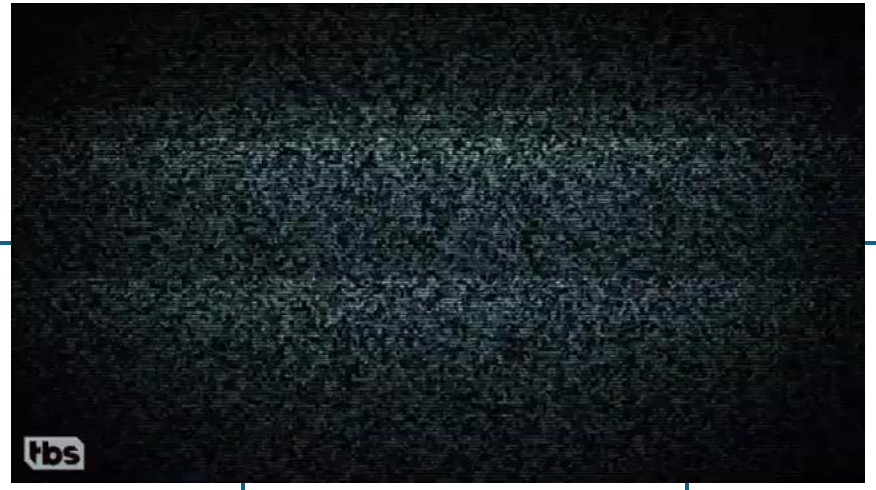
Shot2Story

A New Benchmark for Comprehensive Understanding of Multi-shot Videos

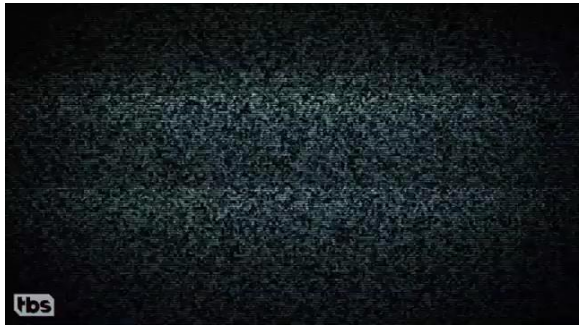
Mingfei Han
ReLER Lab, AAII
University of Technology Sydney

Video clip – What does the video convey?

