Inria — INVENTEURS DU MONDE NUMÉRIQUE
CZECH TECHNICAL UNIVERSITY IN PRAGUE
ENS | PSL

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Millions Narrated Video Clips

Antoine Miech*, Dimitri Zhukov*, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic

The HowTo100M dataset

23 k — wikiHow visual tasks

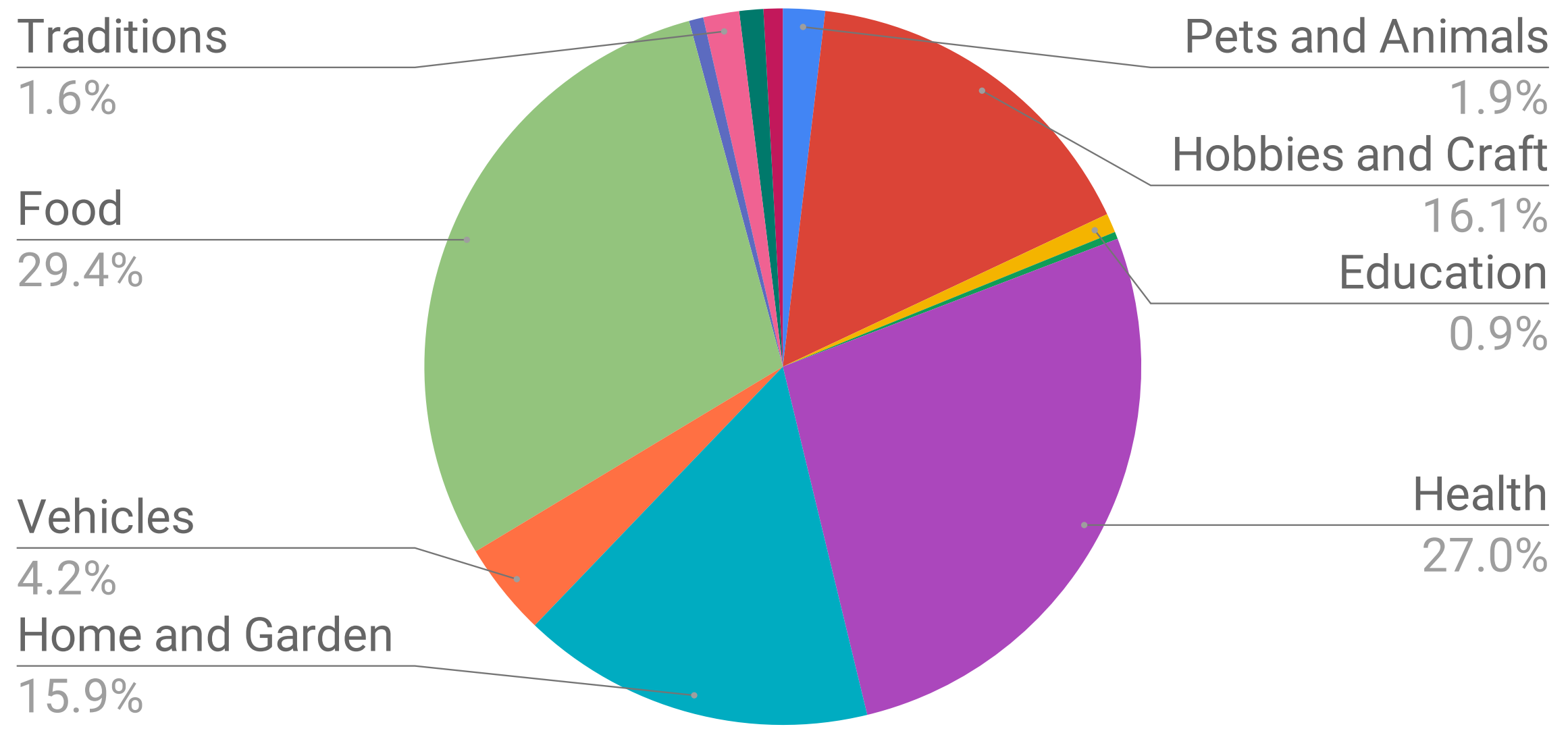
1.2 M — YouTube videos

136 M — Short clips

ASR — Supervision*

*ASR: Automated Speech Recognition

0 — Manual annotation



Video clips high level category repartition

Samples of videos frames and their associated narrations from the 136M collected video clips

