Learning visual language models for video understanding

Antoine Yang
https://antoyang.github.io/

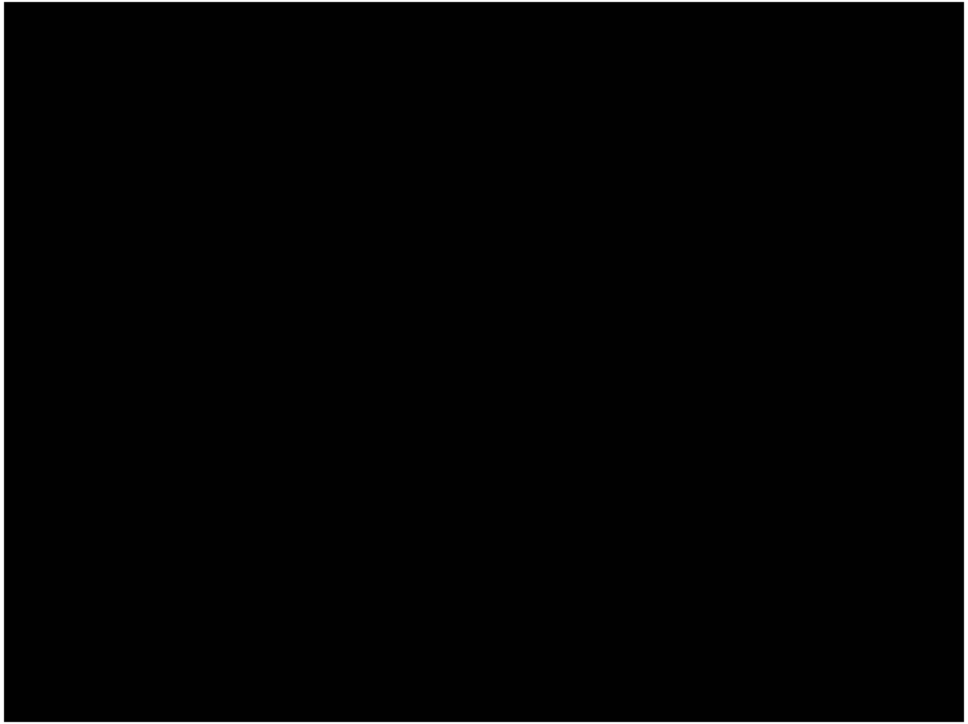
Visual language models

- Language is a fundamental aspect of human communication
- Vision is a fundamental aspect of human perception

-> Developing machines that can process both is crucial e.g. for human-computer interaction, search, customer support, accessibility…



Example of a visually-aware chatbot

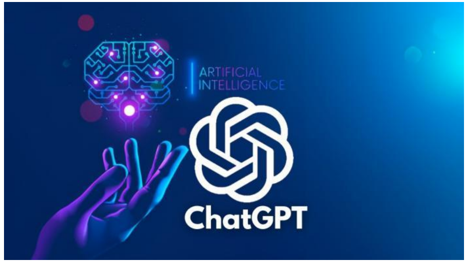


What are they doing? -> Martial arts

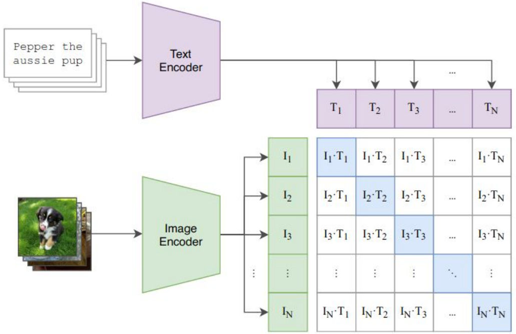
How many men are there? -> 2

What does a machine need to do that?

- Question-answering ability



- Vision-language understanding



Why does the kid trust the man?

Scene understanding is not enough!

Because the man saved his life!

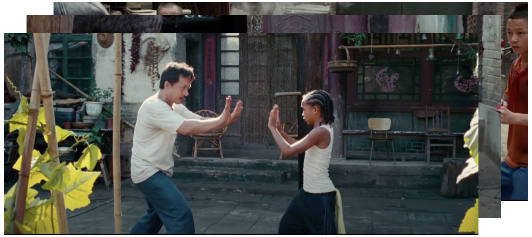
What else do we need?

- Localizing events in time

- Multi-event reasoning

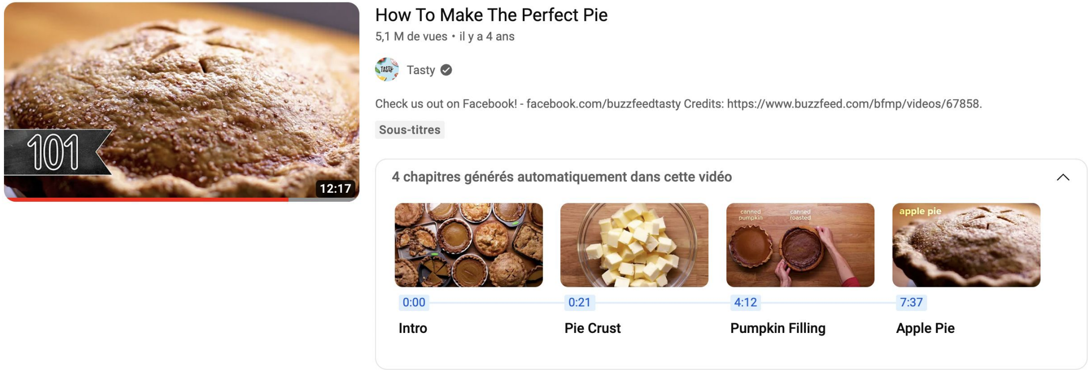
Applications: Beyond answering questions

- Video-to-text summarization



This video is about a kid that learns kung fu. First the kid is attacked by 6 aggressors. A man appears and defeat them, thereby saving the kid’s life. The kid then starts training with the man and becomes stronger day after day. He ends up winning a prestigious competition against his toughest aggressors.