Shot2Story



Shot 0: 0-6.5s



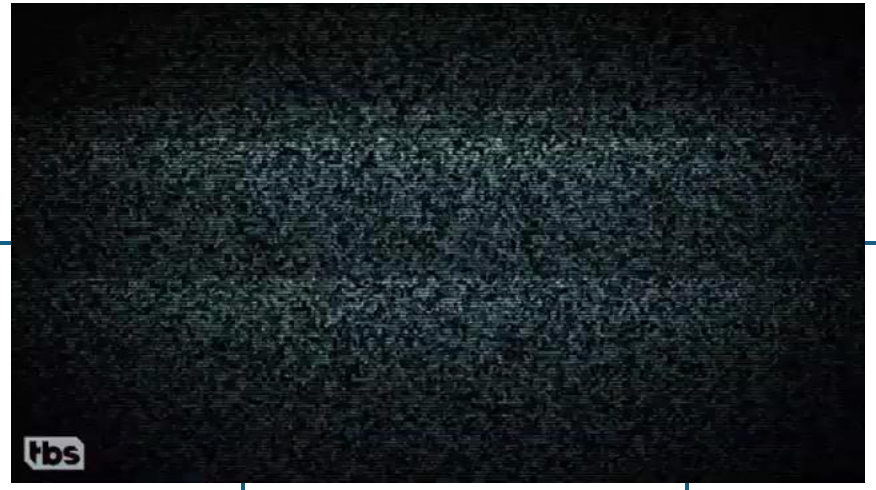
Shot 1: 6.5-10.5s

Shot 2: 10.5-13.5s

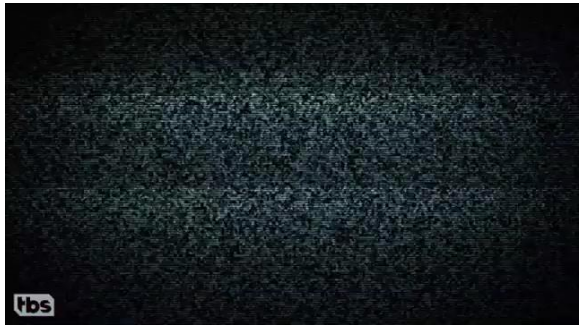
Shot 3: 13.5-15.7s

Video clip – Understand the multi-shot video clip

Shot2Story



Shot 0: 0-6.5s



Shot 1: 6.5-10.5s

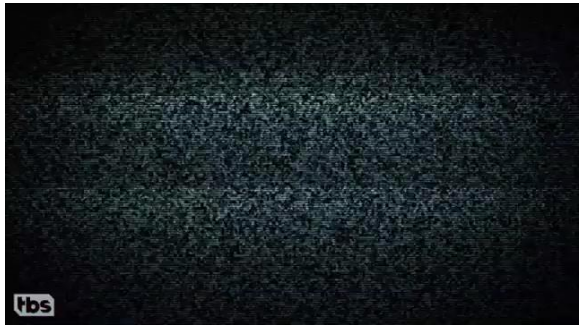
Shot 2: 10.5-13.5s

Shot 3: 13.5-15.7s

Video clip – Understand the multi-shot video clip

Shot2Story

Shot 0: 0-6.5s



Show a man hurting himself cutting an avocado. ***DON'T RISK IT***

Shot 1: 6.5-10.5s

Show scooping the avocado with a spoon. ***SWEET***

Shot 2: 10.5-13.5s

Show a woman raising her thumb and praising sth.

Shot 3: 13.5-15.7s

Show a pre-scooped avocado product

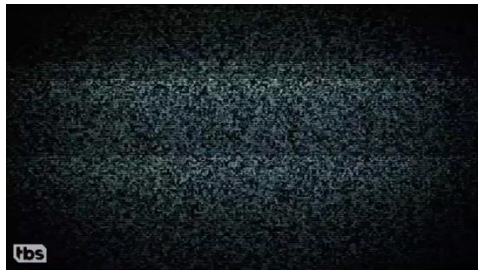
- **Real-world videos are stories**
  - Stories span **multiple** scenes, each contributing crucially to understanding.
  - Understanding videos in ***single shots*** is not enough; we must grasp how shots connect to tell the ***full story***

Shot2Story

To facilitate the research of multi-shot videos, we present

Shot2Story

providing ...

	Shot 0: 0-6.5s	Shot 1: 6.5-10.5s	Shot 2: 10.5-13.5s	Shot 3: 13.5-15.7s
				
Single-shot visual captions	It's a man in a kitchen chopping something with a knife. The man is ...	It's a close-up of a person scooping avocado out with a spoon on a wooden cutting...	It's a woman in a pink shirt with a cast on her arm. She holds her thumb up ...	It's a bowl of guacamole with chips on the side. There is a plastic container of ...
Single-shot narration captions	The background voice says don't risk getting injured from cutting up avocados.	The background voice says the product is refrigerated and pre-scooped for ...	N/A	N/A

Multi-shot video summary

The video begins with a man in a kitchen, wearing a T-shirt with the number 19 on it. He is chopping something with a knife, but he seems to hurt himself and appears in pain. This scene is presented in black and white for dramatic effect. As a cautionary message, the yellow words "DON'T RISK IT" appear on the screen. The video then transitions to a bowl of guacamole with chips ...

Multi-shot video QA

Q: What overarching message is the video conveying?
A: The video conveys that manual avocado preparation is risky and promotes a safer pre- scooped product.

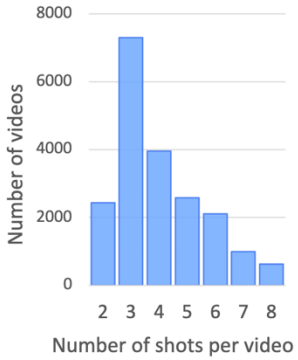
Q: What is the visual narrative structure of the video?
A: The video starts with an injury from manual chopping, introduces the pre-scooped product, and ends with an endorsement.

