



*Inria*

# Learning from Narrated Videos

Jean-Baptiste Alayrac  
[jbalayrac.com](http://jbalayrac.com)

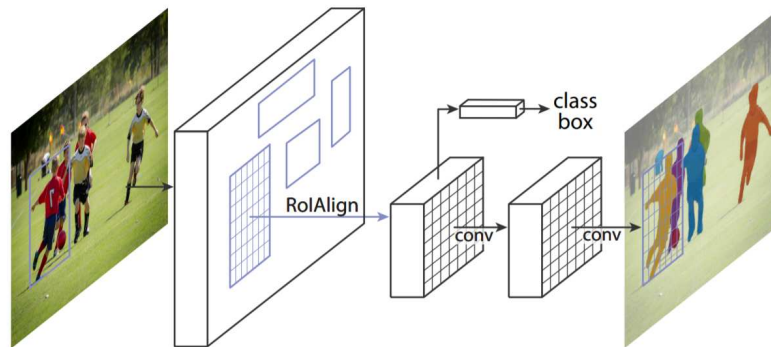
3rd Workshop on YouTube-8M  
Large-Scale Video Understanding  
28/10/2019

# Success of Supervised Learning



## Pose estimation

[Towards Accurate Multi-person Pose Estimation in the Wild, *Papandreou, Zhu, Kanazawa, Toshev, Tompson, Bregler and Murphy*, CVPR17]



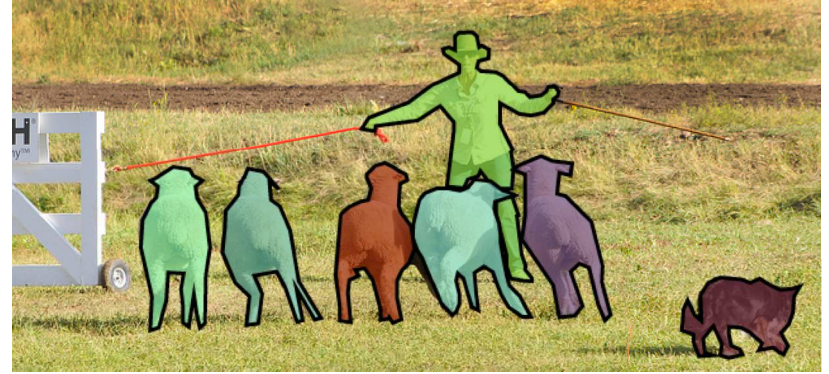
## Image Segmentation

[Mask R-CNN, *He, Gkioxari, Dollár, and Girshick*, ICCV17]

# Issues of Supervised Learning

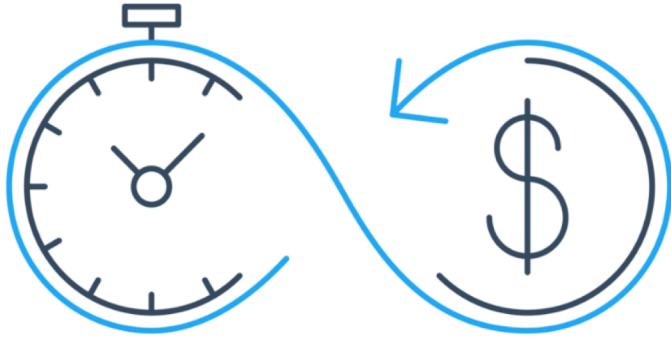


Labels are expensive



Agreement: definition? granularity?

# Issues of Supervised Learning



Labels are expensive



Even more problematic for videos!

# Weakly supervised learning

*Use weaker and readily available source of supervision*



**#dog #bike**

Training info: **image level label**

*[Barnard et al'03], [Joulin et al'10], [Deselaers et al'12], [Song et al'14], [Wang et al'14], [Cinbis et al'15], [Oquab et al'15], [Kantorov et al'16], [Bilen and Vedaldi'16]...*

# Weakly supervised learning

*Use weaker and readily available source of supervision*



Training info: **video narration (ASR)**

*[Alayrac et al'16/17], [Malmaud et al, 15], [Sener et al'15], [Huang et al'17], [Zhou et al'17], [Kuehne et al'17],... ..*

# What are instructional videos?



- Depict complex, **goal-oriented** human activities (e.g. *how to change a car tire*)
- **Multimodal:** video and language
- Can be obtained at **scale** (e.g. on YouTube), without manual annotation

# Glossary

- **Tasks:** a complex human activity involving interacting with objects and/or performing multiple small actions.

*Example: "make pancakes", "change a car tire", ...*

- **Steps:** an atomic action composing a task.

*Example: "crack egg", "remove tire", ...*



# Overview of the talk

## 1) Leveraging the structure of narrated videos

### Making Meringue

*Pour* egg

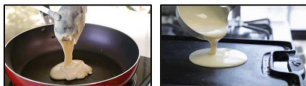
Add sugar

Whisk *mixture*



### Making Pancakes

*Pour* mixture



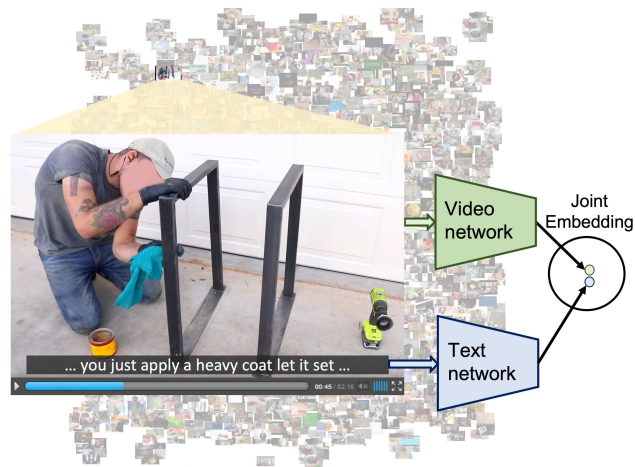
### Making Lemonade

*Pour* water



[Cross-task weakly supervised learning from instructional videos](#), Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, Josef Sivic, *CVPR2019*

## 2) Leveraging the scale of narrated videos



[HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips](#), Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic, *ICCV2019*

