VidChapters-7M: Video Chapters at Scale

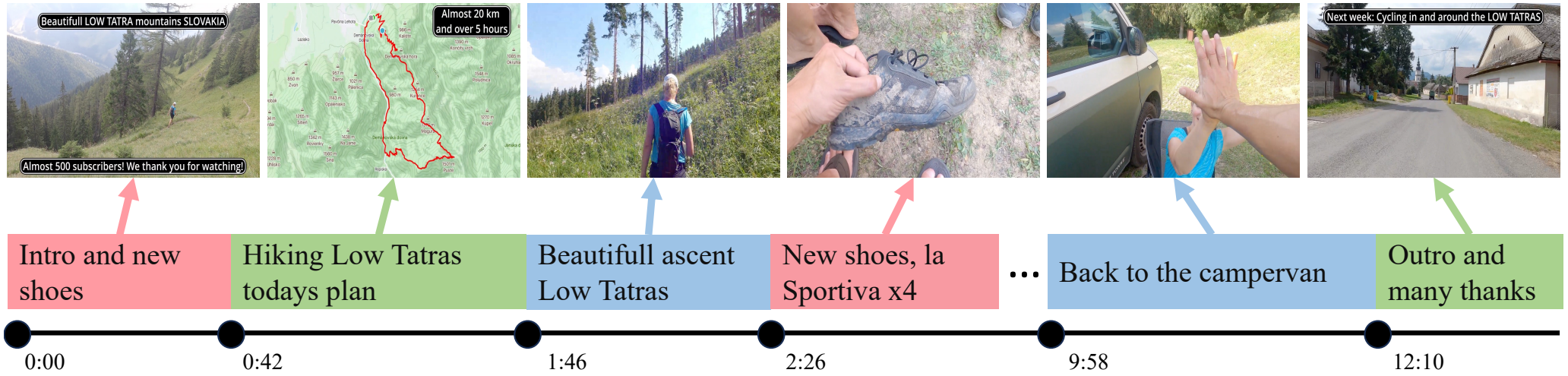
Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, Cordelia Schmid