- Improved navigation with automatically generated video chapters



Collaborators

Antoine Miech
(DeepMind)

Josef Sivic
(CIIRC CTU Prague)

Ivan Laptev
(Inria)

Cordelia Schmid
(Inria / Google)

Arsha Nagrani
(Google)

Paul Hongsuck
Seo (Google)

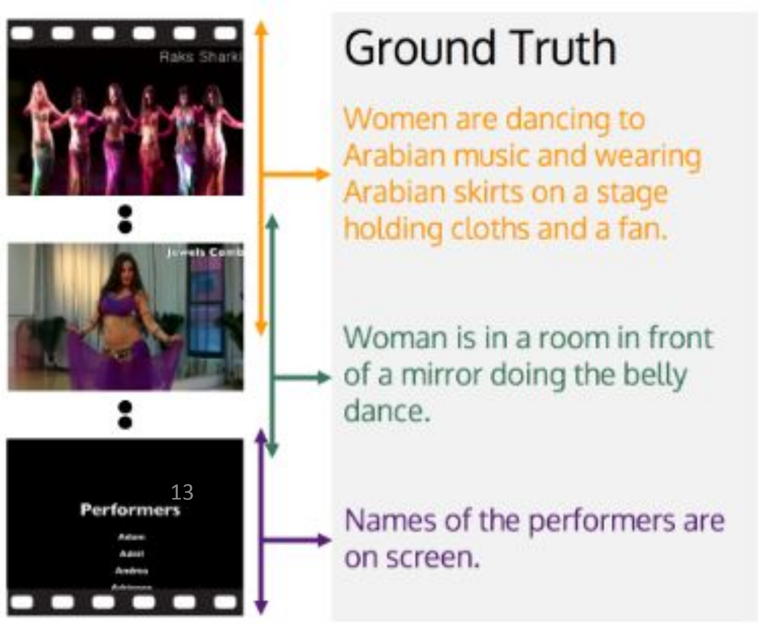
Jordi Pont-Tuset
(Google)

Jean Zay (IDRIS)

Google

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.

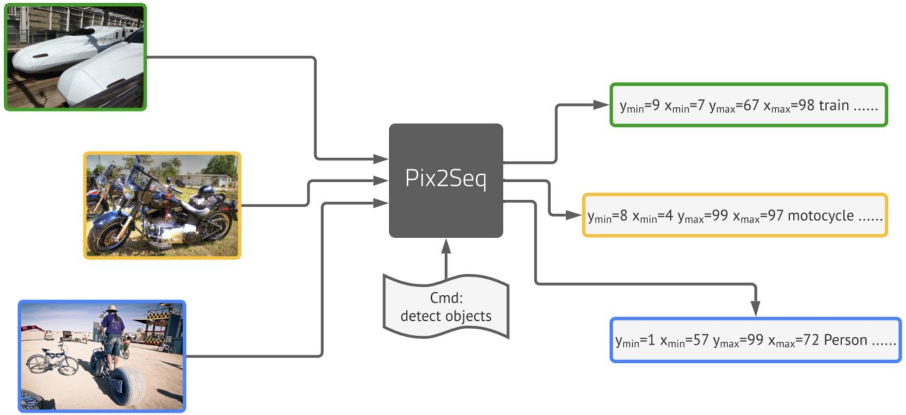


Example from the ActivityNet-Captions dataset [Krishna 2017].

[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

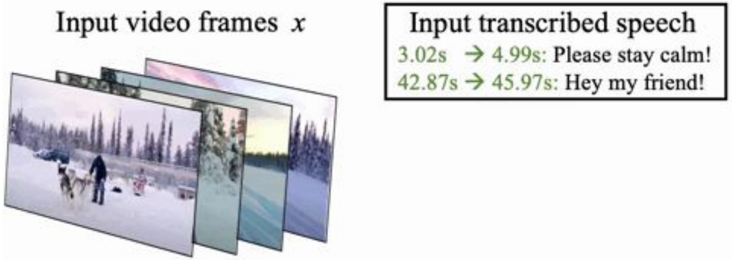
- Pix2seq [Chen 2022] casts object detection as sequence generation.
- Spatial coordinates are quantized and tokenized.



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

The Vid2Seq model

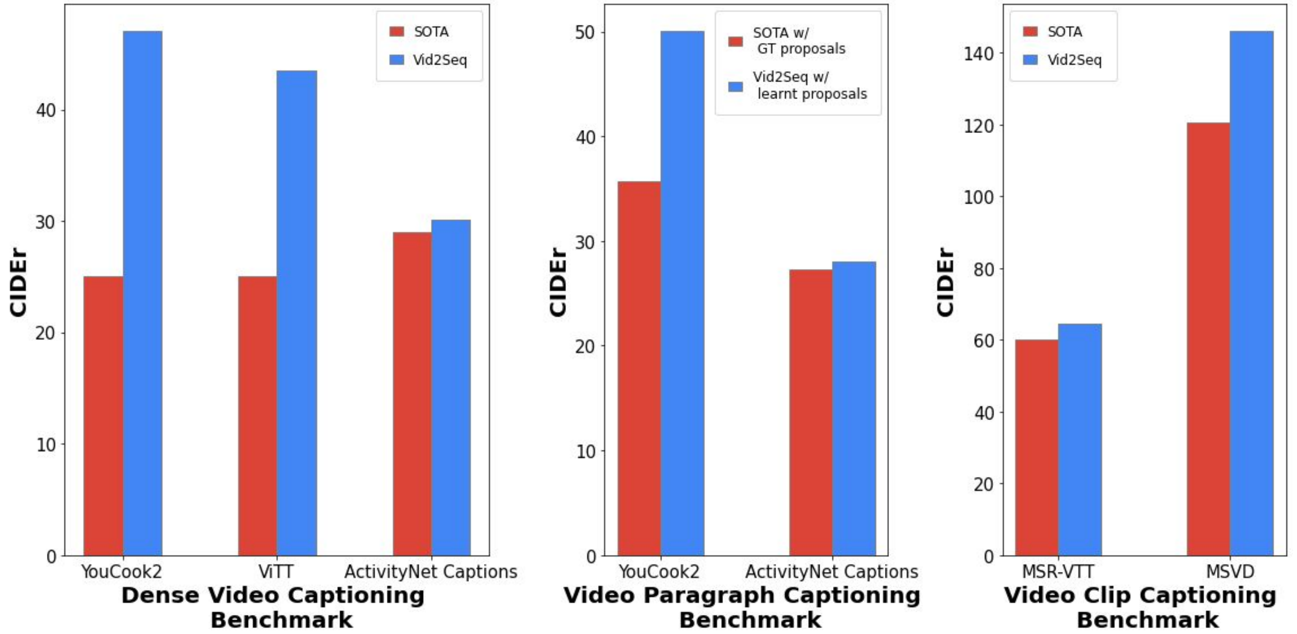
- Formulates dense video captioning as a sequence-to-sequence problem.
- Time is quantized and jointly tokenized with the text.
- **Model architecture:** visual encoder, text encoder and text decoder.