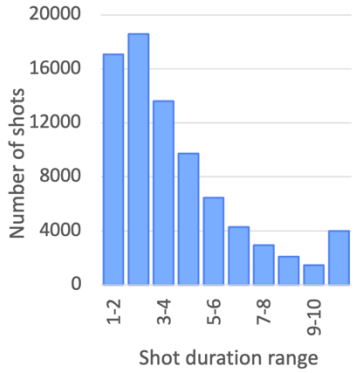
To facilitate the research of multi-shot videos, we present

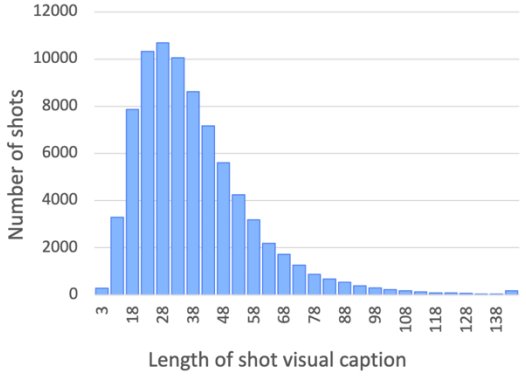
Shot2Story

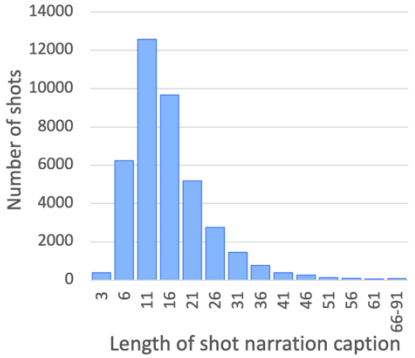
providing ...

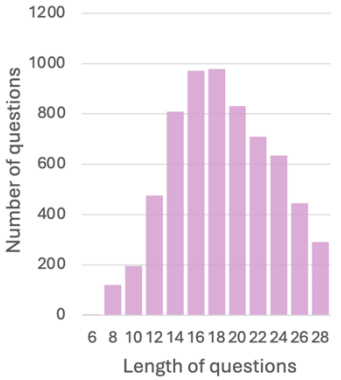
- 188k single-shot visual captions, 96k narration (audio) captions
- 11K multi-shot video question-answering pairs