Motivation & Contributions

Annotating videos is:
- Expensive
- Takes a lot of time
- Hard to scale

Few large-scale annotated video description datasets exist:
- MSR-VTT: ~10k clips / 200k captions
- LSMDC: ~ 100k clips / captions

We collect 136M narrated video clips sourced from 1.2M publicly available instructional videos.

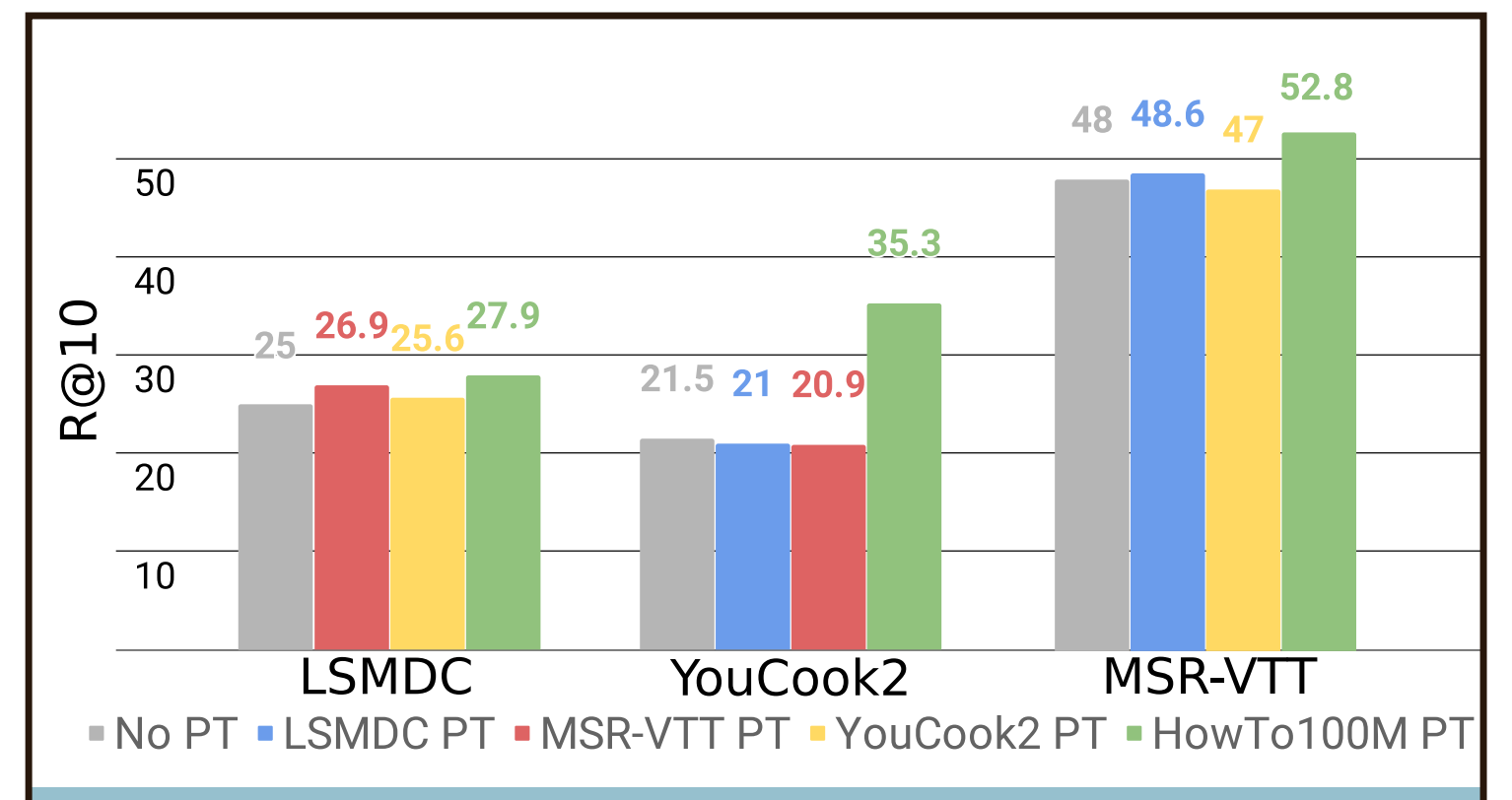
- We train a text-video joint embedding without any manually annotated clip-caption pairs.
- Our model can outperform models trained on manually annotated datasets.