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 speech tokens.

Vid2Seq is 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 has competitive event localization performance without task-specific design.

Model	YouCook2		ViTT		ActivityNet Captions	
	Recall	Precision	Recall	Precision	Recall	Precision
SoTA	20.7	20.6	32.2	32.1	**59.0**	**60.3**
Vid2Seq	**27.9**	**27.8**	**42.6**	**46.2**	52.7	53.9

[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.

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.

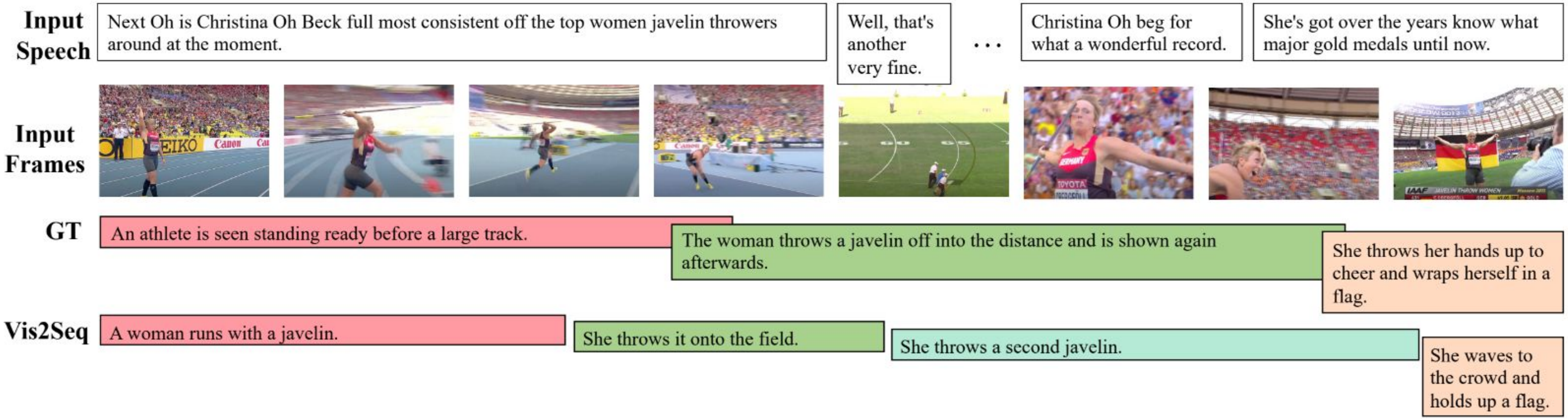
Captioning	Pretraining	YouCook2			ActivityNet Captions		
		Recall	Precis	F1	Recall	Precis.	F1
✗	✗	17.8	19.4	17.7	47.3	57.9	52.0
✓	✗	17.2	20.6	18.1	42.5	**64.1**	49.2
✗	✓	25.7	21.4	22.8	52.5	53.0	51.1
✓	✓	**27.9**	**27.8**	**27.3**	**52.7**	53.9	**52.4**

Data and model scaling.

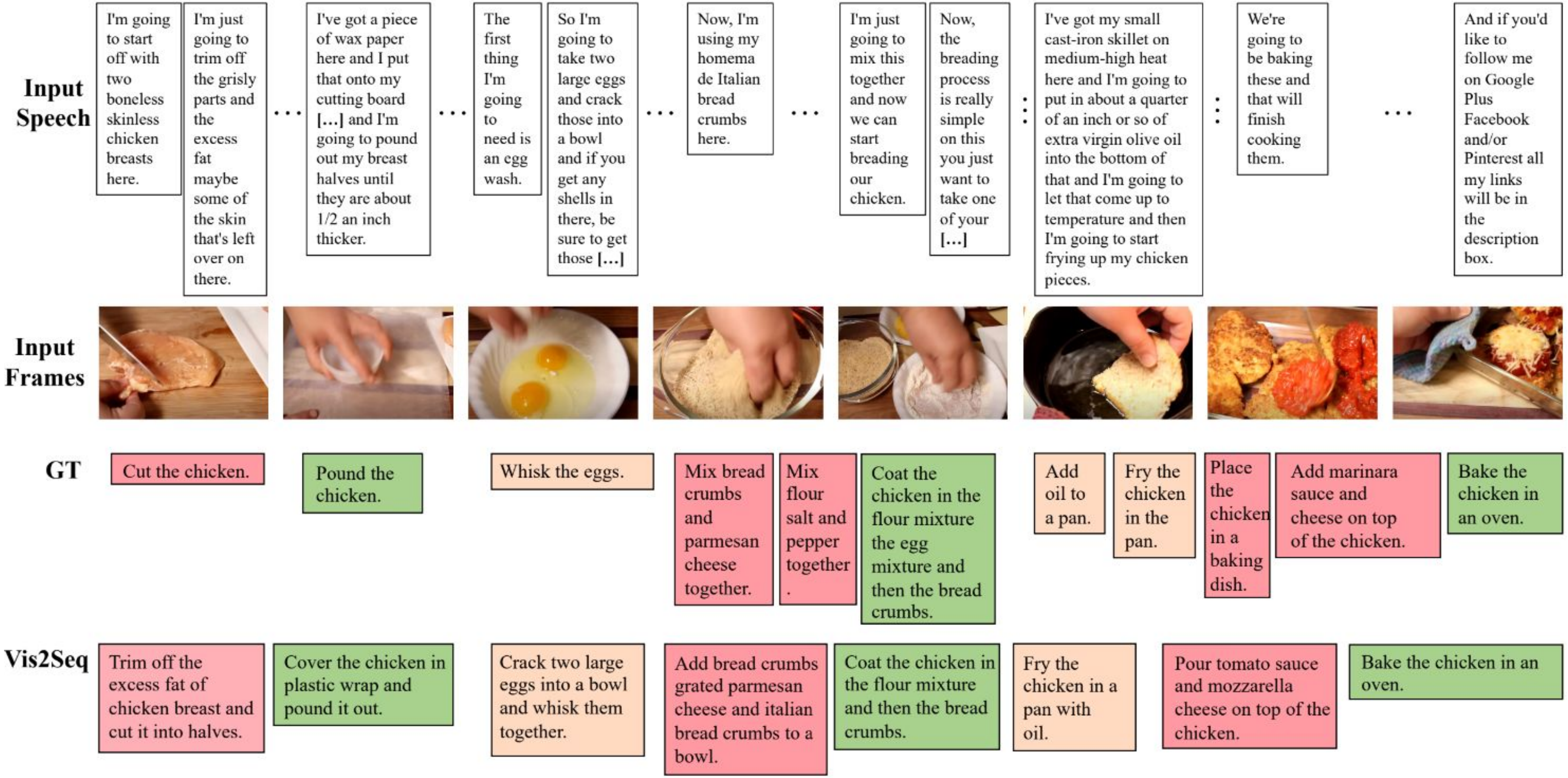
Language Model	Pretraining		YouCook2			ActivityNet Captions		
	# Videos	Dataset	SODA	CIDEr	F1	SODA	CIDEr	F1
T5-Small	15M	YTT	6.1	31.1	24.3	5.5	26.5	52.2
T5-Base	0	-	4.0	18.0	18.1	5.4	18.8	49.2
T5-Base	15K	YTT	6.3	35.0	24.4	5.1	24.4	49.9
T5-Base	150K	YTT	7.3	40.1	26.7	5.4	27.2	51.3
T5-Base	1M5	YTT	7.8	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

