
A blue off-road truck is driving on a sand dune and jumping into the air.
There are ants tunneling under a thick carpet of moss.
A person is holding a long haired dachshund in their arms.
There is a river flowing through a forest and the water is flowing downstream.
A group of basketball players are practicing their shots on the court.
A rocket launches into space on the launch pad.
Someone is frying dough balls in a pan with oil.
A person is kneading dough and putting jam on it.
Panda-70M
A person is driving a boat on a river with rocks and waterfalls.
A woman is playing golf at an outdoor driving range.
It is a rally car driving on a dirt road in the countryside, with people watching from the side of the road.
The waves are crashing on the beach and the water is foamy.
A rhino and a lion are fighting in the dirt.
A blue toyota tacoma truck is parked in a parking lot surrounded by trees.
A person is making a pie crust on a table.
A large pile of lava blocking a road.

Animal

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

"A rhino and a lion are fighting in the dirt."

Scenery

"There is a road in the desert with mountains in the background. The sky is blue, and there are clouds in it."

"An aerial view of a freeway with a lot of traffic."

Food

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

"A pink frosted cake sitting on a table."

Sports Activity

"A woman is playing golf at an outdoor driving range."

"A group of basketball players are practicing their shots on the court."

Vehicle

"It is a rally car driving on a dirt road in the countryside, with people watching from the side of the road."

"A toyota hilux driving on a dirt road."

Tutorial and Narrative

"A person is making a green clay model of a monster using different tools."

"A person is welding a piece of metal using a welding torch, and the metal is glowing red hot."

News and TV Shows

"A rocket launches into space on the launch pad."

"A young man and woman are standing in front of a body of water with a city skyline in the background."

Gaming and 3D Rendering

"A man in a spartan armor kneeling down."

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

Existing Video-Caption Datasets

High-quality manual caption but with limited samples

ActivityNet
with 10K videos
[Heilbron et al. 2015]

"A man was sitting inside a room. He is holding a bowl of noodles and broth. He is drinking the broth from a bowl."

MSR-VTT
with 10K videos
[Xu et al. 2016]

"We can see the tents. A woman in a purple-pink jacket bends down. Someone in pink bends down. The camera turns to the left to show the sun. The sun can be seen."

Large-scale video dataset but with ASR caption

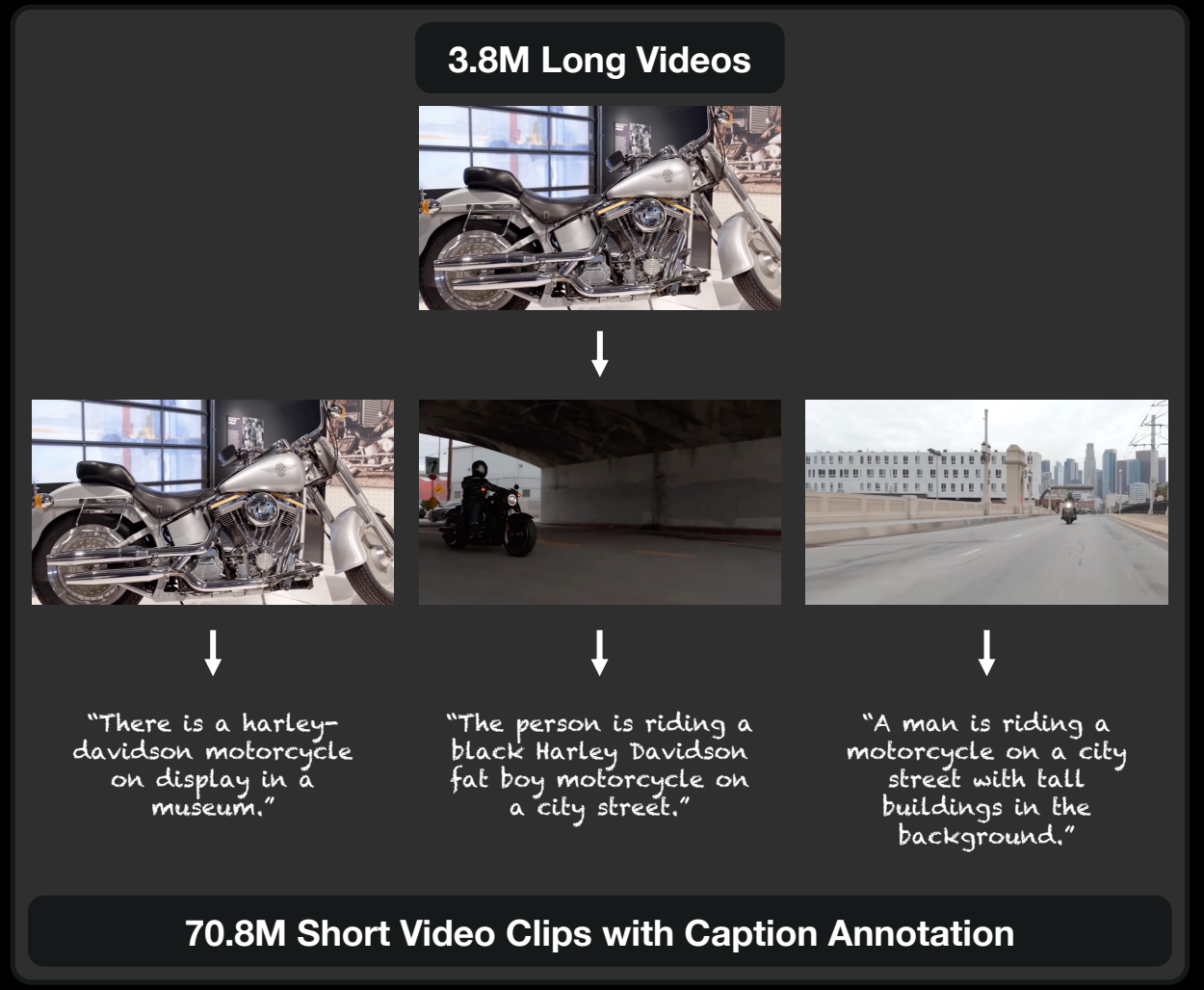
YT-Temporal-180M
with 180M videos
[Zellers et al. 2021]

