PhD Defense
Learning Visual Language Models for Video Understanding

Antoine Yang
https://antoyang.github.io/
23rd of November 2023

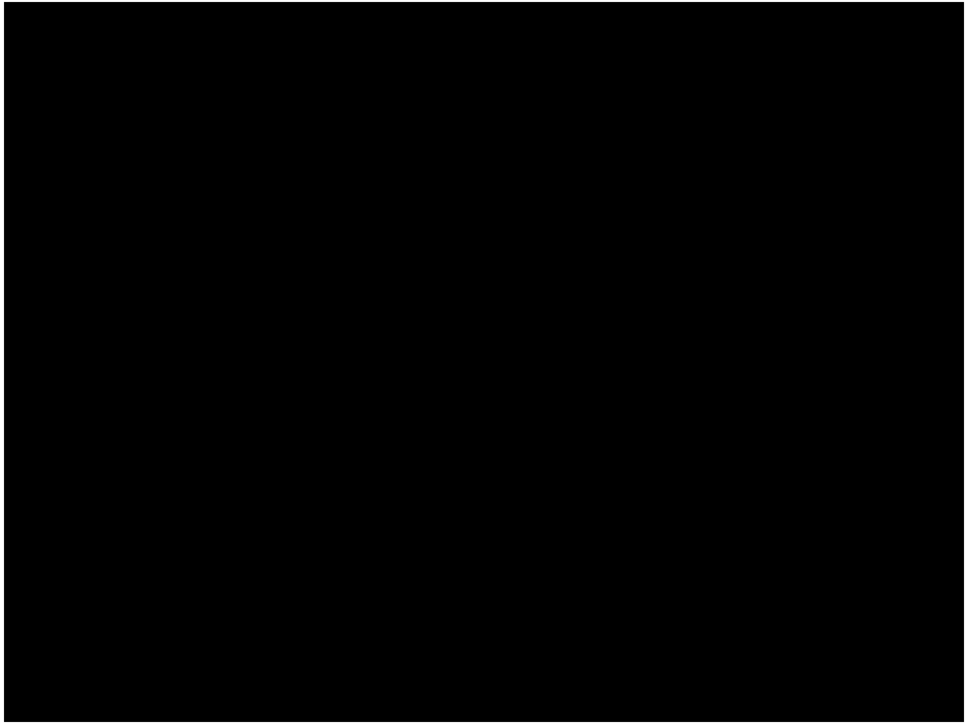
Visual language models

- 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

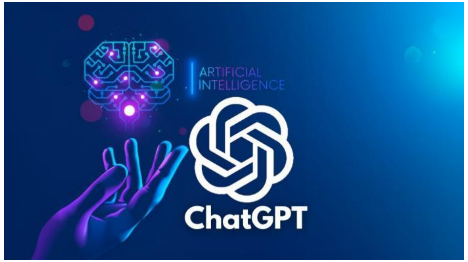


What are they doing? -> Martial arts

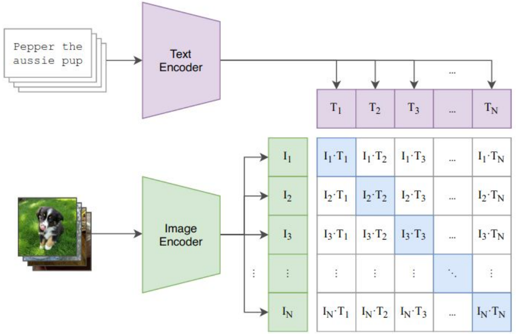
How many men are there? -> 2

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!

Caption: two people in a garden doing martial arts.

Because the man saved his life!

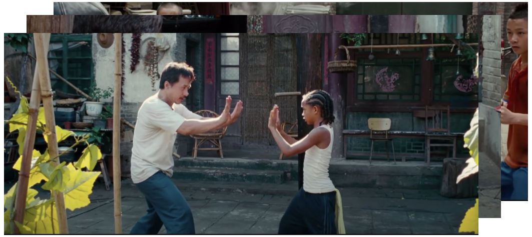
What else do we need?

- Localizing events in time

- Multi-event reasoning

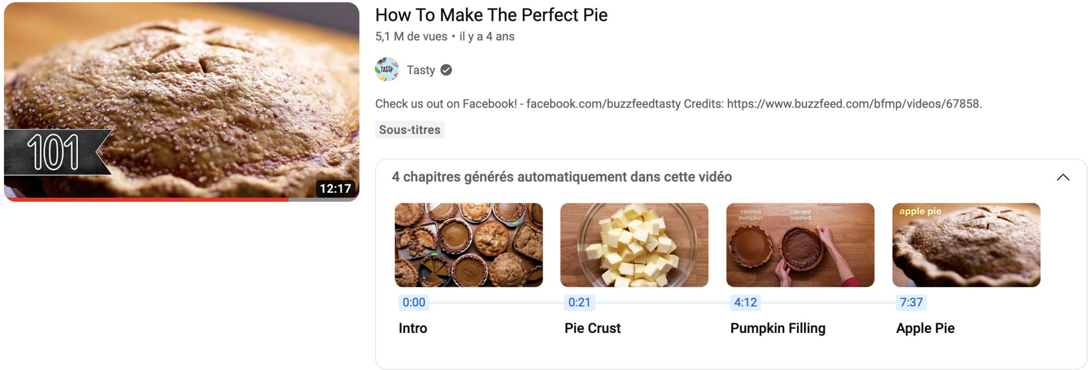
Applications: Beyond answering questions

- 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



Contributions

- **Video Question Answering**
  - Just Ask: Learning to Answer Questions from Millions of Narrated Videos (ICCV’21 Oral + TPAMI)
  - Zero-Shot Video Question Answering via Frozen Bidirectional Language Models (NeurIPS’22)
- **Spatio-Temporal Video Grounding**
  - TubeDETR: Spatio-Temporal Video Grounding with Transformers (CVPR’22 Oral)
- **Dense Video Captioning**
  - Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning (CVPR’23)
  - VidChapters-7M: Video Chapters at Scale (NeurIPS’23 D&B)

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

Video Question Answering (VideoQA)

Open-Ended Question: Where are the men?

Answer: **Track**

Multiple-Choice Question: What are the lined up men doing?

Proposal 1: **Running**

Proposal 2: Talking

Proposal 3: Shaving

Challenges

- Videos, questions and answers are highly diverse.
- Manual annotation is expensive.
- Yet prior approaches (e.g. [Le 2020]) are fully-supervised.

[Jang 2017]

Question: How many times does the cat lick?

Answer: **7 times**

Question: What does the cat do 3 times?

Answer: **put head down**

Question: What is the color of the bulldog?

Answer: **brown**

[Jang 2017] TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering, Yunseok Jang et al, CVPR 2017.
[Le 2020] Hierarchical Conditional Relation Networks for Video Question Answering, Thao Minh Le et al, CVPR 2020.

Learning from narrated videos

- Paired (video, speech) data is easy to obtain at scale and helps learning text-video retrieval or action recognition [Miech 2020].
- But (video, speech) data differs from (video, question, answer) data.

[Miech 2019]

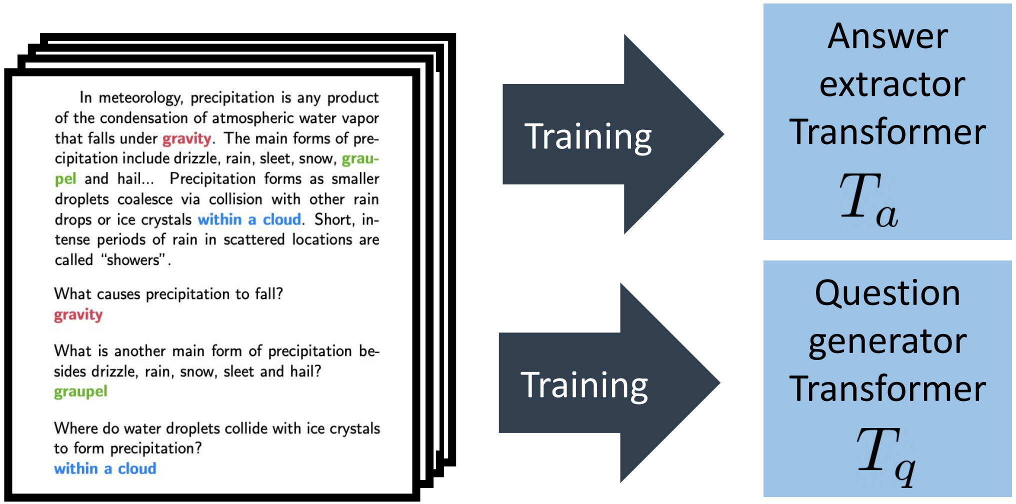
[Miech 2019] HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, Antoine Miech et al, ICCV 2019.

