JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, Cordelia Schmid

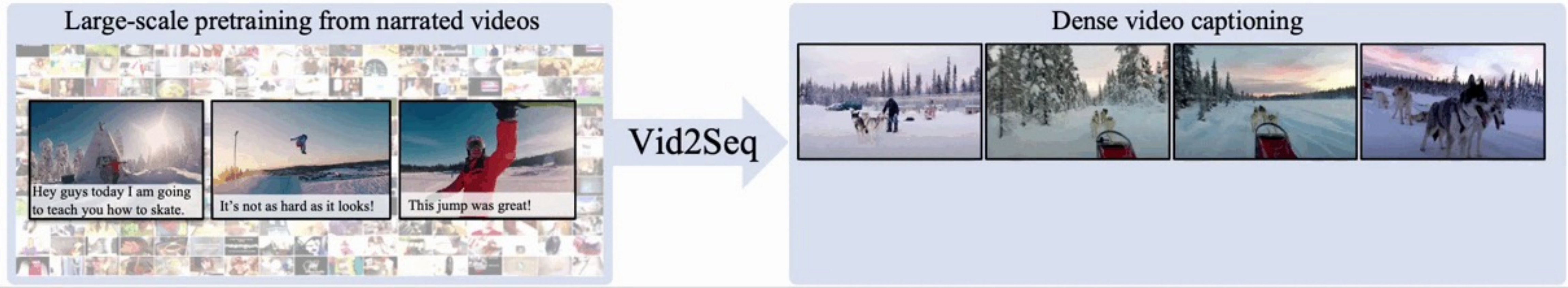
Poster: Wed-AM-237

Project page: https://antoyang.github.io/vid2seq.html

Paper: https://arxiv.org/abs/2302.14115

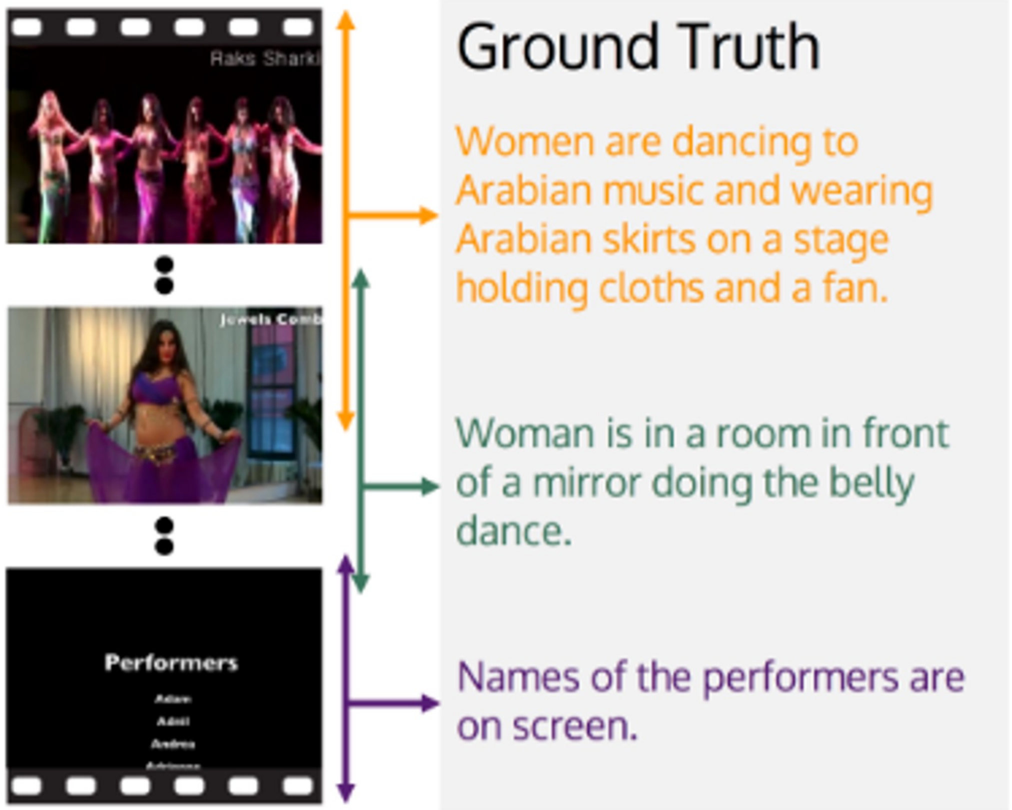
Vid2Seq overview

- Vid2Seq is a visual language model for dense video captioning.
- Vid2Seq is pretrained on millions of unlabeled narrated videos.
- Vid2Seq achieves SoTA on various captioning tasks.


Large-scale pretraining from narrated videos
Hey guys today I am going to teach you how to skate.
It's not as hard as it looks!
This jump was great!
Vid2Seq
Dense video captioning

Dense Video Captioning

- **Task:** generate temporally localized captions for all events in an untrimmed minutes-long video.
