

# Multimodal Perception and Interaction with Transformers

Francois Yvon, Camille Guinaudeau, Marc Evrard  
Univ Paris Saclay (LISN CNRS)

James L. Crowley  
Grenoble Institut Polytechnique, Univ Grenoble Alpes

# Multimodal Perception with Transformers

## Plan:

### Transformers in Natural Language Processing (François Yvon, 1h30)

- Text classification and language models
- The Transformer architecture
- Encoder-Decoder architecture for Neural Machine translation

### Transformers in Speech (Marc Evrard, 45 minutes)

- Speech representation
- Speech Transformer
- Speech Recognition Transformers

### Transformers in Vision (Camille Guinaudeau, 45 minutes)

- From CNN to Vision Transformer
- Vision Transformers
- Multi-Modal Transformer and Temporal encoding

### Conclusions (James Crowley, 15 minutes)

- Research Challenges and Data Sets

# Research Challenges and Data Sets

- Ego-Centric Perception: Kitchen activities
  - EPIC-Kitchens 55 (2018)
  - EPIC-Kitchens 100 (2021)
- Visual Question and Answering (VQA)
- Vision and Language Navigation (VLN)
- Social-IQ

# Egocentric Perception of Non-scripted Daily activity



## Egocentric Perception of Non-scripted Daily activity

Data Sets: Epic Kitchens <https://epic-kitchens.github.io/2021>

## Key References

Damen, D., et al. (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 720-736). (Also appeared in PAMI 2020).

Damen, D., et al., EPIC-KITCHENS-55 - 2020 Challenges Report, CVPR 2019.

Damen, D., et al., EPIC-KITCHENS-200 - Rescaling Egocentric Vision, 2021

# EPIC: Egocentric Perception of Non-scripted Daily activity



**EPIC Kitchens-55:** a large-scale egocentric video benchmark recorded by 32 participants in their native kitchen environments. Videos depict **nonscripted** daily activities accompanied by Audio Narration. 55 hours of video (11.5M frames). Ground truth labeling for 39.6K action segments and 454.2K object bounding boxes. Narrations (speech and text) added post-recording by participants

Damen, D., et al. (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 720-736)


# EPIC-55 Research Challenges: Object Detection Challenge



**Object Detection:** 125 Visual object classes and 331 Noun classes, grouped into grouped into 19 super categories

**Evaluation Metrics:** mean average precision (mAP) metric from PASCAL VOC, using IoU thresholds of 0.05, 0.5 and 0.75 similar to MS-COCO

# EPIC-55 Research Challenges: Action Recognition Challenge

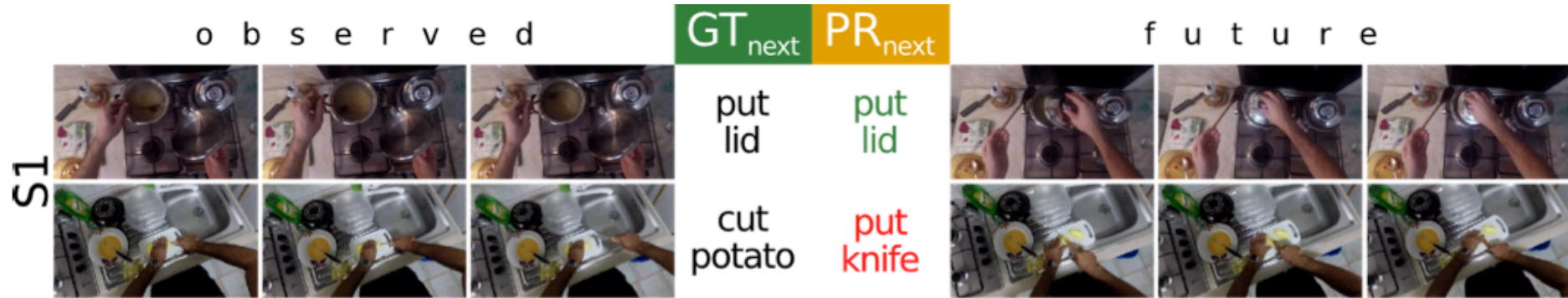
S1		GT <sub>action</sub> mix pasta	PR <sub>action</sub> mix pasta		GT <sub>action</sub> dry hand	PR <sub>action</sub> dry hand		GT <sub>action</sub> wash spoon	PR <sub>action</sub> wash bowl		GT <sub>action</sub> fill kettle	PR <sub>action</sub> wash tap
		GT <sub>action</sub> wash cup	PR <sub>action</sub> wash cup		GT <sub>action</sub> cut tomato	PR <sub>action</sub> cut tomato		GT <sub>action</sub> turn heat	PR <sub>action</sub> adjust heat		GT <sub>action</sub> cut vegetable	PR <sub>action</sub> put knife

**Action Recognition Challenge:** Given an action segment, classify the segment into its action class, where classes are defined (verb, noun), with 26 verbs and 70 noun classes.