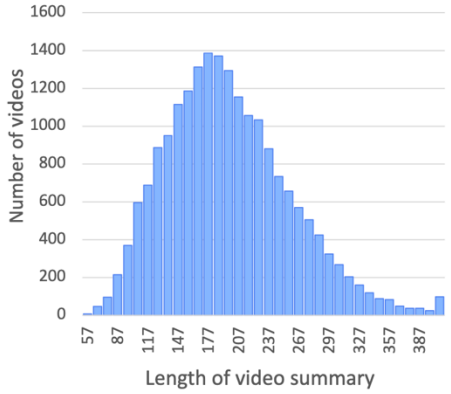


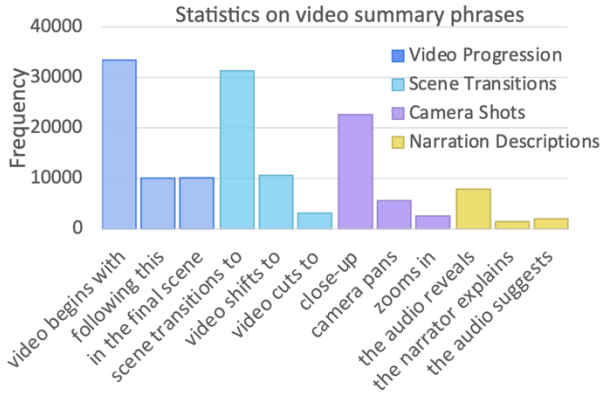


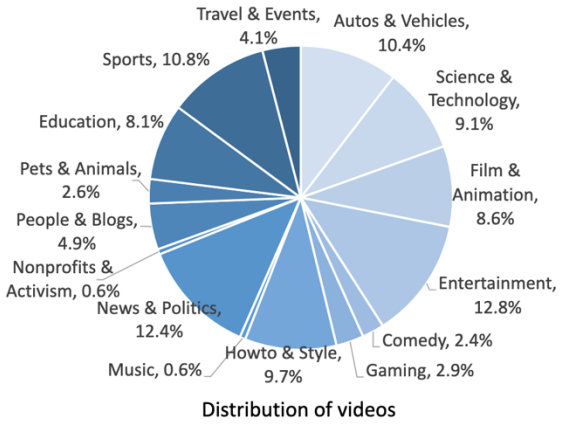


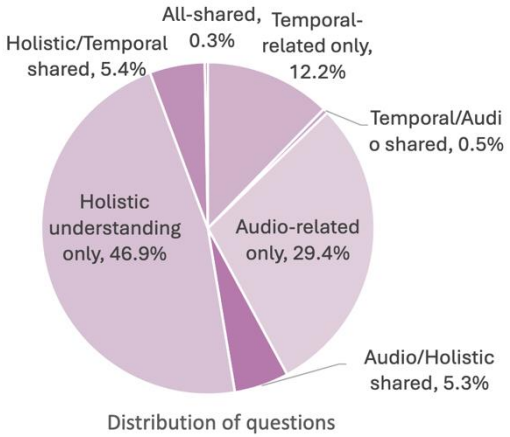