[Miech 2020] End-to-End Learning of Visual Representations from Uncurated Instructional Videos, Antoine Miech et al, CVPR 2020.

Leveraging language models

Let’s apply question generation language models [Raffel 2020] trained on text-only annotations [Rajpurkar 2016] to the narration.

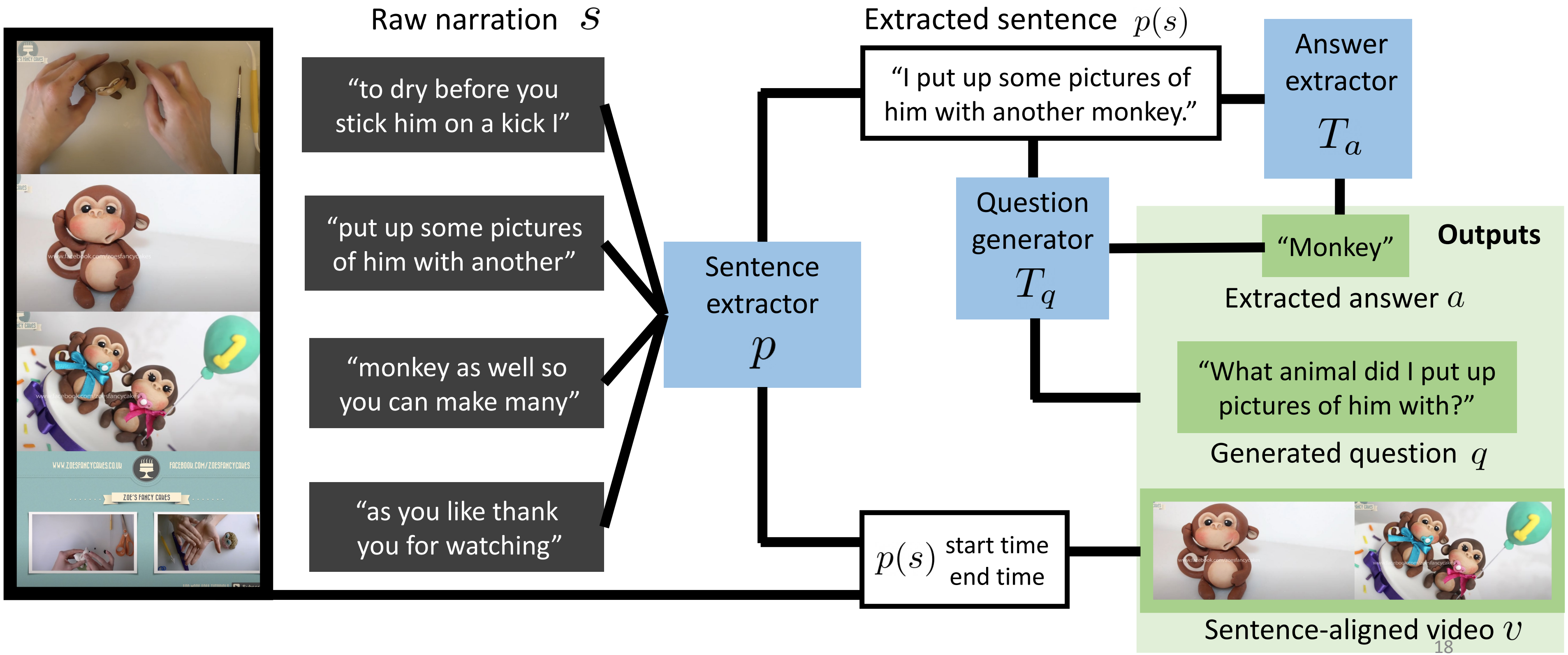
Manually annotated QA text corpus



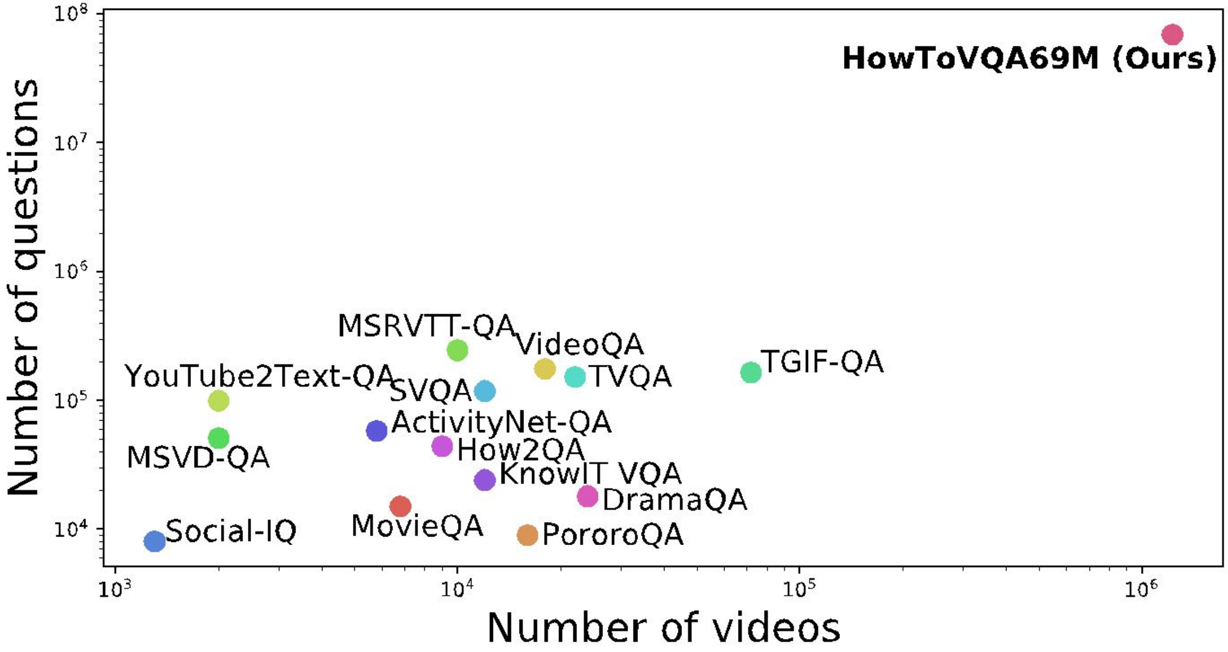
[Rajpurkar 2016]

[Rajpurkar 2016] SQuAD: 100,000+ Questions for Machine Comprehension of Text, Pranav Rajpurkar et al, EMNLP 2016.
[Raffel 2020] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Colin Raffel et al, JMLR 2020.