## Evaluation Metrics:

- (1) Aggregate metrics: top- 1 and top-5 accuracy for cv, cn and (cv,cn) – we refer to these as ‘verb’, ‘noun’ and ‘action’.
- (2) Per-class metric: precision and recall for classes with more than 100 samples in training

# EPIC-55 Research Challenges: Action Anticipation Challenge



**Action Anticipation Challenge:** Given an action segment, predict the action class by observing the video segment *preceding* the action.

## Evaluation Metrics:

- (1) Aggregate metrics: top- 1 and top-5 accuracy for cv, cn and (cv,cn) – we refer to these as ‘verb’, ‘noun’ and ‘action’.
- (2) Per-class metric: precision and recall for classes with more than 100 samples in training



# EPIC-55 Results: CVPR June 2019

D. Damen, E. Kazakos, W. Price, J. Ma, H. Doughty, A. Furnari, G. M. Farinella,  
 EPIC-KITCHENS-55- 2020 Challenges Report, at CVPR 2019, Los Angeles, June 2019

## Object Detection Challenge:

Rank	Team	Submissions		Few Shot Classes (%)			Many Shot Classes (%)			All Classes (%)			
		Entries	Date	IoU >0.05	IoU >0.5	IoU >0.75	IoU >0.05	IoU >0.5	IoU >0.75	IoU >0.05	IoU >0.5 ▲	IoU >0.75	
SI	1	<b>hutom</b>	51	05/30/20	<b>47.44</b>	<b>35.75</b>	<b>14.32</b>	<b>60.77</b>	<b>46.50</b>	<b>15.60</b>	<b>58.27</b>	<b>44.48</b>	<b>15.36</b>
	2	<b>DHARI</b>	27	05/29/20	<b>54.98</b>	<b>32.40</b>	<b>14.55</b>	<b>68.74</b>	<b>43.88</b>	<b>15.38</b>	<b>66.15</b>	<b>41.72</b>	<b>15.23</b>
	3	<b>FB AI</b>	69	04/01/20	<b>26.55</b>	<b>19.01</b>	<b>8.22</b>	<b>58.44</b>	<b>46.22</b>	<b>15.61</b>	<b>52.44</b>	<b>41.10</b>	<b>14.22</b>
	4	CVG Lab Uni Bonn	23	05/12/20	39.36	26.66	7.89	53.50	41.28	12.46	50.84	38.53	11.60
	5	VCL	61	05/18/20	33.23	23.16	5.00	50.78	37.91	9.79	47.48	35.13	8.89
	6	[2] (baseline)	-	09/03/18	30.63	20.28	2.75	49.55	37.39	9.82	45.99	34.18	8.49

