

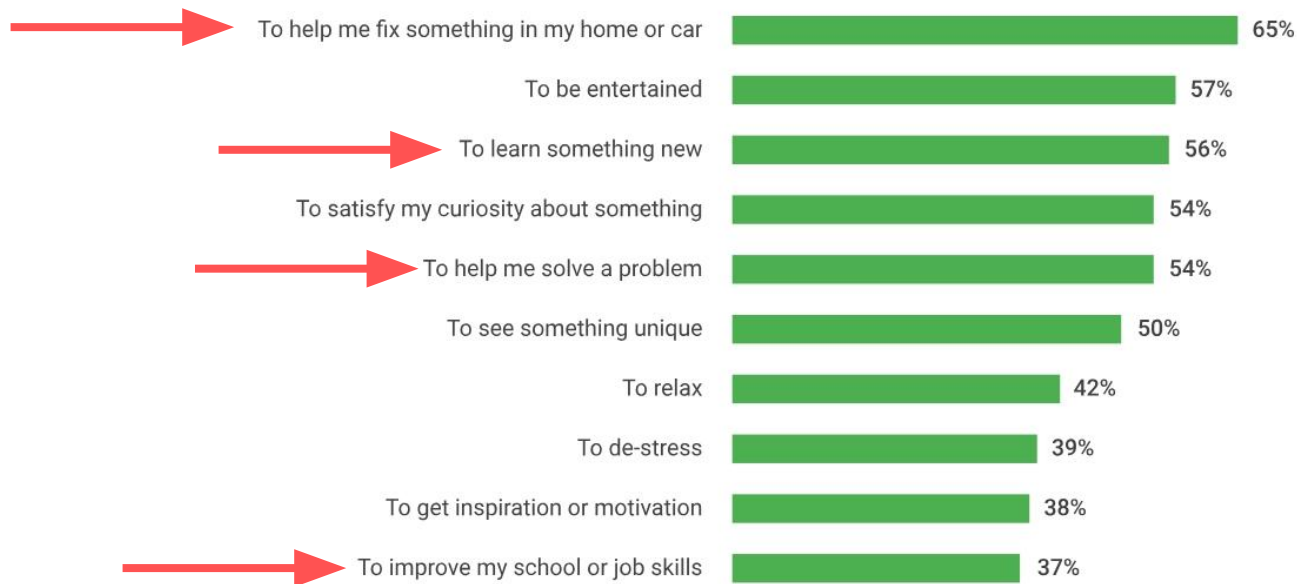
A Case Study on

# Combining ASR and Visual Features for Generating Instructional Video Captions

---

Jack Hessel, Bo Pang, Zhenhai Zhu, Radu Soricut

## Why people turn to YouTube

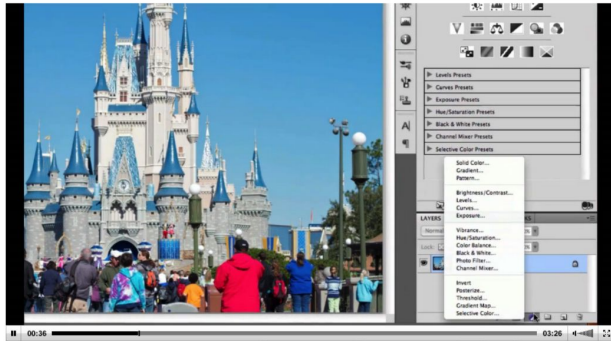


### Think with Google

2and2/Google, "The Values of YouTube" Study, Oct. 2017 (n of 1,006 consumers between the ages of 18-54, with 918 monthly YouTube users). Respondents were asked to choose which platforms they turn to for a range of needs.

# How best to present instructional videos?

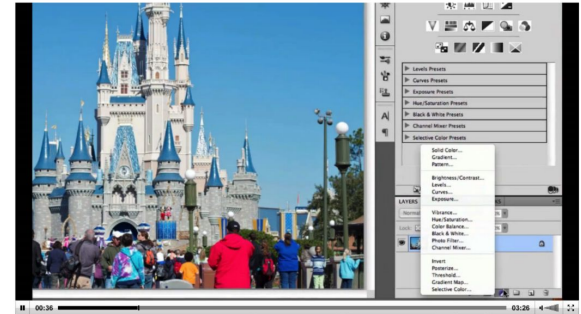
Photoshop: Vintage Effect



*Raw Instructional Video*



Photoshop: Vintage Effect



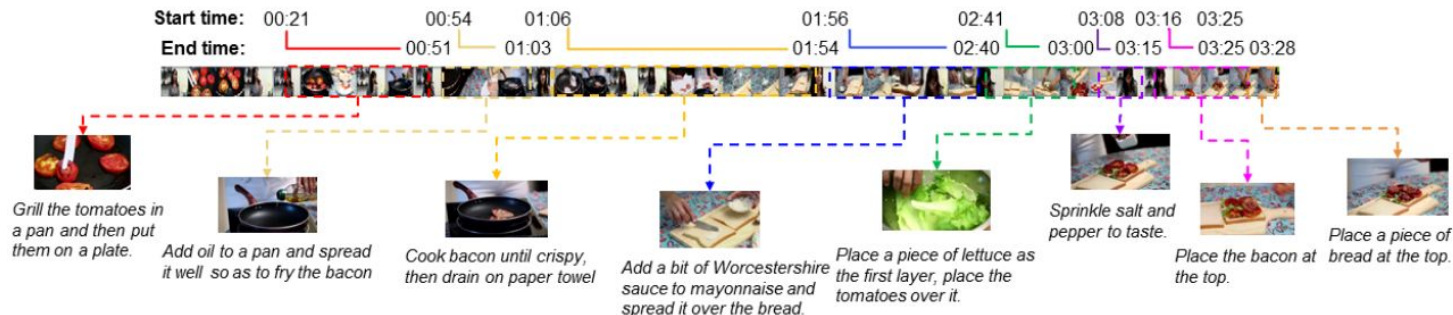
*Structured Representation*

# YouCook2 and Segment Captioning

2K videos  
89 different recipes  
176 hours of video



8 steps/video  
English annotation  
8 words/step



# Current segment captioning performance?


BLEU-4 METEOR ROUGE-L CIDEr

# Current segment captioning performance?

BLEU-4 METEOR ROUGE-L CIDEr

Zhou et al. (2018c)	3.84	11.6	27.4	.38
Sun et al. (2019b)	4.07	11.0	27.5	.50
Sun et al. (2019a)	4.31	11.9	29.5	.53

Video-only models  
from computer  
vision community



# Current segment captioning performance?

BLEU-4 METEOR ROUGE-L CIDEr

Zhou et al. (2018c)	3.84	11.6	27.4	.38
Sun et al. (2019b)	4.07	11.0	27.5	.50
Sun et al. (2019a)	4.31	11.9	29.5	.53
Human Estimate	15.2	25.9	45.1	3.8


Video-only models  
from computer  
vision community



# Current segment captioning performance?

	BLEU-4	METEOR	ROUGE-L	CIDEr
Constant Prediction	2.70	10.3	21.7	.15
Zhou et al. (2018c)	3.84	11.6	27.4	.38
Sun et al. (2019b)	4.07	11.0	27.5	.50
Sun et al. (2019a)	4.31	11.9	29.5	.53
Human Estimate	15.2	25.9	45.1	3.8

Video-only models  
from computer  
vision community





# Current segment captioning performance?

	BLEU-4	METEOR	ROUGE-L	CIDEr
Constant Prediction	2.70	10.3	21.7	.15
Zhou et al. (2018c)	3.84	11.6	27.4	.38
Sun et al. (2019b)	4.07	11.0	27.5	.50
Sun et al. (2019a)	4.31	11.9	29.5	.53
Human Estimate	15.2	25.9	45.1	3.8

Video-only models  
from computer  
vision community

"heat some oil in a pan and add salt and  
pepper to the pan and stir."

Bridging the gap with new signal sources?

Automatic Speech Recognition! (ASR)

# Prior work has shown ASR can be useful for video understanding tasks

[Gupta and Mooney, 2010; Motwani and Mooney, 2012; Regneri et al., 2013; Naim et al., 2015; Malmaud et al., 2015; Sener et al. 2015; Alayrac et al., 2016; Hendricks et al., 2017; Kuehne et al., 2017; Huang et al., 2017, 2018; Hahn et al., 2018; inter alia]



"next, in the hot pan I place the tomatoes, and you only want to flip these things once, I'd say..."

