ShareGPT4Video:
Improving Video Understanding and Generation with Better Captions

Lin Chen[1,4] , Xilin Wei[4], Jinsong Li[2,4],Xiaoyi Dong[2,4], Pan Zhang[4], Yuhang Zang[4],
Zehui Chen[1], Haodong Duan[4], Bin Lin[3], Zhenyu Tang[3],
Li Yuan[3], Yu Qiao[3], Dahua Lin[2,4,5] Feng Zhao[1], Jiaqi Wang[4]

University of Science and Technology of China[1]
The Chinese University of Hong Kong[2]
Peking University[3] Shanghai AI Laboratory[4]
CPII under InnoHK

Introduction

- Motivation
  - Mismatch between rich semantic and temporal video content and brief captions in existing data

Panda: The hotel room has a beautiful view of the eiffel tower and breakfast is set up on the table.

Video ChatCaptioner: In the video, a woman is sitting at a table indoors. She is alone at the table and there are no visible objects on the table. The woman appears to be sitting still and there is no visible movement outside the window. There are no other people in the room and no visible decoration or artwork on the walls. The room is well lit and there are no plants or flowers present.

Introduction

- Challenges
  - Inter-frame precise temporal change understanding

Multi-frame with Timestamps Caption: **The video maintains a concentrated focus on food preparation from slicing and chopping to blending and dressing, ending with the assembly of a vibrant and texturally diverse salad.In the video, a man with tattooed forearms meticulously prepares a salad in a well-equipped kitchen. He begins by slicing a cucumber, then moves on to cutting a red onion, chopping garlic, and slicing a green, leafy herb. Throughout these tasks, the tomatoes, cucumber slices, and chopped onions remain neatly arranged, highlighting his organized approach.**

After completing the chopping, he arranges the ingredients on a cutting board and transitions to combining them in a large glass bowl. He mixes the ingredients together to ensure an even distribution of flavors before preparing the salad dressing. He pours olive oil into a smaller bowl, adds lemon juice, mustard, black pepper and powder resembling salt, and then whisk vigorously for an emulsified consistency.

Returning to chopping, he prepares a shallot or red onion next to the oil-mustard mixture. Finally, he adds croutons from a metal bowl into the salad, signaling the completion of the preparation process.

Throughout the video, the kitchen setting remains consistent as the man focuses on crafting a meticulous and flavorful dish. The video maintains a concentrated focus on the food preparation, from slicing and chopping to blending and dressing, ultimately resulting in the assembly of a vibrant and texturally diverse salad.

Temporal Mismatch ☹

Introduction

- Challenges
  - Intra-frame detailed content description

Image Grid Caption: The video depicts a detailed salad preparation process by a man with tattooed forearms in a well-equipped kitchen. He begins by slicing a cucumber on a cutting board, then chopping garlic. Each ingredient remains separate during preparation, highlighting an organized approach.