- **Prior approaches (e.g. [Wang 2021]):** are task specific and trained only on manually annotated datasets.



Ground Truth

Women are dancing to Arabian music and wearing Arabian skirts on a stage holding cloths and a fan.

Woman is in a room in front of a mirror doing the belly dance.

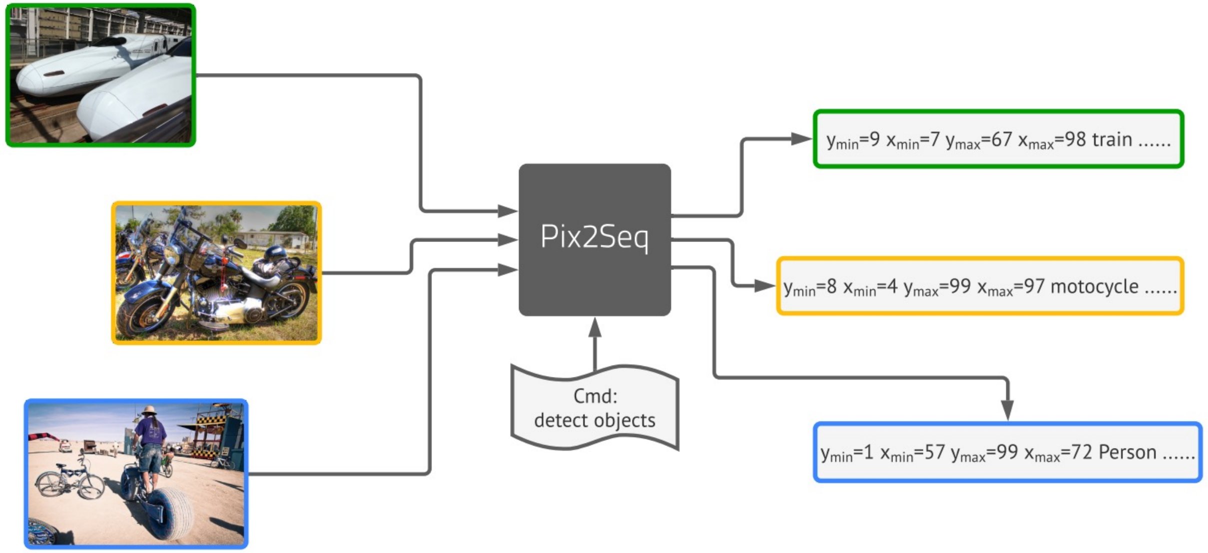
Names of the performers are on screen.

Example from the ActivityNet-Captions dataset [Krishna 2017].

[Krishna 2017] Dense-Captioning Events in Videos, Ranjay Krishna et al, ICCV 2017.
[Wang 2021] End-to-End Dense Video Captioning with Parallel Decoding, Teng Wang et al, ICCV 2021.

Localization as language modeling

- Pix2seq [Chen 2022] casts object detection as sequence generation.
- Spatial coordinates are quantized and tokenized.