# Raw ASR versus recipe step:

Automatic Speech Recognition

Target Recipe Step

## Raw ASR versus recipe step:

“knob of ginger and cut off a little bit and then just zest it on my cool duster here and i'm also going to put in some garlic”

cut up ginger and grate into the bowl

Automatic Speech Recognition

Target Recipe Step

# Raw ASR versus recipe step:

“knob of ginger and cut off a little bit and then just zest it on my cool duster here and i'm also going to put in some garlic”

“best quality olive oil I can find”

Automatic Speech Recognition

cut up ginger and grate into the bowl

heat some olive oil in a sauce pan

Target Recipe Step

# Raw ASR versus recipe step:

“knob of ginger and cut off a little bit and then just zest it on my cool duster here and i'm also going to put in some garlic”

“best quality olive oil I can find”

“...”

Automatic Speech Recognition

cut up ginger and grate into the bowl

heat some olive oil in a sauce pan

drain on paper towels and serve warm

Target Recipe Step

# Improving performance with ASR?

“... the subtitles or action sequences automatically generated by machines, e.g., YouTube’s ASR system, are inaccurate and require manual intervention...”

- Zhou et al. 2018



# ASR Retrieval baseline...

... my video and today we will first mix together three cups of **flour**, two teaspoons of **baking soda**...

## Training recipe steps...

Fry the onion  
Preheat the oven to 350 degrees  
Steam the edamame  
Add the chicken to the pan  
Mix the **flour** and **baking soda**  
Mix the onions and garlic  
Let simmer for 20 minutes  
Boil the corn and add to the pan

Mix the flour and baking soda

BLEU-4 METEOR ROUGE-L CIDEr

---

CNST	2.70	10.03	21.69	0.15
Sun et al. (2019a)	4.31	11.91	29.47	0.53

---

BLEU-4 METEOR ROUGE-L CIDEr

---

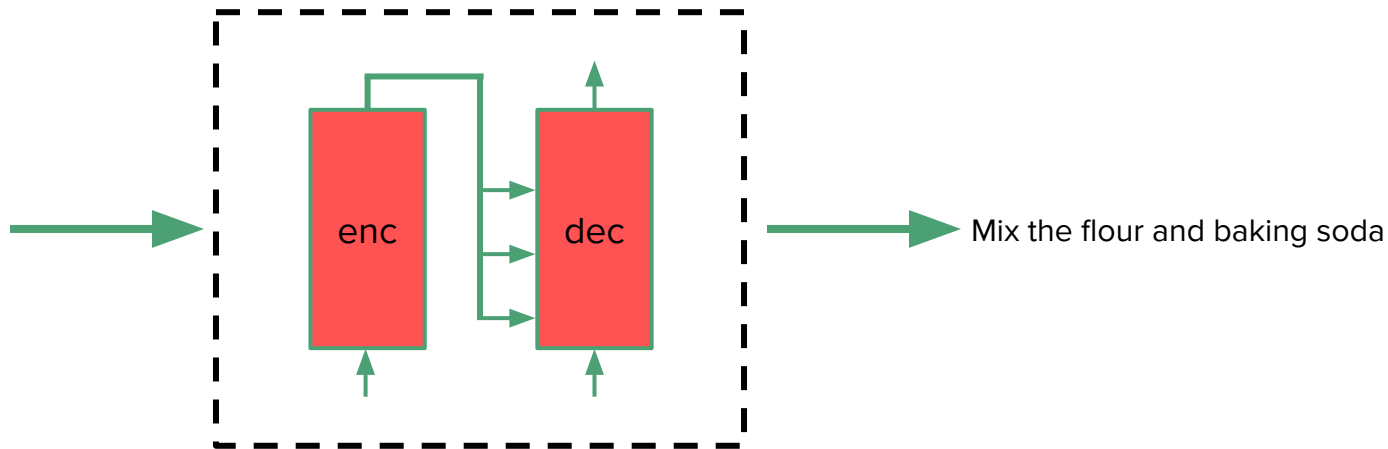
CNST	2.70	10.03	21.69	0.15
Sun et al. (2019a)	4.31	11.91	29.47	0.53

---

RET	5.68	14.29	28.06	0.80
-----	------	-------	-------	------

# ASR Transformer

... my video and today we will first mix together three cups of flour, two teaspoons of baking soda...



*\*hyperparameters selected separately for each of 10 cross-validation splits according to dev ROUGE-L...*

BLEU-4 METEOR ROUGE-L CIDEr

---

CNST	2.70	10.03	21.69	0.15
Sun et al. (2019a)	4.31	11.91	29.47	0.53

---

RET	5.68	14.29	28.06	0.80
-----	------	-------	-------	------

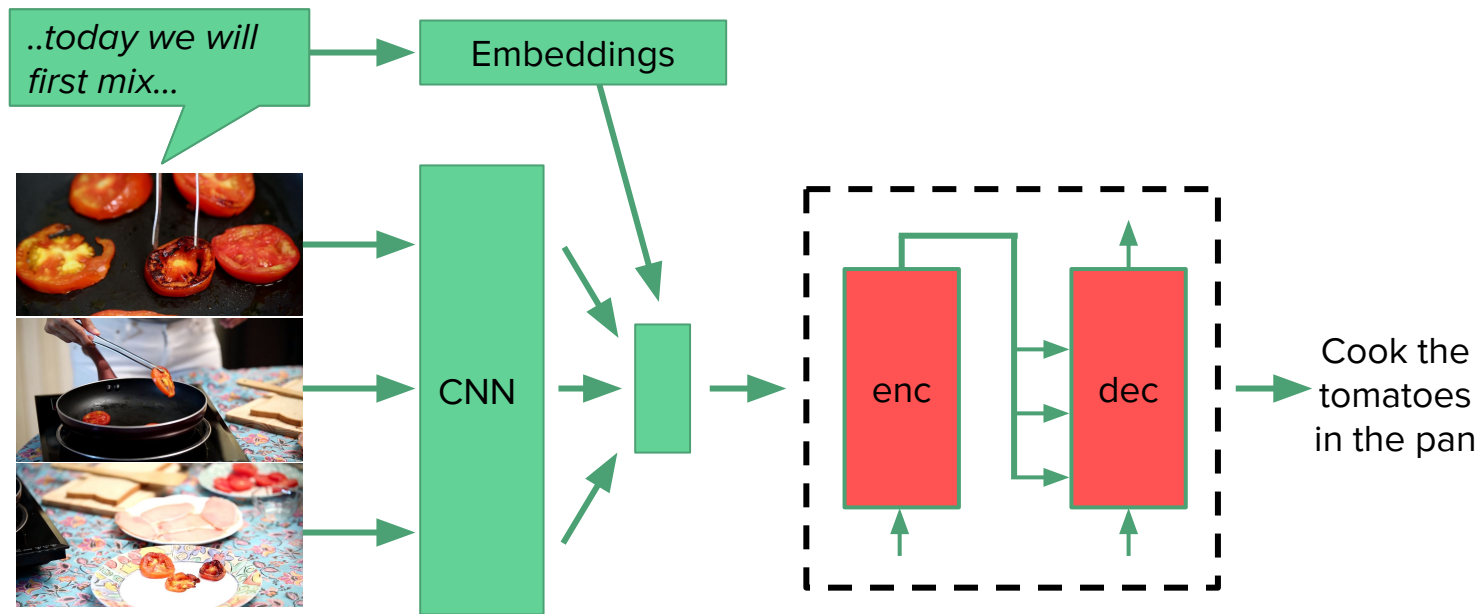
BLEU-4 METEOR ROUGE-L CIDEr

---

CNST	2.70	10.03	21.69	0.15
Sun et al. (2019a)	4.31	11.91	29.47	0.53
RET	5.68	14.29	28.06	0.80
AT	<u>8.55</u>	16.93	35.54	1.06