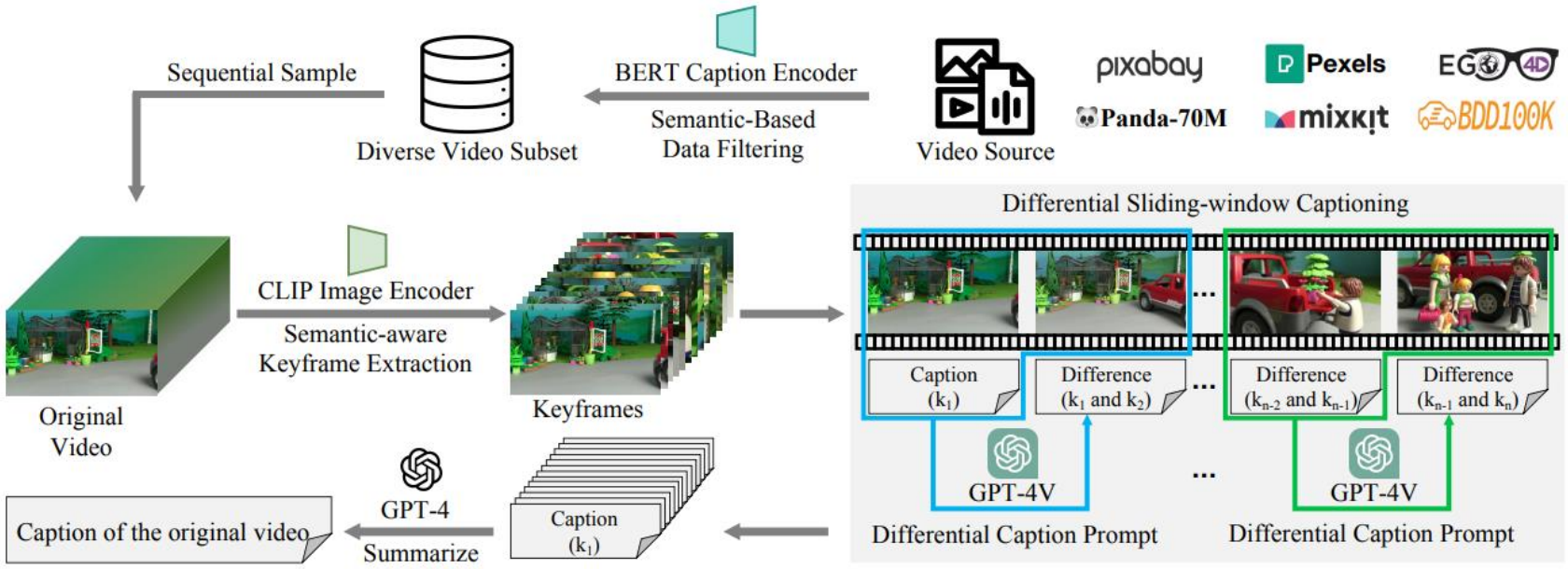
Afterward, he arranges the ingredients in a large glass bowl and mixes them together evenly. He then prepares a salad dressing by combining olive oil, lemon juice and whisking vigorously for an emulsified consistency.

Lastly, he adds croutons from a metal bowl to the salad, indicating the final step. Throughout the process, the kitchen setting remains consistent, with the man's actions focused on creating a meticulous and flavorful dish. The video maintains a concentrated focus on food preparation from slicing and chopping to blending and dressing, ending with the assembly of a vibrant and texturally diverse salad.

Description Omission

Methodology

- Semantic-Based Data Filtering
- Semantic-aware Key-frame Extraction
- Differential Sliding-window Captioning