Text-to-Video retrieval

Method	Trainset	R@1	R@5	R@10	Median R
Random	None	0.03	0.15	0.3	1675
HGLMM FV CCA [21]	YouCook2	4.6	14.3	21.6	75
Ours	YouCook2	4.2	13.7	21.5	65
Ours	HowTo100M	6.1	17.3	24.8	46
Ours	PT: HowTo100M FT: YouCook2	**8.2**	**24.5**	**35.3**	**24**

YouCook2

- Our off-the-shelf model trained on HowTo100M already outperforms the same model trained on YouCook2.
- Fine-tuning our model on YouCook2 yields further improvements.



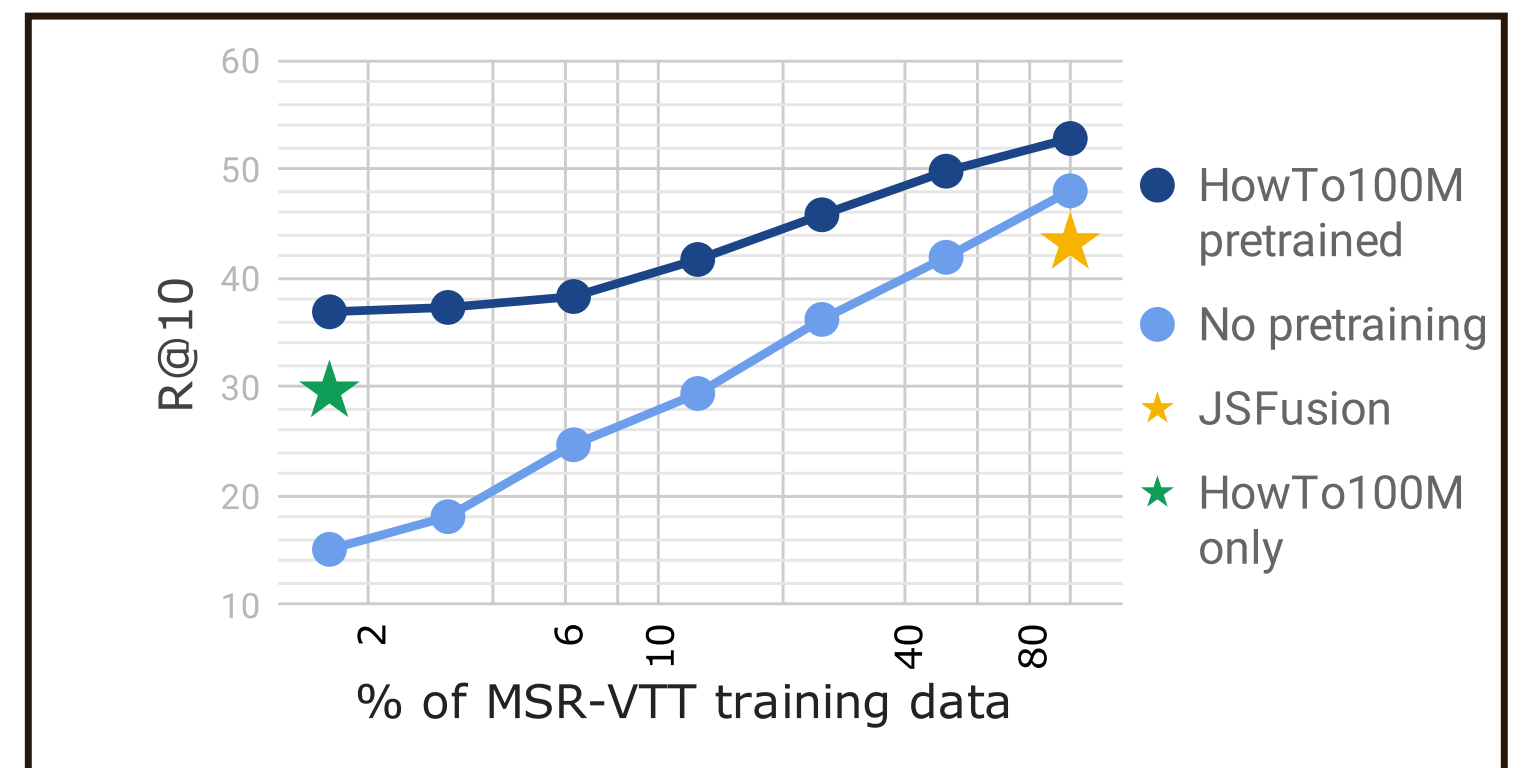
Cross-Dataset finetuning

- HowTo100M is the best pretraining video description dataset when compared to other annotated video description datasets..