# 1

## Cross-Task Weakly Supervised Learning from Instructional Videos, *CVPR19*



D. Zhukov\*



D. Fouhey



G. Cinbis



I. Laptev



J. Sivic

**How much can we  
leverage the structure in  
narrated videos and what  
can we get from that?**

# What do we mean by structure here?

## Task: Make Meringue



**Structure within task**

# What do we mean by structure here?

## Task: Make Meringue



**Structure within task**

## Task: Making Pancakes

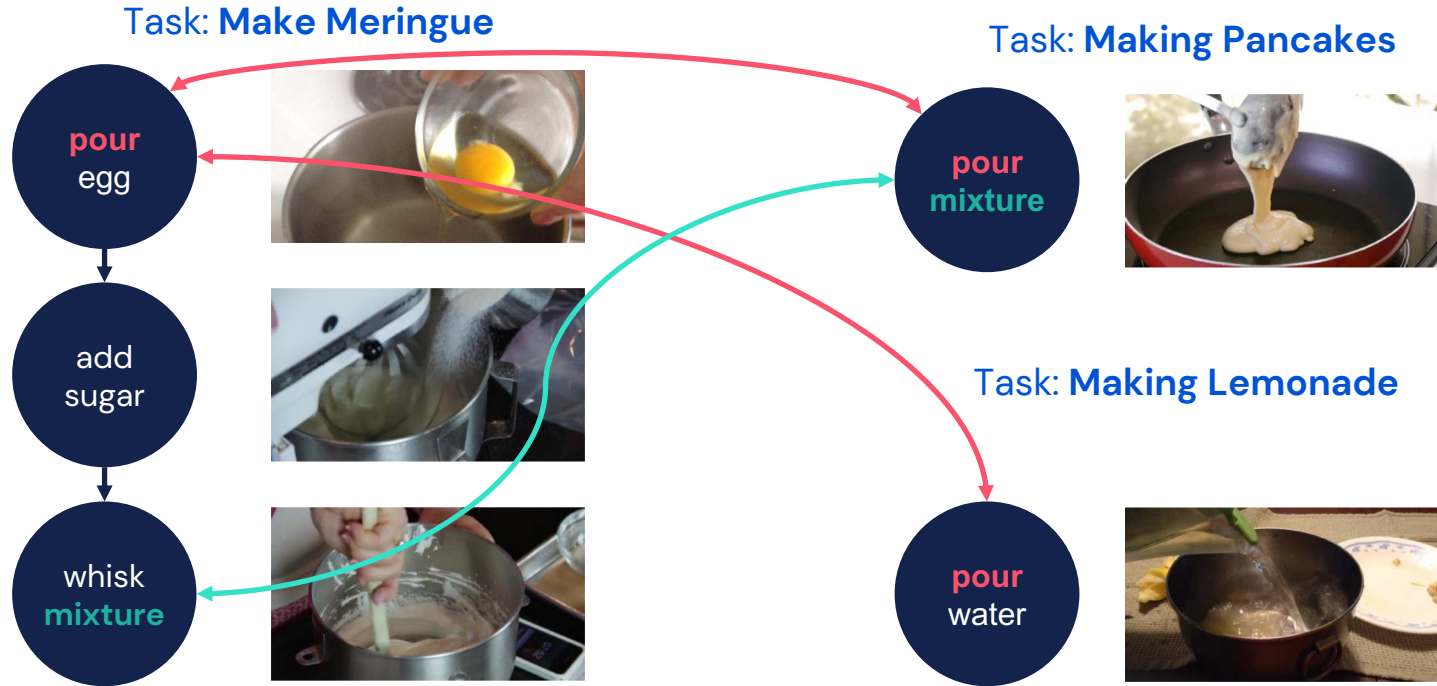


## Task: Making Lemonade



**Structure across task**

# What do we mean by structure here?



Structure within task

Structure across task

# Weakly supervised learning of step visual models

## Input

- A set of **tasks**  
ex: "Make Meringue", "Make Pancakes",  
"Change a car tire", ...

- For each task, a **list of steps**:

*Make Pancake*

- 1) pour egg
- 2) add milk
- 3) whisk mixture

- For each task, a **set of narrated videos**:



*"... now we pour the egg..."*

## Output

*What and When?*

- A visual **classifier** for each step

- **Localize** each steps in all videos



# Our assumptions

- **Temporal ordering.** Steps always occur in the order given by the list of steps.
- **At least once.** We assume that for each video, each step occurs once.
- **Video and narration.** Correlation between video and language.



# The approach

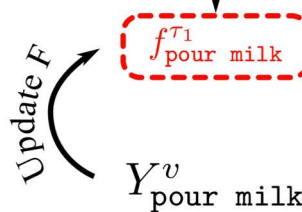
TL;DR: We jointly (i) **learn step classifiers over pretrained visual features** and (ii) **localize where the steps happen** in the video.

**Formally:** This is done by an alternate optimization between the parameters of the step classifier (**F**) and the localization variable (**Y**) under specific constraints that reflects our assumptions.

Video  $v \in \tau_1$

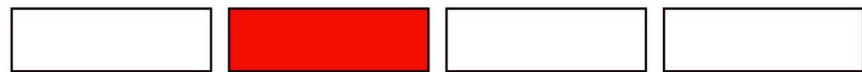


Alternate  
Optimization



**pour milk?**

Constraints



Time

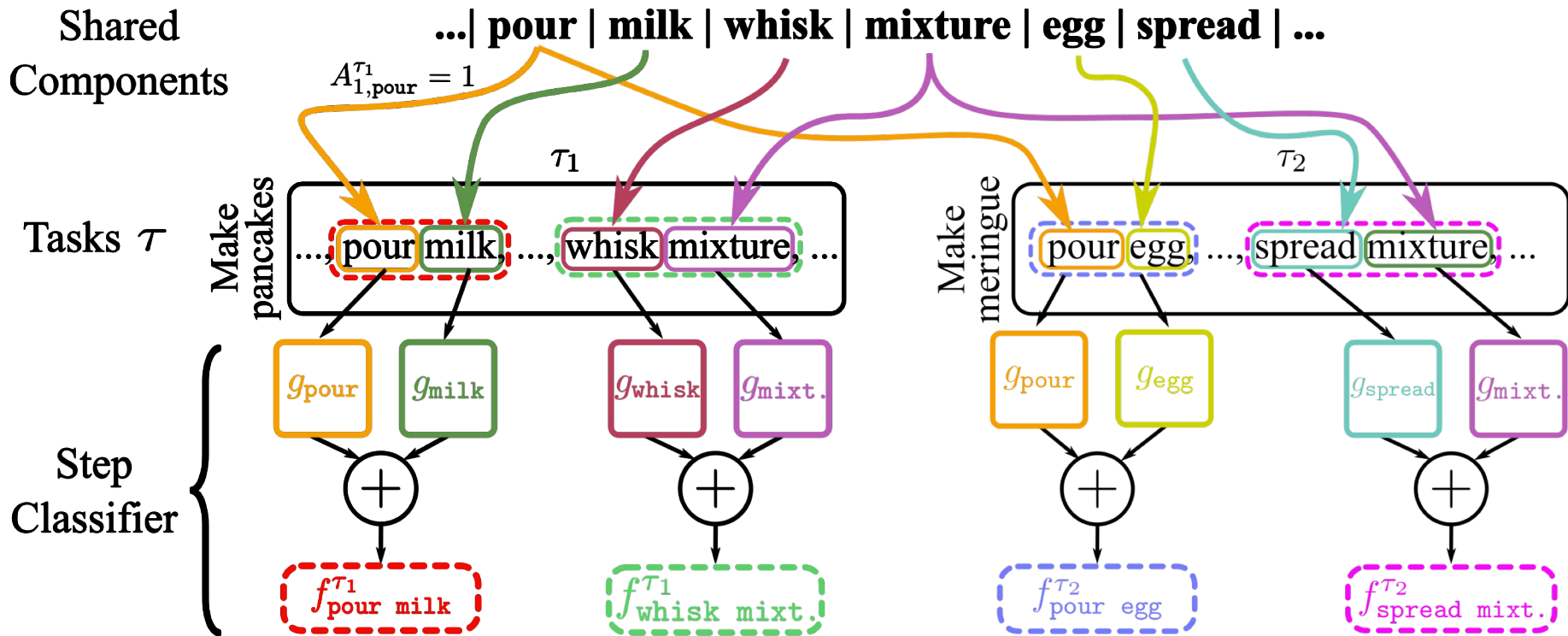
Narration "[...] now I'm gonna **pour some milk** into the bowl and [...]"

$$\min_{Y \in \mathcal{C}, F \in \mathcal{F}} \sum_{\tau} \sum_{v \in \mathcal{V}(\tau)} h(X^v, Y^v; F)$$