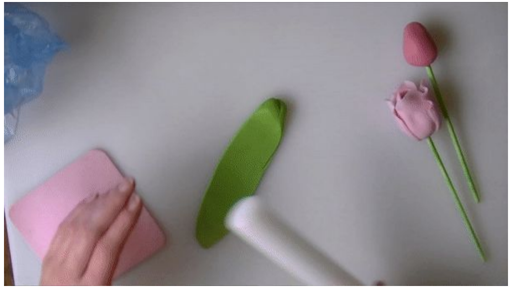
Generating video-question-answer triplets



HowToVQA69M: a large-scale VideoQA training dataset



Noise in HowToVQA69M



Speech: So you bring it to a point and we'll, just cut it off at the bottom.
Generated question: What do we do at the bottom?
Generated answer: cut it off

≈ 30%

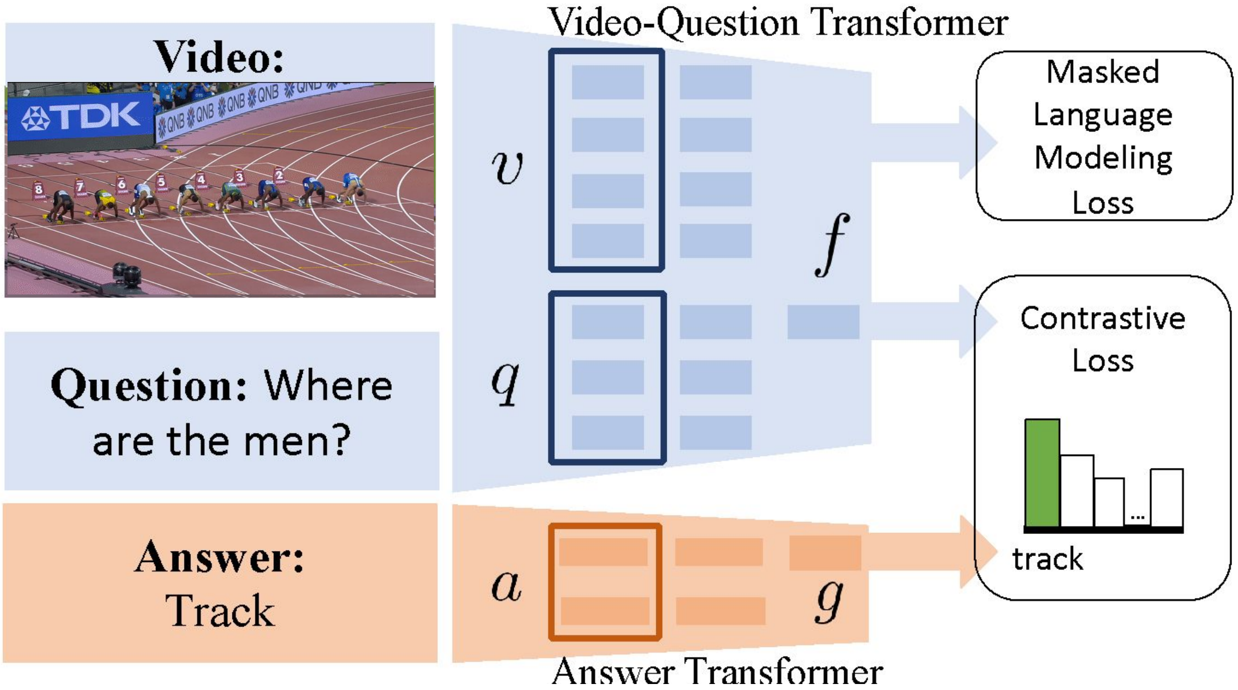
Speech: So you bring it to a point and we'll, just cut it off at the bottom.
Generated question: What color did you peel on the other side?
Generated answer: orange

Wrong QA Generation
≈ 31%

Speech: You can't miss this…
Generated question: What can't you do?
Generated answer: miss

Weak video-speech correlation
≈ 39%

VQA-T model and training procedure



iVQA: a new manually collected, open-ended VideoQA benchmark

- 10K videos from HowTo100M, each annotated with a question about objects and scenes.

5 different answers collected per question.

Question: What shape is the handcraft item in the end?

Answers	shell	✓	2 annotators
	spiral	✓	2 annotators
	heart	✓	1 annotator

Reduced language bias.

Question: What is the chef wearing over her shirt?

Answer: apron

Easy to guess without watching => excluded

VQA-T can do zero-shot VideoQA with *no manual supervision of visual data.*

The VQA-T model pretrained on HowToVQA69M outperforms its text-only variant and its variant pretrained on HowTo100M directly.

Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
Random	∅	0.09	0.02	0.05	0.05	25.0
QA-T	HowToVQA69M	4.4	2.5	4.8	11.6	38.4
VQA-T	HowTo100M	1.9	0.3	1.4	0.3	46.2
VQA-T	HowToVQA69M	**12.2**	**2.9**	**7.5**	**12.9**	**51.1**

Qualitative zero-shot results on iVQA

Demo: http://videoqa.paris.inria.fr/ & https://www.youtube.com/watch?v=8ZjnbehPzmE

Question: What is the man cutting?
GT answer: pipe
QA-T (HowToVQA69M): onion
VQA-T (HowTo100M): knife holder
Ours: pipe

Question: What is the largest object at the right of the man?
GT answer: wheelbarrow
QA-T (HowToVQA69M): statue
VQA-T (HowTo100M): trowel
Ours: wheelbarrow

Question: What fruit is shown in the end?
GT answer: watermelon
QA-T (HowToVQA69M): pineapple
VQA-T (HowTo100M): slotted spoon
Ours: watermelon

Qualitative zero-shot results on iVQA

VQA-T achieves SoTA after finetuning.

Method	Pretraining Data	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HCRN [Le 2020]	∅	35.6	36.1	-	-
HERO [Li 2020]	HowTo100M	-	-	-	74.1
ClipBERT [Lei 2021]	COCO + VG	37.4	-	-	-
CoMVT [Seo 2021]	HowTo100M	39.5	42.6	38.8	82.3
Just Ask (∅)	∅	39.6	41.2	36.8	80.8
Just Ask	HowToVQA69M	**41.5**	**46.3**	**38.9**	**84.4**

[Le 2020] Hierarchical Conditional Relation Networks for Video Question Answering, Thao Minh Le et al, CVPR 2020.
[Li 2020] HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training, Linjie Li et al, EMNLP 2020.
[Lei 2021] Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling, Jie Lei et al, CVPR 2021.
[Seo 2021] Look Before you Speak: Visually Contextualized Utterances, Paul Hongsuck Seo et al, CVPR 2021.

HowToVQA69M pretraining improves generalization to rare answers.

Results on subsets of iVQA corresponding to four quartiles with Q1 and Q4 corresponding to samples with most frequent and least frequent answers:

Pretraining Data	Finetuning	Q1	Q2	Q3	Q4
∅	✓	38.4	16.7	5.9	2.6
HowTo100M	✓	46.7	22.0	8.6	3.6
HowToVQA69M	✗	9.0	8.0	9.5	7.7
HowToVQA69M	✓	**47.9**	**28.1**	**15.6**	**8.5**

Neural QA generation improves over rule-based QA generation.

Generation method	Finetuning	iVQA	ActivityNet-QA	How2QA
[Heilman 2010] (rule-based)	✗	7.4	1.1	41.7
Just Ask (neural)	✗	**12.2**	**12.9**	**51.1**
[Heilman 2010] (rule-based)	✓	31.4	38.5	83.0
Just Ask (neural)	✓	**35.4**	**38.9**	**84.4**

ASR: And then just squeeze it through like that.
Question (Heilman et al): What do then just squeeze through like that?
Answer (Heilman et al): it
Question (ours): How do you do it?
Answer (ours): squeeze it through

ASR: It is a staple in a lot of asian kitchens.
Question (Heilman et al): What is it?
Answer (Heilman et al): a staple in a lot of asian kitchens
Question (ours): In what type of kitchens is it a staple?
Answer (ours): asian kitchens

ASR: And you want it over a very low heat.
Question (Heilman et al): What do you want it over?
Answer (Heilman et al): over a very low heat
Question (ours): What kind of heat do you want it to be over?
Answer (ours): low heat

[Heilman 2010] Good Question! Statistical Ranking for Question Generation, Michael Heilman et al, ACL 2010.

Our method scales with the size of the pretraining dataset.

Fraction of HowTo100M videos	ZS iVQA	ZS MSVD-QA	Finetuning iVQA	Finetuning MSVD-QA
1%	4.5	3.6	24.2	42.8
10%	9.1	6.2	29.2	44.4
20%	9.5	6.8	31.3	44.8
50%	11.3	7.3	32.8	45.5
100%	**12.2**	**7.5**	**35.4**	**46.3**

Our VideoQA generation approach generalizes to video alt-text descriptions.

