Relational Space-Time Query in Long-Form Videos
Supplementary Material

Xitong Yang[1] Fu-Jen Chu[1] Matt Feiszli[1] Raghav Goyal[1,2] Lorenzo Torresani[1] Du Tran[1]

[1] Meta AI [2] University of British Columbia

1. Implementation Details and Ablation

In this section, we describe the models and the training / inference setup for individual perception modules in our explicit system. We also provide ablation study on some of the design choices.

Activity Module. We adopt the state-of-the-art activity recognition model, MViT-v2 [7], pre-trained on Kinetics-400 [4] as the backbone model. The model takes 32 frames as input with a temporal stride 9 for ReST-ADL (∼9-second window) and 1 for ReST-Ego4D (∼16-second window). We train the model for 10 epochs using SGD with a cosine learning schedule [9], where the initial learning rate is set to be 0.1 and 0.01 for the classifier and the pretrained backbone, respectively.

Detection Module. We obtain object detection results using the publicly available object detectors and follow their own inference setup. We use the top-20 detection results for each frame because the query performance saturates even with more boxes. We experiment with three different object detectors:

- **Mask R-CNN RPN** [2], where we take the results of the region proposal network (RPN) of Mask R-CNN in order to achieve sufficient recall of the objects whose category labels do not fall into the COCO taxonomy [8]. We use the model with a ResNet-101-FPN backbone and the model checkpoint is available on Detectron2 [13].
- **OLN** [6], which involves a classification-free object localization network and is designed for open-world object proposals.
- **GGN [11] (default)**, which leverages pairwise affinities to generate additional training annotations as diversified as pixel diversities allowed for open-world instance segmentation. We generate object detection results using the tightest bounding boxes of the segmentation outputs.

We compare the performance of the three detectors in Tab. 4. Note that we use the ground truth annotation for the activity module to remove the side affect of noisy activity prediction (we cannot use the ground truth annotation for the embedding and interaction modules because the bounding boxes are different). We observe that GNN [11] achieves the best result with stronger open-world object detection capacity and we use it as our default model for the detection module.

Embedding Module. We extract feature embeddings of image crops (obtained by the annotation or object detection) using a pre-trained ResNet-50 model [3]. As the pre-trained weights of a model plays a central role to the quality of feature embedding, we experiment with models trained with three different settings:

- **ImageNet-pretrained**, a standard model pre-trained on ImageNet [1] for image classification.
- **ImageNet+Object365**, an ImageNet-pretrained model followed with fine-tuning on the large-scale Object-365 dataset [10] proposed for object detection.
- **Inst.-contrastive (default)**, a model trained with an additional supervised contrastive loss [5] to enforce discrimination across different instances. The model is finetuned from the ImageNet+Object365 model.

The results using the three embedding models are shown in Tab. 4. We use ground truth annotation for all other modules to remove the impact of noisy predictions. We observe that adding an additional loss for instance discrimination significant improves the embedding performance because the ImageNet-pretrained features are trained for classifying object categories instead of instances. We therefore use this model as the default embedding module. Note that the system achieve 100% recall for object-query because this task only takes into account the predictions of the activity and interaction module, which are ground truth labels in this ablation study.

		Activity-query		Object-query		Time-query	
		R@1x	Rej.	R@3x	Rej.	R@3x	Rej.
Detection	Mask R-CNN RPN	69.39	74.15	44.71	100	70.03	80.70
	OLN	66.71	77.61	43.02	100	65.67	83.07
	GNN (default)	74.41	73.69	45.78	100	73.34	77.55
Embedding	ImageNet-pretrained	71.2	94.1	100	100	71.7	96.8
	ImageNet+Object365	69.4	95.3	100	100	68.2	97.1
	Inst.-contrastive (default)	78.5	92.6	100	100	75.6	95.1

Table 4. Ablation on the detection and embedding module on split-1 of ReST-ADL.

Interaction Module. Taking an image crop of a detected object from the detection module, we use a ResNet-50 model to predict whether the object is being interacted ("active") or not. To train the model, we use the ground truth object annotations in the training set, which include both the bounding boxes of the objects and their interaction labels. We apply simple augmentations to the ground truth bounding boxes, such as horizontal flipping and box jittering, to increase sample diversity, and take a ratio of 1:4 for positive and negative samples per image for training. The model backbone is initialized with ImageNet-pretrained weights and we train the model for 10 epochs using SGD with a cosine learning schedule [9], where the initial learning rate is set to be 0.1 for the classification layer and 0.01 for the backbone layers.

We note that the model is only trained on ground truth object boxes, which means that the model output $P_{active}(o_i)$ should be interpreted as the probability of *an object* being interacted with (*i.e.*, conditioned on detecting an object with the bounding box). That's why the final output of the interaction module is formulated as: $P_{inter}(b_i) = P_{obj}(b_i) \times P_{active}(o_i)$, which takes into account the likelihood a detected bounding box is actually an object.