---

# ASR + Video Transformer



BLEU-4 METEOR ROUGE-L CIDEr

---

CNST	2.70	10.03	21.69	0.15
Sun et al. (2019a)	4.31	11.91	29.47	0.53

---

RET	5.68	14.29	28.06	0.80
AT	<u>8.55</u>	16.93	35.54	1.06



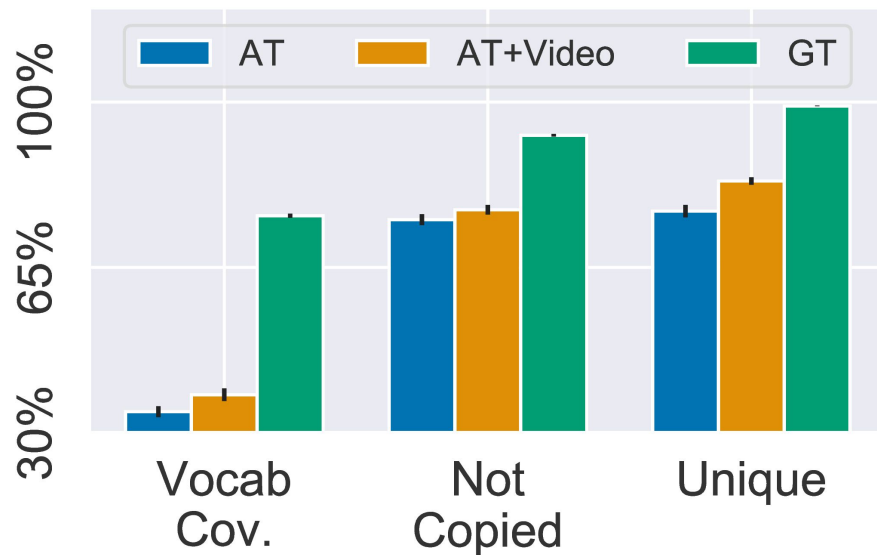
## BLEU-4 METEOR ROUGE-L CIDEr

---

CNST	2.70	10.03	21.69	0.15
Sun et al. (2019a)	4.31	11.91	29.47	0.53
RET	5.68	14.29	28.06	0.80
AT	<u>8.55</u>	16.93	35.54	1.06
AT+Video	<u>9.01</u>	<u>17.77</u>	<b>36.65</b>	<b>1.12</b>

---

# Caption Diversity Statistics



AT+Video captions are slightly more diverse than AT captions!

# What do the generations look like?

## The good!

input



"so i just want to go ahead and remove all of this fat from our chicken... cut it into about one inch pieces so you want pieces"



"... color them and then shape them ... tongs so as not to burn yourself it goes with total tacos in a frying pan ...!"

groundtruth

cut the chicken into pieces

prepare the tortillas and roll them using rolling pin

prediction

cut the chicken into pieces

place the tortilla on the pan and roll

# What do the generations look like?

## The ugly!

input



"get the colored variety the kashmiri variety is very good one and a half tablespoon of coriander"



"..."

groundtruth

add chile powder

place the chicken on the rice

prediction

add the coriander powder coriander powder coriander powder coriander powder coriander powder coriander powder coriander powder and coriander powder to the pan

add the sauce to the pot

What signal does video have that text doesn't?

# What signal does video have that text doesn't?



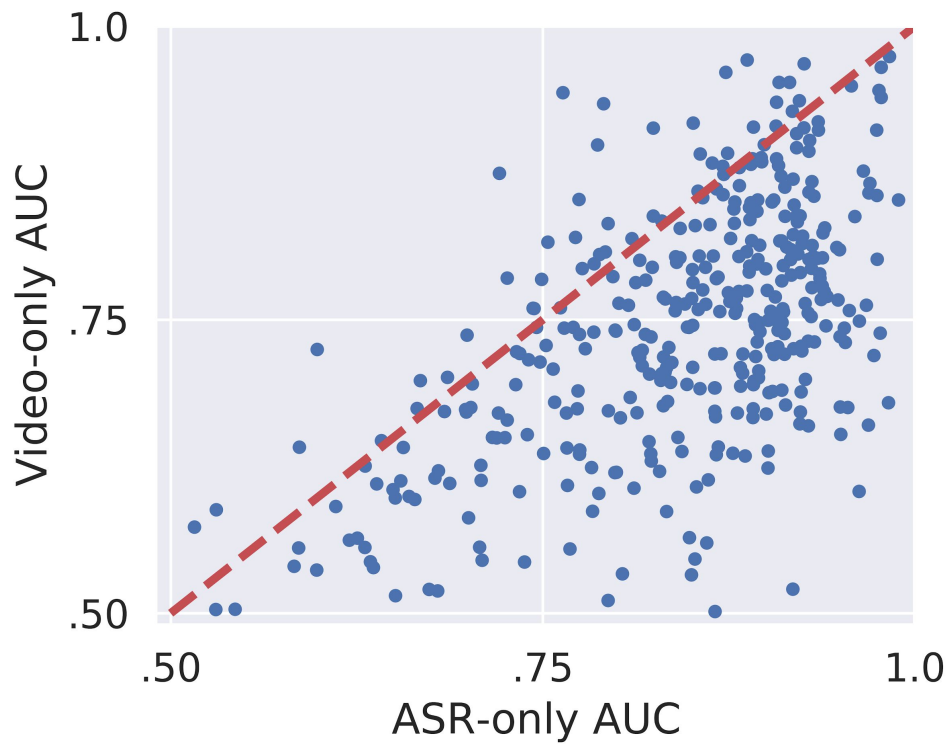
1	grill
0	broil
0	bake
1	tomatoes
0	onions
0	potatoes
0	leeks
1	pan
...	...

... the first step is to cook the tomatoes so let's get started add some oil to the pan and then...

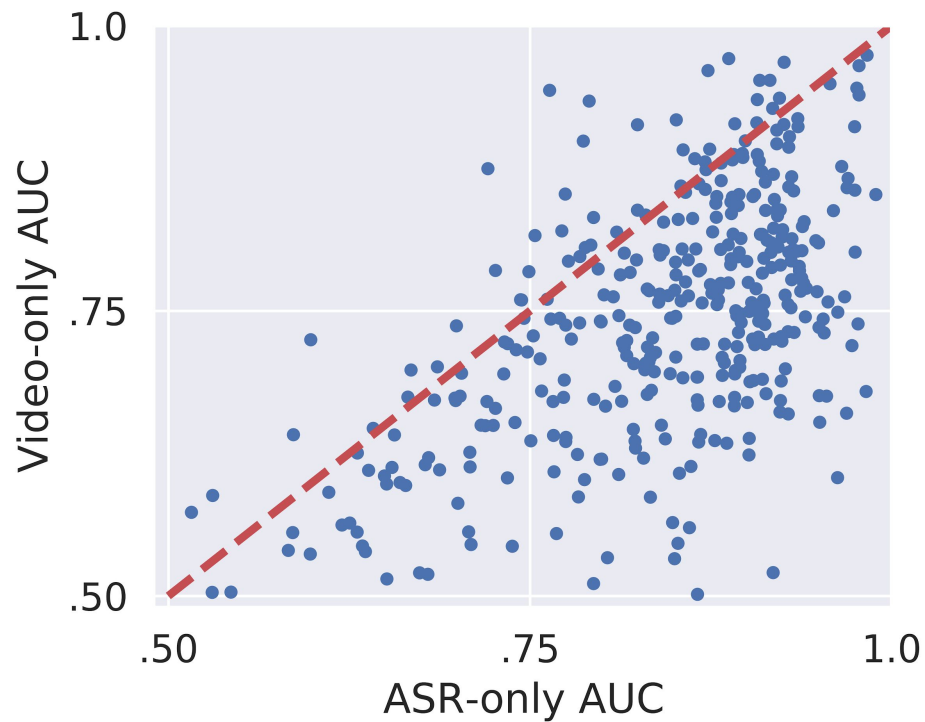


1	grill
0	broil
0	bake
1	tomatoes
0	onions
0	potatoes
0	leeks
1	pan
...	...