Tasks

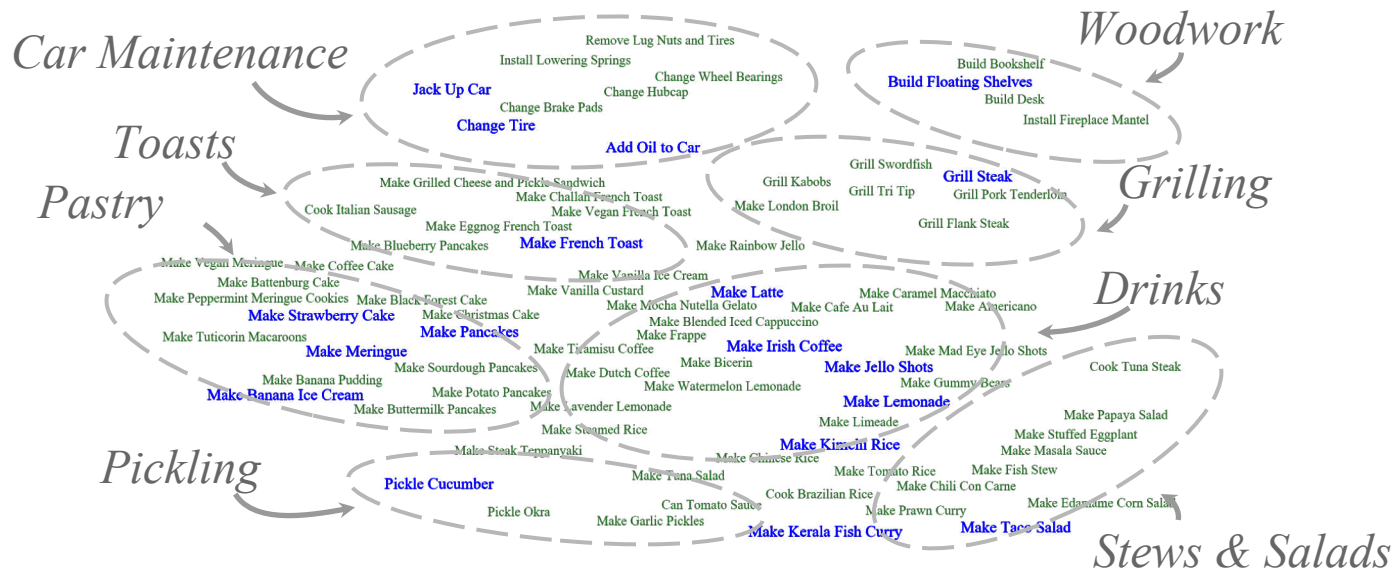
Videos

# Component based model for steps



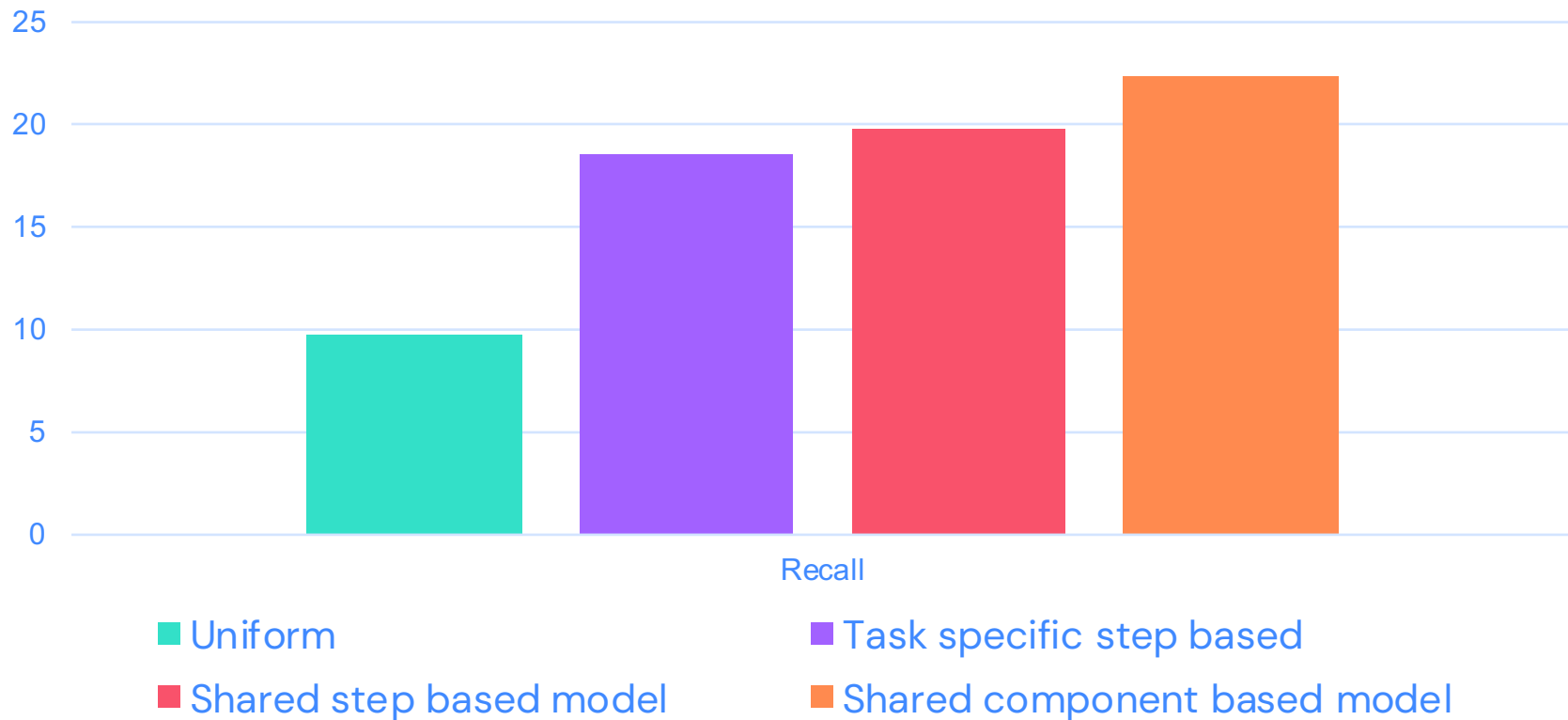
# The CrossTask dataset

- Designed to assess the benefit of sharing knowledge across tasks:
- 18 primary tasks, 2750 videos with full temporal annotation
- 65 related tasks, 1950 videos without annotation
- Diverse set of tasks: Car maintenance, gardening, cooking, home repair



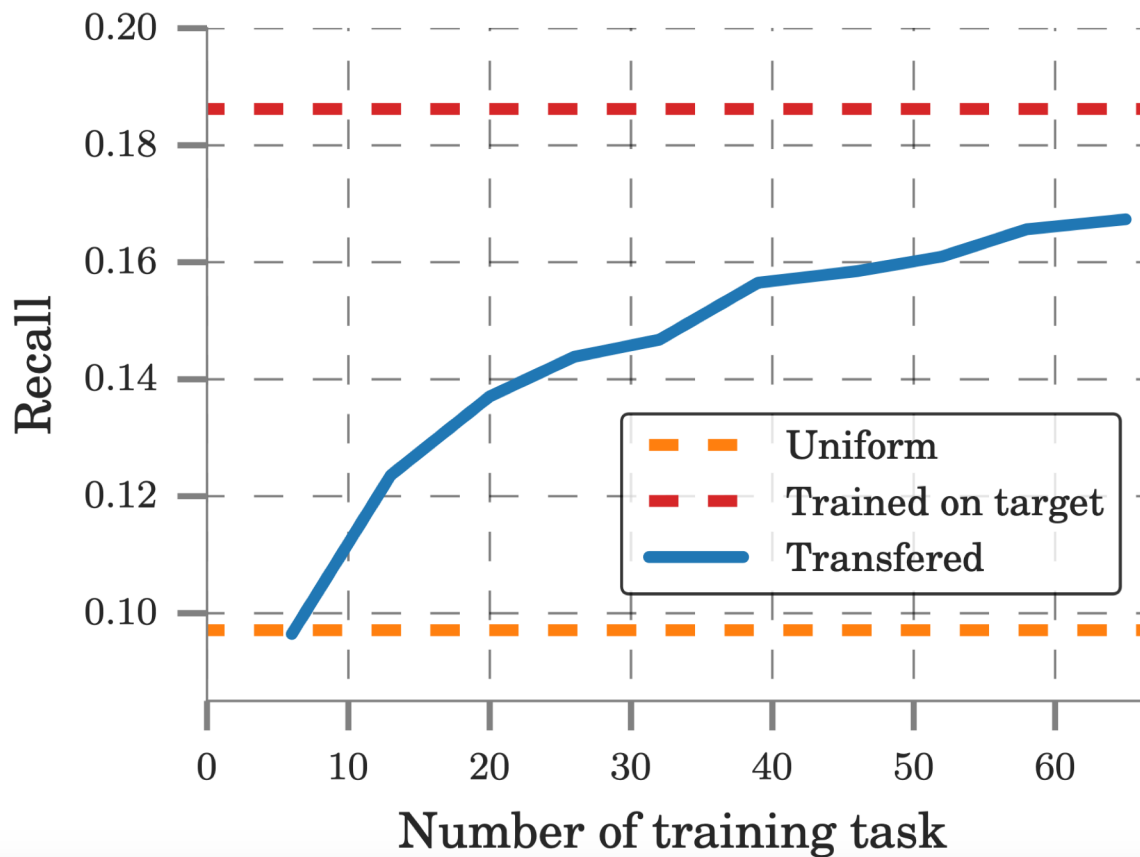
# Results: gains from sharing

## Localization metric (recall) on CrossTask



## Results: novel task transfer

We train only on related task and transfer to the unseen primary task.