Starting from WebVid2M [Bain 2021], we generate WebVidVQA3M, a dataset of 3M VideoQA triplets, with our approach.

Pretraining Data	Fine tuning	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HowToVQA69M	✗	12.2	2.9	7.5	12.9	51.1
WebVidVQA3M	✗	7.3	5.3	12.3	6.2	49.8
HowToVQA69M + WebVidVQA3M	✗	**13.3**	**5.6**	**13.5**	**12.3**	**53.1**
∅	✓	23.0	39.6	41.2	36.8	80.8
HowToVQA69M	✓	**35.4**	41.5	46.3	38.9	84.4
WebVidVQA3M	✓	28.1	41.2	45.4	38.1	82.4
HowToVQA69M + WebVidVQA3M	✓	35.2	**41.8**	**47.5**	**39.0**	**85.3**

[Bain 2021] Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval, Max Bain et al, ICCV 2021.

Conclusion

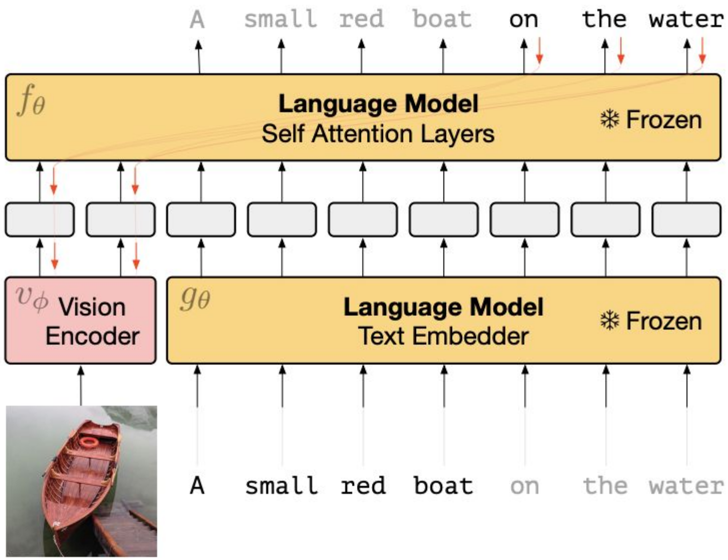
- We automatically generate a large-scale VideoQA dataset, HowToVQA69M, using language models and narrated videos.
- We manually collect iVQA, a new VideoQA benchmark with redundant annotations and reduced language bias.
- We show that a video-question model trained contrastively with an answer model highly benefits from pretraining on HowToVQA69M: The resulting VQA-T model is capable of zero-shot generalization and achieves SoTA results on 4 existing benchmarks after finetuning.

Limitations

- Generating data is expensive (10K GPUH for HowToVQA69M).
- The generation also relies on text QA manual annotations.
- The VQA-T model cannot use the speech modality.

Multi-modal few-shot learning with frozen autoregressive language models

- Frozen autoregressive language models can tackle zero-shot VQA [Tsimpoukelli 2021] without data generation.
- But they require billions of parameters to work well hence are difficult to train and deploy.



[Tsimpoukelli 2021]

[Tsimpoukelli 2021] Multi-modal Few-Shot Learning with Frozen Language Models, Maria Tsimpoukelli et al, NeurIPS 2021.

Bidirectional masked language models (BiLM)

- [Schick 2021] shows light BiLM can compete with large autoregressive language models in text-only tasks using cloze task formulations.
- Can we tackle zero-shot VideoQA with light BiLM?

Autoregressive language models

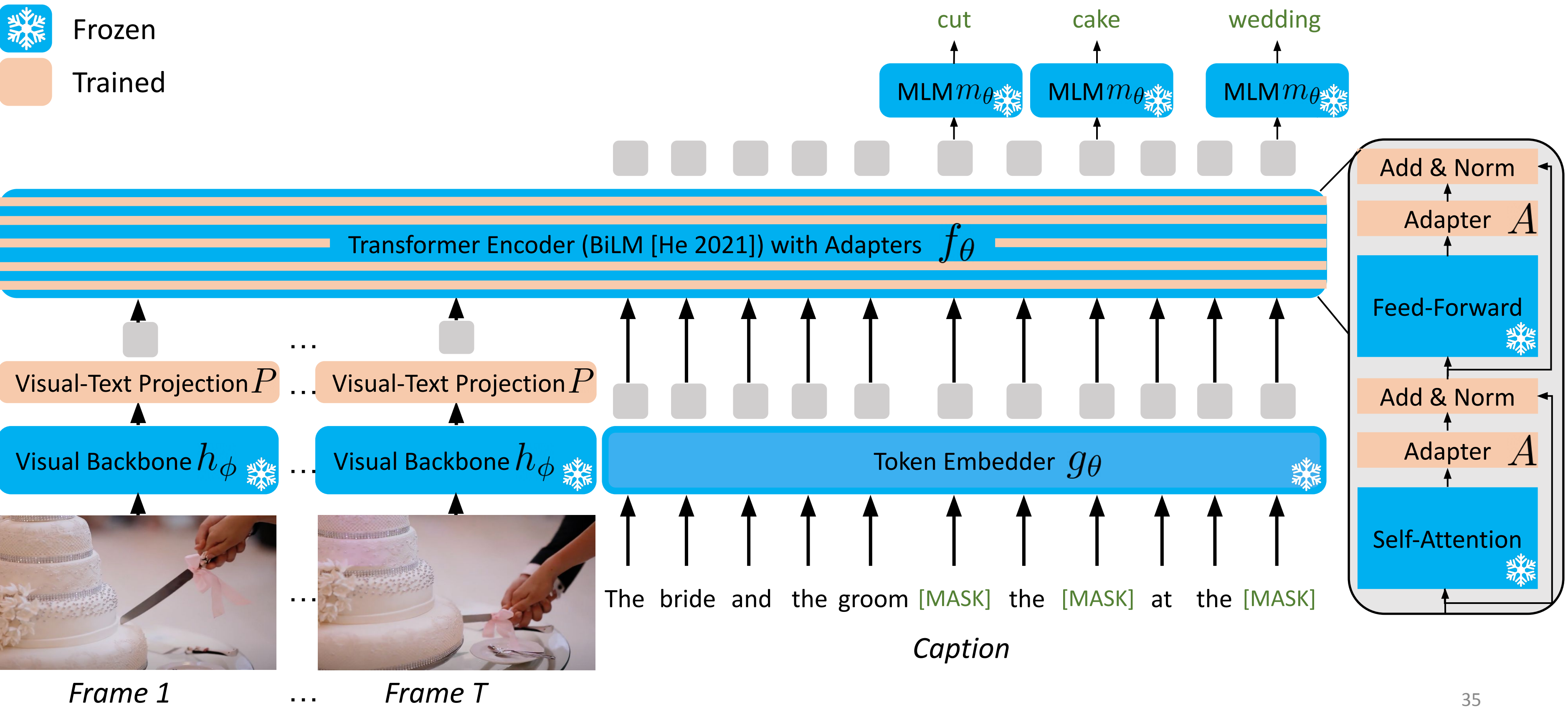
[BOS] -> The
The -> dog
The dog -> is
The dog is -> running
The dog is running -> in
The dog is running in -> the
The dog is running in the -> snow
The dog is running in the snow -> EOS

Bidirectional language models (BiLM)

The dog is [MASK] in the snow -> running

[Schick 2021] It's not just size that matters: Small language models are also few-shot learners, Timo Schick et al, NAACL 2021.

Connecting frozen BiLM and visual backbone



[He 2021] DeBERTa: Decoding-enhanced BERT with Disentangled Attention, Pengcheng He et al, ICLR 2021.

Downstream task adaptation

The **answer embedding module** is initialized from the *frozen* masked language modeling head and maps a [MASK] token to an answer.