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

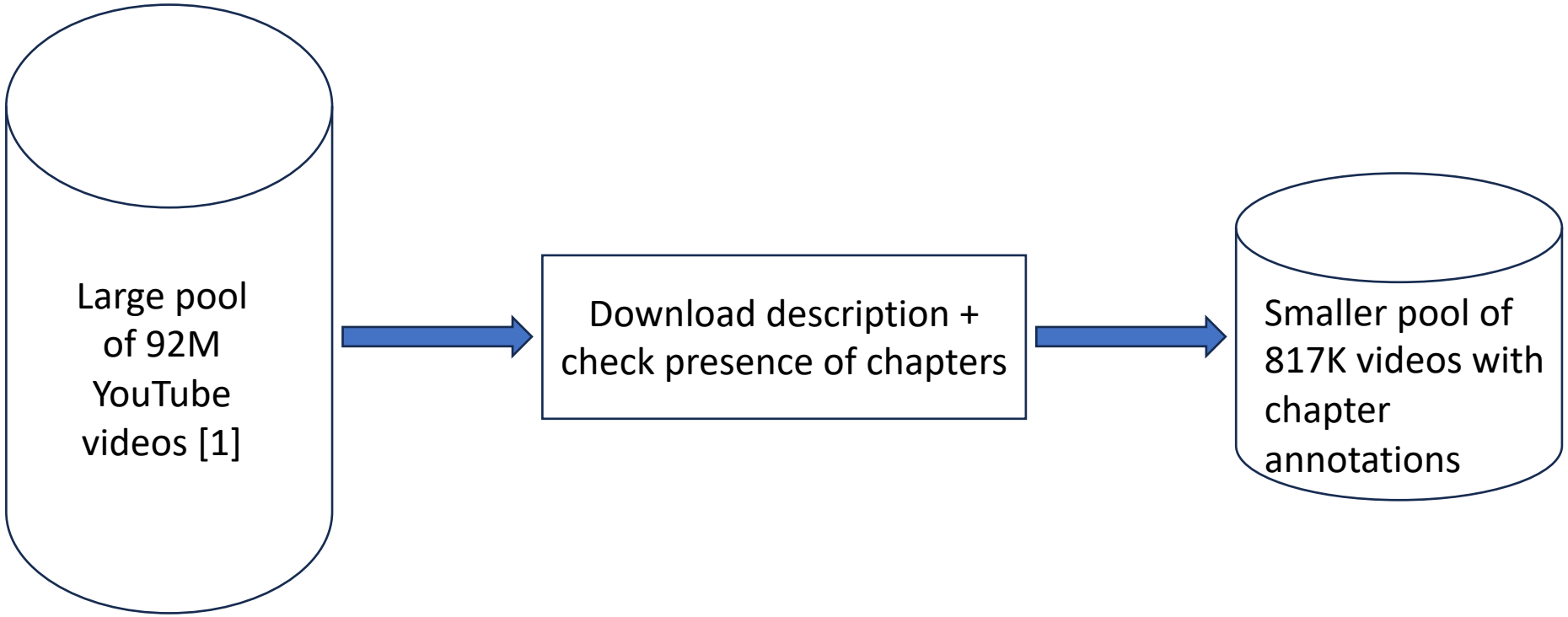
Paper: https://arxiv.org/abs/2309.13952

Video Chapter Generation

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



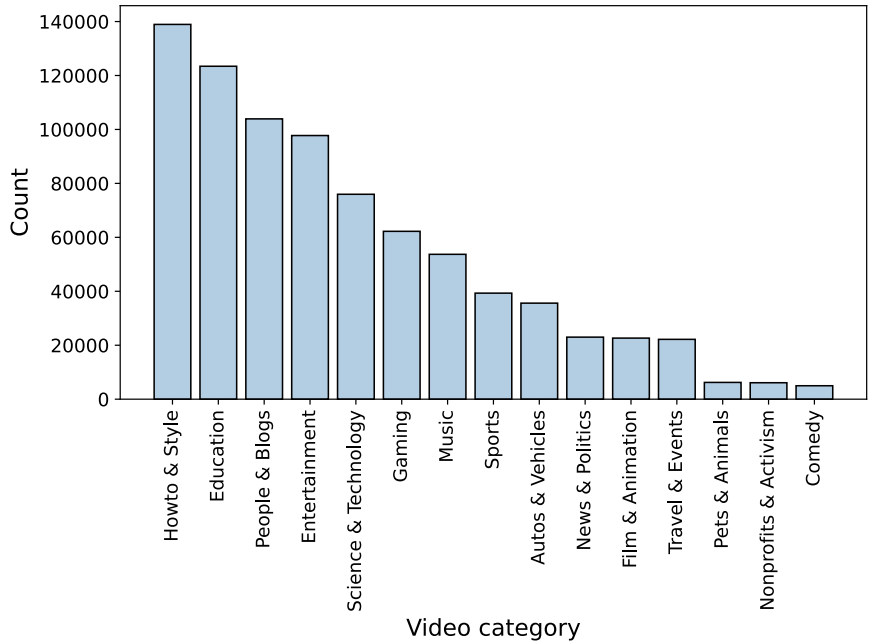
Data collection procedure



[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	Number of videos	Video duration (min)	Number of descriptions	Annotations
HowTo100M [64]	1M	7	136M	Speech transcripts
YT-Temporal-1B [118]	**19M**	6	∼ **900M**	Speech transcripts
HD-VILA-100M [108]	3M	7	103M	Speech transcripts
ActivityNet Captions [42]	20K	3	100K	Dense Captions
YouCook2 [127]	2K	6	15K	Dense Captions
ViTT [36]	8K	4	56K	Dense Captions
Ego4D [29]	10K	**23**	4M	Dense Captions
VidChapters-7M (Ours)	817K	**23**	7M	**Speech transcripts + User-annotated Chapters**