# Qualitative results

## Source Steps From Related Tasks

Cut Steak



Cut Tomato



Add  
Tomato



Add Cherries  
to Cake



## Unseen Task: Make French Strawberry Cake

Cut Strawberry



Add Strawberry  
To Cake



# 2

## HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, *ICCV19*



A. Miech\*



D. Zhukov\*



M. Tapaswi



I. Laptev



J. Sivic

\*equal contribution

**How much can we scale  
Instructional Video  
dataset and what can we  
get from that?**



# The HowTo100M dataset in numbers

- 23K human **tasks** scrapped from WikiHow
- 1.2M unique YouTube **videos** (duration 15 years)
- 136M **clips** with **narration** transcribed into text (mostly from ASR)
- Larger than any existing manually annotated captioning dataset

Dataset	Clips	Captions	Videos	Duration	Source	Year
Charades [48]	10k	16k	10,000	82h	Home	2016
MSR-VTT [58]	10k	200k	7,180	40h	Youtube	2016
YouCook2 [67]	14k	14k	2,000	176h	Youtube	2018
EPIC-KITCHENS [7]	40k	40k	432	55h	Home	2018
DiDeMo [15]	27k	41k	10,464	87h	Flickr	2017
M-VAD [52]	49k	56k	92	84h	Movies	2015
MPII-MD [43]	69k	68k	94	41h	Movies	2015
ANet Captions [26]	100k	100k	20,000	849h	Youtube	2017
TGIF [27]	102k	126k	102,068	103h	Tumblr	2016
LSMDC [44]	128k	128k	200	150h	Movies	2017
How2 [45]	185k	185k	13,168	298h	Youtube	2018
<b>HowTo100M</b>	<b>136M</b>	<b>136M</b>	<b>1.221M</b>	<b>134,472h</b>	Youtube	2019

# How to collect HowTo100M?

## Step 1 : WikiHow

wikiHow to do anything...

HELP US EXPLORE LOG IN MESSAGES

We're trying to help everyone on the planet learn how to do anything. Join us.

How to Make No Bake C

Join wikiHow

- Facebook
- Google
- Civic
- Email

Have an account? [Log In](#)

Random Article Write An Article

wikiHow Worldwide

wikiHow in other languages:  
English, español, Čeština, Deutsch, Français, हिन्दी, Bahasa Indonesia, Italiano, 日本語, Nederlands, Português, Русский, العربية, ไทย, Türkçe, Tiếng Việt, 한국어, 中文. You can also help start a new version of wikiHow in your language.

**Result: list of 130k tasks**

...

How to be healthy

How to cook quinoa in a Rice Cooker

How to Sew an Apron

How to Break a Chain

How to April Fool your Girlfriend

...

**Annotation cost: 0**

# How to collect HowTo100M?

## Step 2 : Filter task by verb to keep visual tasks

Result: list of 23k tasks

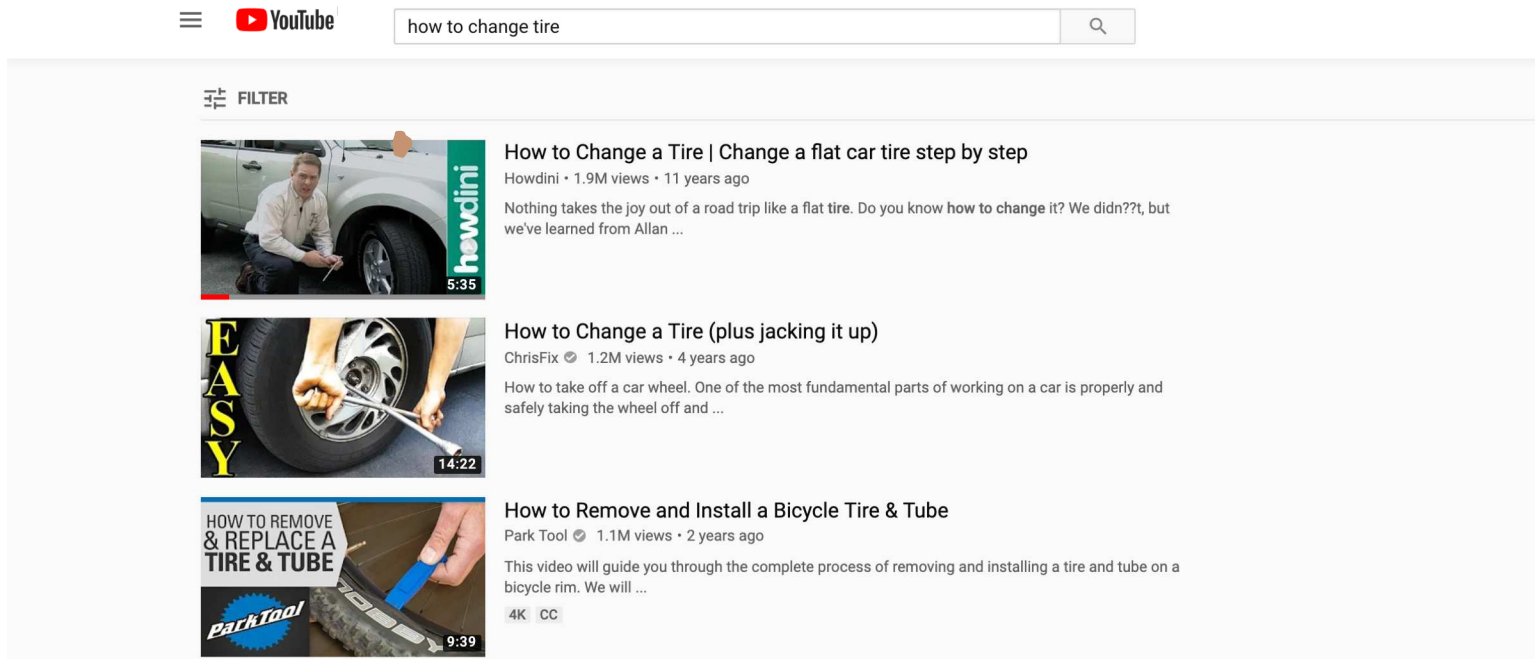
- ...
- ~~How to **Be** healthy~~
- ✓ How to **Cook** quinoa in a Rice Cooker
  - ✓ How to **Sew** an Apron
  - ✓ How to **Break** a Chain
- ~~How to April **Feel** your Girlfriend~~
- ...

Annotation cost: 8 hours for Antoine

# How to collect HowTo100M?

## Step 3 : YouTube queries for videos with captions

Result: 1.2 M unique videos



The screenshot shows a YouTube search interface. At the top, the YouTube logo is on the left, and a search bar contains the text 'how to change tire' with a magnifying glass icon on the right. Below the search bar, there is a 'FILTER' button. The search results are listed below, each with a video thumbnail, title, channel name, view count, and upload date. The first result is 'How to Change a Tire | Change a flat car tire step by step' by 'Howdini' with 1.9M views and 11 years ago. The second result is 'How to Change a Tire (plus jacking it up)' by 'ChrisFix' with 1.2M views and 4 years ago. The third result is 'How to Remove and Install a Bicycle Tire & Tube' by 'Park Tool' with 1.1M views and 2 years ago.