- Open-ended VideoQA:

“[CLS] Question: <Question>? Answer: [MASK]. Subtitles: <Subtitles> [SEP]”

- Multiple-choice VideoQA:

“[CLS] Question: <Question>? Is it ’<Answer Candidate>’? [MASK]. Subtitles: <Subtitles> [SEP]”

- Video-conditioned fill-in-the-blank:

“[CLS] <Sentence with a [MASK] token>. Subtitles: <Subtitles> [SEP]”

Bidirectional models perform better with less parameters than autoregressive models.

We find that the suffix (specific to BiLM) is crucial to performance.

Method	Language Model	LM Params	Train time (GPUH)	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA
Autoregressive	GPT-Neo-1.3B	1.3B	200	6.6	4.2	10.1	17.8	14.4
	GPT-Neo-2.7B	2.7B	360	9.1	7.7	17.8	17.4	20.1
	GPT-J-6B	6B	820	21.4	9.6	26.7	24.5	37.3
Bidirectional	BERT-Base	**110M**	**24**	12.4	6.4	11.7	16.7	23.1
	BERT-Large	340M	60	12.9	7.1	13.0	19.0	21.5
	DeBERTa-V2-XLarge	890M	160	**27.3**	**16.8**	**32.2**	**24.7**	**41.0**

FrozenBiLM is SoTA for zero-shot VideoQA.

Method	Training Data	LSMDC	iVQA	MSRVTT -QA	MSVD -QA	Activity Net-QA	TGIF-QA	How2 QA	TVQA
Random	-	0.1	0.1	0.1	0.1	0.1	0.1	25.0	20.0
ViT-L/14 [Radford 21]	CLIP	1.2	9.2	2.1	7.2	1.2	3.6	47.7	26.1
Just Ask [Yang 2022]	HowToVQA69M + WebVidVQA3M	-	13.3	5.6	13.5	12.3	-	53.1	-
Reserve [Zellers 22]	YT-Temporal-1B	31.0	-	5.8	-	-	-	-	-
FrozenBiLM	WebVid10M [Bain 2021]	**51.5**	**26.8**	**16.7**	**33.8**	**25.9**	**41.9**	**58.4**	**59.7**
GPT-4 [OpenAI 23]	???	45.7							

[Radford 2021] Learning Transferable Visual Models From Natural Language Supervision, Alec Radford et al, NeurIPS 2021.
[Yang 2022] Learning to Answer Visual Questions from Web Videos, Antoine Yang et al, TPAMI 2022.
[Zellers 22] MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound, Rowan Zellers et al, CVPR 2022.
[Bain 2021] Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval, Max Bain et al, ICCV 2021.
[OpenAI 2023] https://openai.com/research/gpt-4

Qualitative zero-shot results (open-ended)

More examples: https://www.youtube.com/watch?v=4aLSUvSirOA

Question: What is the man holding at the start of the video?
GT answer: guitar, electric guitar
Just Ask: typewriter
UnFrozenBiLM: beer
FrozenBiLM (text-only): scissors
FrozenBiLM: guitar

Question: What item hanging on the wall features a tree?
GT answer: quilt
Just Ask: christmas tree
UnFrozenBiLM: fabric
FrozenBiLM (text-only): tree
FrozenBiLM: quilt

Question: Which category of sports does this sport belong to?
GT answer: surfing
Just Ask: second
UnFrozenBiLM: swimming
FrozenBiLM (text-only): 1
FrozenBiLM: surfing

Qualitative zero-shot results (multiple-choice)

Question: When did the chef flipped over the layer of rice and seaweed?
GT answer: A0
A0: after she sprinkled sesame
A1: after she added cucumber
A2: after she added fish
A3: after she cut the cucumbers

UnFrozenBiLM: A3
FrozenBiLM (text-only):A1
FrozenBiLM: A0

Qualitative zero-shot results (fill-in-the-blank)

Sentence: Each singer in the front row ____ a huge toad.
GT answer: holds
UnFrozenBiLM: plays
FrozenBiLM (text-only): wears
FrozenBiLM: holds

Sentence: Someone ____ him to the truck and across the street.
GT answer: chases
UnFrozenBiLM: follow
FrozenBiLM (text-only): drags
FrozenBiLM: chases

Sentence: A woman wraps food in newspapers and brings it over to their ____.
GT answer: table
UnFrozenBiLM: man
FrozenBiLM (text-only): home
FrozenBiLM: table

Qualitative zero-shot results

FrozenBiLM is competitive in fully-supervised setting.

Method	Trained Params	LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
Just Ask [Yang 2022]	157M	-	35.4	41.8	47.5	39.0	-	85.3	-
SiaSamRea [Yu 2021]	-	-	41.6	45.5	-	39.8	60.2	84.1	-
MERLOT [Zellers 2021]	223M	52.9	-	43.1	-	41.4	**69.5**	-	78.7
Reserve [Zellers 2022]	644M	-	-	-	-	-	-	-	**86.1**
UnFrozenBiLM	890M	58.9	37.7	45.0	53.9	43.2	66.9	**87.5**	79.6
FrozenBiLM	**30M**	**63.5**	**39.6**	**47.0**	**54.8**	**43.2**	68.6	86.7	82.0

[Yang 2022] Learning to Answer Visual Questions from Web Videos, Antoine Yang et al, TPAMI 2022.
[Yu 2021] Learning from Inside: Self-driven Siamese Sampling and Reasoning for Video Question Answering, Weijiang Yu et al, NeurIPS 2021.
[Zellers 2021] MERLOT: Multimodal Neural Script Knowledge Models, Rowan Zellers et al, NeurIPS 2021.
[Zellers 2022] MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound, Rowan Zellers et al, CVPR 2022.

FrozenBiLM is efficient in few-shot settings.

Supervision	LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
0% (zero-shot)	51.5	26.8	16.7	33.8	25.9	41.9	58.4	59.2
1% (few-shot)	56.9	31.1	36.0	46.5	33.2	55.1	71.7	72.5
10% (few-shot)	59.9	35.3	41.7	51.0	37.4	61.2	75.8	77.6
100% (fully-supervised)	**63.5**	**39.6**	47.0	**54.8**	**43.2**	**68.6**	**86.7**	**82.0**

FrozenBiLM benefits from both visual and speech inputs.

Visual	Speech	LSMDC	iVQA	MSRVTT-QA	MSVD-QA	Activity-QA	TGIF-QA	How2QA	TVQA
✗	✗	47.9	11.0	6.4	11.3	22.6	32.3	29.6	23.2
✗	✓	49.8	13.2	6.5	11.7	23.1	32.3	45.9	44.1
✓	✗	50.9	26.2	**16.9**	33.7	25.9	41.9	41.9	29.7
✓	✓	**51.5**	**26.8**	16.7	**33.8**	**25.9**	**41.9**	**58.4**	**59.2**

Benefits of freezing with adapter training

Freeze	Adapter	LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
✗	✗	37.1	21.0	**17.6**	31.9	20.7	30.7	45.7	45.6
✓	✗	50.7	27.3	16.8	32.2	24.7	41.0	53.5	53.4
✓	✓	**51.5**	**26.8**	16.7	**33.8**	**25.9**	**41.9**	**58.4**	**59.2**

Conclusion

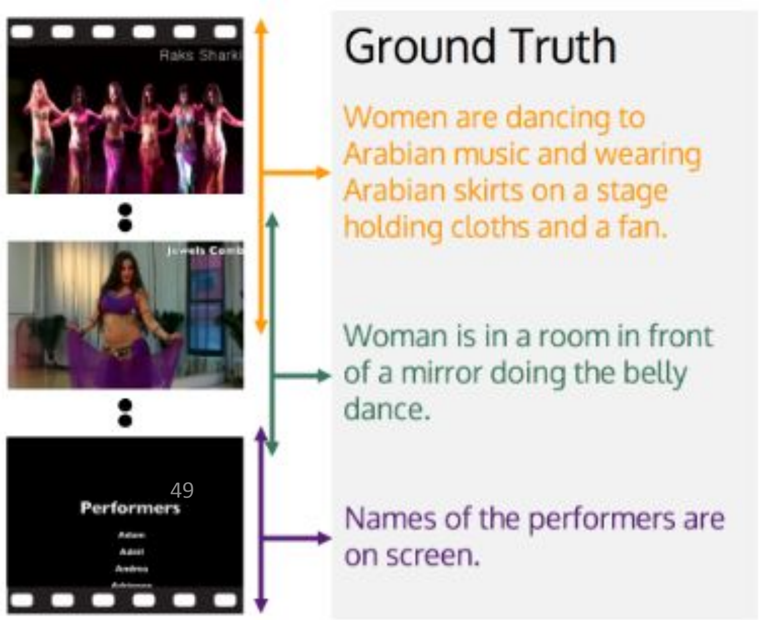
- We present FrozenBiLM, a framework that handles multi-modal inputs using frozen bidirectional language models and enables zero-shot VideoQA through masked language modeling.
- We show the superiority of FrozenBiLM over prior autoregressive language models for zero-shot VideoQA.
- FrozenBiLM largely improves the SoTA in zero-shot VideoQA on 8 benchmarks, shows competitive performance in the fully-supervised setting and strong results in few-shot settings.