Pix2Seq

Cmd: detect objects

y_{min}=9 x_{min}=7 y_{max}=67 x_{max}=98 train ......

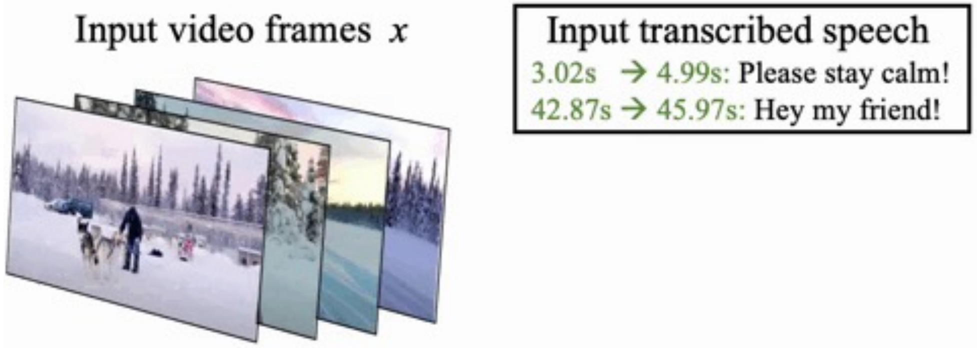
y_{min}=8 x_{min}=4 y_{max}=99 x_{max}=97 motocycle ......

y_{min}=1 x_{min}=57 y_{max}=99 x_{max}=72 Person ......

[Chen 2022] Pix2seq: A Language Modeling Framework for Object Detection, Ting Chen et al, ICLR 2022.

The Vid2Seq model

- Formulates dense video captioning as a sequence-to-sequence problem.
- Time is quantized and jointly tokenized with the text.
- **Model architecture:** visual encoder, text encoder and text decoder.