DiffSW

Keyframes

00:00

DiffSW

Keyframes

00:00

Caption of Keyframe 1

DiffSW

Keyframes

00:00

02:00

Caption of Keyframe 1

DiffSW

Keyframes

00:00

02:00

Caption of Keyframe 1

Difference between Keyframe1 & Keyframe2

DiffSW

Keyframes

00:00

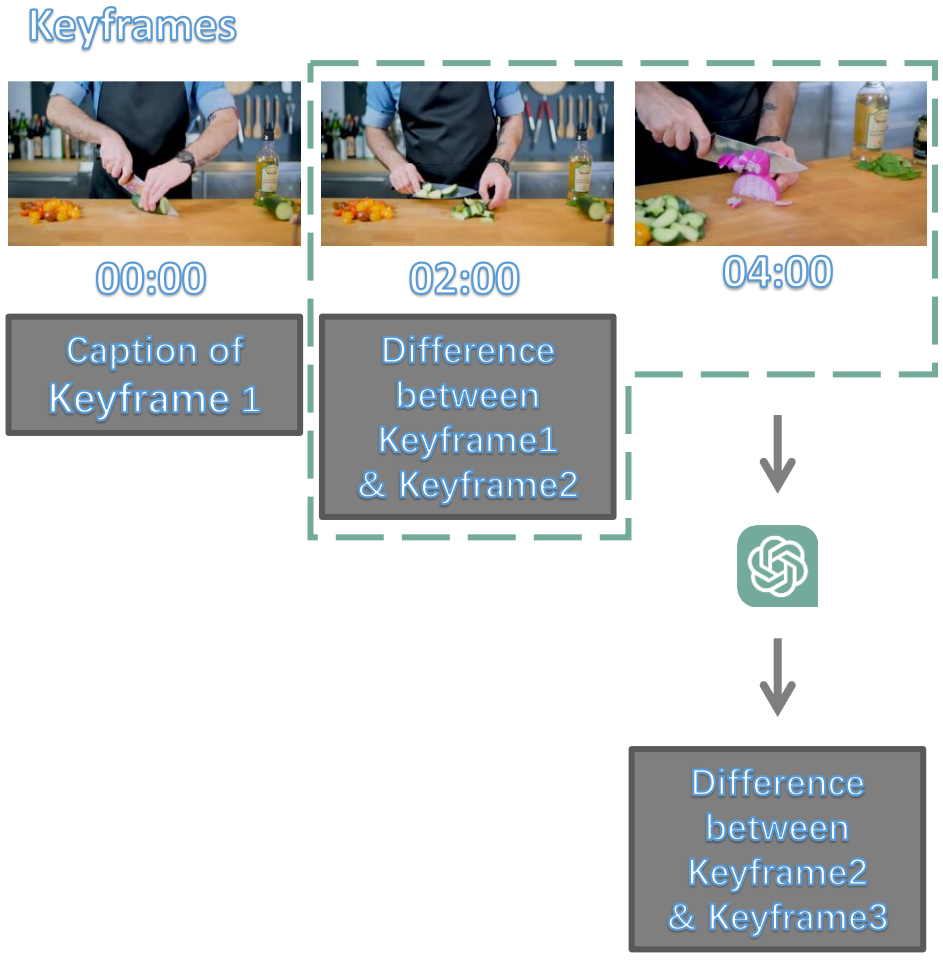
02:00

04:00

Caption of Keyframe 1

Difference between Keyframe1 & Keyframe2

DiffSW


Keyframes
00:00
02:00
04:00
Caption of Keyframe 1
Difference between Keyframe1 & Keyframe2
Difference between Keyframe2 & Keyframe3

DiffSW

Keyframes

00:00

Caption of Keyframe 1

02:00

Difference between Keyframe1 & Keyframe2

04:00

Difference between Keyframe2 & Keyframe3

DiffSW

Keyframes


00:00

Caption of Keyframe 1


02:00

Difference between Keyframe1 & Keyframe2


04:00

Difference between Keyframe2 & Keyframe3


06:00

Difference between Keyframe3 & Keyframe4


32:00

Difference between Keyframe14 & Keyframe15


34:00

Difference between Keyframe15 & Keyframe16


38:00

Difference between Keyframe16 & Keyframe17


40:00

DiffSW

Keyframes








00:00 | 02:00 | 04:00 | 06:00 | 32:00 | 34:00 | 38:00 | 40:00

Caption of Keyframe 1

Difference between Keyframe1 & Keyframe2

Difference between Keyframe2 & Keyframe3

Difference between Keyframe3 & Keyframe4

Difference between Keyframe14 & Keyframe15

Difference between Keyframe15 & Keyframe16

Difference between Keyframe16 & Keyframe17

Difference between Keyframe17 & Keyframe18

DiffSW

Keyframes


00:00

Caption of Keyframe 1


02:00

Difference between Keyframe1 & Keyframe2


04:00

Difference between Keyframe2 & Keyframe3


06:00

Difference between Keyframe3 & Keyframe4


32:00

Difference between Keyframe14 & Keyframe15


34:00

Difference between Keyframe15 & Keyframe16


38:00

Difference between Keyframe16 & Keyframe17


40:00

Difference between Keyframe17 & Keyframe18

DiffSW

Keyframes

00:00 | 02:00 | 04:00 | 06:00 | 32:00 | 34:00 | 38:00 | 40:00

Caption of Keyframe 1	Difference between Keyframe1 & Keyframe2	Difference between Keyframe2 & Keyframe3	Difference between Keyframe3 & Keyframe4	Difference between Keyframe14 & Keyframe15	Difference between Keyframe15 & Keyframe16	Difference between Keyframe16 & Keyframe17	Difference between Keyframe17 & Keyframe18