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



Qualitative results



Qualitative results

Conclusion

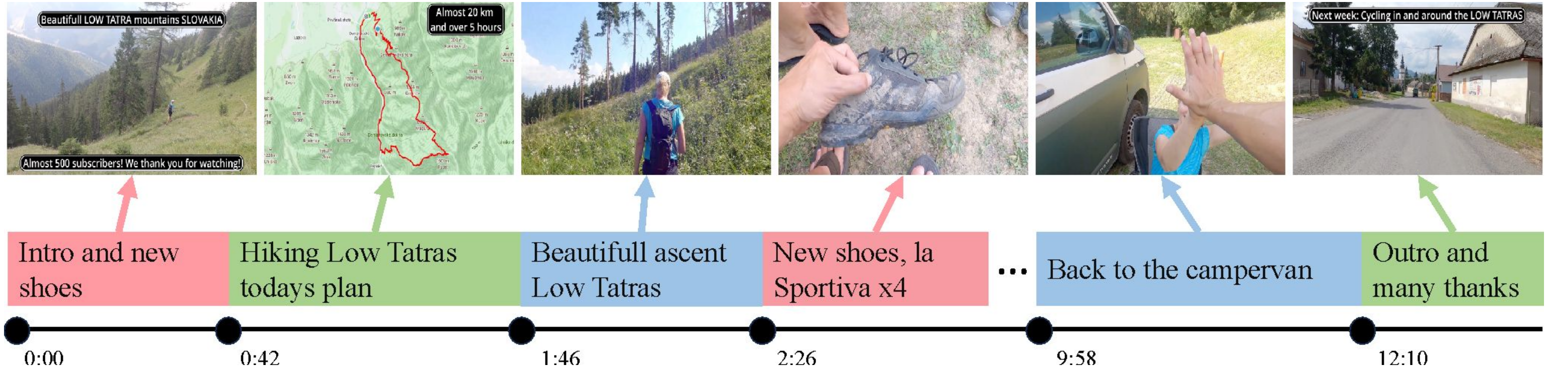
- Vid2Seq is a visual language model for dense video captioning.
- Vid2Seq can be effectively pretrained on unlabeled narrated videos at scale.
- The pretrained Vid2Seq model improves the SoTA on 3 dense video captioning datasets, 2 video paragraph captioning datasets, 2 video clip captioning datasets,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

Limitations

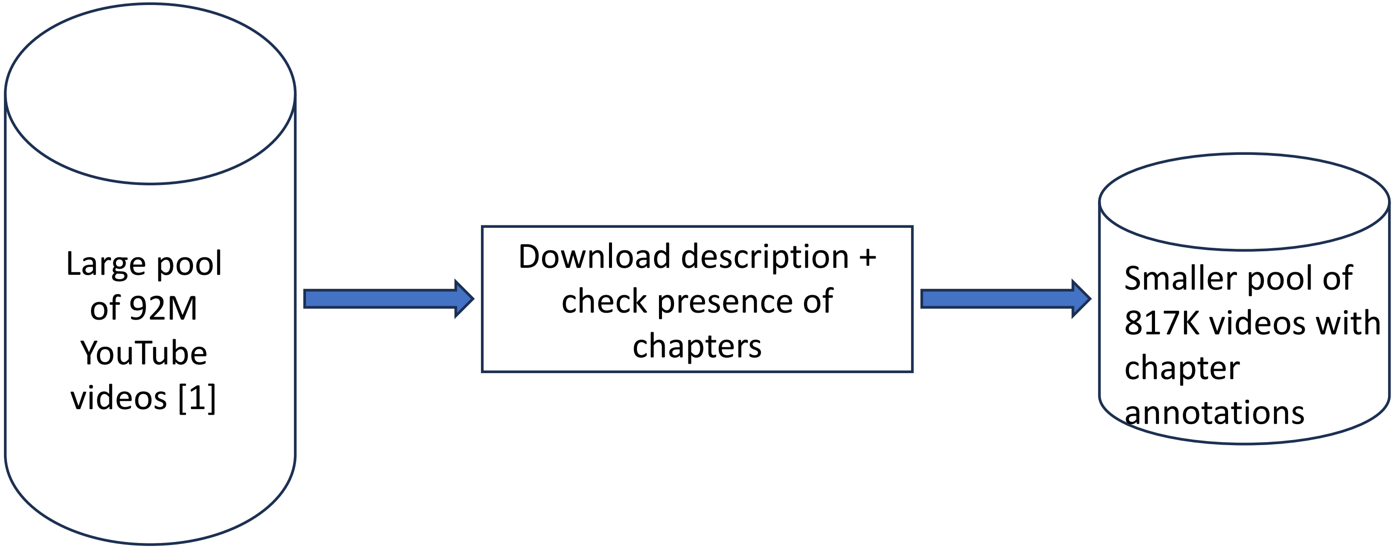
- Vid2Seq cannot use raw audio inputs (beyond speech transcripts).
- Does Vid2Seq generalize to other tasks, e.g. VideoQA or temporal action localization?
- Pretraining gains are subject to video domain -> Vid2Seq event localization performance is below task-specific approaches on ActivityNet Captions.