Method	Trainset	R@1	R@5	R@10	Median R
Random	None	0.1	0.5	1.0	500
C+LSTM+SA+FC7 [47]	MSR-VTT	4.2	12.9	19.9	55
VSE-LSTM [20]	MSR-VTT	3.8	12.7	17.1	66
SNUVL [59]	MSR-VTT	3.5	15.9	23.8	44
Kaufman et al. [18]	MSR-VTT	4.7	16.6	24.1	41
CT-SAN [60]	MSR-VTT	4.4	16.6	22.3	35
JSFusion [58]	MSR-VTT	10.2	31.2	43.2	13
Ours	HowTo100M	7.5	21.2	29.6	38
Ours	MSR-VTT	12.1	35.0	48.0	12
Ours	PT: HowTo100M FT: MSR-VTT	**14.9**	**40.2**	**52.8**	**9**

MSR-VTT

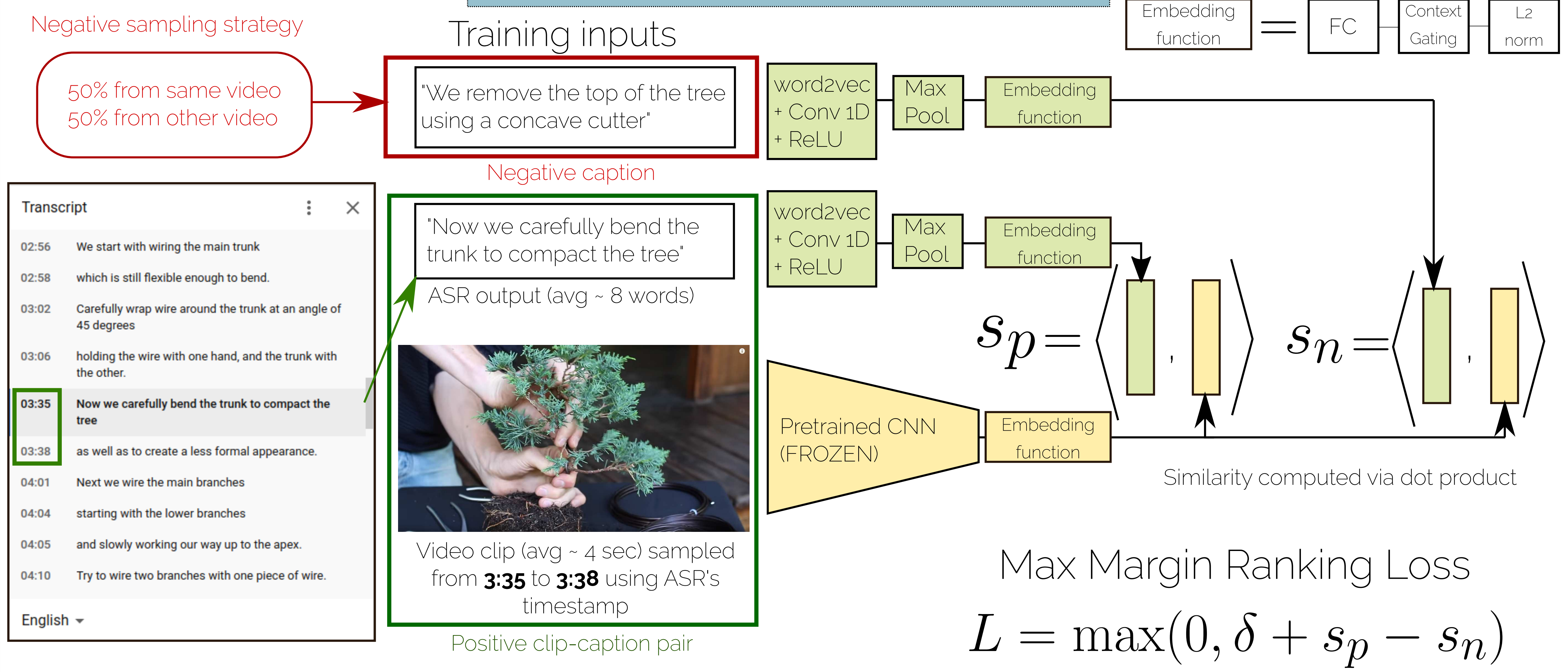
- Our off-the-shelf model trained on HowTo100M shows good performance on MSR-VTT.
- Fine-tuning our model on MSR-VTT outperforms state-of-the-art by a large margin.