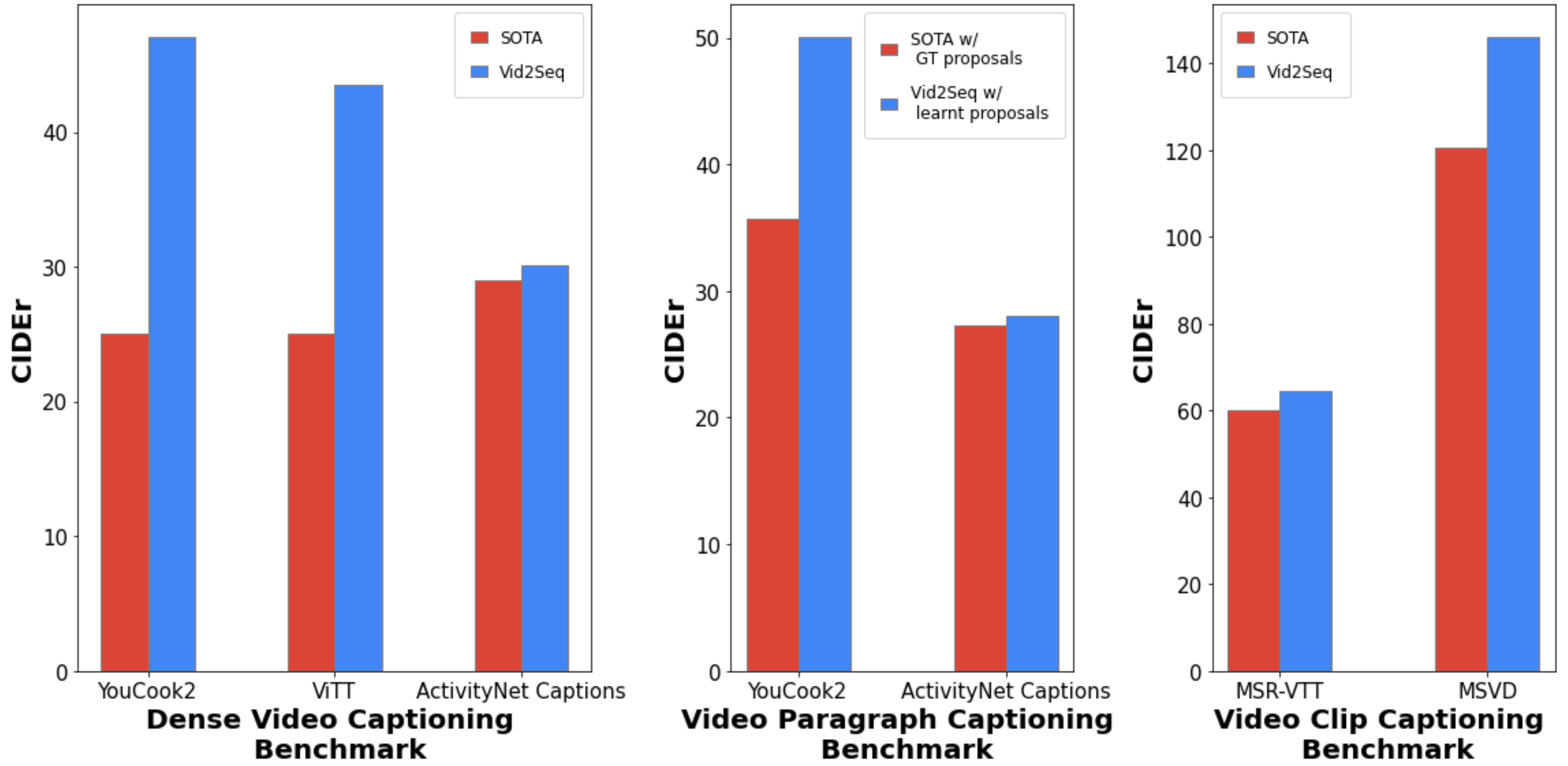

Input video frames x
Input transcribed speech
3.02s → 4.99s: Please stay calm!
42.87s → 45.97s: Hey my friend!

Pretraining Vid2Seq on untrimmed narrated videos

- Speech is also cast as a single sequence of text and time tokens.
- **Generative objective:** given visual inputs, predict speech.
- **Denoising objective:** given visual inputs and noisy speech, predict masked tokens.

Input video frames x

Vid2Seq improves the SoTA on video captioning tasks.



[Wang 2021] End-to-End Dense Video Captioning with Parallel Decoding, Teng Wang et al, ICCV 2021.
[Zhu 2022] End-to-end Dense Video Captioning as Sequence Generation, Wanrong Zhu et al, COLING 2022.
[Lei 2020] MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning, Jie Lei et al, ACL 2020.
[Seo 2022] End-to-end Generative Pretraining for Multimodal Video Captioning, Paul Hongsuck Seo et al, CVPR 2022.
[Lin 2022] SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning, Kevin Lin et al, CVPR 2022.

Vid2Seq generalizes well to few-shot settings.

We also find that pretraining is crucial for few-shot generalization.

Data	**YouCook2**			**ViTT**			**ActivityNet Captions**		
	SODA	CIDEr	METEOR	SODA	CIDEr	METEOR	SODA	CIDEr	METEOR
1%	2.4	10.1	3.3	2.0	7.4	1.9	2.2	6.2	3.2
10%	3.8	18.4	5.2	10.7	28.6	6.0	4.3	20.0	6.1
50%	6.2	32.1	7.6	12.5	38.8	7.8	5.4	27.5	7.8
100%	**7.9**	**47.1**	**9.3**	**13.5**	**43.5**	**8.5**	**5.8**	**30.1**	**8.5**

Benefits of pretraining on untrimmed videos

Unlike standard video captioning pretrained models, Vid2Seq is pretrained on *untrimmed* narrated videos (where speech sentences are split by the time tokens).

