λ_{adv}) and show the performance on the target dataset.

VI?		λ_{adv}	dvc_eval			SODA		
PT	FT		B4	M	C	M	C	tIoU
		0	1.68	8.91	52.5	8.91	37.3	59.0
✓		0.01	**2.20**	**9.45**	52.4	8.99	**39.9**	55.0
✓		0.1	2.06	9.44	**55.2**	**9.02**	39.5	**56.0**
✓		1	1.70	9.29	50.5	8.75	36.4	55.1
	✓	0.01	1.47	8.75	49.8	8.72	35.8	58.8
	✓	0.1	**1.77**	**8.89**	**53.0**	**8.91**	**37.2**	59.1
	✓	1	1.50	8.67	49.4	8.67	35.6	**59.5**
✓	✓	0.01	2.46	**9.60**	53.1	8.99	39.3	55.4
✓	✓	0.1	**2.66**	9.19	**59.0**	**9.27**	**45.2**	**58.1**
✓	✓	1	1.58	9.30	49.7	8.67	35.3	54.9

jects on the middle row). EgoHOS suffers from generalizing to novel real-life environments where diverse object types and shapes could be present. The SAM inference (middle) can segment any kind of object with higher generalization. Our refinement (right) computes the overlap between the two results and outputs the most overlapped segments from the SAM predictions. This enables us to obtain further refined results even in crowded cooking environments (*e.g*., middle row).

4. Discussions

Scripted *vs*. unscripted: Scripted and unscripted captures each have pros and cons concerning data realism and annotation quality. While unscripted videos, such as Ego4D [3] and EPIC-KITCHENS [2], reflect actual activities, these videos could include ambiguity in captions from human annotators, affecting the consistency of caption content and granularity. Such inconsistency complicates cross-domain evaluation. Our scripted approach not only aligned the content and granularity between datasets, but also instructed participants to maintain action coherency in Sec. 1, enabling natural step transitions in captured videos.

Unsupervised methods: Zero-shot generalization and unsupervised adaptation remain challenging in video captioning, as evidenced by the source-only results shown in Tab. 3 of the main paper. Our benchmark provides supervised baselines and evaluations on egocentric videos, setting the stage for future studies to develop unsupervised methods.

Overcoming recipe class gap: As shown in Sec. 1, the recipe class distribution is not perfectly aligned between YC2 and EgoYC2. In addition to focusing on the view gap addressed in the main paper, resolving category shift [1, 7, 10], the gap in the output (label) space, will be an important future challenge.

Comparison with Ego-Exo4D: We provide the comparison with a recently released Ego-Exo4D dataset [4], featuring synchronized egocentric and exocentric videos with textual annotations. In capture setups, the work follows the strong assumption of time-synchronized and calibrated scenarios, while our captures between YC2 and EgoYC2 are based on a relaxed assumption; they are not synchronized and not captured in the same environment. Regarding its text annotations [1], the knowledge of the coherency between

[1] https://docs.ego-exo4d-data.org/annotations/atomic_descriptions/

steps is not explicitly modeled, as each description is instructed to be annotated independently. In contrast, our procedural captions are intended to model the necessary steps to accomplish a target task, which inherently includes inter-step relationships in the captions.

References

[1] Z. Cao, L. Ma, M. Long, and J. Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–155, 2018. 3

[2] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling egocentric vision. *International Journal of Computer Vision (IJCV)*, 130(1):33–55, 2022. 3

[3] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M.g Xu, E. Zhongcong Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. Soo Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, Lo Torresani, M.i Yan, and J. Malik. Ego4D: Around the world in 3, 000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2022. 3

[4] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, A. Kumar, V. Baiyya, S. Bansal, B. Boote, E. Byrne, Z. Chavis, J. Chen, F. Cheng, F. Chu, S. Crane, A. Dasgupta, J. Dong, M. Escobar, C. Forigua, A. Gebreselasie, S. Haresh, J. Huang, M. M. Islam, S. Jain, R. Khirodkar, D. Kukreja, K. J. Liang, J. Liu, S. Majumder, Y. Mao, M. Martin, E. Mavroudi, T. Nagarajan, F. Ragusa, S. K. Ramakrishnan, L. Seminara, A. Somayazulu, Y. Song, S. Su, Z. Xue, E. Zhang, J. Zhang, A. Castillo, C. Chen, X. Fu, R. Furuta, C. Gonzalez, P. Gupta, J. Hu, Y. Huang, Y. Huang, W. Khoo, A. Kumar, R. Kuo, S. Lakhavani, M. Liu, M. Luo, Z. Luo, B. Meredith, A. Miller, O. Oguntola, X. Pan, P. Peng, S. Pramanick, M. Ramazanova, F. Ryan, W. Shan, K. Somasundaram, C. Song, A. Southerland, M. Tateno, H. Wang, Y. Wang, T. Yagi, M. Yan, X. Yang, Z. Yu, S. C. Zha, C. Zhao, Z. Zhao, Z. Zhu, J. Zhuo, P. Arbelaez, G. Bertasius, D. Damen, J. Engel, G. M. Farinella, A. Furnari, B. Ghanem, J. Hoffman, C. V. Jawahar, R. Newcombe, H. S. Park, J. M. Rehg, Y. Sato, M. Savva, J. Shi, M. Z. Shou, and M. Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2024. 3

[5] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 2

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 3

[7] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2507–2516, 2019. 3

[8] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7190–7198, 2018. 2

[9] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(11):2740–2755, 2019. 2

[10] Y. Xu, J. Yang, H. Cao, Z. Chen, Q. Li, and K. Mao. Partial video domain adaptation with partial adversarial temporal attentive network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9312–9321, 2021. 3

[11] L. Zhang, S. Zhou, S. Stent, and J. Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 127–145, 2022. 2, 3

[12] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7590–7598, 2018. 1, 2