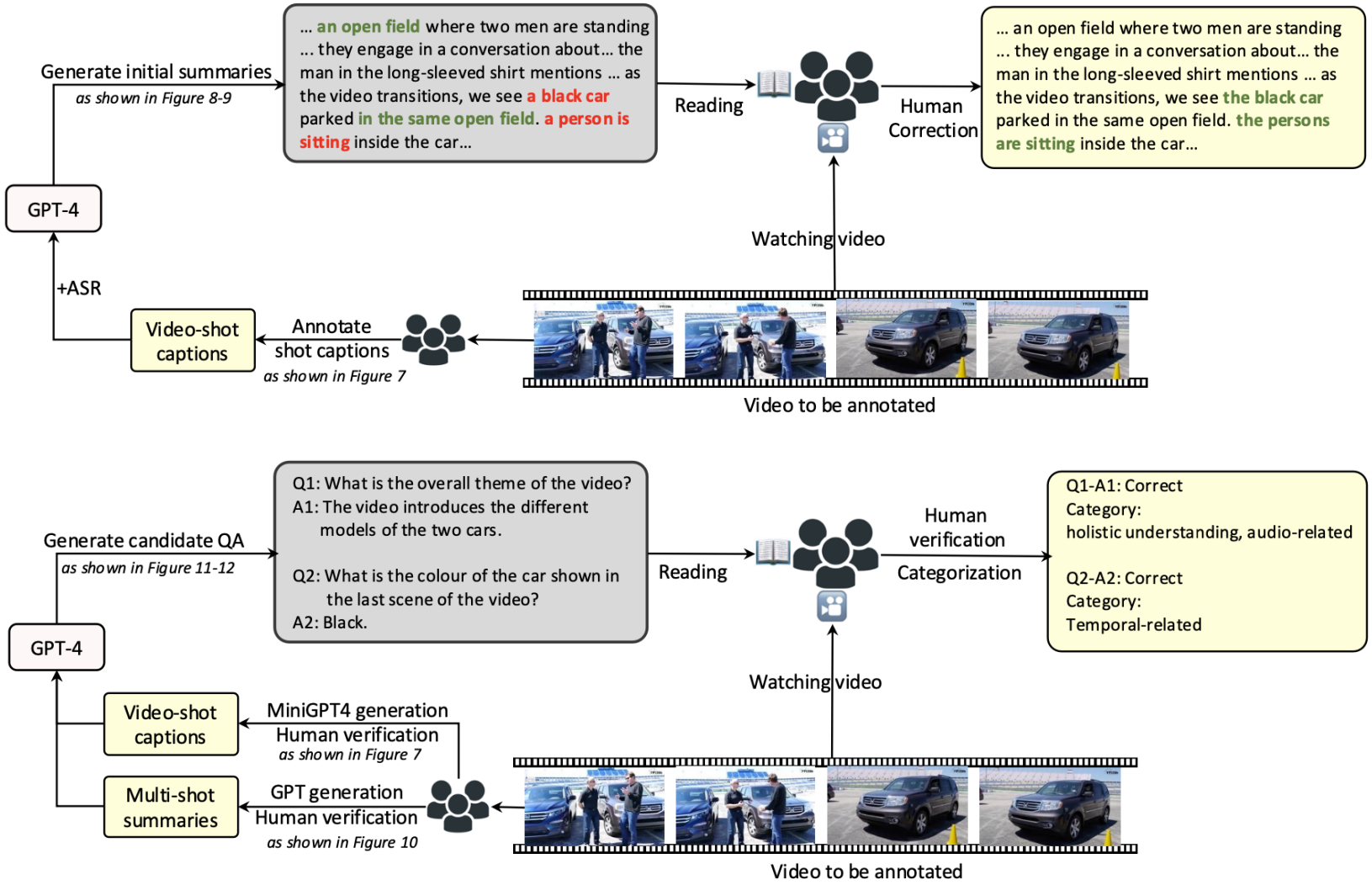






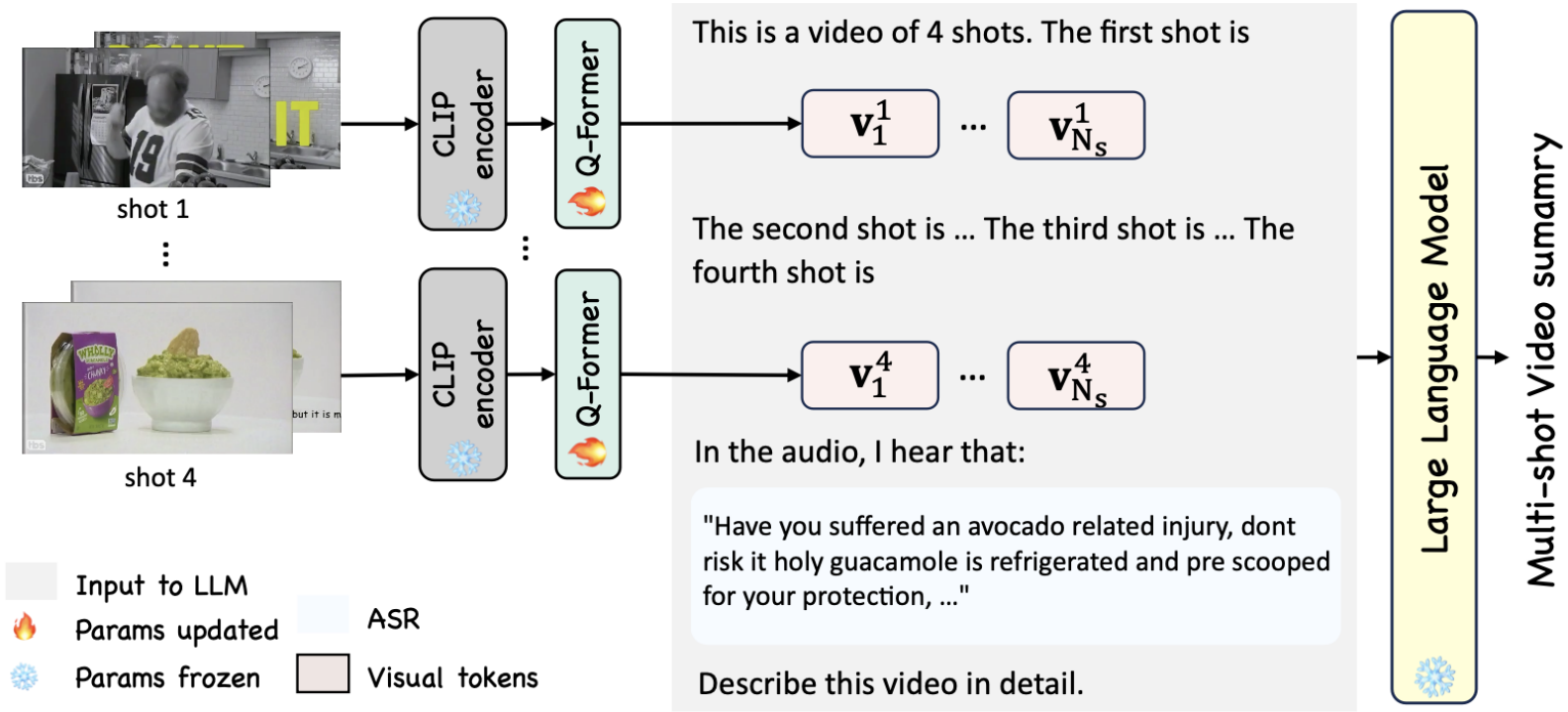
Shot2Story: Video-language benchmark for multi-shot videos