↓

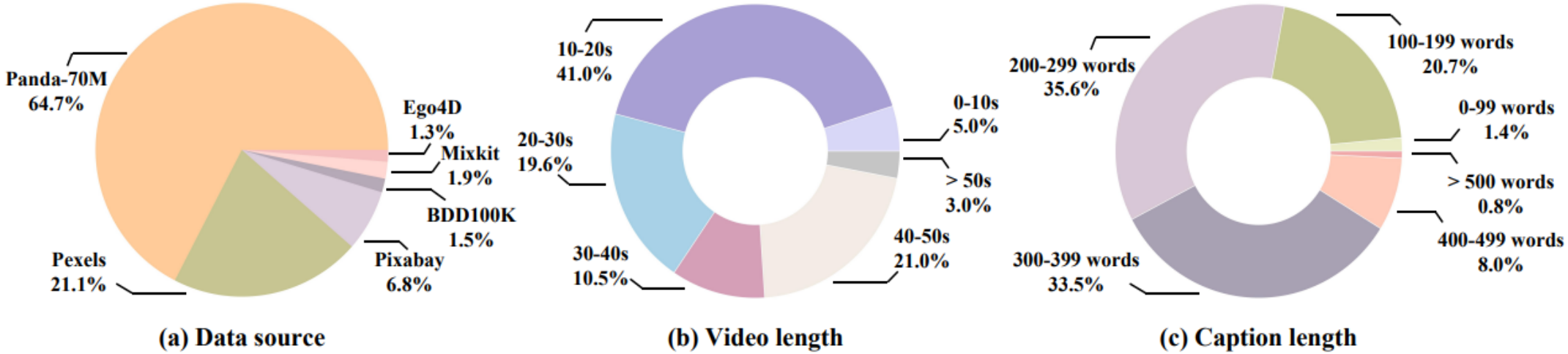
Summarize

↓

Caption of the video

Experiments

- Analysis of the ShareGPT4Video dataset



Experiments

- Video Understanding

Table 1: **The gain from high-quality captions is universal among model architectures and scales.** We report the baseline based on their public checkpoints. The best results are **bold**.

Model	VideoBench	MVBench	TempCompass	Avg.
VideoLLaVA-7B [40]	34.5	43.0	50.6	42.7
VideoLLaVA-7B+Ours	**35.2**	**43.6**	**52.7**	**43.8**
LLaMA-VID-7B [38]	36.5	41.3	48.1	42.0
LLaMA-VID-7B+Ours	**38.2**	**43.2**	**50.6**	**44.0**
LLaMA-VID-13B [38]	48.3	43.3	51.4	47.7
LLaMA-VID-13B+Ours	**52.4**	**44.2**	**53.3**	**50.0**

Table 2: **Combined with VQA data, detailed captions can benefit LVLMs more compared to short captions.** The baseline (first row) utilizes only 153K VQA data. The best results are in **bold**.

Caption	Unlock ViT	VideoBench	MVBench	TempCompass	Avg.
–	×	37.3	47.2	57.2	47.2
short	×	36.9	47.5	56.1	46.8
short	✓	37.5	47.9	56.9	47.4
detailed	×	40.7	50.3	60.7	50.6
detailed	✓	**41.2**	**51.2**	**61.5**	**51.3**

Table 3: **Comparison with SOTA methods on TempCompass.** With 7B parameters, ShareGPT4Video-8B outperforms competitors in 19 out of 20 dimensions, despite these competitors using larger training data or more parameters. The best results are **bold** and the second-best results are underlined.

Model	Multi-Choice QA					Yes/No QA					Caption Matching					Caption Generation					Avg.
	AC	DI	SP	EV	AT	AC	DI	SP	EV	AT	AC	DI	SP	EV	AT	AC	DI	SP	EV	AT	
Valley-7B [47]	47.0	29.3	32.5	18.9	29.9	58.1	52.0	52.5	50.3	52.9	65.0	53.8	52.6	53.0	53.8	54.0	31.0	32.7	34.2	41.4	33.4
PandaGPT-13B [14]	35.5	27.8	29.3	31.8	30.9	53.0	49.6	50.8	53.7	52.2	56.6	51.4	44.3	55.0	49.0	23.7	25.7	26.0	29.8	32.6	40.4
VideoLLaMA-13B [83]	54.1	24.5	28.1	32.8	28.5	68.1	46.0	48.8	51.8	50.9	73.1	47.4	47.1	52.0	48.3	54.3	21.3	13.9	38.5	33.9	43.3
VideoChatGPT-7B [50]	47.0	31.6	28.4	37.1	30.9	52.5	50.0	49.5	51.0	50.0	64.6	48.6	47.8	49.3	48.6	40.9	28.4	24.5	31.8	33.9	42.4
mPLUG-Owl-7B [76]	66.6	29.3	32.2	34.8	35.4	64.4	50.6	51.2	51.3	52.0	56.9	45.3	46.4	49.3	49.0	46.5	28.2	30.4	31.2	36.5	44.5
VideoLLaVA-7B [36]	70.4	32.2	38.2	41.4	39.9	74.3	51.8	50.3	49.2	51.1	88.2	53.8	**61.9**	57.0	58.3	50.8	28.7	23.2	38.2	33.6	49.9
LLaMA-VID-7B [38]	58.6	29.9	29.3	30.5	26.0	63.0	48.8	49.2	48.4	52.7	72.7	45.6	52.2	49.0	49.0	53.0	28.0	21.9	35.5	35.9	44.2
ShareGPT4Video-8B	**87.6**	**34.6**	**47.5**	**62.9**	**64.2**	**75.2**	**53.8**	**58.6**	**66.5**	**65.6**	**93.3**	**58.1**	58.8	**75.0**	**75.3**	**79.8**	**32.6**	**36.6**	**50.8**	**53.4**	**61.5**