# Per-word performance

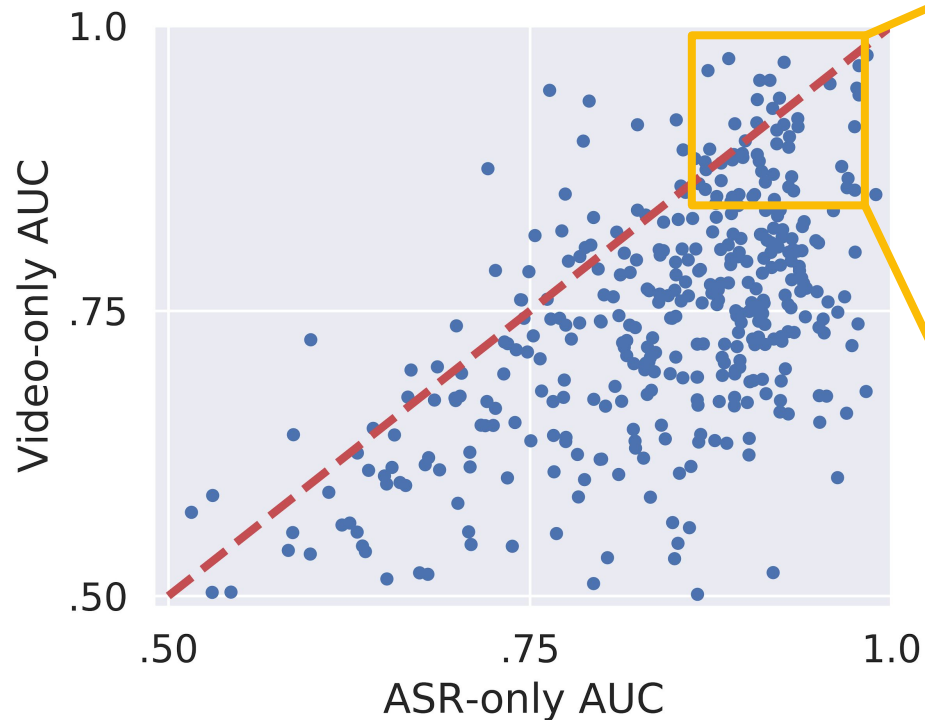


# Per-word performance





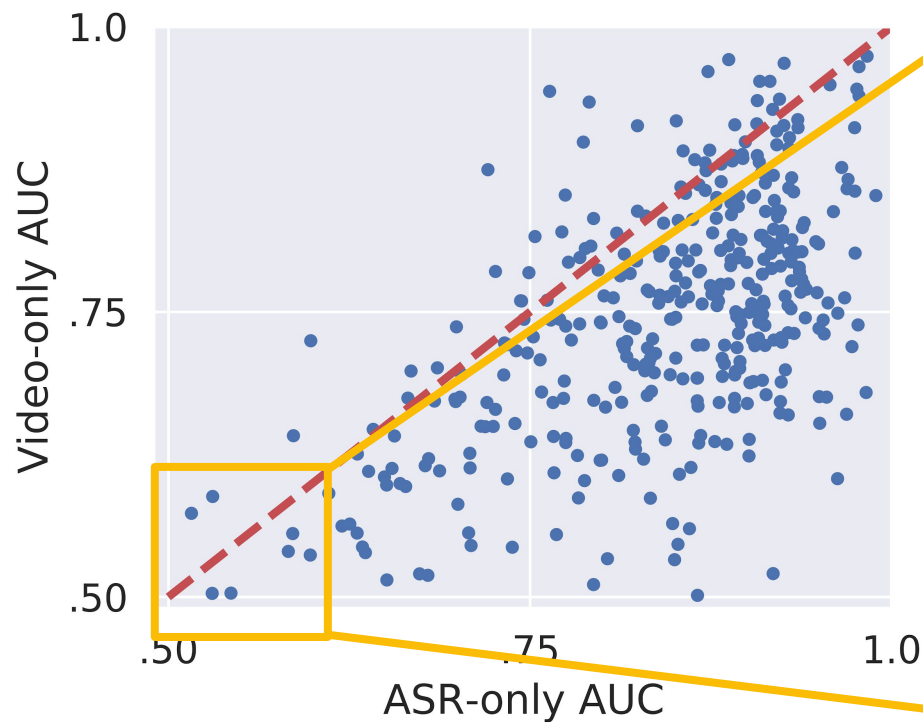
# Per-word performance



## Universally Easy:

knead	97.8
nori	97.1
yeast	96.1
mozzarella	95.8
lettuce	95.3
pancake	94.7
wrapper	94.3
patty	93.4
dal	93.0
grill	92.9
pizza	92.9
oven	92.7
bake	92.3
processor	92.3
peel	92.1
mint	92.1
cayenne	92.1
avocado	91.9
broccoli	91.8
burger	91.7

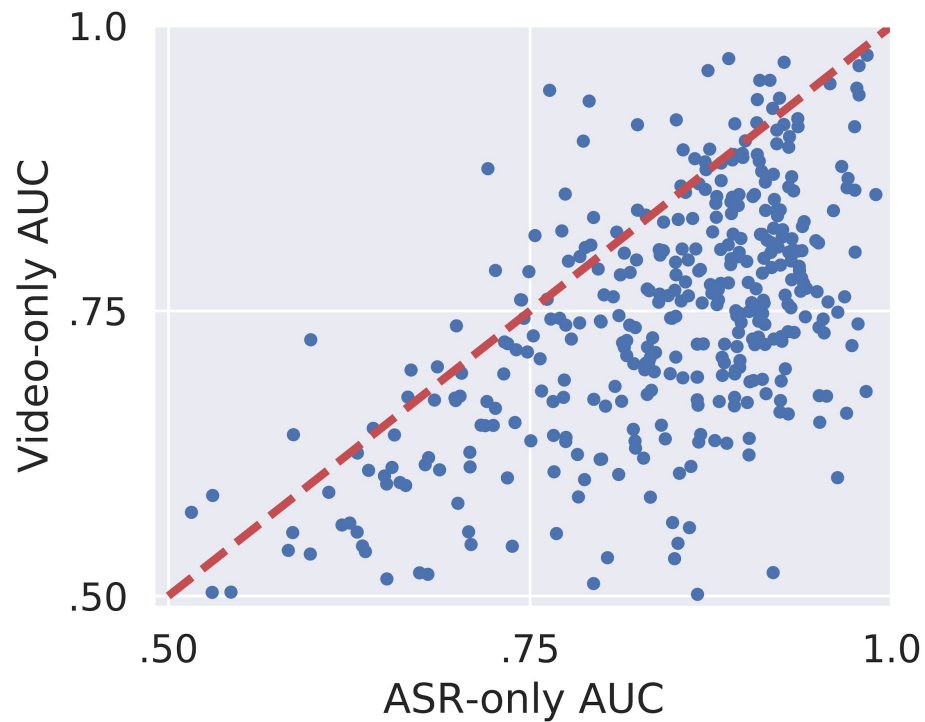
# Per-word performance



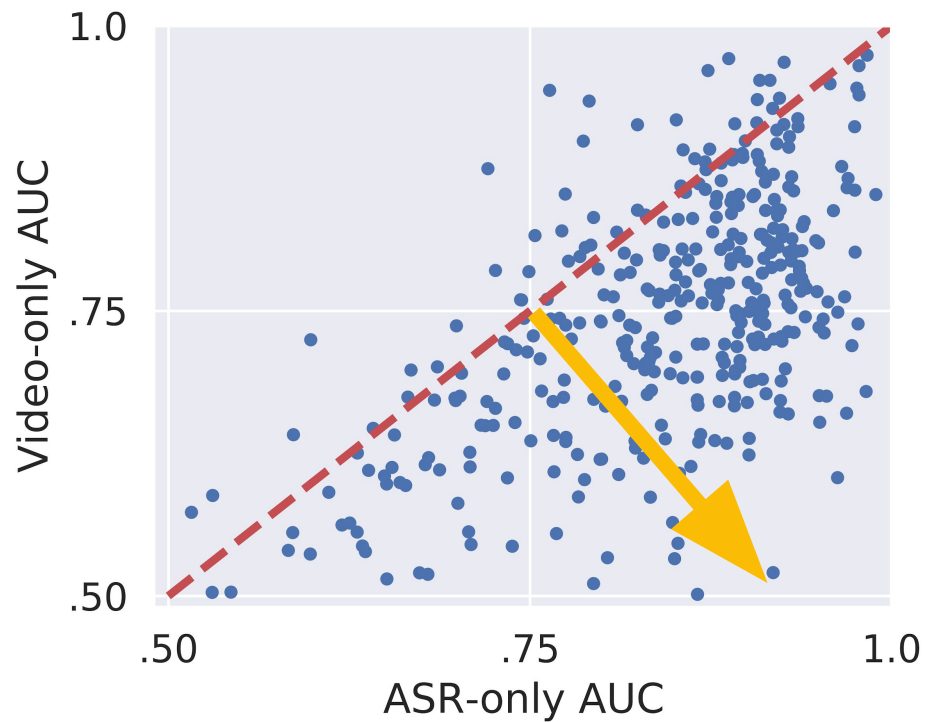
## Universally Hard:

4	43.8
bit	51.7
about	52.1
prepare	52.3
mixed	54.5
then	54.5
spoon	54.9
or	55.9
it	56.2
ready	56.7
3	57.1
few	58.3
more	58.5
some	58.8
marinated	58.9
of	59.1
separate	59.3
take	59.5
2	59.7
little	59.9

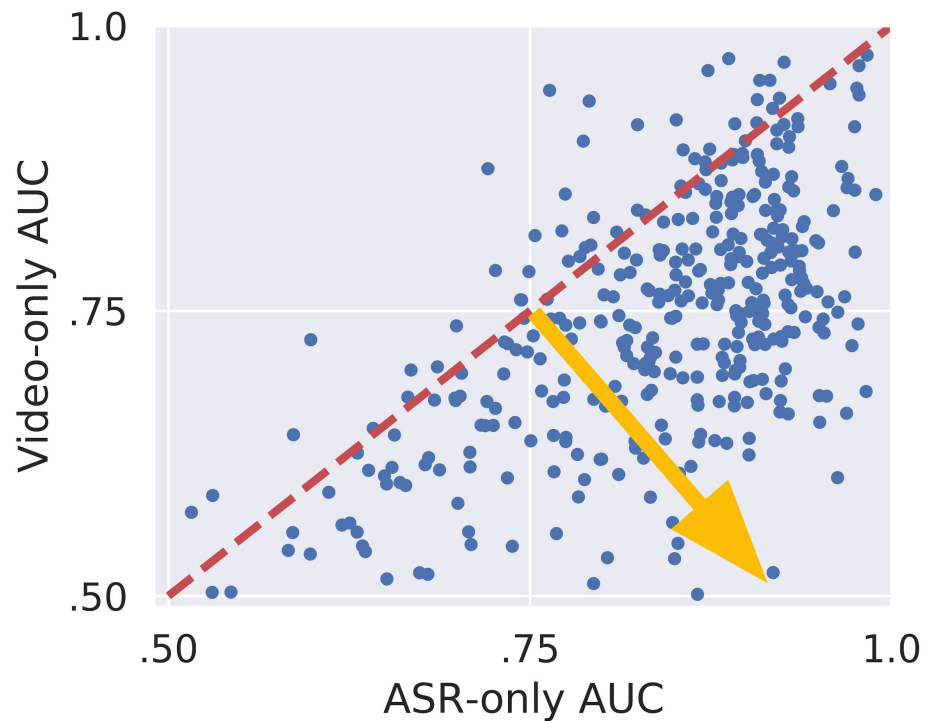
# Per-word performance



# Per-word performance



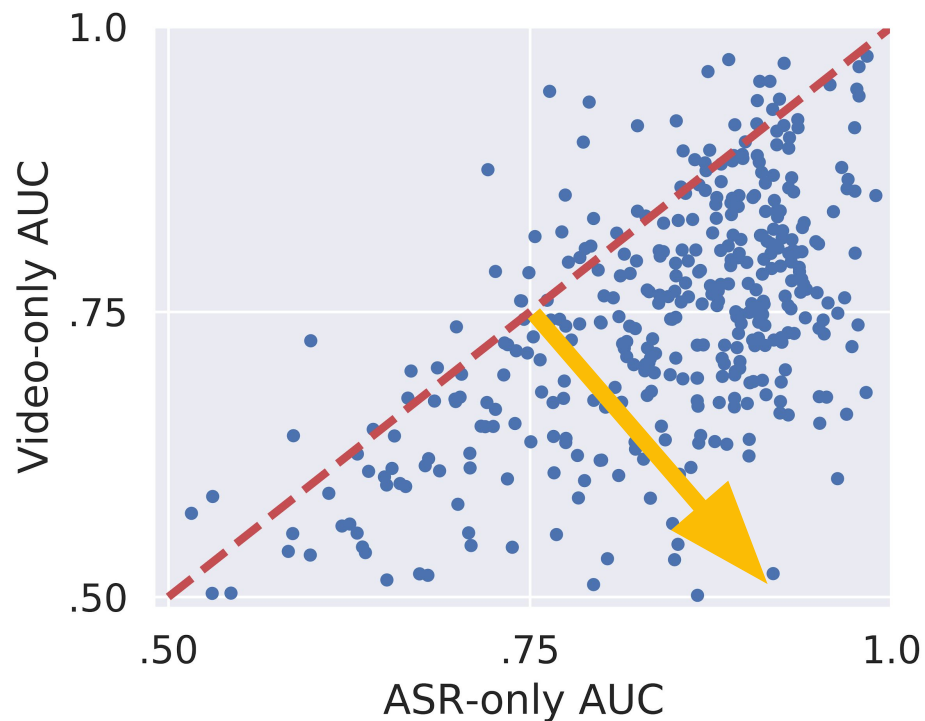
# Per-word performance



## ASR Better:

fat	39.7
turn	36.4
sea	35.9
white	31.7
chilies	30.8
dried	30.6
beer	30.3
pancetta	30.0
mustard	29.8
spice	28.4
sliced	28.3
cinnamon	28.0
warm	27.8
cream	27.4
yellow	27.0
pepper	26.9
olive	26.3
salt	26.1
soda	25.3
thyme	25.2

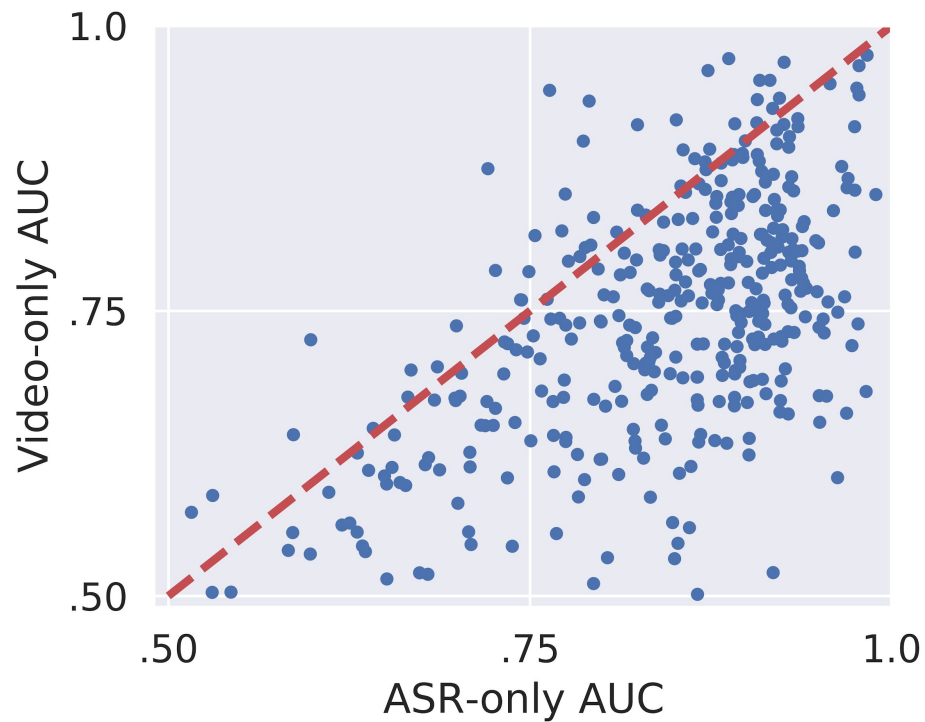
# Per-word performance



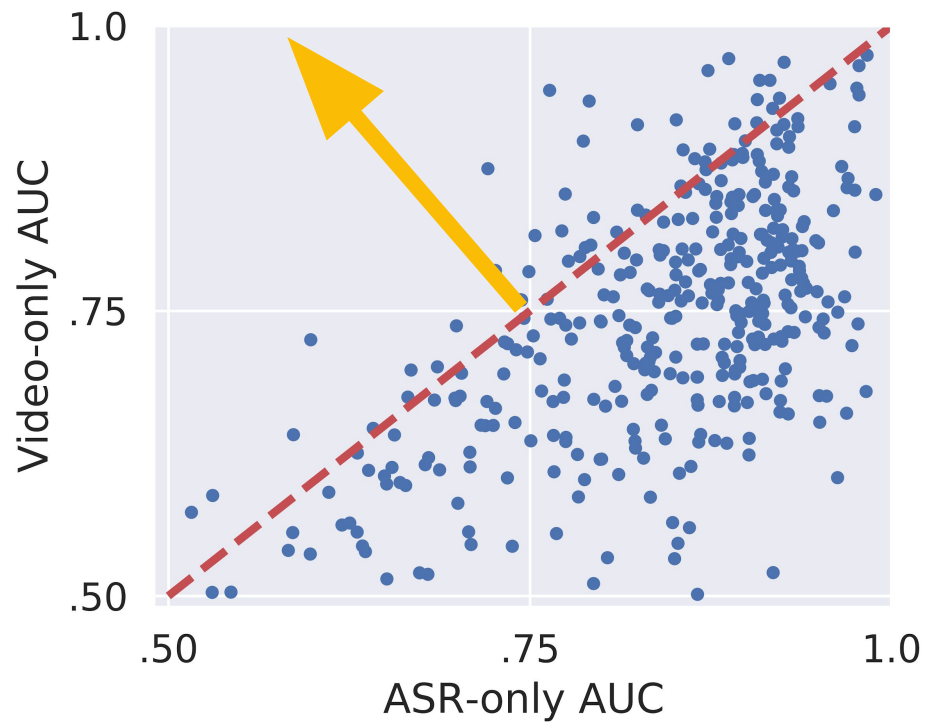
## ASR Better:

fat	39.7
turn	36.4
<b>sea</b>	35.9 (salt)
<b>white</b>	31.7 (pepper)
chilies	30.8
<b>dried</b>	30.6 (chillies)
beer	30.3
pancetta	30.0
mustard	29.8
spice	28.4
sliced	28.3
cinnamon	28.0
warm	27.8
cream	27.4
yellow	27.0
pepper	26.9
<b>olive</b>	26.3 (oil)
salt	26.1
soda	25.3
thyme	25.2

# Per-word performance

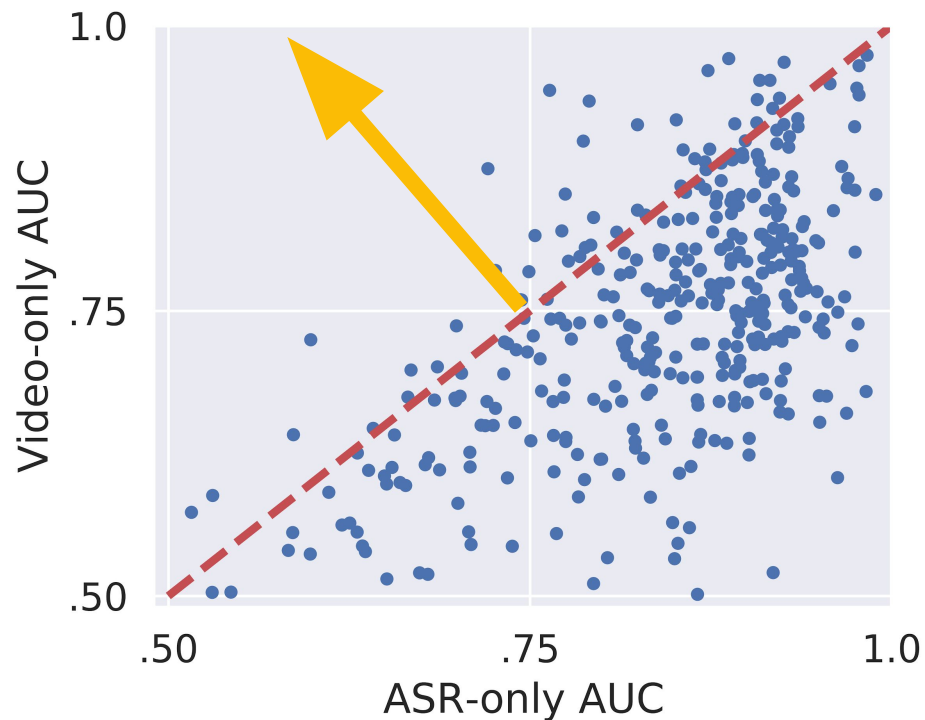


# Per-word performance





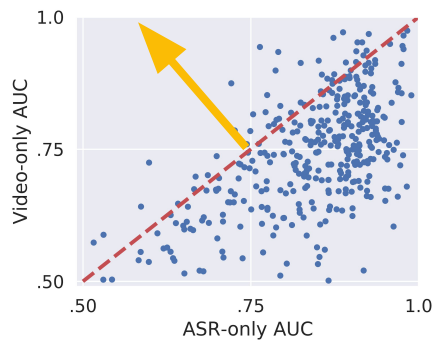
# Per-word performance



## Video Better:

sandwich	18.0
stove	15.4
tuna	14.3
again	12.6
mince	11.2
wok	8.9
burger	8.8
pizza	8.4
serve	7.8
4	6.6
mussels	6.6
tray	6.3
bowl	5.9
or	5.8
mixed	5.7
now	5.5
plate	4.8
dal	4.3
pancake	4.2
beef	3.8

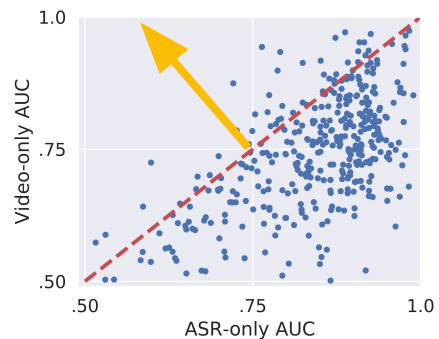
# Per-word performance



## Video Better:

sandwich	18.0
stove	15.4
tuna	14.3
again	12.6
mince	11.2
wok	8.9
burger	8.8
pizza	8.4
serve	7.8
4	6.6
mussels	6.6
tray	6.3
bowl	5.9
or	5.8
mixed	5.7
now	5.5
plate	4.8
dal	4.3
pancake	4.2
beef	3.8

# Per-word performance



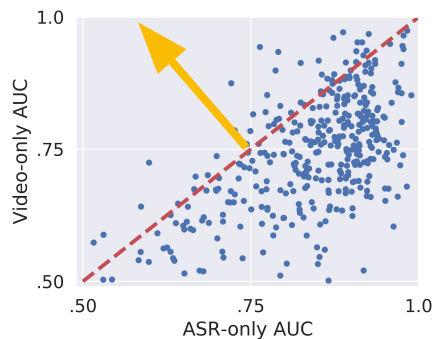
## Video Better:

sandwich	18.0
stove	15.4
tuna	14.3
again	12.6
mince	11.2
wok	8.9
burger	8.8
pizza	8.4
serve	7.8
4	6.6
mussels	6.6
tray	6.3
bowl	5.9
or	5.8
mixed	5.7
now	5.5
plate	4.8
dal	4.3
pancake	4.2
beef	3.8

Concrete, frequent, and regularly **unstated**

# Per-word performance

Concrete, frequent, and regularly unstated



## Video Better:

sandwich	18.0
<b>stove</b>	<b>15.4</b>
tuna	14.3
again	12.6
mince	11.2
<b>wok</b>	<b>8.9</b>
burger	8.8
pizza	8.4
<b>serve</b>	<b>7.8</b>
4	6.6
mussels	6.6
tray	6.3
<b>bowl</b>	<b>5.9</b>
or	5.8
mixed	5.7
now	5.5
<b>plate</b>	<b>4.8</b>
dal	4.3
pancake	4.2
beef	3.8

... that's  
perfection in my  
book right there  
that's...

“put the dish on a **plate** and  
**serve**”

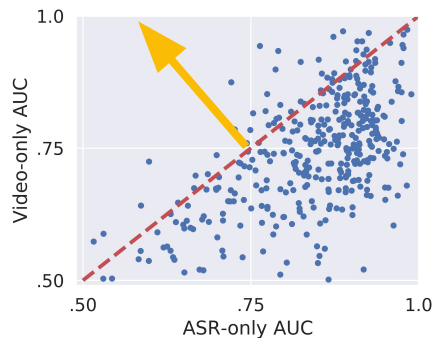
... this knife is  
so sharp...

“cut the tomatoes into pieces add it  
to the **bowl**”

... that's done  
take it off the  
heat and set...