Pretraining input		**YouCook2**			**ActivityNet Captions**		
Untrimmed	Time tokens	SODA	CIDEr	F1	SODA	CIDEr	F1
✗	✗	4.0	18.0	18.1	5.4	18.8	49.2
✓	✗	5.5	27.8	20.5	5.5	26.5	52.1
✓	✓	**7.9**	**47.1**	**27.3**	**5.8**	**30.1**	**52.4**

Effect of pretraining losses and modalities

The visual inputs only model benefits from the generative objective.

The denoising objective helps the model with visual+speech inputs.

Finetuning Input		Pretraining losses		YouCook2			ActivityNet Captions		
Visual	Speech	Generative	Denoising	SODA	CIDEr	F1	SODA	CIDEr	F1
✓	✗	No pretraining		3.0	15.6	15.4	5.4	14.2	46.5
✓	✓	No pretraining		4.0	18.0	18.1	5.4	18.8	49.2
✓	✗	✓	✗	5.7	25.3	23.5	**5.9**	**30.2**	51.8
✓	✓	✓	✗	2.5	10.3	15.9	4.8	17.0	48.8
✓	✓	✓	✓	**7.9**	**47.1**	**27.3**	5.8	30.1	**52.4**

Captioning helps localization after pretraining.

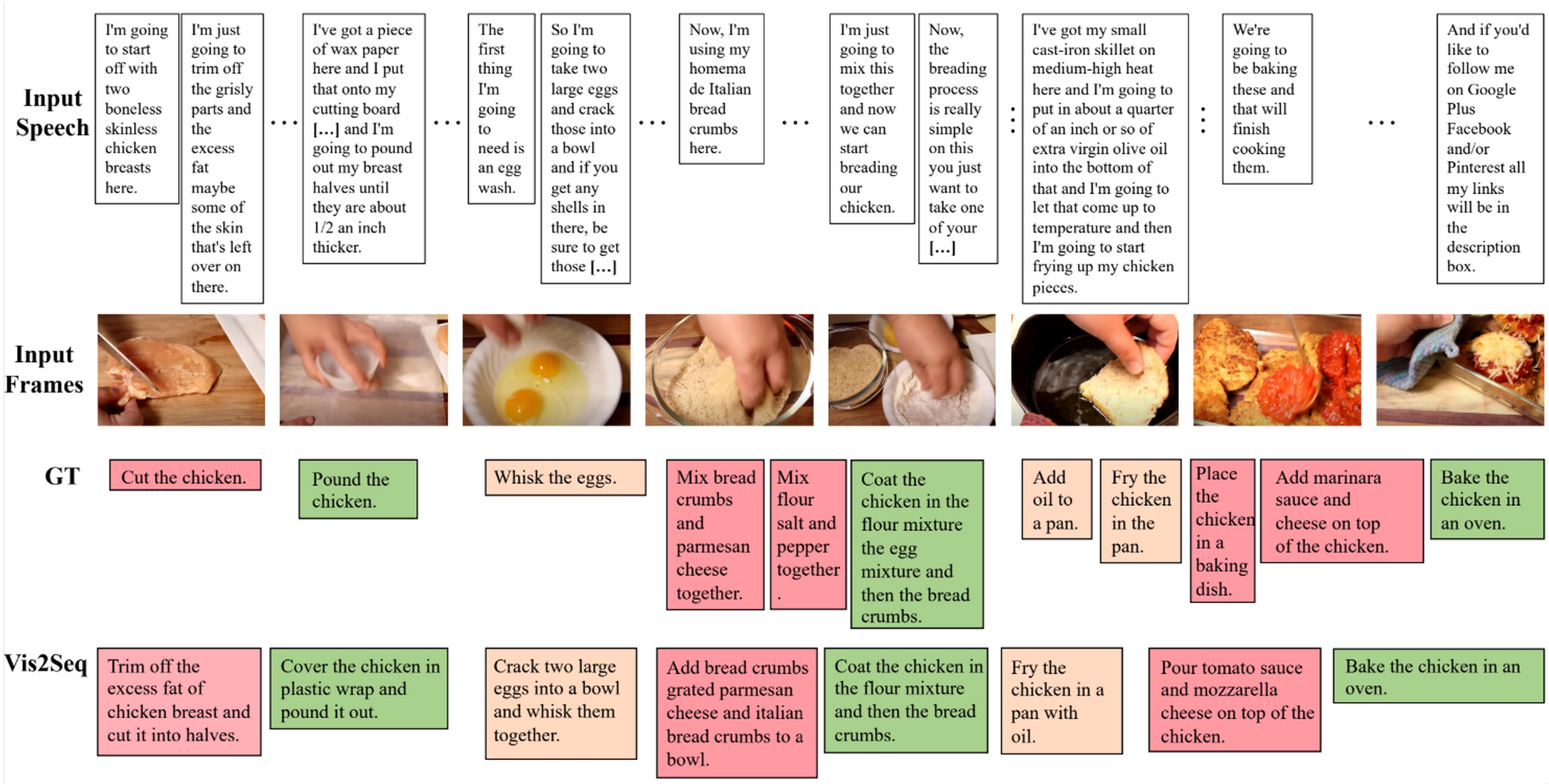
Contextualizing the noisy speech boundaries with their semantic content is important.