## Action Recognition Challenge:

Rank	Team	Submissions		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall			
		Entries	Date	VERB	NOUN	ACTION ▲	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	
SI	1	<b>UTS-Baidu</b>	14	05/28/20	<b>70.41</b>	<b>52.85</b>	<b>42.57</b>	<b>90.78</b>	<b>76.62</b>	<b>63.55</b>	<b>60.44</b>	<b>47.11</b>	<b>24.94</b>	<b>45.82</b>	<b>50.02</b>	<b>26.93</b>
	2	<b>NUS-CVML</b>	18	05/29/20	<b>63.23</b>	<b>46.45</b>	<b>41.59</b>	<b>87.50</b>	<b>70.49</b>	<b>64.11</b>	<b>51.54</b>	<b>42.09</b>	<b>25.37</b>	<b>40.99</b>	<b>42.69</b>	<b>26.98</b>
		<b>UTS-Baidu</b>	16	05/30/19	<b>69.80</b>	<b>52.27</b>	<b>41.37</b>	<b>90.95</b>	<b>76.71</b>	<b>63.59</b>	<b>63.55</b>	<b>46.86</b>	<b>25.13</b>	<b>46.94</b>	<b>49.17</b>	<b>26.39</b>
	3	<b>SAIC-Cambridge</b>	34	05/27/20	<b>69.43</b>	<b>49.71</b>	<b>40.00</b>	<b>91.23</b>	<b>73.18</b>	<b>60.53</b>	<b>60.01</b>	<b>45.74</b>	<b>24.95</b>	<b>47.40</b>	<b>46.78</b>	<b>25.27</b>
	3	<b>FBK-HuPBA</b>	50	05/29/20	<b>68.68</b>	<b>49.35</b>	<b>40.00</b>	<b>90.97</b>	<b>72.45</b>	<b>60.23</b>	<b>60.63</b>	<b>45.45</b>	<b>21.82</b>	<b>47.19</b>	<b>45.84</b>	<b>24.34</b>
4	<b>GT-WISC-MPI</b>	12	01/30/20	68.51	49.96	38.75	89.33	72.30	58.99	51.04	44.00	23.70	43.70	47.32	23.92	
5	<b>G-Blend</b>	14	05/28/20	66.67	48.48	37.12	88.90	71.36	56.21	51.86	41.26	20.97	44.33	44.92	21.48	

## Action Anticipation Challenge

Rank	Team	Submissions		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall			
		Entries	Date	VERB	NOUN	ACTION ▲	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	
SI	1	<b>NUS-CVML</b>	18	05/29/20	<b>37.87</b>	<b>24.10</b>	<b>16.64</b>	<b>79.74</b>	<b>53.98</b>	<b>36.06</b>	<b>36.41</b>	<b>25.20</b>	<b>9.64</b>	<b>15.67</b>	<b>22.01</b>	<b>10.05</b>
	2	<b>VI-I2R</b>	28	05/23/20	<b>36.72</b>	<b>24.61</b>	<b>16.02</b>	<b>80.39</b>	<b>54.90</b>	<b>37.11</b>	<b>31.03</b>	<b>26.02</b>	<b>8.68</b>	<b>15.28</b>	<b>22.03</b>	<b>8.70</b>
	3	<b>Ego-OMG</b>	16	05/26/20	<b>32.20</b>	<b>24.90</b>	<b>16.02</b>	<b>77.42</b>	<b>50.24</b>	<b>34.53</b>	<b>14.92</b>	<b>23.25</b>	<b>4.03</b>	<b>15.48</b>	<b>19.16</b>	<b>5.36</b>
	4	<b>UNIPD-UNICT</b>	16	05/26/20	36.73	24.26	15.67	79.87	53.76	36.31	35.86	25.16	7.42	14.12	21.30	7.62
	5	<b>GT-WISC-MPI</b>	20	11/12/19	36.25	23.83	15.42	79.15	51.98	34.29	24.90	24.03	6.93	15.31	21.91	7.88

# EPIC Kitchens-100

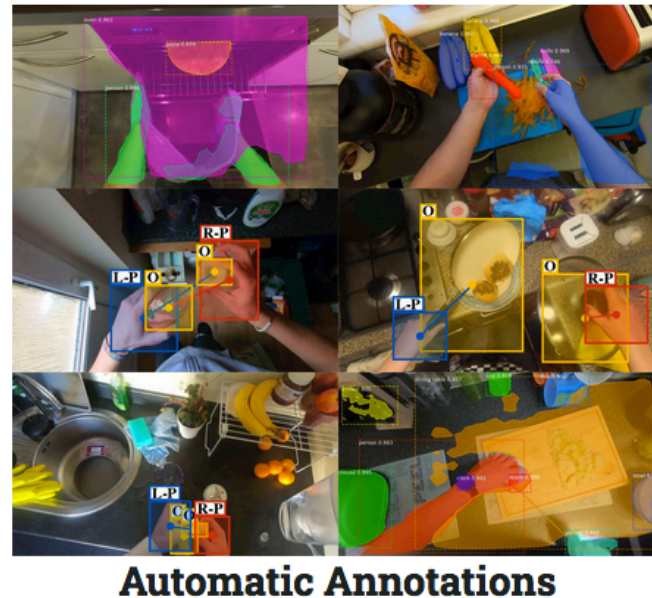
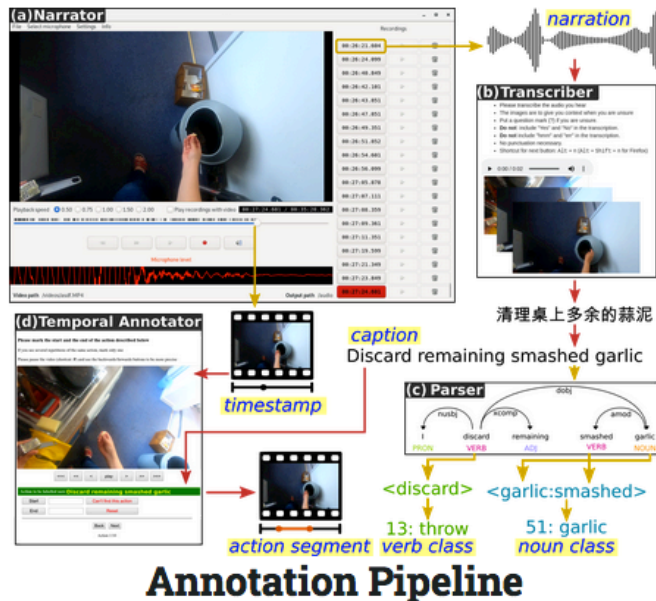