YouTube

how to change tire

FILTER

**How to Change a Tire | Change a flat car tire step by step**  
Howdini • 1.9M views • 11 years ago  
Nothing takes the joy out of a road trip like a flat tire. Do you know how to change it? We didn't, but we've learned from Allan ...

**How to Change a Tire (plus jacking it up)**  
ChrisFix • 1.2M views • 4 years ago  
How to take off a car wheel. One of the most fundamental parts of working on a car is properly and safely taking the wheel off and ...

**How to Remove and Install a Bicycle Tire & Tube**  
Park Tool • 1.1M views • 2 years ago  
This video will guide you through the complete process of removing and installing a tire and tube on a bicycle rim. We will ...


4K CC

Annotation cost: 0

# How to collect HowTo100M?

## Step 4 : Create clips

Result: 136M narrated clips



you want to do is put on a flat spare  
then you'll remove the tire careful

Transcript

- 00:18 down then give it a couple good wrap
- 00:21 make sure it's full of air worst thing
- 00:24 you want to do is put on a flat spare
- 00:25 then you'll remove the tire careful this**
- 00:28 can be quite heavy well we have our
- 00:30 spare tire in our Jack we've set our
- 00:32 reflective warning signal out
- 00:34 we've also dropped the wheel on the
- 00:36 opposite side of the flat tire so we're
- 00:38 ready to start changing it the first
- 00:40 thing we'll do is jack the vehicle up
- 00:42 you'll have to loosen the jack a little

English (auto-generated)

How to Change a Tire | Change a flat car tire step by step

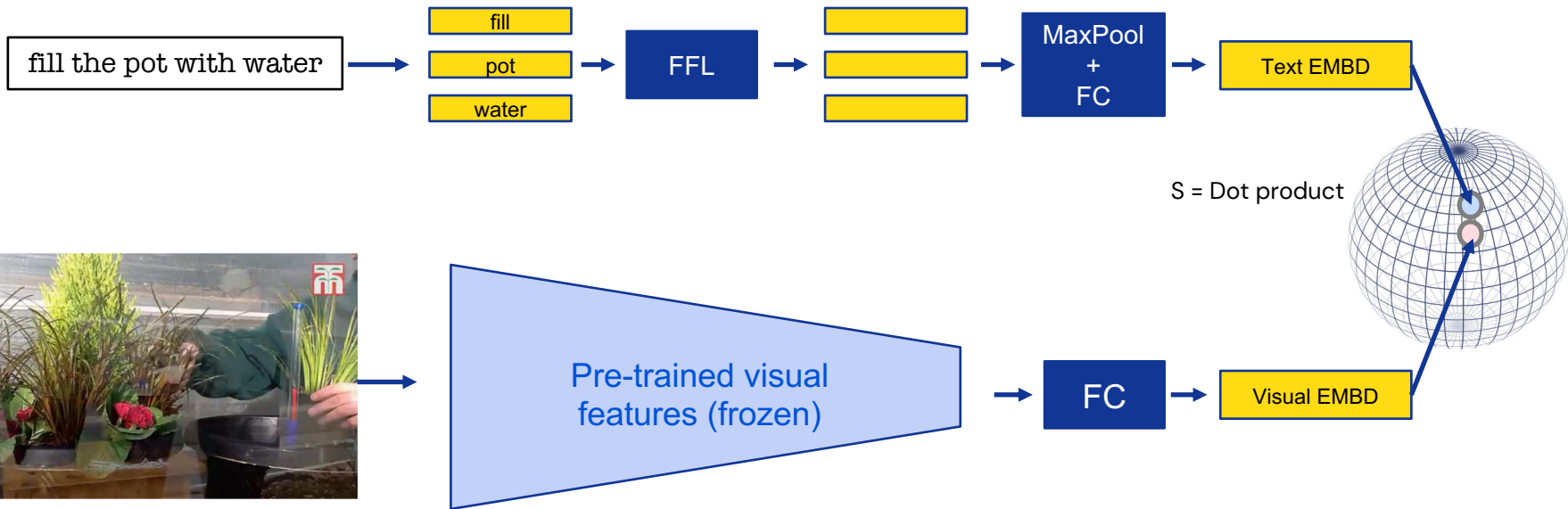
2,216,300 views • Jan 31, 2008

👍 14K 🗨️ 799 ➦ SHARE ⚙️ SAVE ⋮

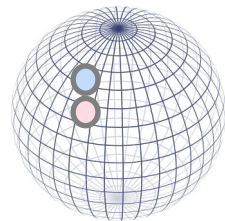
Annotation cost: 0

# Learning a visual-text embedding on HowTo100M

Pre-trained word2vec  
word embeddings (dim=300)  
(No stop words)



# Learning a visual-text embedding on HowTo100M



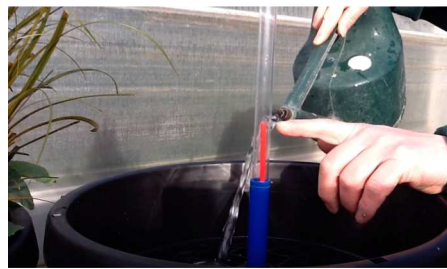
$$S_{i,j} = S(X_i, Y_j) \text{ (dot product)}$$

$$\forall(i, j), j \neq i, S_{i,i} > S_{i,j}, S_{i,i} > S_{j,i}$$

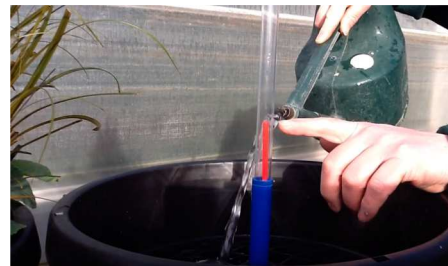
$$L = \frac{1}{B} \sum_{i=1}^B \sum_{j \neq i} \left[ \max(0, m + S_{i,j} - S_{i,i}) + \max(0, m + S_{j,i} - S_{i,i}) \right]$$



... fill pot water ...



... fill pot water ...



... these nice plants...