Video Chapter Generation

- **Goal:** improve navigation in long videos.
- **Task:** segment a long video into segments and generate a chapter title for each.



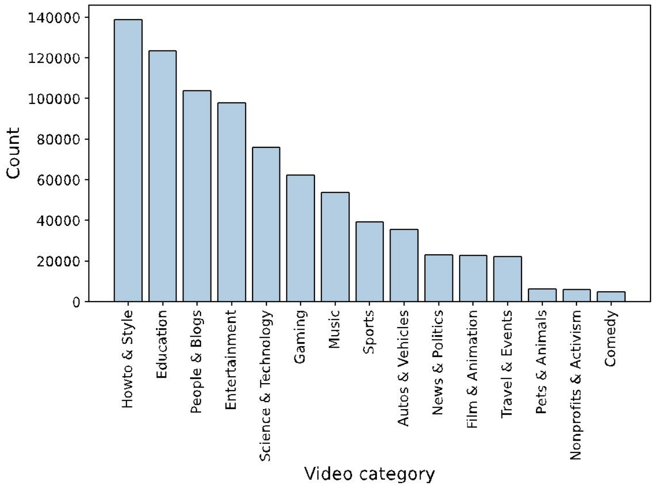
Data collection procedure



[1] MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound, R. Zellers et al, CVPR 2022.

Data statistics

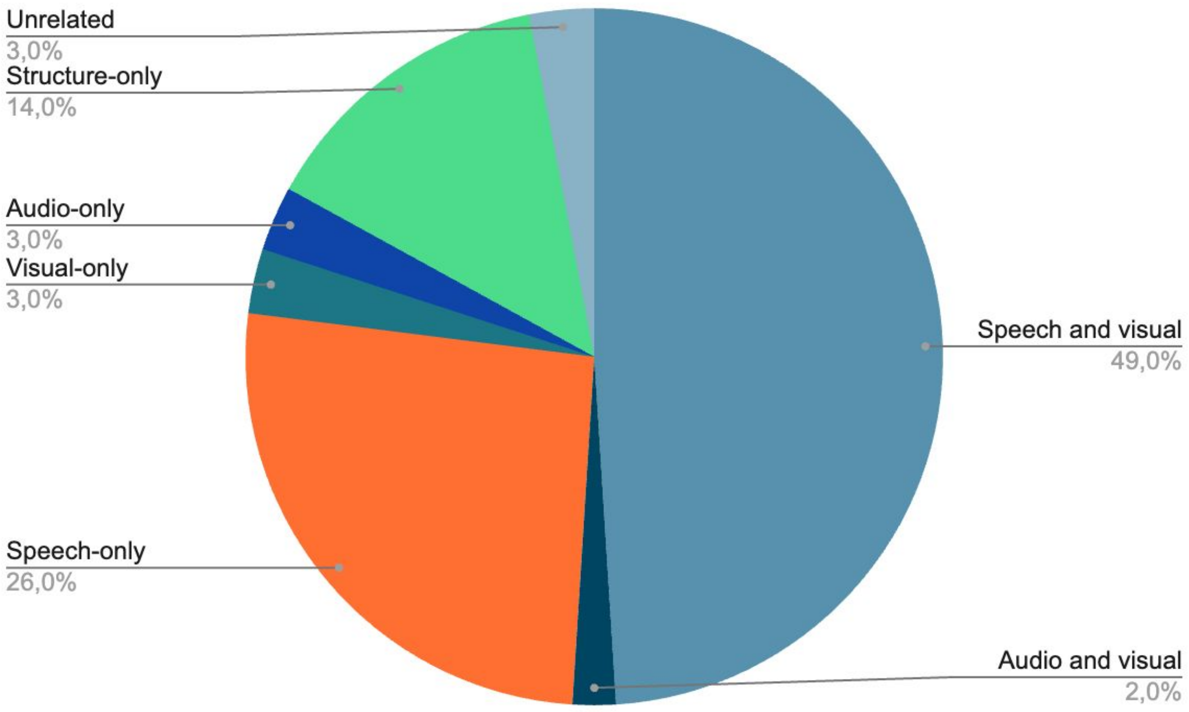
- 817K videos & 7M chapters
- 8 chapters per video (avg)
- Chapter duration (avg): 142s
- Video duration (avg): 1354s
- 97% videos with ASR
- 93% videos in English