**EPIC Kitchens-100:** 100 hours, 20M frames, 90K actions in 700 variable-length videos, capturing long-term unscripted activities in 45 environments, using head-mounted cameras. Annotated with denser and more complete annotations of fine-grained actions (54% more actions per minute, +128% more action segments)

Ground truth labeling for 39.6K action segments and 454.2K object bounding boxes. Narrations (speech and text) added post-recording by participants

Damen, D., et al. (2021). ReScaling egocentric vision: The epic-kitchens dataset. IJCV 2021

# EPIC Kitchens-100 Data Collection



45 participants in 4 cities collected video over 2 to 4 days using GoPro Hero7 black. Videos are narrated off-line in native language using "Pause and talk" to provide synchronized audio-visual recording. Narratives are translated English with Amazon Mechanical Turk, spell checked and transformed to verbs/nouns.


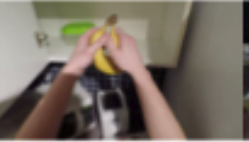



<https://epic-kitchens.github.io/2021>

# EPIC-Kitchens-100: Five research challenges

Five research challenges

- 1) Action Recognition
- 2) Action Detection
- 3) Action Anticipation
- 4) Cross-modal retrieval
- 5) Domain adaptation

# EPIC-100 Research Challenges: Action Recognition Challenge

							
GT	dry hand	slice chilli	clean pan	take banana	open bag	squeeze lemon	apply spreads
TSN	} dry hand }	} slice chilli }	} clean pan }	take corn	put bag	insert lemon	put bread
TRN				take corn	take bag	take kiwi	put bread
TBN				take potato	put bag	squeeze lemon	put fork
TSM				take bag	take bag	squeeze kiwi	put plate
SlowFast				take juicer	take bag	squeeze kiwi	put yoghurt

**Action Recognition Challenge:** Given an action segment, classify the segment into its action class. Data contains 53 action classes with 128 instances

## Evaluation Metrics:

- (1) Aggregate metrics: top-1 and top-5 accuracy for cv, cn and (cv,cn) – we refer to these as ‘verb’, ‘noun’ and ‘action’.
- (2) Per-class metric: precision and recall for classes with more than 100 samples in training

# EPIC-100: Action Detection Challenge



## Action Detection Challenge

**Action Detection:** Given a video, detect Action instances with Start Time, Stop time, verb, noun and action class.

**Data:** 100 hours of audio-video recording, 4053 action classes, 89977 action instances, average 128.5 actions/video and 53.2 classes/video, 28% overlap

**Evaluation Metrics:** mean average precision (mAP) metric. Temporal segments are matched with Intersection over Union from 0.1 to 0.5

# EPIC-100: Action Anticipation Challenge

Observed						
Future						
GT	get tomato	put glass	open bin	roll dough	flip fish	turn-on microwave
RU-LSTM (Top 5)	get tomato	put glass	open bag drawer box cupboard cloth	knead take put dough squeeze mix	take spoon mix oil open onion put courgette pour lid	take cupboard open button close alarm press kettle set spoon

**Action Anticipation Challenge:** Given an action segment, predict the (Verb, Noun, Action) classes by observing a segment preceding the action segment by 1 second.