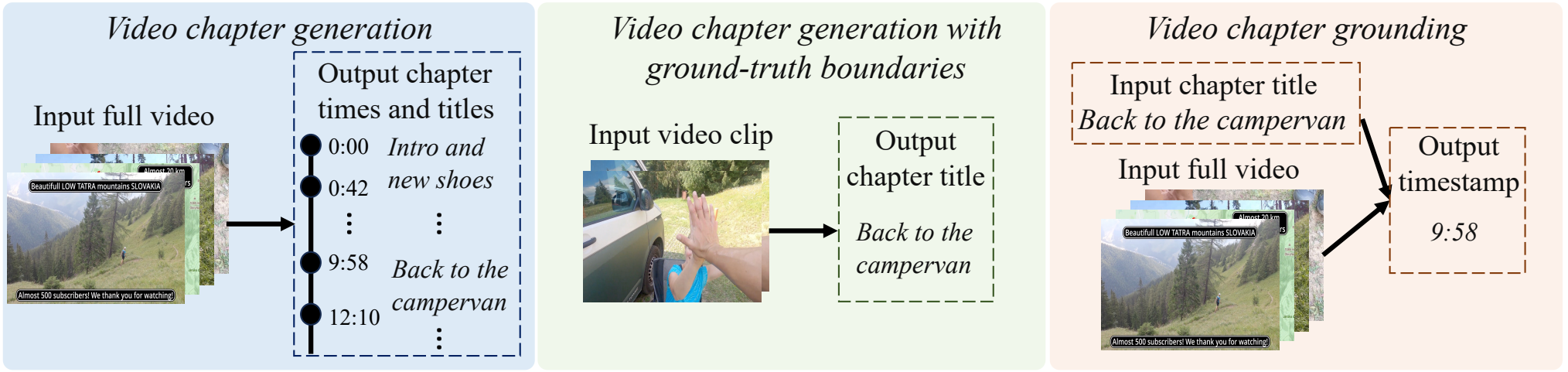
Table 1: **Comparison of VidChapters-7M with existing datasets**. We consider open-sourced video datasets that contain dense natural language descriptions aligned over time. VidChapters-7M is much larger than current dense video captioning datasets. Compared to datasets with ASR (top 3 rows), it is smaller in the total number of videos but contains longer videos with richer annotations (chapters).

Manual assessment

Type of chapter titles	Percentage
Speech and visual	49
Audio and visual	2
Speech-only	26
Visual-only	3
Audio-only	3
Structure-only	14
Unrelated	3

Table 2: **Manual assessment of the informativeness of chapter titles in the VidChapters-7M dataset over a random sample of 100 videos.** Video chapter titles can be based on speech and vision; audio and vision; vision, audio or speech alone; or only on the structure of the video (*e.g.* "step 1", "step 2" etc). In a small number of cases, video chapters are unrelated to the video content.

New benchmarks



Video chapter generation

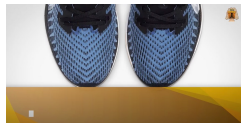
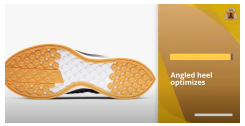
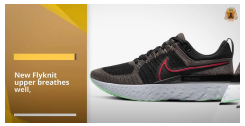
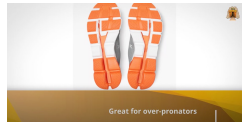
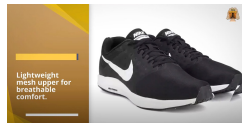
Method	Modalities	Pretraining Data	Finetuned	S	B1	B2	B3	B4	C	M	RL
Text tiling [32] + Random	Speech	∅	✗	0.4	0.6	0.2	0.1	0.0	0.8	0.7	0.6
Text tiling [32] + LLaMA [93]	Speech	Text mixture	✗	0.2	0.4	0.1	0.1	0.0	0.5	0.3	0.4
Shot detect [92] + BLIP-2 [51]	Visual	129M image-texts	✗	0.6	0.7	0.3	0.1	0.1	0.2	0.6	0.8
Vid2Seq [114]	Speech+Visual	C4 + HowTo100M	✗	0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.1
PDVC [101]	Visual	∅	✓	6.8	9.4	3.7	1.4	0.9	35.8	9.4	11.4
Vid2Seq [114]	Speech	C4	✓	10.2	9.5	6.7	4.0	2.7	48.8	8.5	11.0
Vid2Seq [114]	Speech	C4 + HowTo100M	✓	10.5	9.9	7.0	4.2	2.9	50.7	8.7	11.4
Vid2Seq [114]	Visual	C4	✓	3.1	2.3	1.5	0.6	0.5	10.9	2.2	2.9
Vid2Seq [114]	Visual	C4 + HowTo100M	✓	5.5	4.5	2.8	1.2	0.9	21.1	4.1	5.5
Vid2Seq [114]	Speech+Visual	C4	✓	10.6	9.9	7.0	4.2	2.8	51.3	8.8	11.6
Vid2Seq [114]	Speech+Visual	C4 + HowTo100M	✓	**11.4**	**10.9**	**7.7**	**4.6**	**3.1**	**55.7**	**9.5**	**12.6**