# Evaluation procedure

→ Text to video retrieval: **YouCook2, MSRVTT, LSMDC**

🔍 Answering the phone



→ Action localization: **CrossTask**

loose bolt



jack car



remove wheel

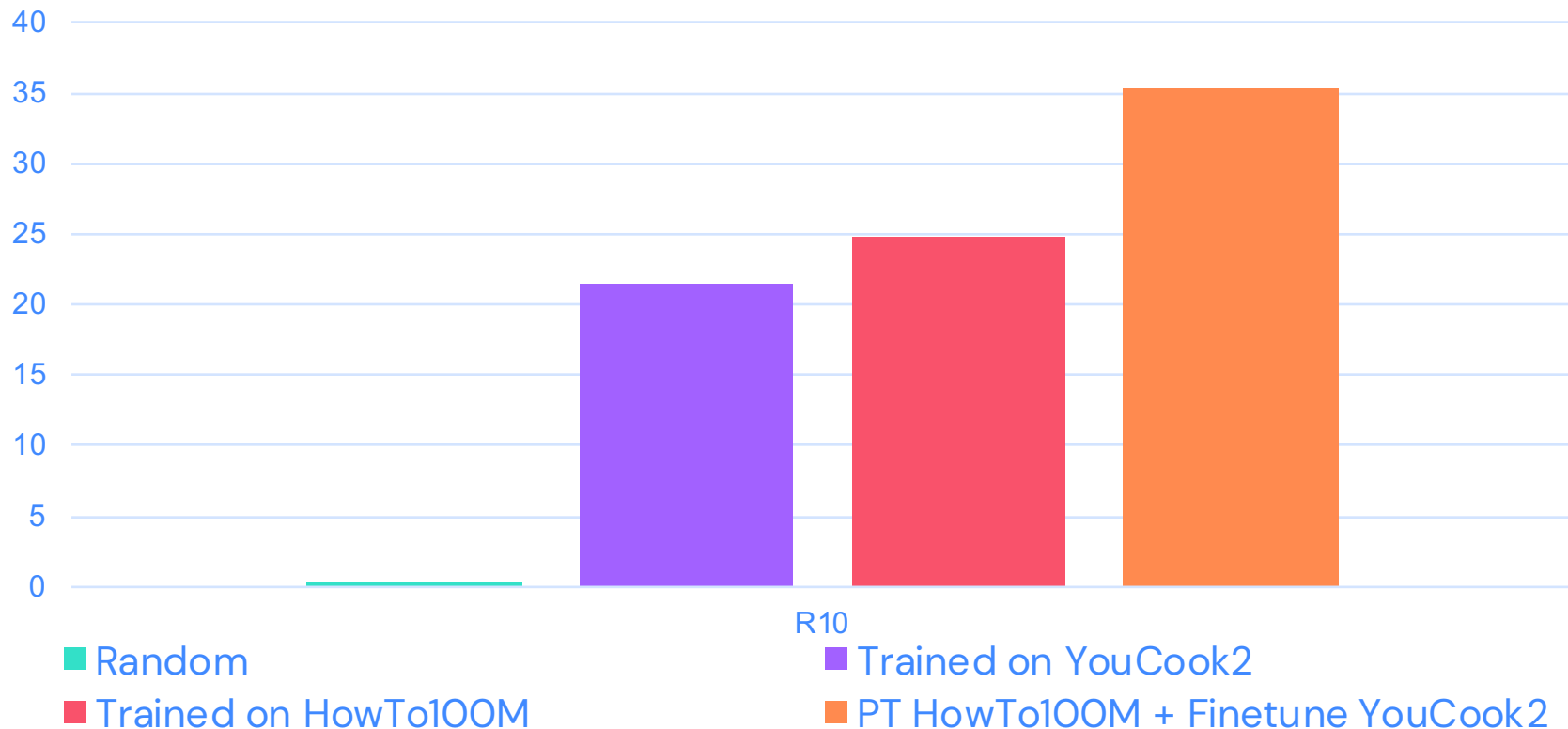




**Beauty of having a joint text and video embedding:**  
*In both cases, we can evaluate without finetuning!*

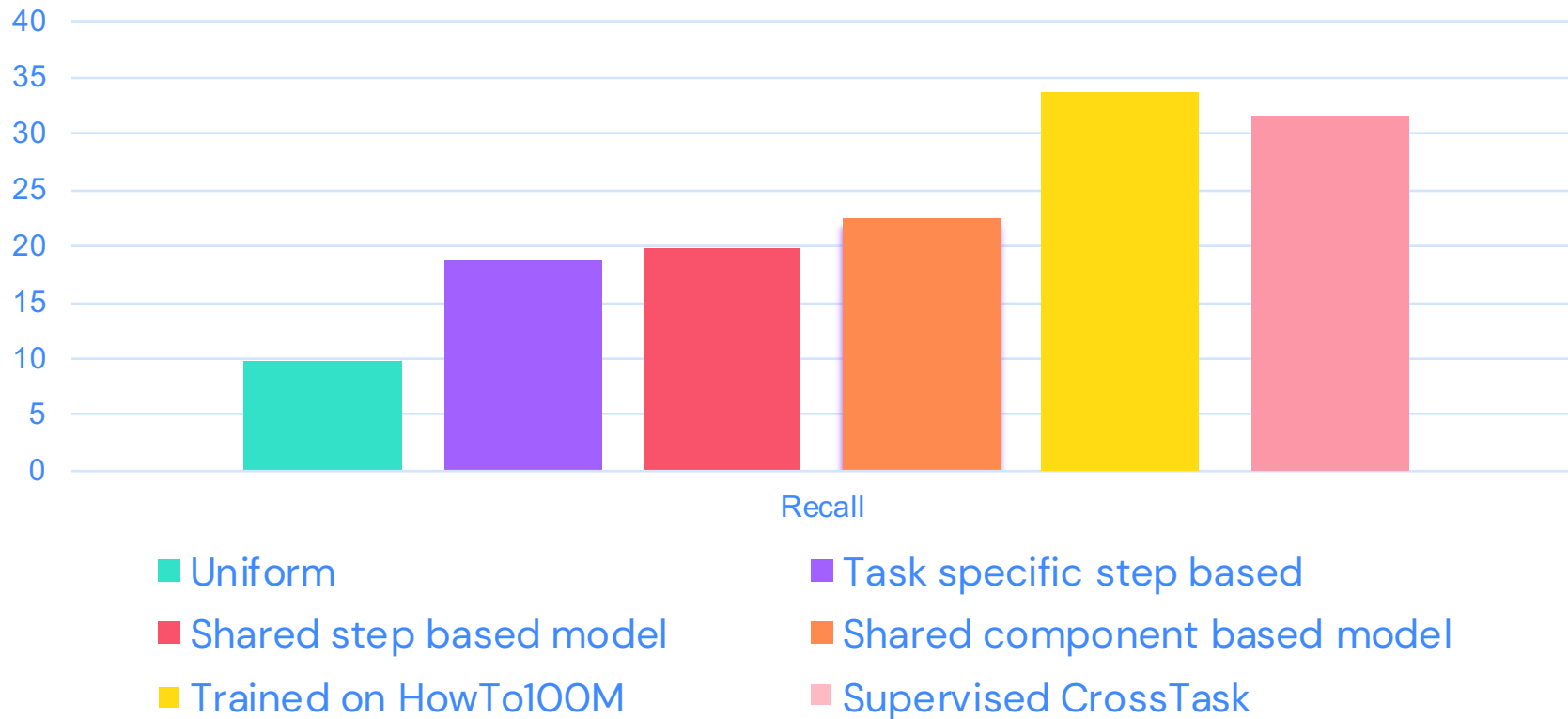
# Within domain: YouCook2 retrieval (YouTube cooking videos)