## Evaluation Metrics:

- (1) Aggregate metrics: top- 1 and top-5 accuracy for (Verb, Noun, Action) classes
- (2) Per-class metric: precision and recall for (Verb, Noun, Action) classes

# EPIC-100: Cross-Modal Action Retrieval Challenge



**Cross-Modal Action Retrieval Challenge:** Given an query segment, rank segments in a gallery set that are semantically relevant

**Text to video:** Query is text caption, gallery contains videos

**Video to text:** Query is video: gallery contains text captions.

**Evaluation Metrics:**

(1) Normalized Discounted Cumulative Gain (nDCG). Given query  $x_r$ , and a gallery  $C_r$

$$nDCG(x_i, C_r) = \frac{DCG(x_i, C_r)}{IDCG(x_i, C_r)}$$

Where:

$$DCG(x_i, C_r) = \sum_{j=1}^{|C_r|} \frac{\mathcal{R}(x_i, c_j)}{\log(j+1)}$$

$$IDCG(x_i, C_r) = DCG(x_i, \hat{C}_r)$$



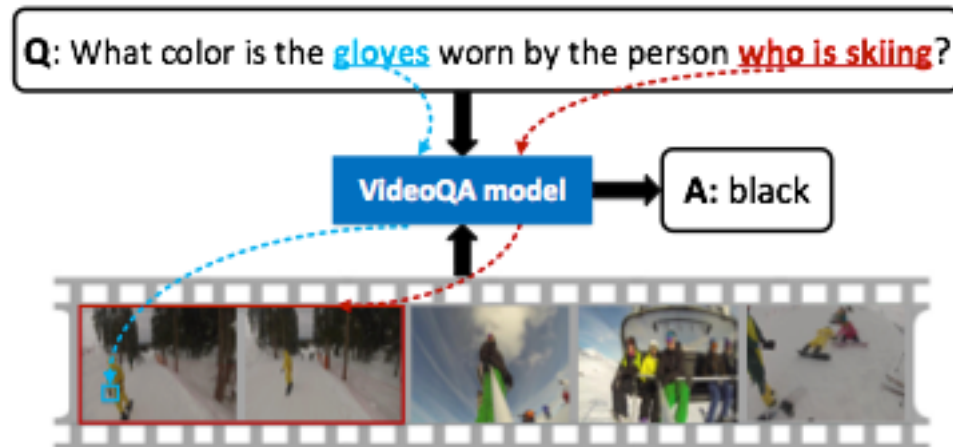
# EPIC-100: Domain Adaptation Challenge

							
GT	dry hand	slice chilli	clean pan	take banana	open bag	squeeze lemon	apply spreads
TSN	{ dry hand }	{ slice chilli }	{ clean pan }	take corn	put bag	insert lemon	put bread
TRN				take corn	take bag	take kiwi	put bread
TBN				take potato	put bag	squeeze lemon	put fork
TSM				take bag	take bag	squeeze kiwi	put plate
SlowFast				take juicer	take bag	squeeze kiwi	put yoghurt

**Unsupervised Domain Adaptation Challenge:** Given a labeled source domain (kitchen) from 2018 learn to adapt to an unlabeled target domain from 2020. Source and Targets are from the 16 participants who provided recordings from both 2018 and 2020.

**Evaluation Metrics:** Same as with action recognition - Given an action segment, classify the segment into its action class, where classes are defined (verb, noun), with 26 verbs and 70 noun classes.

# Visual Question and Answering



**VisualQA Problem:** Generate natural language answer to a question about a video

Image from Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. (2019, July). Activitynet-QA: A dataset for understanding complex web videos via question answering. AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9127-9134).

# VQA Datasets:

Datasets	Video source	QA pairs generation	QA tasks	# Videos	# QA pairs	Average video length
MSVD-QA (Xu et al. 2017)	MSVD	Automatic	OE	1,970	50,505	10s
MSRVTT-QA (Xu et al. 2017)	MSRVTT	Automatic	OE	10,000	243,680	15s
TGIF-QA (Jang et al. 2017)	TGIF	Automatic & Human	OE & MC	56,720	103,919	3s
MovieQA (Tapaswi et al. 2016)	Movies	Human	MC	6,771	6,462	200s
Video-QA (Zeng et al. 2017)	Jukinmedia	Automatic	OE	18,100	174,775	45s
ActivityNet-QA (Ours)	ActivityNet	Human	OE	5,800	58,000	180s

**VisualQA Problem:** Generate natural language answer to a question about a video  
As the videos are collected, Question-Answer Pairs are generated for each video.