Table 3: **Video chapter generation (global metrics) on VidChapters-7M test set.** Here, finetuned refers to finetuning on the VidChapters-7M train set, and speech refers to transcribed speech (ASR).

Method	Modalities	Pretraining Data	Finetuned	R@5s	R@3s	R@0.5	R@0.7	P@5s	P@3s	P@0.5	P@0.7
Text tiling [32]	Speech	∅	✗	9.4	5.8	23.6	8.9	12.6	7.9	26.0	8.8
Shot detect [92]	Visual	∅	✗	31.2	27.4	24.9	12.5	33.2	29.7	18.0	8.7
Vid2Seq [114]	Speech+Visual	C4 + HowTo100M	✗	10.7	9.5	5.8	0.2	23.3	18.5	1.9	0.8
PDVC [101]	Visual	∅	✓	21.1	17.8	31.2	22.5	**45.3**	**40.2**	**47.2**	**26.9**
Vid2Seq [114]	Speech	C4	✓	**37.8**	**29.5**	44.6	26.1	29.0	23.0	38.0	23.4
Vid2Seq [114]	Speech	C4 + HowTo100M	✓	36.7	28.9	46.5	27.2	29.5	23.3	40.4	24.8
Vid2Seq [114]	Visual	C4	✓	35.3	26.4	23.6	8.7	17.9	13.6	17.2	7.1
Vid2Seq [114]	Visual	C4 + HowTo100M	✓	33.5	25.0	33.0	14.5	19.5	14.7	26.2	12.5
Vid2Seq [114]	Speech+Visual	C4	✓	36.3	28.6	45.8	26.9	29.9	23.8	40.9	24.9
Vid2Seq [114]	Speech+Visual	C4 + HowTo100M	✓	36.4	28.5	**48.2**	**28.5**	30.3	24.0	43.1	26.4

Table 4: **Video chapter generation (segmentation metrics) on VidChapters-7M test set.**

Qualitative examples: speech helps

Input Speech	If you are looking for the best Nike running shoes, here is a collection you have got to see.	Number 1. Most Popular. Zoom Pegasus Turbo 2. A souped-up, speed-oriented version of the Pegasus, the Peg Turbo keeps the winning combo of Zoomex and React foams found in the first version.	Unfortunately, the new thin mesh upper has issues. Its minimal heel support means you have to cinch the laces down for a secure fit, but the tongue isn't thick or long enough to prevent the laces from causing irritation.	Number 2. Nike Men's Running Shoes. The new trend in stability shoes is less interference, and the Infinity Run follows that principle by providing comfort, support, and a smooth ride without messing up your natural movement. […]	Number 3. On Women's CloudFlyer Running Shoes. Provide your foot with the cushion it deserves with the On CloudFlyer. Utilizing plush clouds built from zero-gravity foam and a wider CloudTek platform, this daily trainer provides supreme cushioning in a more stable package.	In order to reduce over pronation, the shoe features firmer medial elements that redirect force to the lateral side of the runner's foot. Paired with an even stiffer speed board, the shoe promotes a quicker heel-to-toe transfer that helps get the runner through their pronated phase.	Number 4. Nike Downshifter Men's 7 Running Shoe. The Downshifter 7 Running Shoes from Nike are designed to be lightweight, sturdy and durable, all the while providing you with optimum performance, making them a worthy investment. […]	Number 5. Nike Men's Trail Running Shoes. Made of a breathable mesh upper and a sturdy EVA sole, these Quest running shoes from Nike should pretty much be a staple in every man's shoe closet. Fly-wire cables offer your feet a secure fit, while the soft yet responsive foam is supportive […]
Input Frames								
Ground -Truth		1. Zoom Pegasus Turbo 2		2. Nike Men's Running Shoes	3. ON Women's Cloudflyer Running Shoes		4. Nike Downshifter Men's 7 Running Shoe	5. Nike Men's Trail Running Shoes
Vid2Seq (HTM +VC, no speech)		1. nike nexus running shoe.		2. nike nexus running shoe.	3. nike nexus running shoe.		4. nike nexus running shoe.	5. nike nexus running shoe.
Vid2Seq (HTM +VC)		1. zoom pegasus turbo 2.		2. nike men's running shoes.	3. on women's cloudflyer running shoes.		4. nike downshifter men's 7 running shoe.	5. nike men's trail running shoes.