“heat it to a boil and remove it from  
**stove** to cool it down”

# Per-word performance



## Video Better:

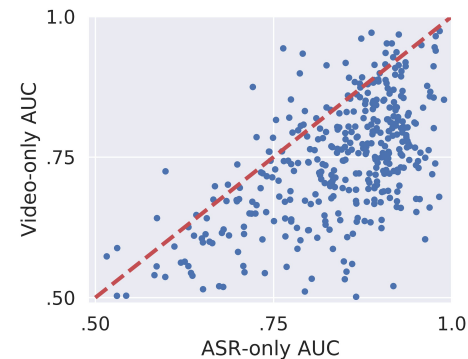
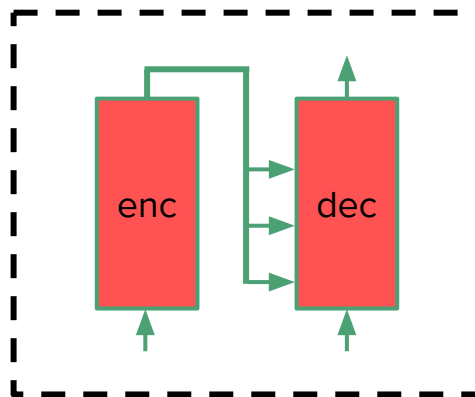
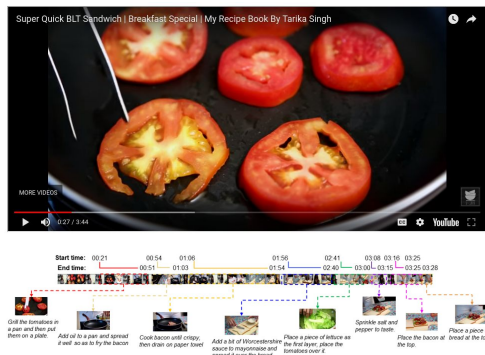
sandwich 18.0  
**stove 15.4**  
tuna 14.3  
again 12.6  
mince 11.2  
**wok 8.9**  
burger 8.8  
pizza 8.4  
**serve 7.8**  
4 6.6  
mussels 6.6  
tray 6.3  
**bowl 5.9**  
or 5.8  
mixed 5.7  
now 5.5  
**plate 4.8**  
dal 4.3  
pancake 4.2  
beef 3.8

Concrete, frequent, and regularly **unstated**

$$\frac{P(\text{appears in caption but not ASR})}{P(\text{appears in caption})}$$

	Unstated % (percentile)
stove	78.1% (90.2)
wok	89.6% (97.9)
plate	75.6% (86.6)
sandwich	74.6% (85.6)
serve	79.0% (91.2)
bowl	73.1% (82.7)

# Thanks!

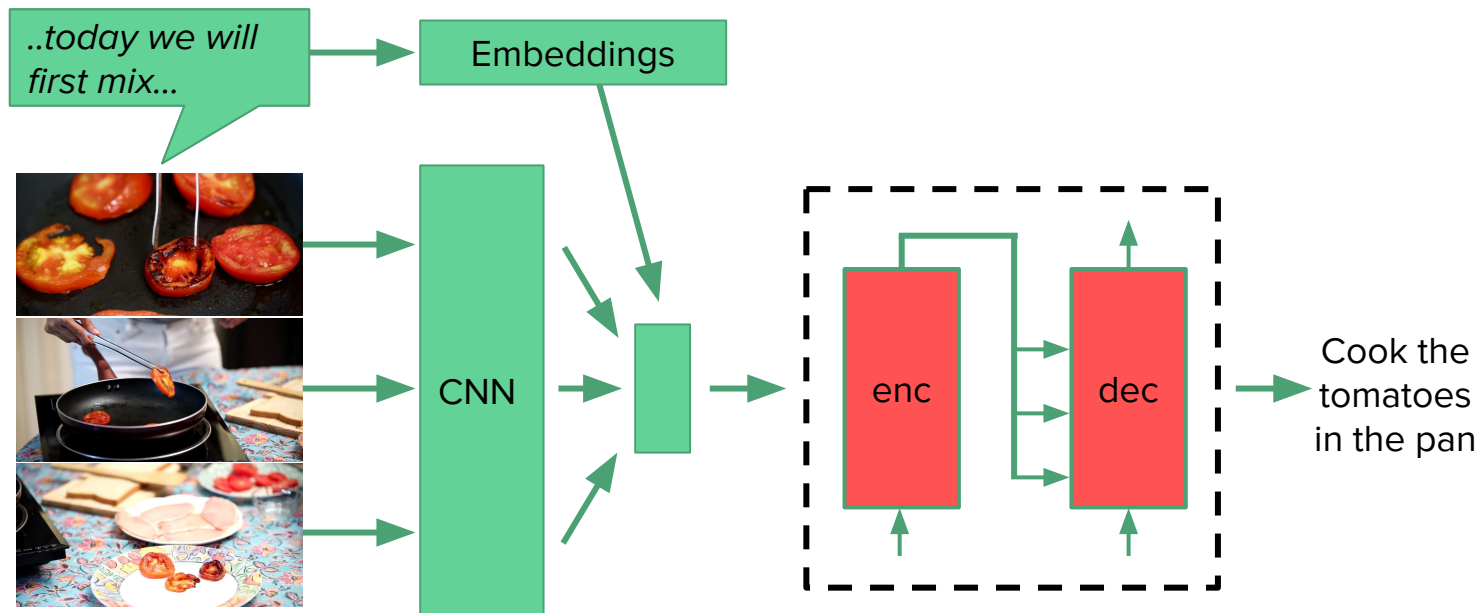


Contact:  
jmhessel@gmail.com  
@jmhessel on Twitter

<http://www.cs.cornell.edu/~jhessel/>

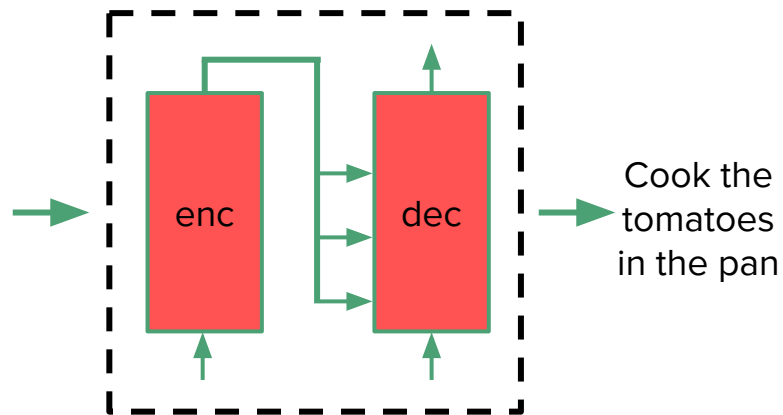
Extras:

# Is object detection "enough"?



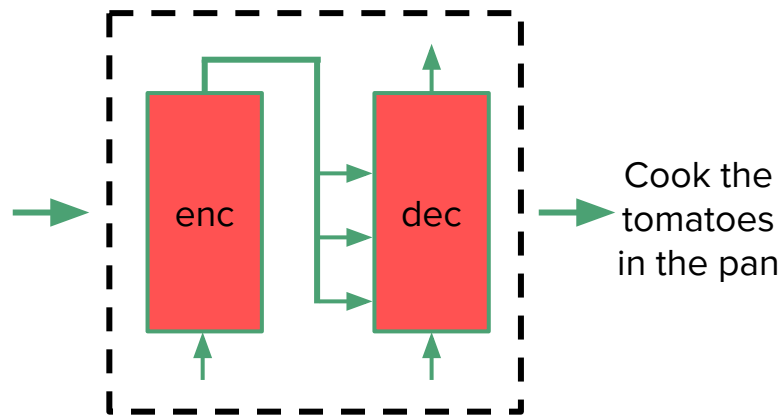


Is object detection "enough"?