Most data sets exploit narrative descriptions or captions provided with the video.  
Activity net uses crowdsourcing to generate QA pairs.

Table from Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. (2019, July). Activitynet-QA: A dataset for understanding complex web videos via question answering. AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9127-9134).

# HowTo100M: 100 Million Narrated Video Clips



Dataset of narrated instructional videos where content creators teach complex tasks with an explicit intention of explaining the visual content on screen.

**Includes 136M video clips with captions** sourced from 1.2M Youtube videos (15 years of video) showing **23k activities** from domains such as cooking, hand crafting, personal care, gardening or fitness. Each video is associated with a narration available as subtitles automatically downloaded from Youtube.

**Challenges:** text based action localization and text-to-video retrieval

Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *IEEE International Conference on Computer Vision, CVPR 2019*, pp. 2630-2640.

# HowToVQA69M: Question-answer triplets for HowTo100M



**Speech:** Fold them in half again, to make a triangle.

**Generated Question:** How do you make a triangle?

➔ **Generated Answer:** Fold them in half again



**Speech:** The sound is amazing on this piano.

**Generated Question:** What kind of instrument is the sound of?

➔ **Generated Answer:** Piano

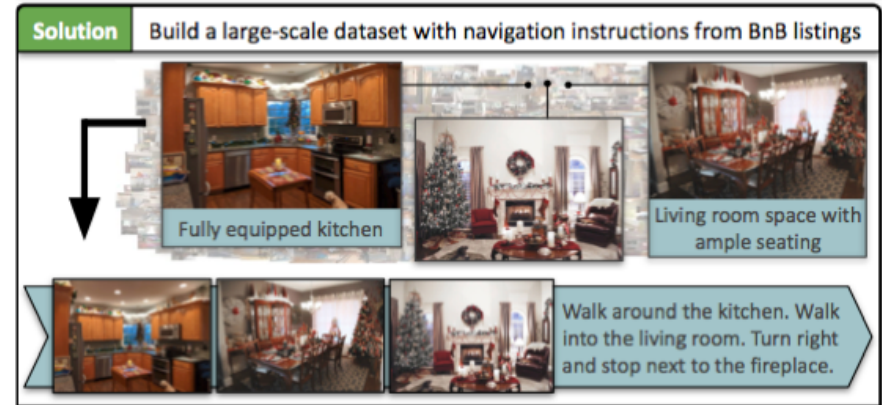
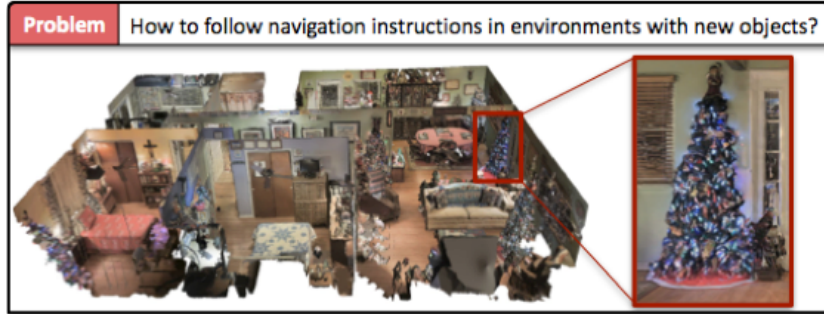
A large dataset with 69 Million video-question-answer triplets generated using transformers to automatically generate questions for videos in HowTo100M.

**Approach:** Use transformers trained on a question-answering text to generate a non-scripted questions and corresponding open-vocabulary answers from text using the HowTo100M data set.

**Challenge:** Given a video and a question, Generate a natural language answer.

Yang, A., Miech, A., Sivic, J., Laptev, I. and Schmid, C., 2021. Just ask: Learning to answer questions from millions of narrated videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1686-1697).