Qualitative examples: vision helps

Video chapter generation given ground-truth boundaries

Method	Modalities	Pretraining Data	Finetuned	B1	B2	B3	B4	C	M	RL
Random	Speech	$\emptyset$	✗	2.4	1.3	0.9	0.7	10.4	2.2	4.4
LLaMA [93]	Speech	Text mixture	✗	0.0	0.0	0.0	0.0	0.0	0.1	0.2
BLIP-2 [51]	Visual	129M image-texts	✗	3.1	1.5	0.9	0.7	12.4	2.2	4.5
Vid2Seq [114]	Speech+Visual	C4 + HowTo100M	✗	2.0	1.2	0.9	0.6	0.9	0.3	0.6
Vid2Seq [114]	Speech	C4 + HowTo100M	✓	21.0	15.5	12.1	10.0	105.3	11.5	24.5
Vid2Seq [114]	Visual	C4 + HowTo100M	✓	10.1	5.6	3.5	2.4	47.1	5.1	14.7
Vid2Seq [114]	Speech+Visual	C4	✓	21.6	15.7	12.3	10.0	110.8	11.5	26.0
Vid2Seq [114]	Speech+Visual	C4 + HowTo100M	✓	**23.5**	**17.2**	**13.4**	**11.0**	**120.5**	**12.6**	**28.3**

Table 5: **Chapter title generation given ground-truth boundaries on VidChapters-7M test set.**

Video chapter grounding

Method	Modalities	Pretraining Data	Finetuned	R@10s	R@5s	R@3s	R@1s	R@0.3	R@0.5	R@0.7	R@0.9
Random	Speech	∅	✗	3.1	1.8	1.2	0.6	0.7	0.3	0.1	0.0
BERT [19]	Speech	BookCorpus + Wikipedia	✗	9.0	6.8	5.4	2.9	0.6	0.3	0.1	0.0
CLIP [72]	Visual	400M image-texts	✗	8.1	5.2	3.7	1.4	10.7	5.2	2.3	0.5
Moment-DETR [45]	Visual	5.4K narrated videos [45]	✗	3.2	1.6	1.1	0.5	11.3	3.6	0.8	0.1
Moment-DETR [45]	Visual	∅	✓	**21.8**	**15.5**	**12.4**	**8.3**	**37.4**	**27.3**	**17.6**	**6.4**