Captioning	Pretraining	YouCook2			ActivityNet Captions		
		Recall	Precis.	F1	Recall	Precis.	F1
✗	✗	17.8	19.4	17.7	47.3	57.9	52.0
✓	✗	17.2	20.6	18.1	42.5	**64.1**	49.2
✗	✓	25.7	21.4	22.8	52.5	53.0	51.1
✓	✓	**27.9**	**27.8**	**27.3**	**52.7**	53.9	**52.4**

Data and model scaling.

Language Model	Pretraining		YouCook2			ActivityNet Captions		
	# Videos	Dataset	SODA	CIDEr	F1	SODA	CIDEr	F1
T5-Small	15M	YTT	6.1	31.1	24.3	5.5	26.5	52.2
T5-Base	0	-	4.0	18.0	18.1	5.4	18.8	49.2
T5-Base	15K	YTT	6.3	35.0	24.4	5.1	24.4	49.9
T5-Base	150K	YTT	7.3	40.1	26.7	5.4	27.2	51.3
T5-Base	1M5	YTT	7.8	45.5	26.8	5.6	28.7	52.2
T5-Base	1M	HTM	**8.3**	**48.3**	26.6	5.8	28.8	**53.1**
T5-Base	15M	YTT	7.9	47.1	**27.3**	5.8	30.1	52.4

Qualitative results



Input Speech

I'm going to start off with two boneless skinless chicken breasts here.

I'm just going to trim off the grisly parts and the excess fat maybe some of the skin that's left over on there.

…

I've got a piece of wax paper here and I put that onto my cutting board [...] and I'm going to pound out my breast halves until they are about 1/2 an inch thicker.

…

The first thing I'm going to need is an egg wash.

So I'm going to take two large eggs and crack those into a bowl and if you get any shells in there, be sure to get those [...]

…

Now, I'm using my homema de Italian bread crumbs here.

…

I'm just going to mix this together and now we can start breading our chicken.

Now, the breading process is really simple on this you just want to take one of your [...]

⋮

I've got my small cast-iron skillet on medium-high heat here and I'm going to put in about a quarter of an inch or so of extra virgin olive oil into the bottom of that and I'm going to let that come up to temperature and then I'm going to start frying up my chicken pieces.

⋮

We're going to be baking these and that will finish cooking them.

…

And if you'd like to follow me on Google Plus Facebook and/or Pinterest all my links will be in the description box.

Input Frames

GT

Cut the chicken.

Pound the chicken.

Whisk the eggs.

Mix bread crumbs and parmesan cheese together.

Mix flour salt and pepper together.

Coat the chicken in the flour mixture the egg mixture and then the bread crumbs.