Limitations

- FrozenBiLM cannot use raw audio inputs (beyond speech transcripts).
- Does FrozenBiLM generalize well to more complex text generation tasks such as video captioning like autoregressive models?
- FrozenBiLM, like most visual language models, cannot tackle localization tasks.

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.

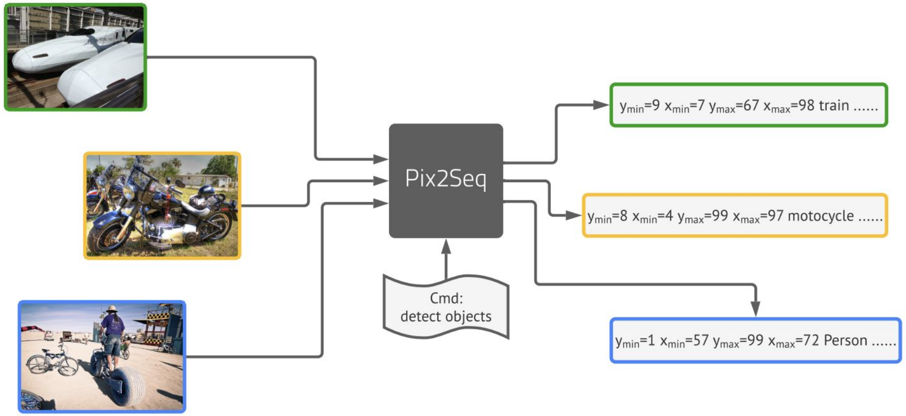


Example from the ActivityNet-Captions dataset [Krishna 2017].

[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

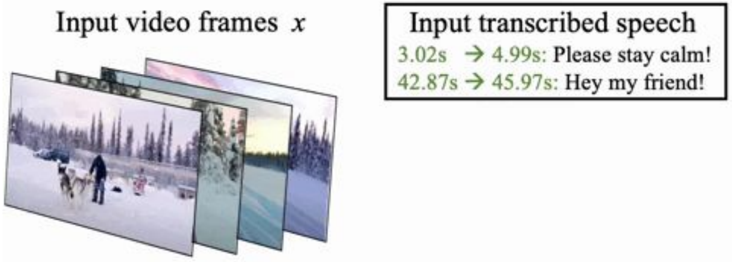
- Pix2seq [Chen 2022] casts object detection as sequence generation.
- Spatial coordinates are quantized and tokenized.



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

The Vid2Seq model

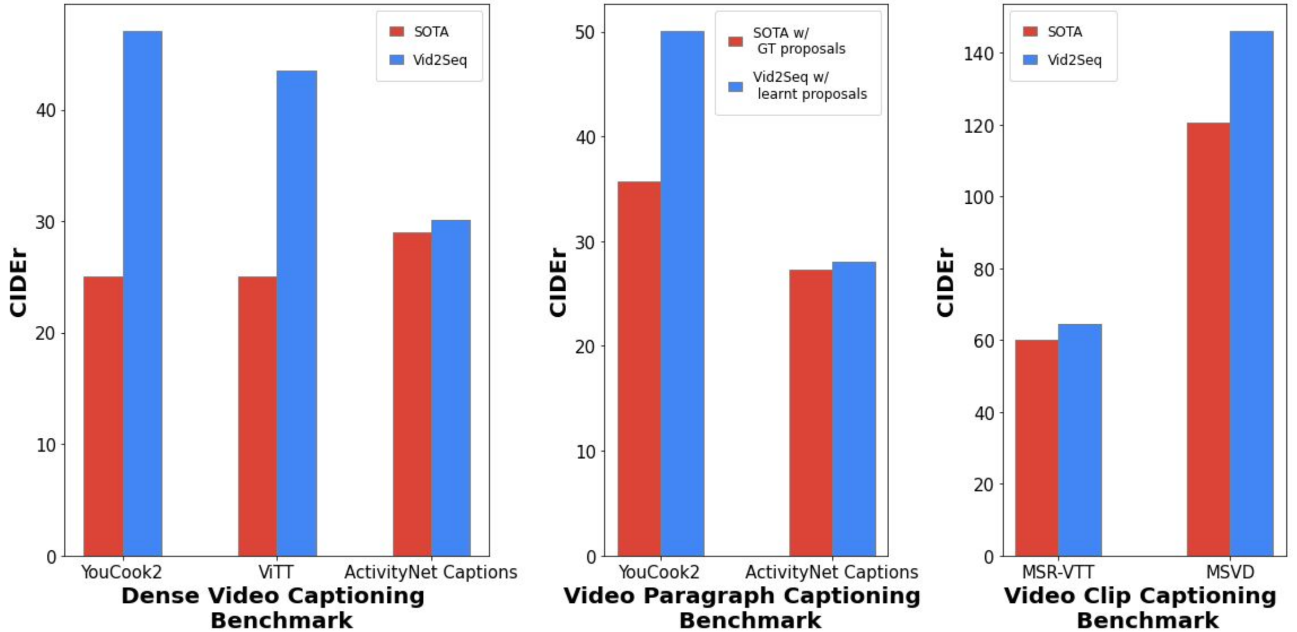
- Formulates dense video captioning as a sequence-to-sequence problem.
- Time is quantized and jointly tokenized with the text.
- **Model architecture:** visual encoder, text encoder and text decoder.



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.

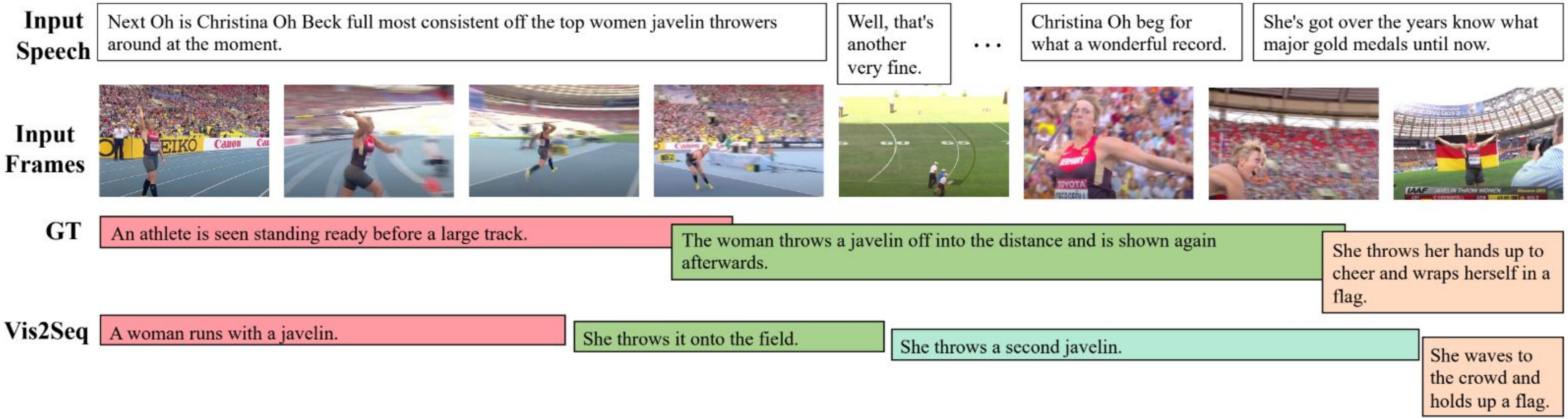
Captioning	Pretraining	YouCook2			ActivityNet Captions		
		Recall	Precis	F1	Recall	Precis.	F1
✗	✗	17.8	19.4	17.7	47.3	57.9	52.0
✓	✗	17.2	20.6	18.1	42.5	**64.1**	49.2
✗	✓	25.7	21.4	22.8	52.5	53.0	51.1
✓	✓	**27.9**	**27.8**	**27.3**	**52.7**	53.9	**52.4**

Data and model scaling.