MSR-VTT low data regime

- Our HowTo100M pretrained model performs well when trained on only few annotated videos (low data regime).
- We outperform previous SoTA by only using 30 % of annotated data.

Joint embedding model



"We remove the top of the tree using a concave cutter"

"Now we carefully bend the trunk to compact the tree"

Transcript

02:56 We start with wiring the main trunk
02:58 which is still flexible enough to bend.
03:02 Carefully wrap wire around the trunk at an angle of 45 degrees
03:06 holding the wire with one hand, and the trunk with the other.
03:35 Now we carefully bend the trunk to compact the tree
03:38 as well as to create a less formal appearance.
04:01 Next we wire the main branches
04:04 starting with the lower branches
04:05 and slowly working our way up to the apex.
04:10 Try to wire two branches with one piece of wire.

English

Video clip (avg ~ 4 sec) sampled from **3:35** to **3:38** using ASR's timestamp

$s_p = \langle \cdot , \cdot \rangle$ $\quad s_n = \langle \cdot , \cdot \rangle$

Similarity computed via dot product

Max Margin Ranking Loss

$$L = \max(0, \delta + s_p - s_n)$$

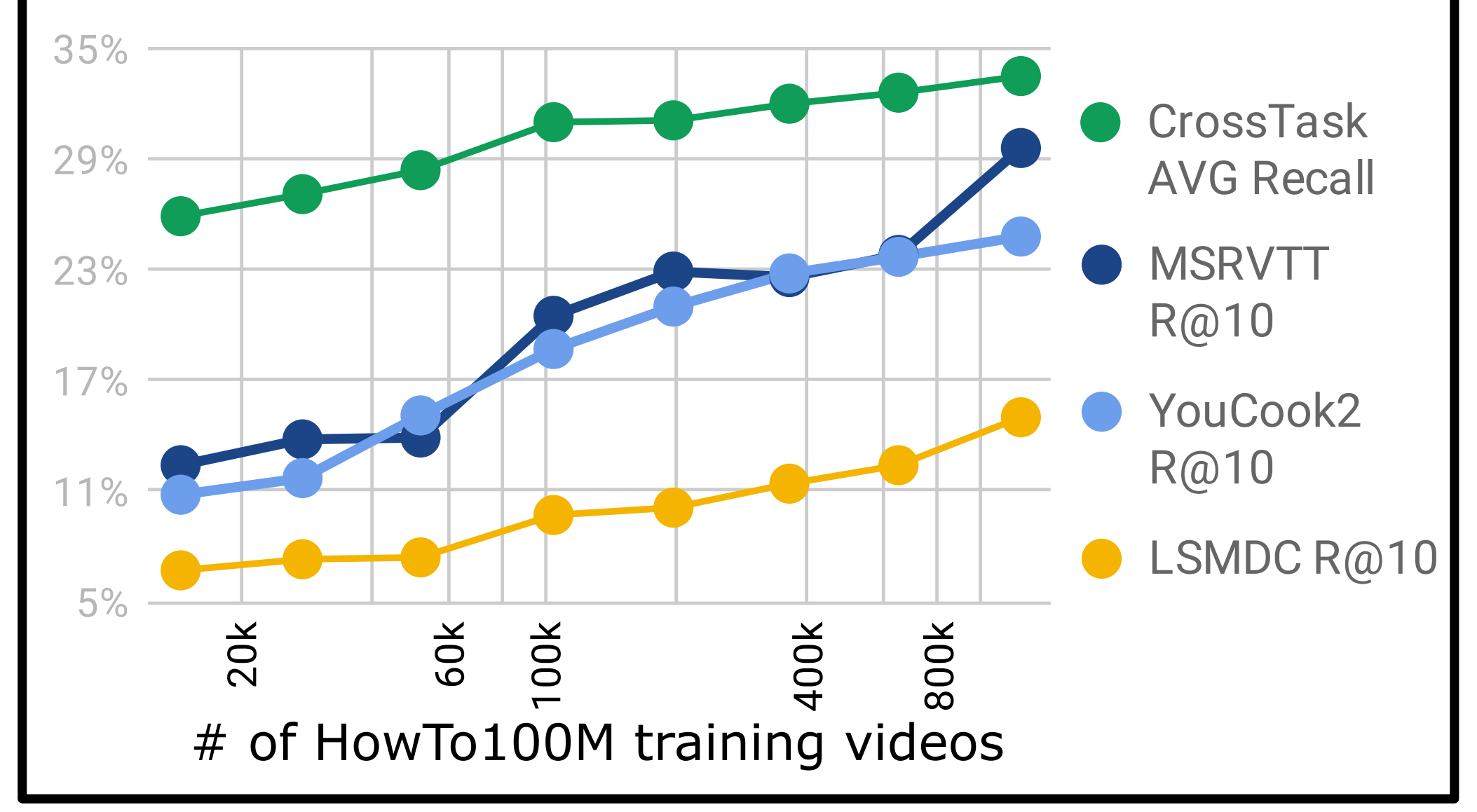
Action localization

	Make Kimchi Rice	Pickle Cucumber	Make Banana Ice Cream	Grill Steak	Jack Up Car	Make Jello Shots	Change Tire	Make Lemonade	Add Oil to Car	Make Latte	Build Shelves	Make Taco Salad	Make French Toast	Make Irish Coffee	Make Strawberry Cake	Make Pancakes	Make Meringue	Make Fish Curry	Average
Fully-supervised upper-bound [63]	19.1	25.3	38.0	37.5	25.7	28.2	54.3	25.8	18.3	31.2	47.7	12.0	39.5	23.4	30.9	41.1	53.4	17.3	31.6
Alayrac [1]	15.6	10.6	7.5	14.2	9.3	11.8	17.3	13.1	6.4	12.9	27.2	9.2	15.7	8.6	16.3	13.0	23.2	7.4	13.3
Zhukov et al. [63]	13.3	18.0	23.4	23.1	16.9	16.5	30.7	21.6	4.6	19.5	35.3	10.0	32.3	13.8	29.5	37.6	**43.0**	13.3	22.4
Ours trained on HowTo100M only	**33.5**	**27.1**	**36.6**	**37.9**	**24.1**	**35.6**	**32.7**	**35.1**	**30.7**	**28.5**	**43.2**	**19.8**	**34.7**	**33.6**	**40.4**	**41.6**	41.9	**27.4**	**33.6**

Significant improvement over SoTA on the CrossTask[1] Action Step Localization task