Add oil to a pan.

Fry the chicken in the pan.

Place the chicken in a baking dish.

Add marinara sauce and cheese on top of the chicken.

Bake the chicken in an oven.

Vis2Seq

Trim off the excess fat of chicken breast and cut it into halves.

Cover the chicken in plastic wrap and pound it out.

Crack two large eggs into a bowl and whisk them together.

Add bread crumbs grated parmesan cheese and italian bread crumbs to a bowl.

Coat the chicken in the flour mixture and then the bread crumbs.

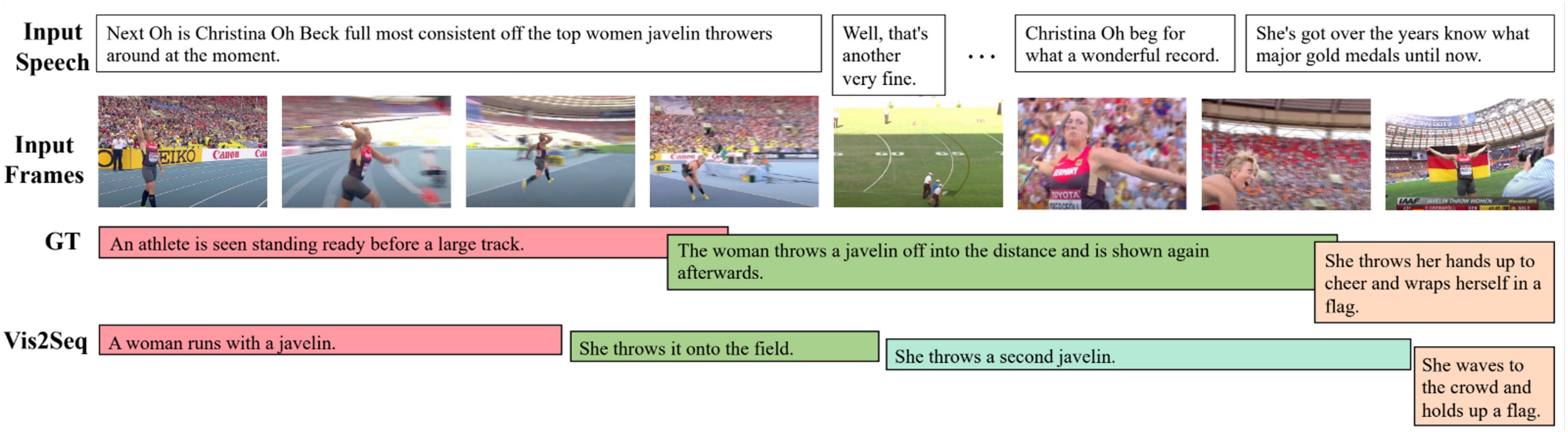
Fry the chicken in a pan with oil.

Pour tomato sauce and mozzarella cheese on top of the chicken.

Bake the chicken in an oven.

Qualitative results

More examples at: https://www.youtube.com/watch?v=3oEHSU5ExsI



Input Speech

Next Oh is Christina Oh Beck full most consistent off the top women javelin throwers around at the moment.

Well, that's another very fine.

. . .

Christina Oh beg for what a wonderful record.

She's got over the years know what major gold medals until now.

Input Frames

GT

An athlete is seen standing ready before a large track.

The woman throws a javelin off into the distance and is shown again afterwards.

She throws her hands up to cheer and wraps herself in a flag.

Vis2Seq

A woman runs with a javelin.

She throws it onto the field.

She throws a second javelin.

She waves to the crowd and holds up a flag.

Conclusion

- Vid2Seq is a visual language model for dense video captioning.
- Vid2Seq can be effectively pretrained on unlabeled narrated videos at scale.
- The pretrained Vid2Seq model improves the SoTA on 3 dense video captioning datasets, 2 video paragraph captioning datasets, 2 video clip captioning datasets, and generalizes well to few-shot setting.