# Vision and Language Navigation



**Task:** Enable a Robot to navigate in realistic environments using natural language instructions.

**Dataset:** BnB: image-caption (IC) pairs from listings from online rental marketplace, with 1.4M indoor images and 0.7M captions. Static image-caption pairs are transformed into visual paths and navigation-like instructions

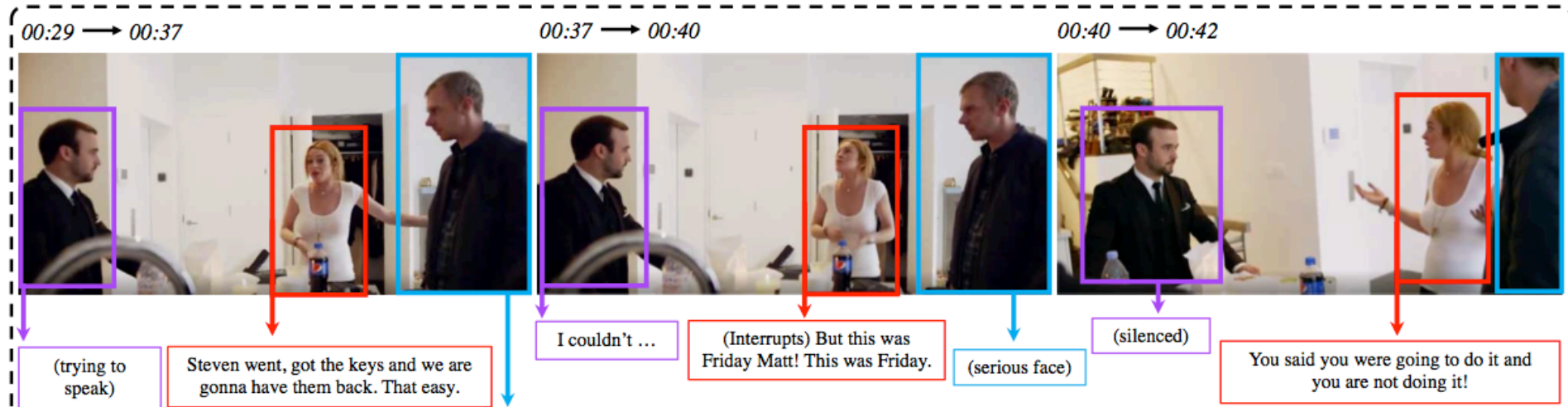
**Challenges:**

**Path Discrimination.** Choose the base path from a set of candidates

**Path Generation:** sequentially predict actions

Guhur, P.L., Tapaswi, M., Chen, S., Laptev, I. and Schmid, C., 2021. Airbert: In-domain Pretraining for Vision-and-Language Navigation. In IEEE International Conference on Computer Vision, ICCV2021, pp. 1634-1643, Oct 2021

# Social-IQ



**DataSet:** 1,250 natural in-the-wild Annotated videos, with 7, 500 questions and 52, 500 correct and incorrect answers, in 3 classes: (easy, intermediate, advanced)

**Challenge:** generate answer for question from video

**Example:**

Q1: How is the discussion between the woman and the man in the white shirt ?

A3: They are having a romantic conversation. <easy>

Zadeh, A., Chan, M., Liang, P.P., Tong, E. and Morency, L.P., Social-IQ: A question answering benchmark for artificial social intelligence. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR2019*, pp. 8807-8817, June 2019

<https://github.com/A2Zadeh/Social-IQ>

# Multimodal Perception with Transformers

Plan:

**Transformers in Natural Language Processing** (François Yvon, 1h30)

- Text classification and language models
- The Transformer architecture
- Encoder-Decoder architecture for Neural Machine translation

**Transformers in Speech** (Marc Evrard, 45 minutes)

- Speech representation
- Speech Transformer
- Speech Recognition Transformers

**Transformers in Vision** (Camille Guinaudeau, 45 minutes)

- From CNN to Vision Transformer
- Vision Transformers
- Multi-Modal Transformer and Temporal encoding

**Conclusions** (James Crowley, 15 minutes)

- Research Challenges and Data Sets



# Multimodal Perception and Interaction with Transformers

Francois Yvon, Camille Guinaudeau, Marc Evrard  
Univ Paris Saclay (LISN CNRS)

James L. Crowley  
Grenoble Institut Polytechnique, Univ Grenoble Alpes