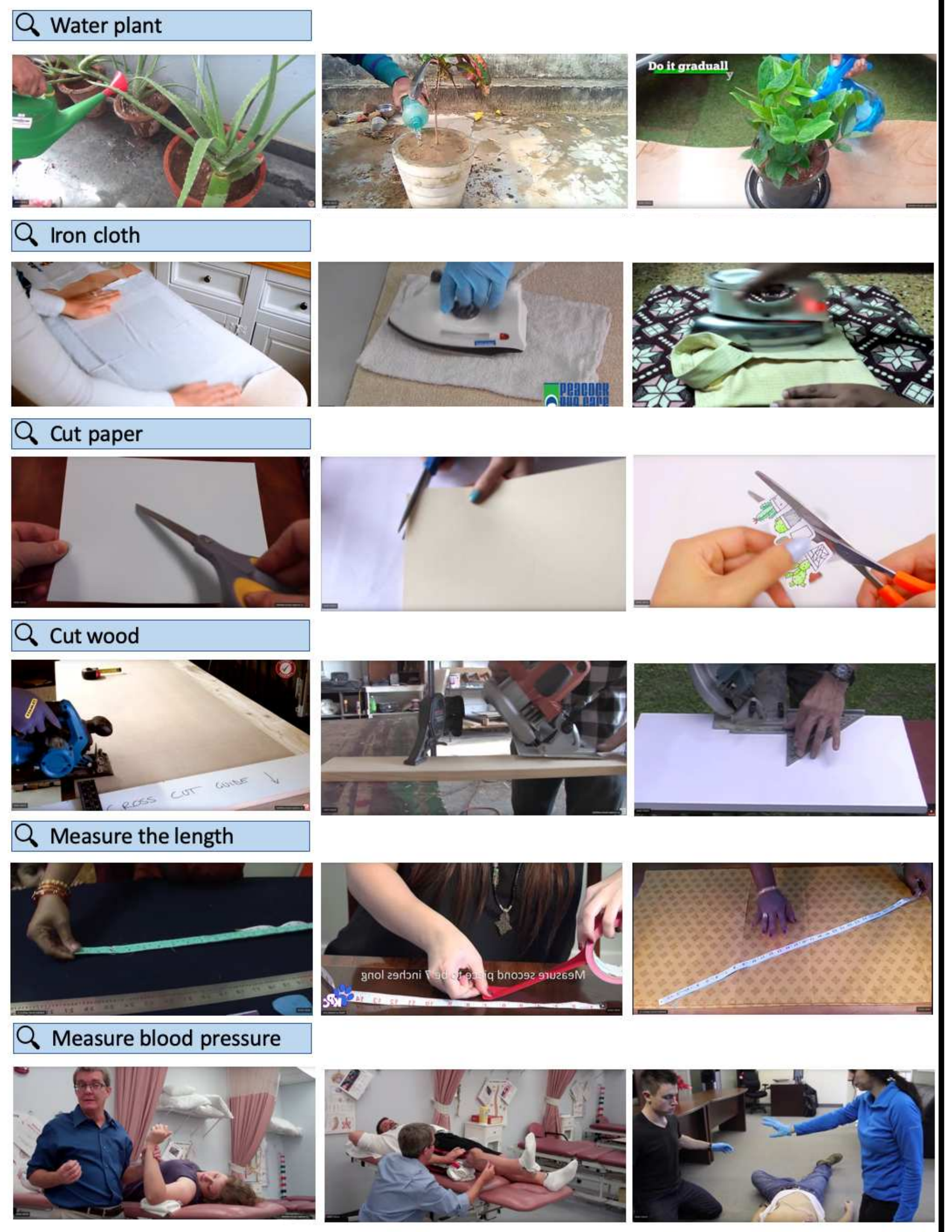
[1] Zhukov et al.. Cross-task weakly supervised learning from instructional videos. CVPR'19

Scale matters



Retrieval examples

- Top 3 retrieved clips given text queries.



Scan me to try the web retrieval demo, get the paper, data, code and trained model !