- **Human involved and rectified text annotations**



Shot2Story: Video-language benchmark for multi-shot videos

- **Baseline**



- Flexible arrangement of shot information, visual tokens and ASR texts
- Adaption to longer videos, benefiting multi-shot training approach

Shot2Story: Video-language benchmark for multi-shot videos

Model	FT Modality	B4	M	R	C
VAST	V+S	10.7	16.1	30.3	33.8
VAST	V+A+S	10.7	16.1	30.4	34.0
MiniGPT4-C	V	9.2	14.7	27.9	25.1
MiniGPT4-C	V+S	11.8	16.7	30.1	35.9
VideoChat2-C	V	8.8	16.1	27.9	23.7
VideoChat2-C	V+S	**13.8**	**18.7**	**32.1**	**43.9**

Video shot captioning task:

- Integrating subtitles (narration) consistently and significantly enhances captioning quality.

Model	FT Modality	B4	M	R	C
Video-ChatGPT w/o ASR (Maaz et al., 2023)	V	4.8	17.3	21.3	1.5
Video-ChatGPT (Maaz et al., 2023)	V+S	3.6	17.8	19.7	1.0
MiniGPT4-SUM-holistic	V+S	7.8	16.9	23.4	2.8
MiniGPT4-SUM-shot w/o ASR	V	10.4	18.5	25.8	4.8
MiniGPT4-SUM-shot	V+S	12.4	19.7	27.6	7.6
VideoChat2-SUM-shot	V+S	**12.7**	**19.8**	**28.3**	**9.0**

Multi-shot Video Summarization task:

- Explicit involvement of shot structure enhances summarization quality of multi-shot videos.

Model	Training data	IT	QA Input	Temporal related	Holistic understanding	Audio related	Overall
LLaMA-VID (Li et al., 2023e)	Cap.+QA	✓	V+T	7.9	9.7	11.4	9.7
Video-ChatGPT (Maaz et al., 2023)	Cap.+QA	✓	V+T	13.1	15.5	14.3	14.2
VideoChat2 (Li et al., 2023c)	Cap.+QA	✓	V+T	15.1	15.4	13	14.5
Video-LLaVA (Lin et al., 2023)	Cap.+QA	✓	V+T	16.4	14.8	11.7	14.3
MiniGPT4-SUM-shot	Summary	✗	T	28.9	31.9	36.7	32.5
VideoChat2-SUM-shot	Summary	✗	T	36.1	41.5	43.8	40.5

Multi-shot Video Question Answering task:

- Using detailed multi-shot summaries only, we achieve enhanced performance across different question types.

Shot2Story: Video-language benchmark for multi-shot videos

- Zero-shot VQA with video summaries generated by our model

Model	Training Data	IT	QA Input	MSRVTT QA	ActivityNet QA
VideoChat (Li et al., 2023b)	Cap.+QA	✓	V+T	45.0	26.5
Video-ChatGPT (Maaz et al., 2023)	Cap.+QA	✓	V+T	49.3	35.2
MovieChat (Song et al., 2023)	Cap.+QA	✓	V+T	52.7	45.7
LLaMA-VID (Li et al., 2023e)	Cap.+QA	✓	V+T	57.7	47.4
VideoChat2 (Li et al., 2023c)	Cap.+QA	✓	V+T	54.1	**49.1**
Video-LLaVA (Lin et al., 2023)	Cap.+QA	✓	V+T	**59.2**	45.3
MiniGPT4-SUM-shot	Summary	✗	T	57.7	45.6
VideoChat2-SUM-shot	Summary	✗	T	58.5	47.1

- Our Shot2Story-based summaries achieve strong **zero-shot generalization**
- **Competitive performance** even without direct QA training data.
- **Summaries alone** (text-based) rival multimodal (visual + text) methods.

Shot2Story: Video-language benchmark for multi-shot videos

- Paper: https://arxiv.org/pdf/2312.10300
- Project page: https://mingfei.info/shot2story/
- Data: https://huggingface.co/datasets/mhan/shot2story
- Code: https://github.com/bytedance/Shot2Story

Shot2Story