Table 4: **Comparison with SOTA methods on VideoBench.** * denotes our evaluation results with the public checkpoints. The best results are **bold** and the second-best results are underlined.

Model	ANet	MSVD	MSRVTT	TGIF	YC2	UCF	MOT	TV	MV	NBA	LE	DM	SQA3D	Avg.
mPLUG-Owl-7B [76]	41.5	42.5	36.3	31.7	27.1	22.8	**27.8**	24.0	30.2	25.1	33.3	51.0	32.0	33.2
Otter-7B [33]	44.3	55.0	**47.0**	34.3	32.7	22.4	16.7	27.7	**37.1**	**34.3**	52.8	48.7	29.7	37.5
Video-LLaMA-7B [83]	39.9	41.2	34.1	31.3	28.9	27.6	16.7	24.8	32.4	26.2	**60.6**	49.1	31.2	32.8
Valley-7B [47]	38.1	32.0	28.0	31.4	29.1	20.3	11.1	23.7	32.6	31.3	41.7	56.5	33.3	34.0
VideoChat-7B [35]	44.6	42.2	37.4	33.7	27.7	22.4	**27.8**	26.2	34.1	28.6	39.9	55.4	31.4	35.4
PandaGPT-7B [14]	45.0	50.4	44.6	29.7	33.0	33.0	16.7	27.9	**37.1**	31.1	41.7	56.0	30.8	37.5
VideoChatGPT-7B [50]	46.6	**57.5**	46.3	35.6	**34.8**	24.1	**27.8**	28.8	36.5	22.5	41.7	**58.2**	37.2	38.5
ChatUniVi-7B [29]	49.0	48.6	41.7	41.3	29.0	28.3	16.7	23.1	33.6	25.7	38.9	53.1	29.1	35.3
VideoLLaVA-7B* [40]	44.1	34.5	30.0	39.4	30.7	19.5	22.2	27.3	33.4	25.6	33.3	50.7	38.9	34.5
LLaMA-VID-7B* [38]	45.2	44.5	39.1	29.1	29.3	27.9	11.1	**34.1**	32.5	28.9	36.1	47.8	36.8	36.5
ShareGPT4Video-8B	**50.8**	45.6	43.0	**42.8**	34.6	**39.7**	22.2	31.9	34.0	30.5	41.7	53.6	**42.9**	**41.2**

Experiments

- Video Generation

Prompt: A drone camera circles around a beautiful historic church built on a rocky outcropping along the Amalfi Coast, the view showcases historic and magnificent architectural details and tiered pathways and patios, waves are seen crashing against the rocks below as the view overlooks the horizon of the coastal waters and hilly landscapes of the Amalfi Coast Italy, several distant people are seen walking and enjoying vistas on patios of the dramatic ocean views, the warm glow of the afternoon sun creates a magical and romantic feeling to the scene, the view is stunning captured with beautiful photography.

Prompt: The video captures the spectacle of a continuous fireworks show against the backdrop of a starry night sky. It commences with a burst of vibrant reds, greens, purples, and yellows that paint the heavens and cast shimmering reflections upon the water below. As the display progresses, the fireworks evolve, transitioning from the initial array to a focus on radiant oranges, yellows, and fiery reds. These explosions form captivating clusters at the heart of the sky, ascending in breathtaking formations accompanied by trailing plumes of smoke, adding a dramatic flourish to the visual narrative. Throughout the duration, the fireworks maintain their dynamic allure, their patterns and positions evolving to underscore the ongoing spectacle. Meanwhile, the mirrored reflections on the water's surface faithfully echo the colors and shapes above, further enhancing the mesmerizing and ever-changing nature of the display.



上海人工智能实验室
Shanghai Artificial Intelligence Laboratory


Thanks, Q & A

Corresponding:
chlin@mail.ustc.edu.cn