Table 6: **Video chapter grounding on VidChapters-7M test set.**

Transfer to dense video captioning

Method	Modalities	Pretraining Data	YouCook2 (val)					ViTT (test)				
			S	C	M	R	P	S	C	M	R	P
PDVC [101]	V	∅	4.4	22.7	4.7	—	—	—	—	—	—	—
E2ESG [130]	T+V	C4 + WikiHow	—	25.0	3.5	20.7	20.6	—	25.0	8.1	32.2	32.1
Vid2Seq [114]	T+V	C4 + HTM	8.3	48.3	9.5	27.1	27.0	—	—	—	—	—
Vid2Seq [114]	T+V	C4 + YT-Temporal-1B	7.9	47.1	9.3	27.9	27.8	13.5	43.5	8.5	42.6	46.2
PDVC†	V	∅	4.8	28.8	5.8	22.6	33.1	9.4	40.6	**16.5**	19.2	37.4
PDVC†	V	VC (Chap.)	5.9	34.7	7.5	28.8	**36.4**	10.1	41.5	16.1	21.3	37.2
Vid2Seq†	T+V	C4 + HTM	8.6	53.2	10.5	29.2	26.2	14.1	44.8	8.7	43.8	44.5
Vid2Seq†	T+V	C4 + VC (ASR+Chap.)	9.8	62.9	11.7	32.5	30.1	**15.1**	**50.9**	9.6	45.1	46.7
Vid2Seq†	T+V	C4 + HTM + VC (ASR)	8.4	50.1	10.3	29.7	26.3	14.3	45.6	8.8	43.7	44.9
Vid2Seq†	T+V	C4 + HTM + 1% of VC (ASR+Chap)	8.8	52.7	10.4	29.3	27.6	13.5	41.6	8.2	44.7	42.1
Vid2Seq†	T+V	C4 + HTM + 10% of VC (ASR+Chap.)	9.9	63.9	12.1	32.4	31.4	14.5	47.4	9.2	45.3	45.9
Vid2Seq†	T+V	C4 + HTM + VC (ASR+Chap.)	**10.3**	**67.2**	**12.3**	**34.0**	31.2	15.0	50.0	9.5	**45.5**	**46.9**

Table 7: **Comparison with the state of the art on the YouCook2 and ViTT dense video captioning benchmarks.** T: Transcribed speech, V: Visual, HTM: HowTo100M [64], VC: VidChapters-7M, Chap.: Chapters. † denote results of our experiments.

Zero-shot dense video captioning

Method	Modalities	Pretraining Data	YouCook2 (val)					ViTT (test)				
			S	C	M	R	P	S	C	M	R	P
Text tiling [32] + Random	T	∅	0.3	0.9	0.3	3.8	6.6	0.3	0.6	0.6	11.6	24.4
Text tiling [32] + LLaMA [93]	T	Text mixture	0.2	0.6	0.2	3.8	6.6	0.2	0.6	0.5	11.6	24.4
Shot detect [92] + BLIP-2 [51]	V	129M image-texts	0.6	1.0	0.5	8.9	5.5	0.2	0.1	0.2	3.1	13.7
Vid2Seq [114]	V	C4 + VC (ASR)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.8
Vid2Seq [114]	V	C4 + VC (Chap.)	0.7	1.1	0.5	21.3	8.6	1.5	1.9	0.6	18.9	10.4
Vid2Seq [114]	T+V	C4 + HTM	0.0	0.1	0.0	0.5	0.6	0.0	0.0	0.0	0.5	1.0
Vid2Seq [114]	T+V	C4 + VC (ASR)	0.1	0.1	0.0	1.1	0.9	0.0	0.0	0.0	0.7	0.6
Vid2Seq [114]	T+V	C4 + VC (Chap.)	0.1	0.2	0.1	0.7	1.4	0.7	1.1	0.3	14.3	12.8
Vid2Seq [114]	T+V	C4 + VC (ASR+Chap.)	3.2	10.2	2.9	20.6	19.7	**9.1**	**30.2**	**6.7**	**33.8**	**40.8**
Vid2Seq [114]	T+V	C4 + HTM + VC (ASR)	0.0	0.1	0.0	1.2	0.9	0.0	0.0	0.0	0.8	0.7
Vid2Seq [114]	T+V	C4 + HTM + 1% of VC (ASR+Chap.)	2.7	7.2	2.1	18.1	17.3	5.5	15.5	4.3	31.3	37.1
Vid2Seq [114]	T+V	C4 + HTM + 10% of VC (ASR+Chap.)	3.2	11.5	3.0	19.4	19.2	6.4	21.6	5.3	31.0	38.2
Vid2Seq [114]	T+V	C4 + HTM + VC (ASR+Chap.)	**3.9**	**13.3**	**3.4**	**22.3**	**20.1**	9.0	28.0	6.5	33.7	40.1

Table 8: **Zero-shot dense video captioning on the YouCook2 and ViTT benchmarks.** T: Transcribed speech, V: Visual, HTM: HowTo100M [64], VC: VidChapters-7M, Chap.: Chapters.