2. More Results on ReST-ADL

Preliminary results with modified TubeDETR. Except for the explicit system, we also report preliminary results using a modified version of TubeDETR [14], which is a state-of-the-art Transformer-based model for spatio-temporal grounding. We make minimum modifications to the original model and therefore we only adapt it for object-query which has a similar output format as spatio-temporal grounding. Specifically, we adapt the model to take activity labels as input, instead of the original language input. A learnable embedding layer is added to encode activity labels to query embeddings. We also add an additional output head for predicting confidence scores of the predicted tubes, which are then used for computing recalls in evaluation.

	activity-query		object-query		time-query	
	R@1x	R@3x	R@1x	R@3x	R@1x	R@3x
Explicit	52.33	70.60	11.22	24.44	40.22	42.27
Random	0.02	0.07	0.02	0.05	0.07	0.17
TubeDETR	–	–	8.33	21.7	–	–
Obj-Trans	41.53	63.93	8.21	16.05	25.33	26.99

Table 5. Comparison the modularized system with the adapted TubeDETR [14] and object-centric Transformer [12] on split-1 of the ReST-ADL benchmark. Results are reported with IoU=0.3.

We compare our modified TubeDETR with the explicit system in Tab. 5. Since TubeDETR is originally proposed for spatio-temporal grounding and is trained on positive queries only, we report results on positive queries and remove the query rejection step in the explicit system for fair comparison. We observe that TubeDETR achieves worse results than our explicit system, especially when the query window becomes larger. This suggests new model designs specific to the new query tasks are needed and we believe our benchmarks will inspire the development of next generations of video models.

Preliminary results with modified object-centric Transformer. We developed an object-centric Transformer (Obj-Trans) similar to the one in [12]. Obj-Trans takes object detection results (bounding boxes) and embeddings of the detected objects and video clips as input, performs self/cross-attention and predicts answers directly. We show the results on ReST-ADL split-1 with "short" window size in Tab. 5. Obj-Trans achieves inferior results compared to our explicit baseline, possibly due to the absence of explicit human-object interaction information and its ineffectiveness in modeling minutes-long videos. We also include the results of random baseline in the table.

	Object-query		Time-query	
	R@1x	R@3x	R@1x	R@3x
Short	7.56 (±0.97)	14.74 (±2.46)	23.60 (±8.52)	26.11 (±9.91)
Medium	7.29 (±1.04)	14.78 (±2.49)	24.36 (±7.28)	25.86 (±7.99)
Long	7.38 (±1.13)	15.66 (±2.85)	21.44 (±5.06)	23.65 (±5.63)

Table 6. Baseline results of the explicit system on ReST-ADL. We provide the results with IoU=0.5 which are not reported in the main paper due to space limit.

More results of the explicit system. As mentioned in the main paper, we provide the results with IoU=0.5 for object-query and time-query on ReST-ADL in Tab. 6.

3. More Details on ReST Benchmarks

In order to specify how challenging each query (both positive and negative) could be, we keep track of the following three attributes for each generated query.

Difficulty. This attribute is used to indicate whether the joint understanding of activities and human-object interactions is necessary to answer this query. The formal definition for different query types are described as follows.

- Activity-query: for positive queries, if the query object occurs but is not interacted with in any activity segments within the query window, the attribute is set to 1; for negative queries, if the query object occurs but is not interacted with within the query window, the attribute is set to 1; for all other cases the attribute is set to 0.
- Object-query: for positive queries, if two or more segments with different activity labels occur in the query window, the attribute is set to 1; for negative queries, if the query activity occurs in the original video (outside of the query window), the attribute is set to 1; for all other cases the attribute is set to 0.
- Time-query: for positive queries, if there exists an activity segment with the same activity label as the query activity but not involving the query object as "active" objects, the attribute is set to 1; for negative queries, if either the query activity or the query object occurs in the query window, the attribute is set to 1; for all other cases the attribute is set to 0.

Visibility. Intuitively, an object / activity is particularly challenging to detect and localize if it only occurs in a short duration. This attribute is used to indicate how long the query activity / object occurs within the query window.

Identical. In egocentric videos, it is possible to observe multiple object instances with almost the same appearance. It is particularly challenging for a vision model to differentiate these object instances and predict the correct answers. We use this attribute to indicate whether a query involves such visually identical objects as the query object.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1

[5] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 1

[6] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 2022. 1

[7] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 1

[9] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1, 2

[10] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1

[11] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. *CVPR*, 2022. 1

[12] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1884–1894, June 2021. 2

[13] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[14] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 2