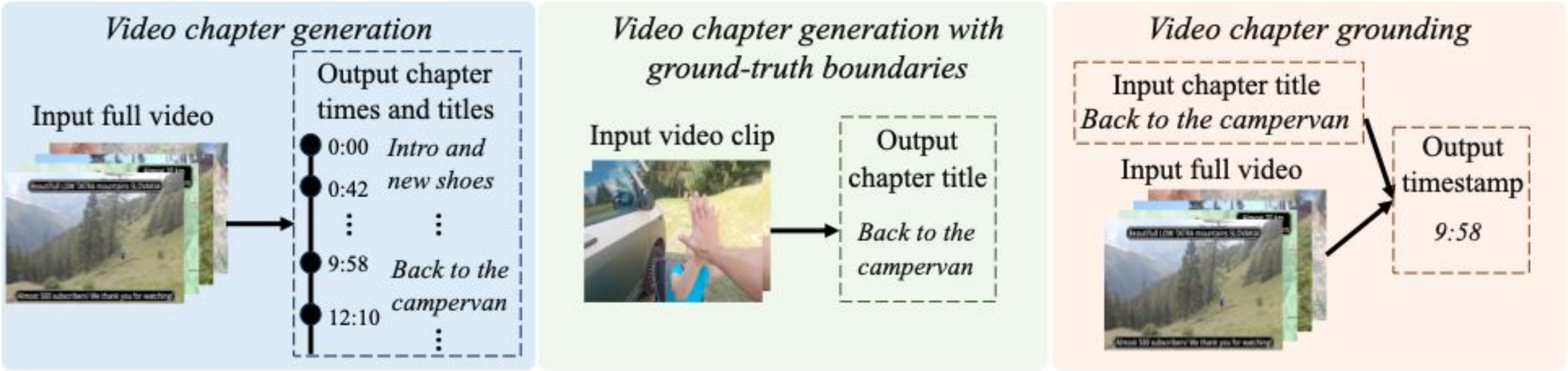
Comparison with other datasets

Dataset	# Videos	Duration (min)	# Descriptions	Annotations
HowTo100M	1M	7	136M	ASR
YT-Temporal-1B	**19M**	6	**900M**	ASR
HD-VILA-100M	3M	7	103M	ASR
ActivityNet Captions	20K	3	100K	Dense captions
YouCook2	2K	6	15K	Dense captions
ViTT	8K	4	56K	Dense captions
Ego4D	10K	**23**	4M	Dense captions
VidChapters-7M	817K	**23**	7M	**ASR+Chapters**

Manual assessment



New benchmarks



Video chapter generation

Model	Modalit ies	PT Data	FT VC	SODA	CIDEr	METEOR	R@3s	P@3s	R@0.7	P@0.7
Text tiling + LLaMA	T	Text mix	No	0.2	0.5	0.3	5.8	7.9	8.9	8.8
Shot detect + BLIP-2	V	129M img-txt	No	0.6	0.2	0.6	27.4	29.7	12.5	8.7
PDVC	V	None	Yes	6.8	35.8	9.4	17.8	**40.2**	22.5	**26.9**
Vid2Seq	T	C4+HTM	Yes	10.5	50.7	8.7	**28.9**	23.3	27.2	24.8
Vid2Seq	V+T	C4	Yes	10.6	51.3	8.8	28.6	23.8	26.9	24.9
Vid2Seq	V+T	C4+HTM	Yes	**11.4**	**55.7**	**9.5**	28.5	24.0	**28.5**	26.4

Video chapter generation given ground-truth boundaries

Model	Modalities	PT Data	FT VC	CIDEr	METEOR
LLaMA	T	Text mix	No	0.0	0.1
BLIP-2	V	129M img-txt	No	12.4	2.2
Vid2Seq	V	C4+HTM	Yes	47.1	5.1
Vid2Seq	T	C4+HTM	Yes	105.3	11.5
Vid2Seq	V+T	C4	Yes	110.8	11.5
Vid2Seq	V+T	C4+HTM	Yes	**120.5**	**12.6**

Video chapter grounding

Model	Modalities	PT Data	FT VC	R@3s	R@0.7
BERT	T	Text mix	No	5.2	0.1
CLIP	V	400M img-txt	No	3.7	2.3
Moment-DETR	V	None	Yes	**12.4**	**17.6**

Transfer to dense video captioning

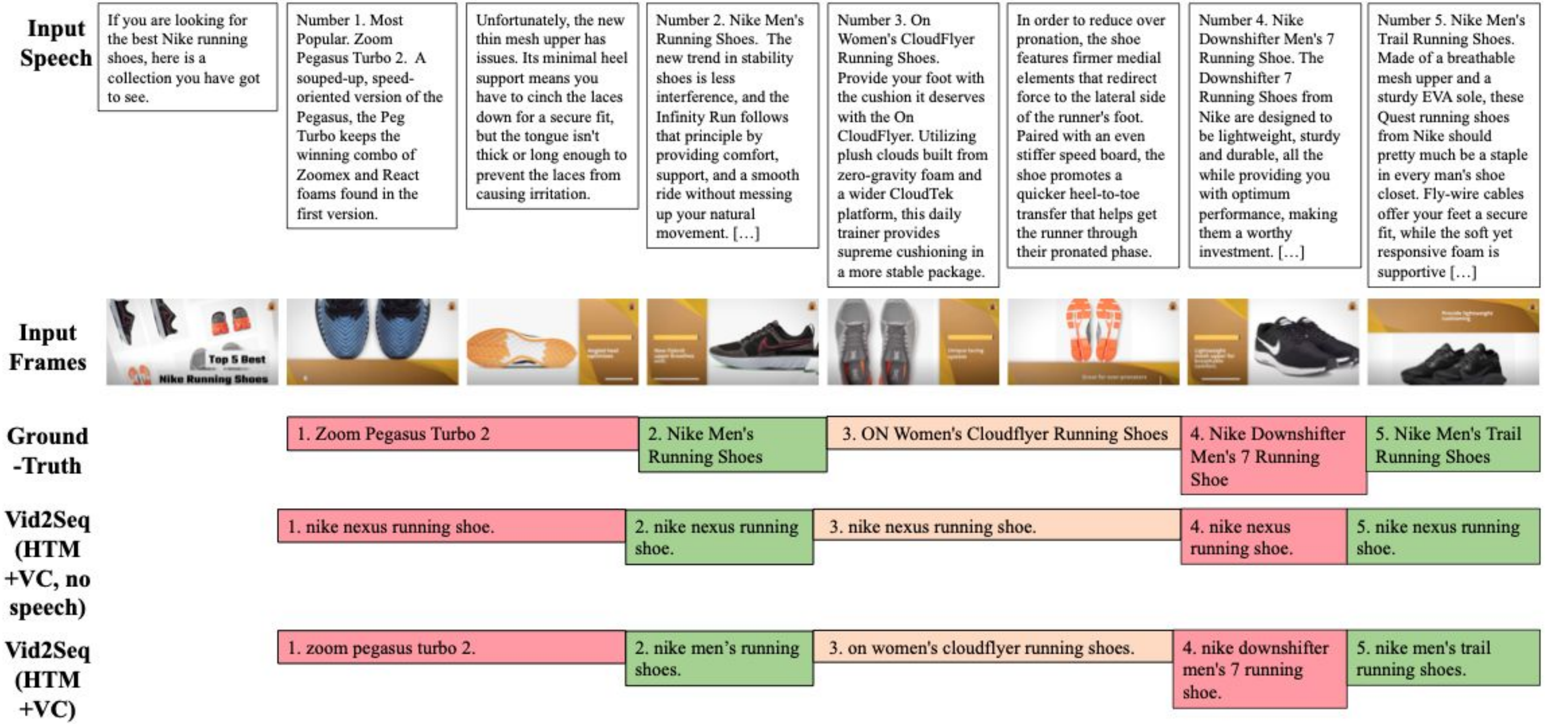
Model	Modalities	PT Data	YouCook2			ViTT		
			SODA	CIDEr	METEOR	SODA	CIDEr	METEOR
SoTA	T+V	C4+YTT	7.9	47.1	9.3	13.5	43.5	8.5
PDVC	V	None	4.8	28.8	5.8	9.4	40.6	**16.5**
PDVC	V	VidChap	5.9	34.7	7.5	10.1	41.5	16.1
Vid2Seq	T+V	C4+HTM	8.6	53.2	10.5	14.1	44.8	8.7
Vid2Seq	T+V	C4+HTM+ 10% VidChap	9.9	63.9	12.1	14.5	47.4	9.2
Vid2Seq	T+V	C4+HTM+ VidChap	**10.3**	**67.2**	**12.3**	**15.0**	**50.0**	9.5

