Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers

Tsai-Shien Chen[1,2], Aliaksandr Siarohin[1], Willi Menapace[1,3], Ekaterina Deyneka[1], Hsiang-wei Chao[1], Byung Eun Jeon[1], Yuwei Fang[1], Hsin-Ying Lee[1], Jian Ren[1], Ming-Hsuan Yang[2], Sergey Tulyakov[1]

[1]Snap Inc., [2]UC Merced, [3]University of Trento

CVPR SEATTLE, WA JUNE 17-21, 2024

Dataset with 70M High-Quality Video-Caption Pairs

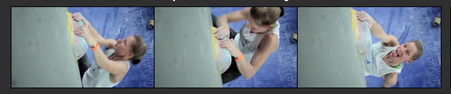
Animal

"A group of dolphins are swimming in the ocean."

Food

"A person is using a chef's knife to chop fresh parsley on a wooden cutting board."

Sports Activity

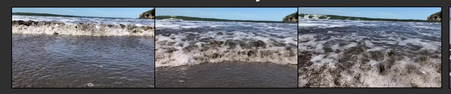


"A female gymnast is practicing her skills on a climbing wall."

Vehicle

"A blue off-road truck is driving on a sand dune and jumping into the air."

Scenery



"The waves are crashing on the beach and the water is foamy."

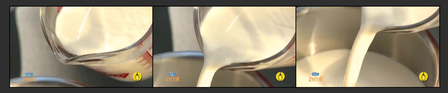
Gaming and 3D Rendering

"A screenshot of a minecraft game showing a snowy landscape."

Long Video Splitting and Captioning

Long Video

Scene N

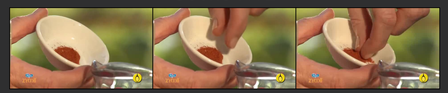


"A white liquid is being poured into a metal bowl."

Scene N+1

"A woman is preparing some food in a kitchen."

Scene N+2



"A person holding a small bowl of red powder."

Dataset Collection Pipeline

Long Video

Splitting
based on shot boundary detection

Stitching
based on semantics similarity

Short Clips

Text Input

[Title] How the Harley-Davidson Fat Boy became the Most Iconic American Motorcycle …
[Subtitles] Fat Boy kind of pioneered that path to make Harley Davidsons even bigger…
[Description] A look at the 30th year anniversary Fat Boy and a in-depth look at what …

Teacher A: **Video-LLaMA** with **video** input only
"A man is riding a motorcycle through an underpass."

Teacher B: **VideoChat** with **video** and **subtitles** inputs
"The speaker is talking about the fat boy what makes Harley Davidsons even …"

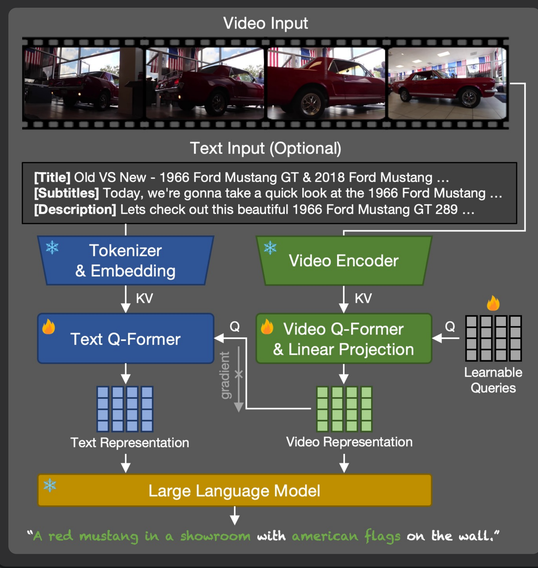
Teacher C: **MiniGPT-4** with **video frame** and **all text** inputs
"The person is riding a black Harley Davidson fat boy motorcycle on a city street."
…

Fine-grained Video-to-Text Retrieval

Caption

"The person is riding a black Harley Davidson fat boy motorcycle on a city street."

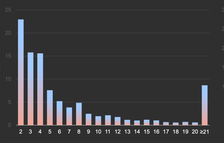
Multimodal Student Captioning Model



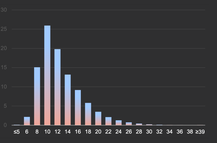
Dataset Statistics

Number of video clips	70,817,169
Total video length	166.8 khr
Average video length	8.5 sec
Average caption length	13.2 words
Resolution	720p↑

Video length (sec)



Caption length (words)



Word Cloud

Performance on Downstream Tasks

Zero-shot Video Captioning

Pretraining Data	B4↑ on MSR-VTT	B4↑ on MSVD
Other 2.5M vid+img	5.8%	12.7%
Panda-2M (Ours)	23.5%	31.2%
Panda-70M (Ours)	**25.4%**	**32.8%**

Zero-shot Text-to-Video / Video-to-Text Retrieval

Pretraining Data	R@1↑ on MSR-VTT	R@1↑ on MSVD
Other 5M vid+img	30.2% / 33.3%	66.3% / **44.4%**
Panda-5M (Ours)	**37.2%** / **36.3%**	**71.2%** / 37.2%

Zero-shot Text-to-Video Generation

Pretraining Data	FVD↓ on UCF101	CLIPSim↑ on MSR-VTT
Other 2.5M vid	499.3	0.2869
Panda-2M (Ours)	**421.9**	**0.2880**

Learn More?

Project Website

Code & Dataset