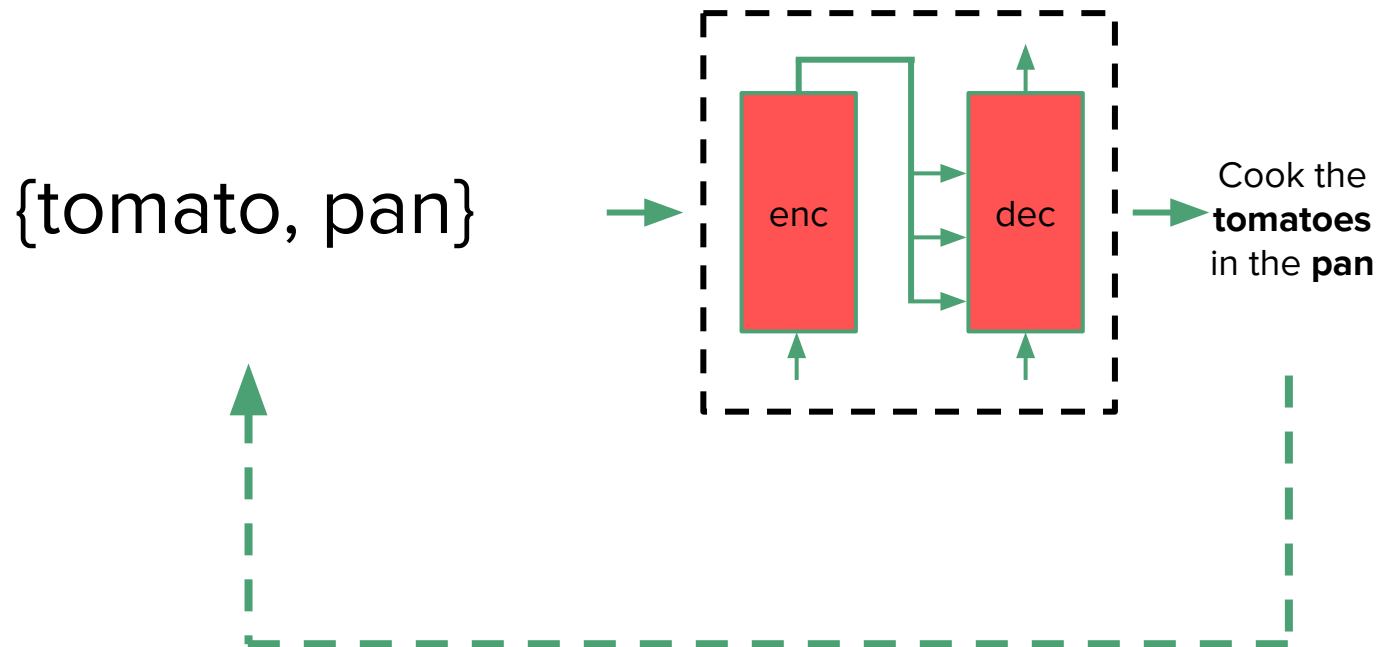


Is object detection "enough"?

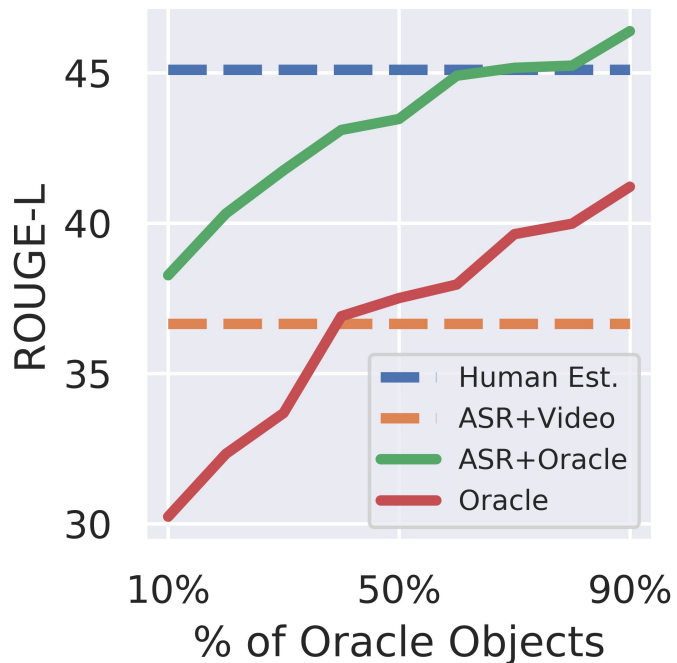
{tomato, pan}



Is object detection "enough"?



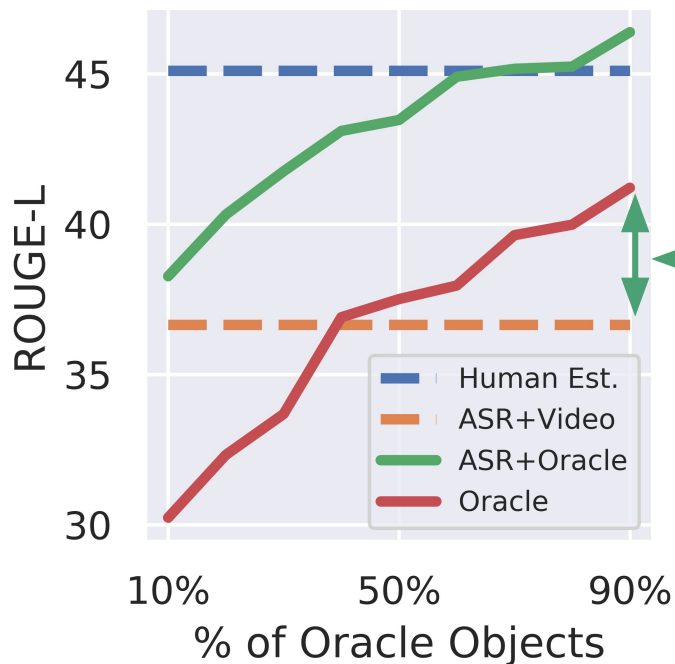
# Is object detection "enough"?



**Stronger oracle**



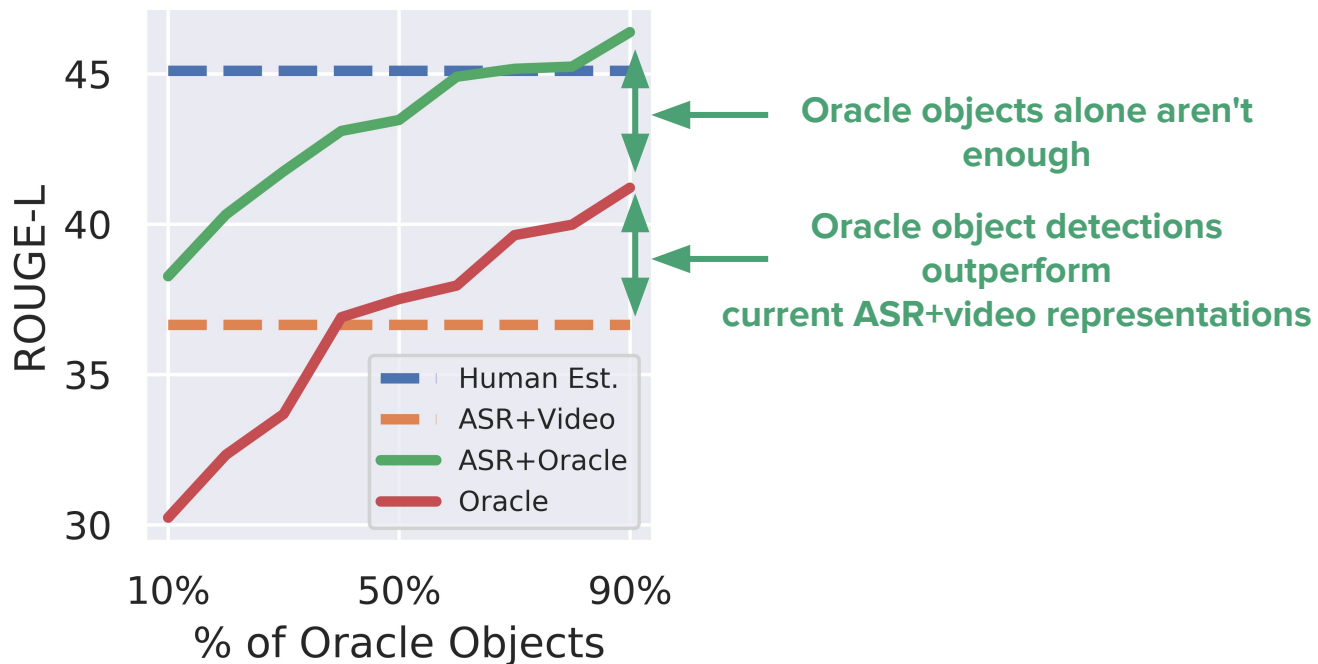
# Is object detection "enough"?



Oracle object detections  
outperform  
current ASR+video representations

**Stronger oracle**

# Is object detection "enough"?



**Stronger oracle**