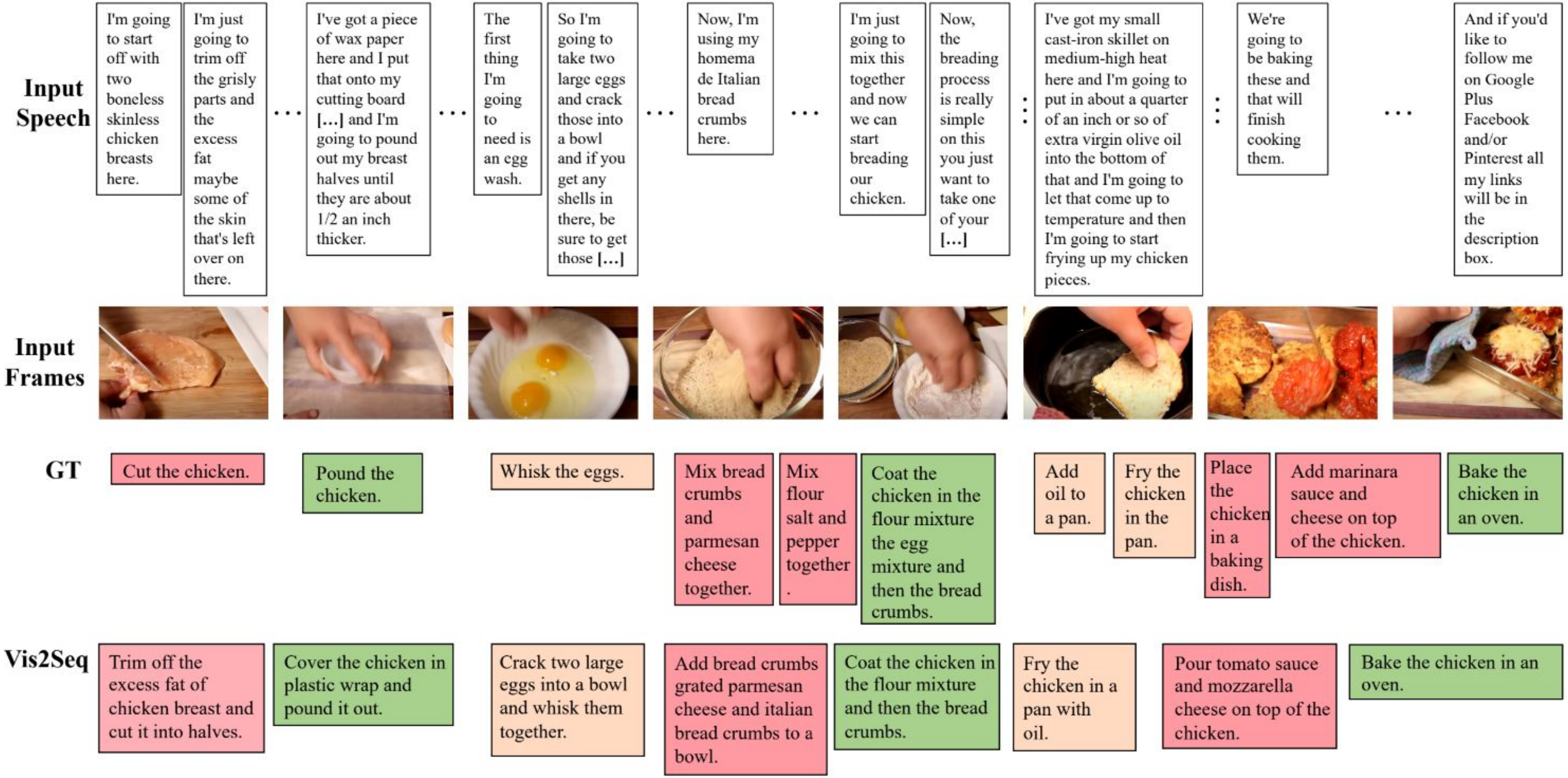
Language Model	Pretraining		YouCook2			ActivityNet Captions		
	# Videos	Dataset	SODA	CIDEr	F1	SODA	CIDEr	F1
T5-Small	15M	YTT	6.1	31.1	24.3	5.5	26.5	52.2
T5-Base	0	-	4.0	18.0	18.1	5.4	18.8	49.2
T5-Base	15K	YTT	6.3	35.0	24.4	5.1	24.4	49.9
T5-Base	150K	YTT	7.3	40.1	26.7	5.4	27.2	51.3
T5-Base	1M5	YTT	7.8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

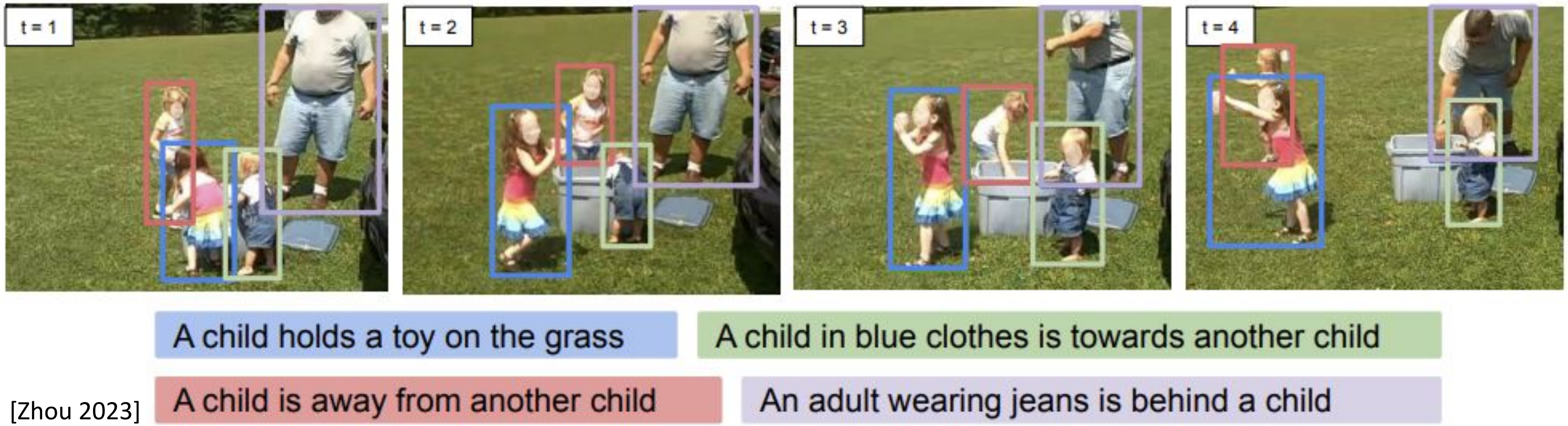
- Vid2Seq cannot use raw audio inputs (beyond speech transcripts).
- Does Vid2Seq generalize to other tasks, e.g. VideoQA or temporal action localization?
- Pretraining gains are subject to video domain -> Vid2Seq event localization performance is below task-specific approaches on ActivityNet Captions.