## YouCook2 (R@10)

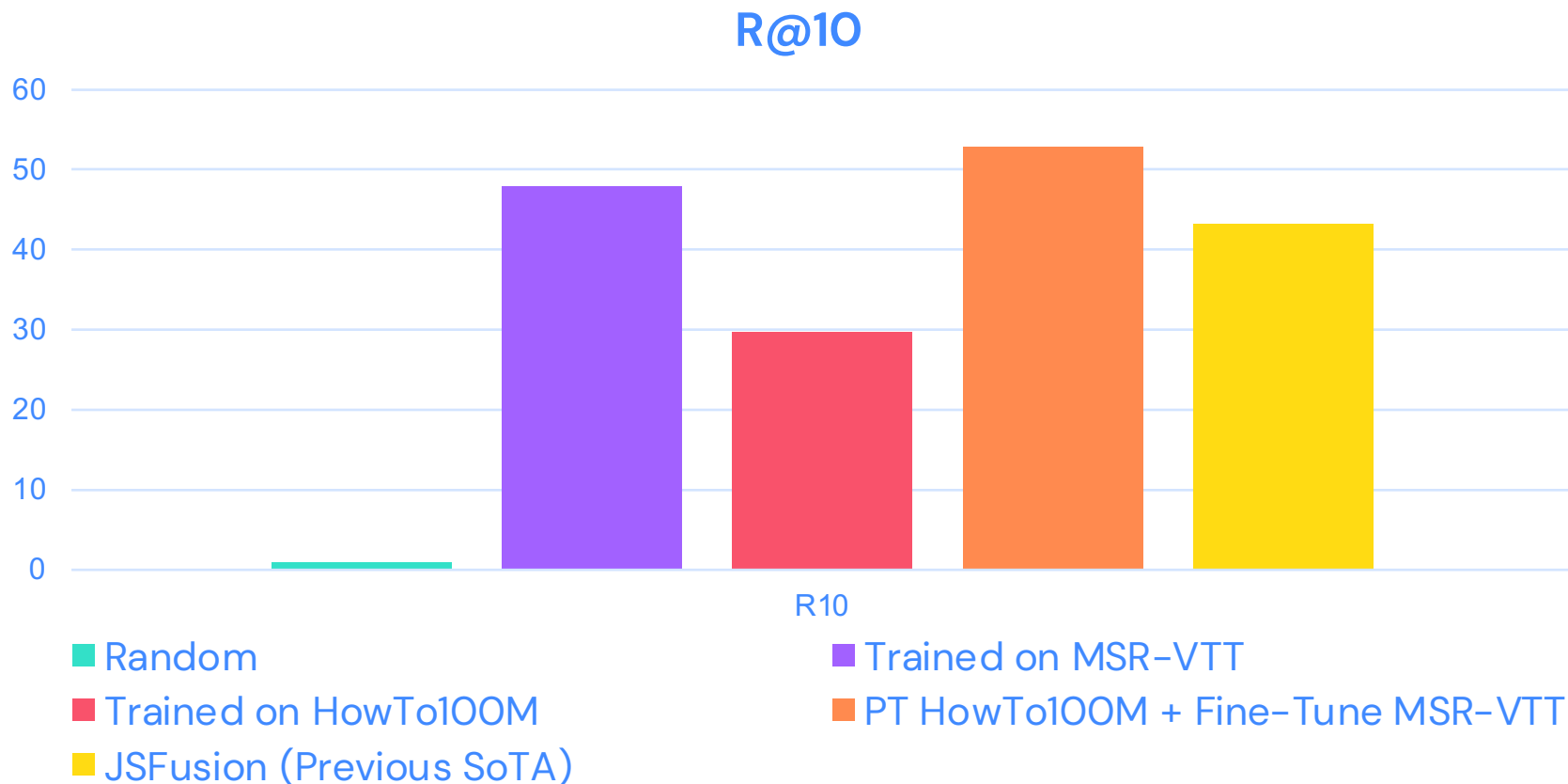


# Within domain: CrossTask action localization

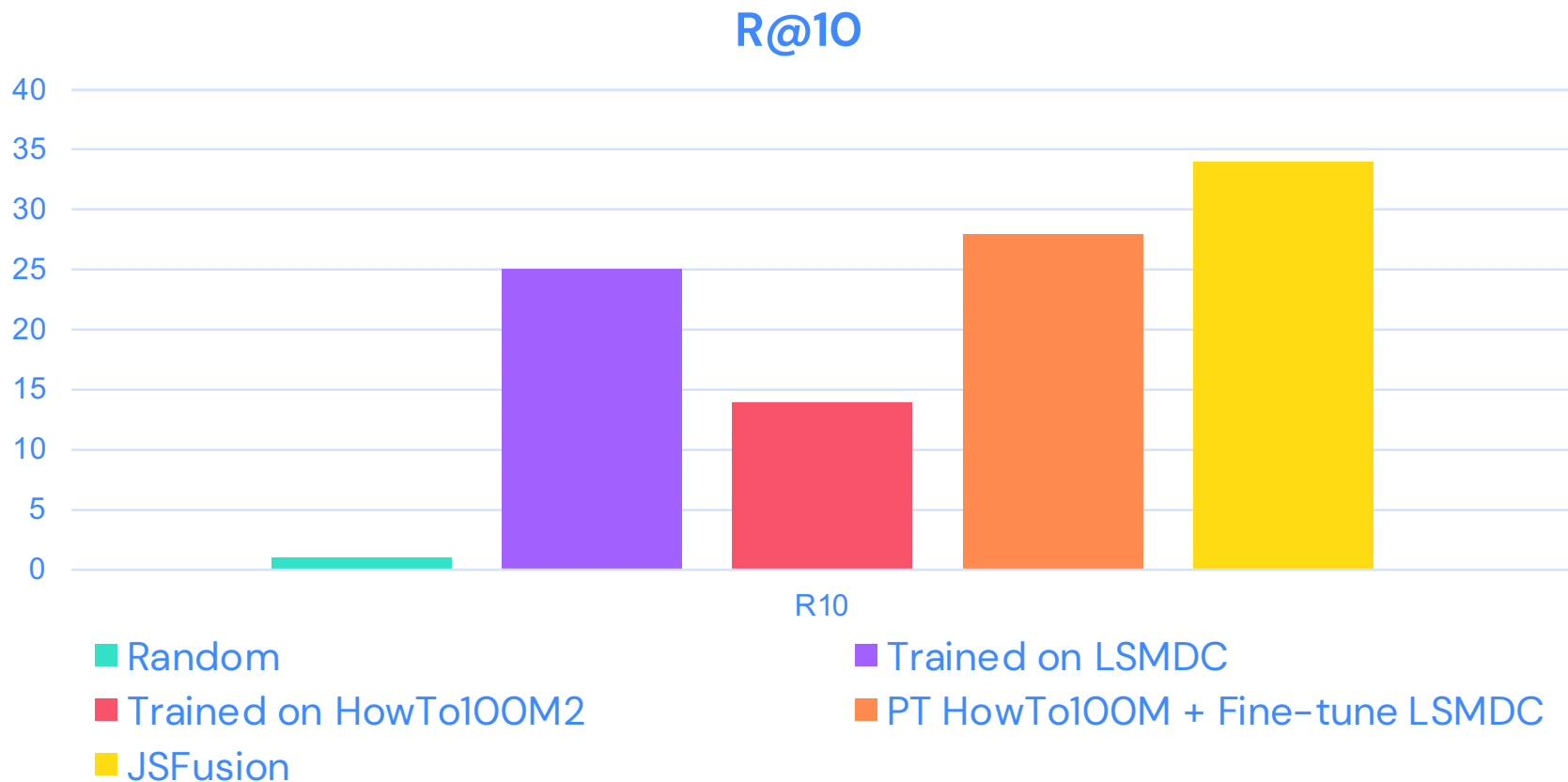
## Localization metric (recall) on CrossTask



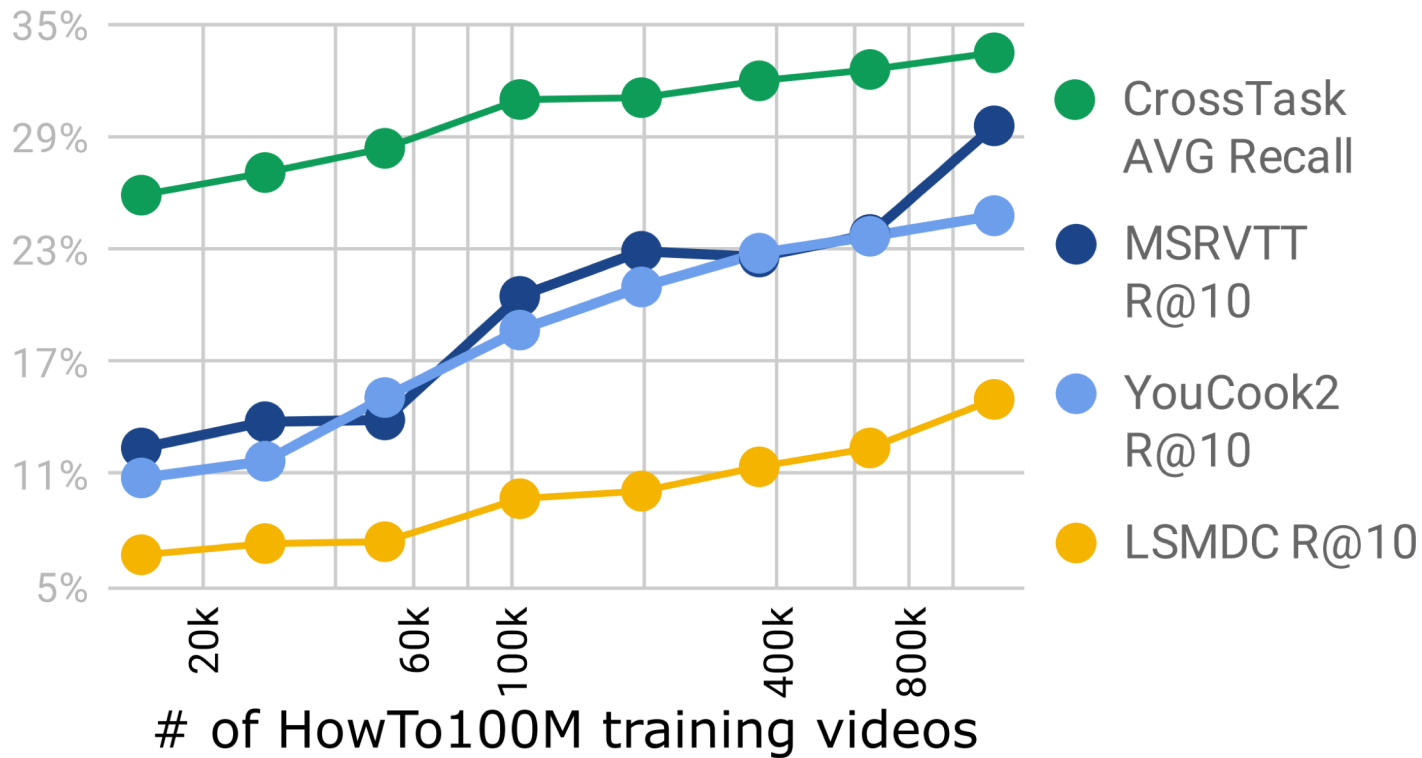
# Out of domain: MSR-VTT (popular & generic YouTube videos)