Zero-shot dense video captioning

Model	Modalities	PT Data	YouCook2			ViTT		
			SODA	CIDEr	METEOR	SODA	CIDEr	METEOR
Text tiling + LLaMA	T	None	0.2	0.6	0.2	0.2	0.6	0.5
Shot Detect + BLIP-2	V	VidChap	0.6	1.0	0.5	0.2	0.1	0.2
Vid2Seq	T+V	C4+HTM	0.0	0.1	0.0	0.0	0.0	0.0
Vid2Seq	T+V	C4+HTM+ 10% VidChap	3.2	11.5	3.0	6.4	21.6	5.3
Vid2Seq	T+V	C4+HTM+ VidChap	**3.9**	**13.3**	**3.4**	**9.0**	**28.0**	**6.5**

Qualitative examples of video chapter generation

Qualitative examples of video chapter generation



Conclusion

- We present VidChapters-7M, a large-scale dataset of user-annotated chapters.
- We benchmark baselines and SoTA video-language models on three tasks built on top of VidChapters-7M, including video chapter generation.
- Pretraining for video chapter generation transfers well to dense video captioning in both zero-shot and finetuning settings, achieving new SoTA on YouCook2 and ViTT.

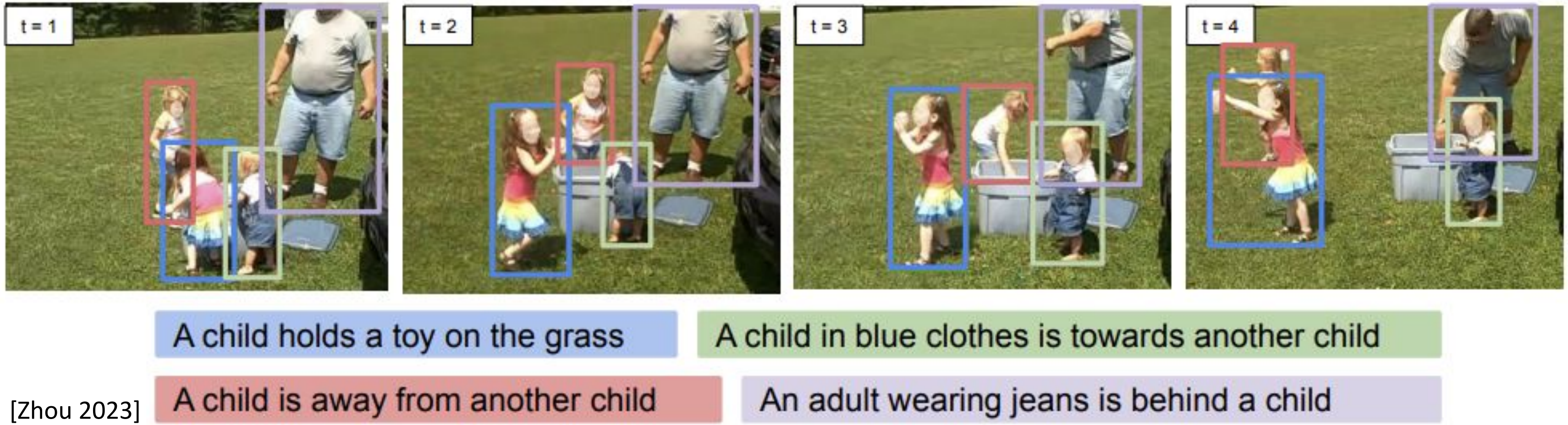
Limitations

- The distribution of VidChapters-7M is inherited from YT-Temporal-1B, which limits its diversity.
- The models evaluated in this work are not specific to chaptering tasks.
- Could this dataset be used to pretrain video-language models for other tasks than dense video captioning?

Gemini 1.5: A Visual Language Model that can understand long videos

Future work - localized dialog

Build flexible visual language models that can dialog about untrimmed videos and also ground their generated text in space and time.



[Zhou 2023]

[Koh 2023] Grounding Language Models to Images for Multimodal Inputs and Outputs, Jing Yu Koh et al, ICML 2023.
[Zhou 2023] Dense Video Object Captioning from Disjoint Supervision, Xingyi Zhou et al, arXiv 2023.