Contributions

- **Video Question Answering**
  - Just Ask: Learning to Answer Questions from Millions of Narrated Videos (ICCV’21 Oral + TPAMI)
  - Zero-Shot Video Question Answering via Frozen Bidirectional Language Models (NeurIPS’22)
- **Spatio-Temporal Video Grounding**
  - TubeDETR: Spatio-Temporal Video Grounding with Transformers (CVPR’22 Oral)
- **Dense Video Captioning**
  - Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning (CVPR’23)
  - VidChapters-7M: Video Chapters at Scale (NeurIPS’23 D&B)

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.

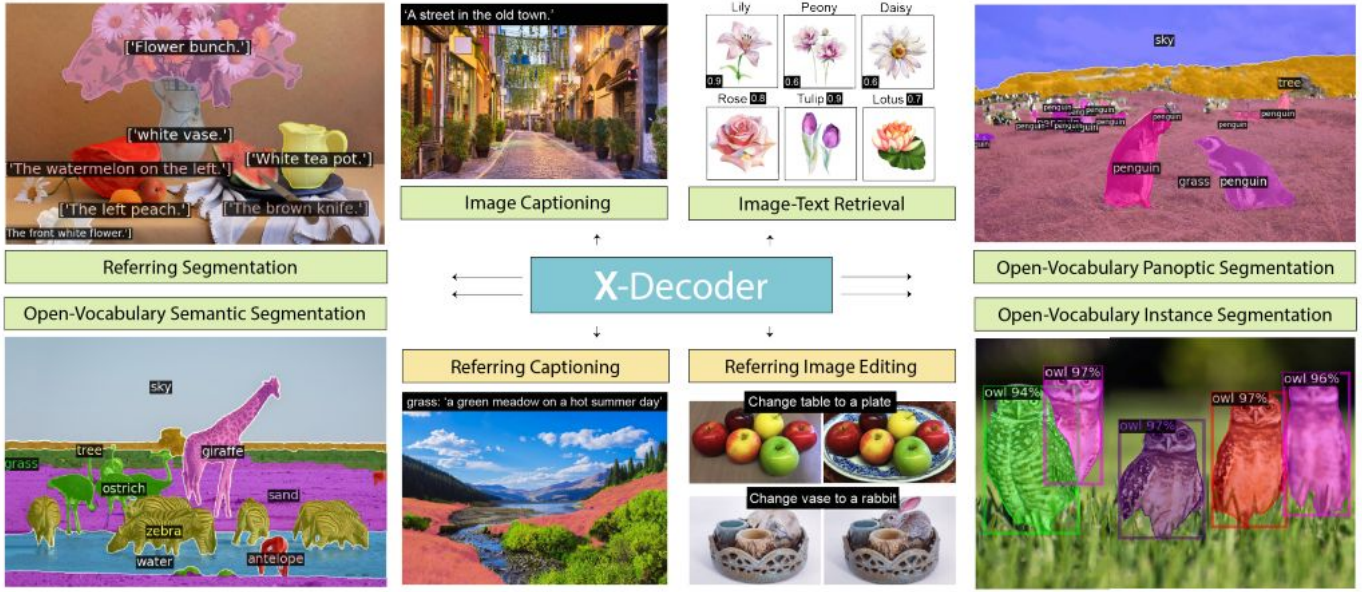


[Zhou 2023]

[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.

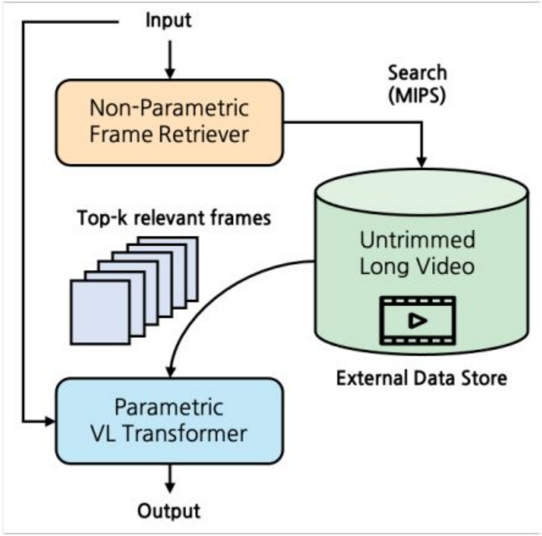


[Zou 2023]

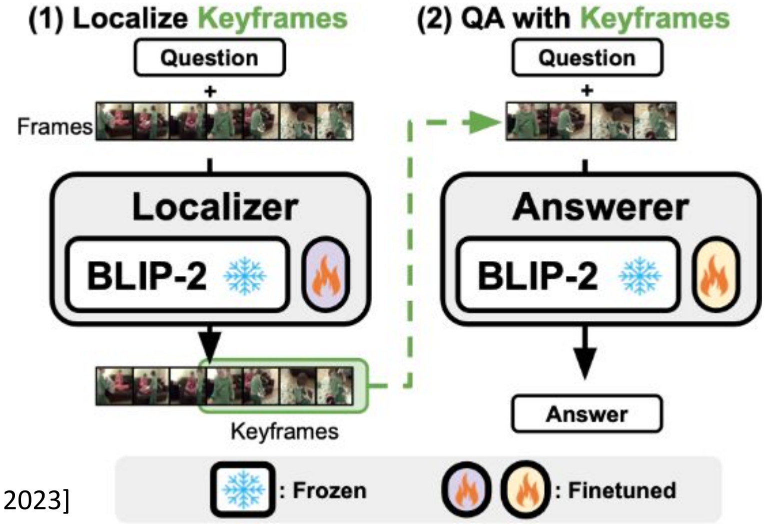
[Zhang 2022] GLIPv2: Unifying Localization and Vision-Language Understanding, Haotian Zhang et al, NeurIPS 2022.
[Zou 2023] Generalized Decoding for Pixel, Image, and Language, Xueyan Zhou et al, CVPR 2023.

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]

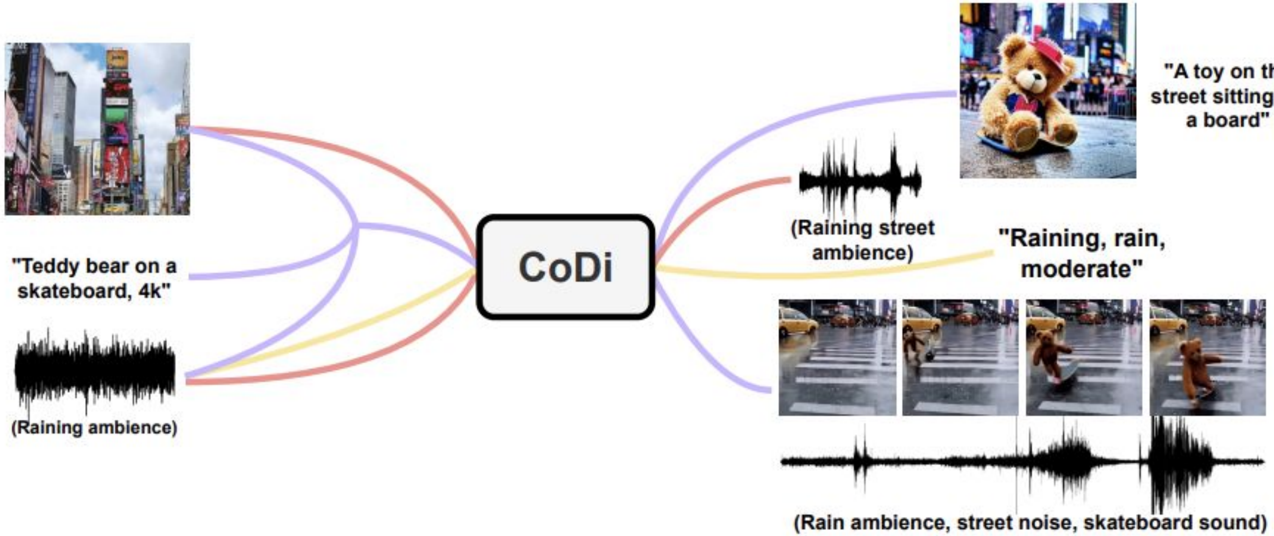
Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the 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.