## Out of domain ++: LSMDC (movies)



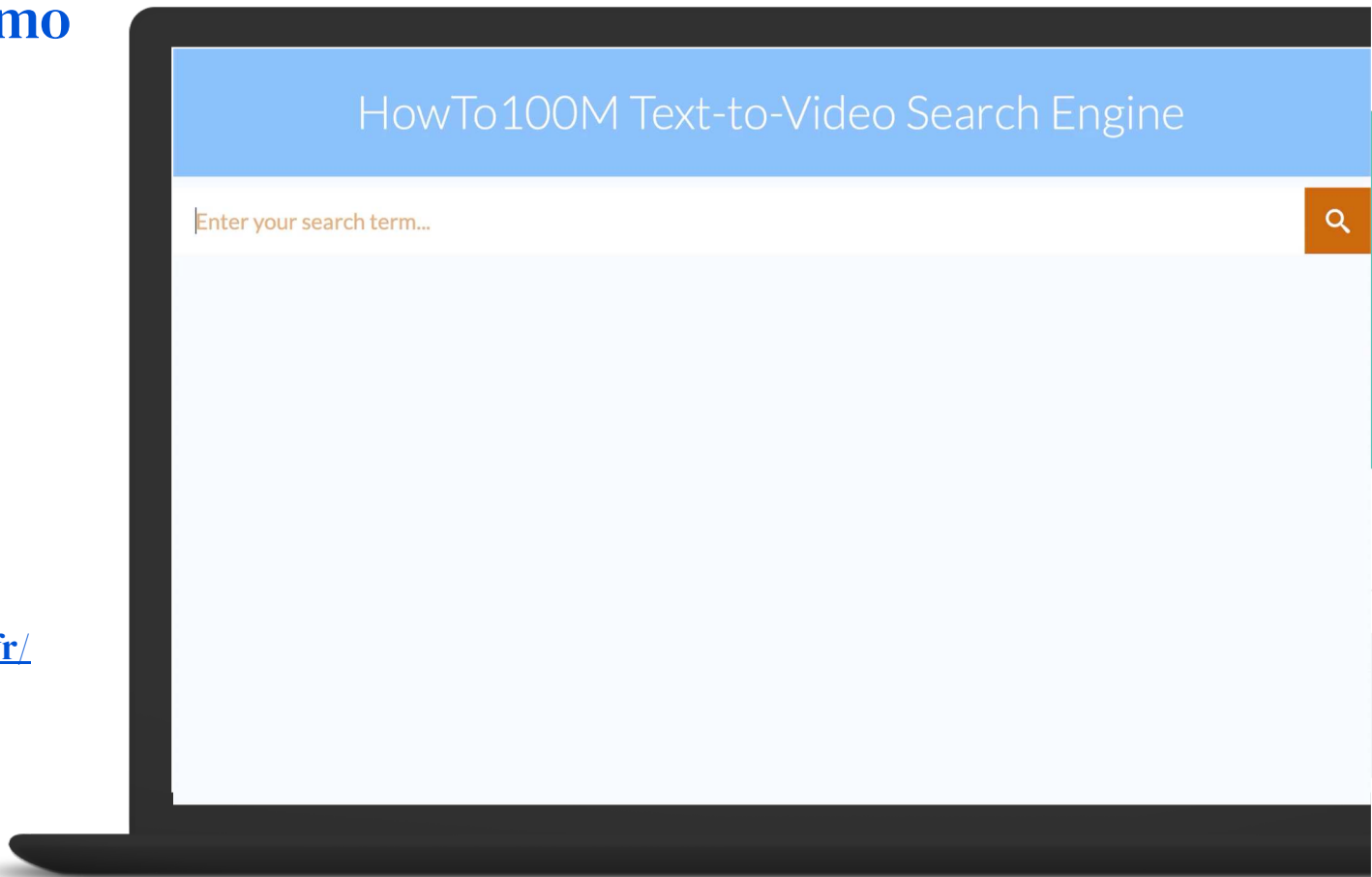
## Coming back to the original question: scale matters!



## Online web demo



<http://howto100m.inria.fr/>



A large white number '3' is centered within a blue wireframe cube. The cube is oriented in a 3D perspective, with its edges visible. The background is a solid dark blue color.

3

A white outline of a rounded rectangular shape is located on the right side of the slide. It has a smooth, curved top and bottom edge and a straight left edge.

Discussion



# Summary

## 1) Leveraging the structure of narrated videos

### Making Meringue

*Pour* egg

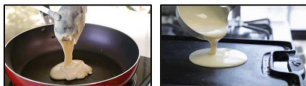
Add sugar

Whisk *mixture*



### Making Pancakes

*Pour* mixture



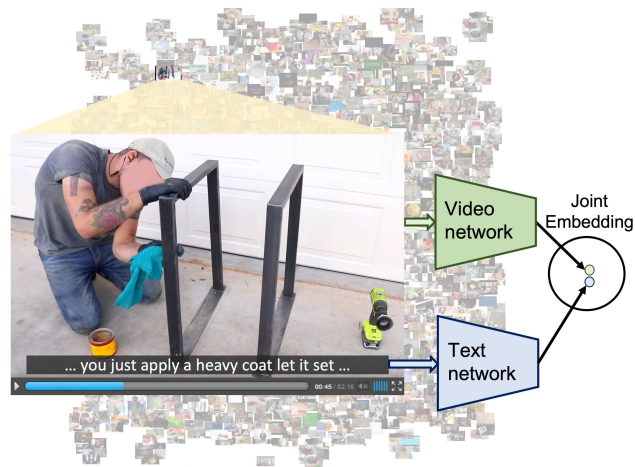
### Making Lemonade

*Pour* water



[Cross-task weakly supervised learning from instructional videos](#), Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, Josef Sivic, *CVPR2019*

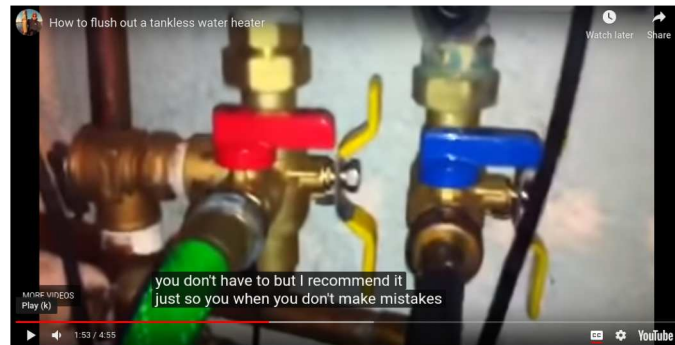
## 2) Leveraging the scale of narrated videos



[HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips](#), Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic, *ICCV2019*

# Future directions

- Dealing with the noise. In 50% of the cases, video and narration are not matching. Something should be done!



- Still relying on pretrained features (obtained from Kinetics or ImageNet) the story is not complete.  
**The dream:** end to end learning directly from HowTo100M.