Future work - unified video model

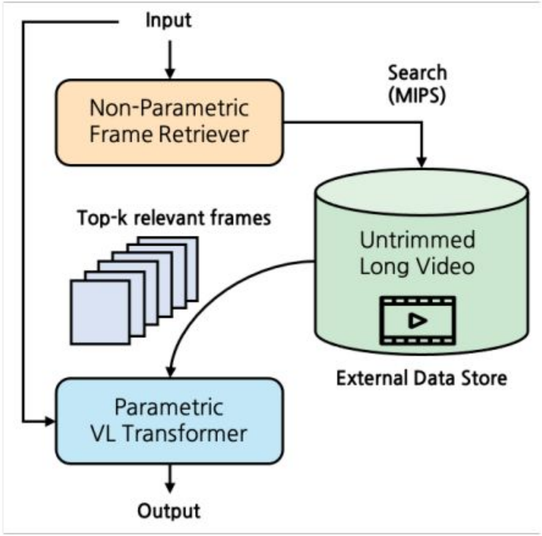
Current video models are still task-specific compared to image models.

[Zou 2023]

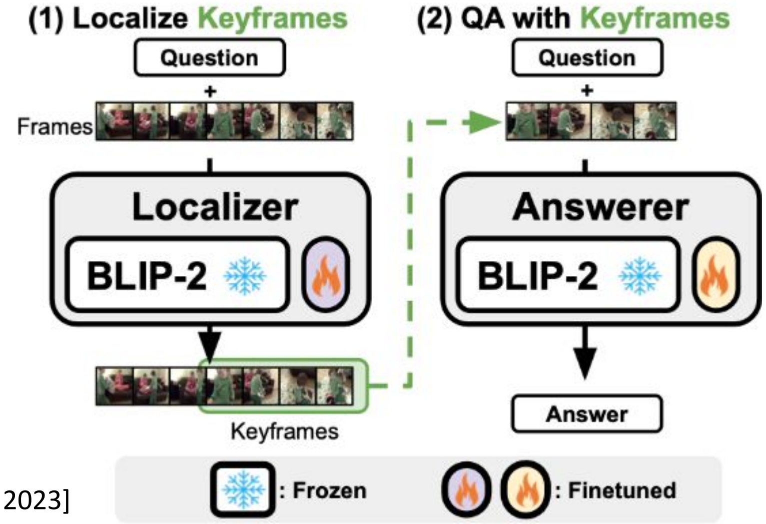
[Zhang 2022] GLIPv2: Unifying Localization and Vision-Language Understanding, Haotian Zhang et al, NeurIPS 2022.
[Zou 2023] Generalized Decoding for Pixel, Image, and Language, Xueyan Zhou et al, CVPR 2023.

Future work - processing long videos

Can we do better than the standard uniform sampling of frames?



[Kim 2023]



[Yu 2023]

[Kim 2023] Semi-Parametric Video-Grounded Text Generation, Sungdong Kim et al, arXiv 2023.
[Yu 2023] Self-Chained Image-Language Model for Video Localization and Question Answering, Shoubin Yu et al, arXiv 2023.

Future work - language models as annotators

Facilitate the collection of video datasets using language models.

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

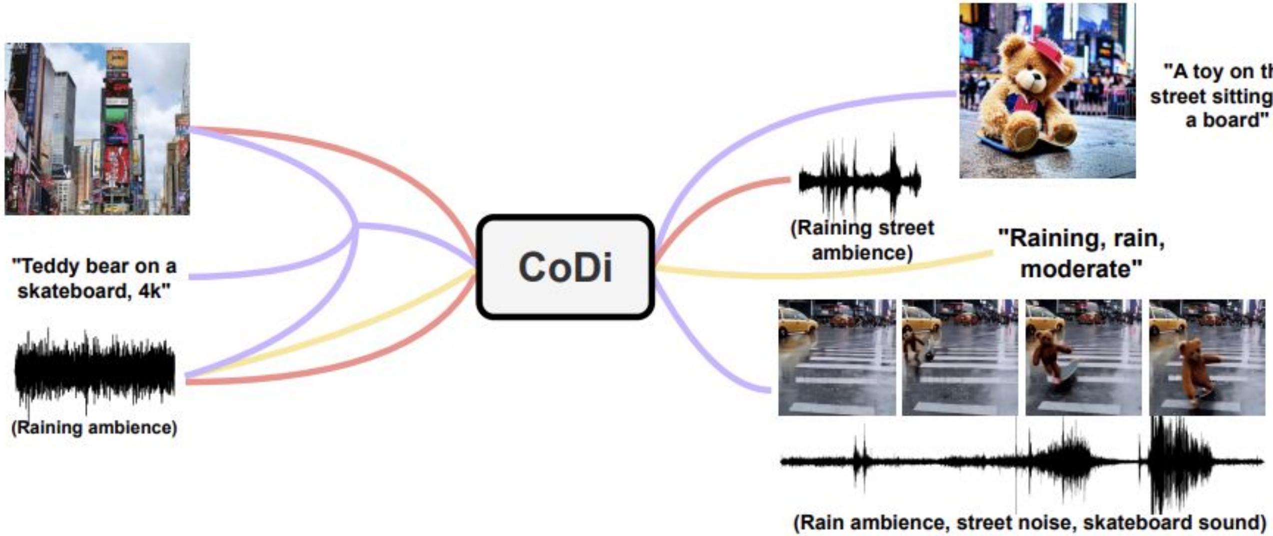
Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

[Liu 2023]

[Liu 2023] Visual instruction tuning, Haotian Liu et al, arXiv 2023.
[Zhang 2023] Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding, Hang Zhang et al, arXiv 2023.

Future work - multi-modality

Build models that can understand more modalities (audio), generate more as well (visual, audio), and learn modalities from one another.



[Tang 2023]

[Girdhar 2023] IMAGEBIND: One Embedding Space To Bind Them All, Rohit Girdhar et al, CVPR 2023.
[Tang 2023] Any-to-Any Generation via Composable Diffusion, Zineng Tang et al, arXiv 2023.