"Today, I'm gonna be going to get some structural energetic therapy done you guys may have never heard of."

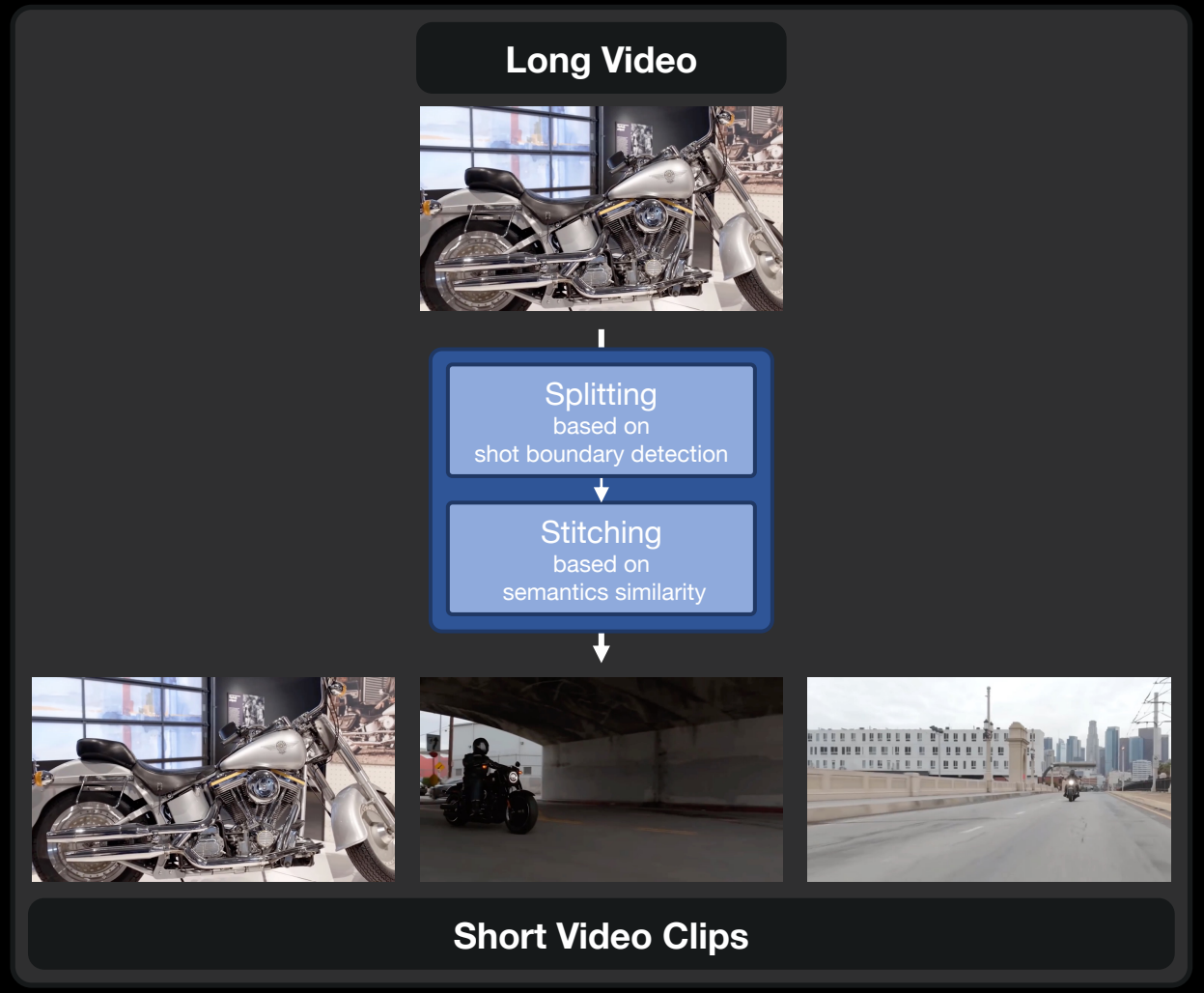
HD-VILA-100M
with 103M videos
[Xue et al. 2022]

"All right. Fire it up. Good luck!"

Dataset Collection Pipeline



Dataset Collection Pipeline: Splitting

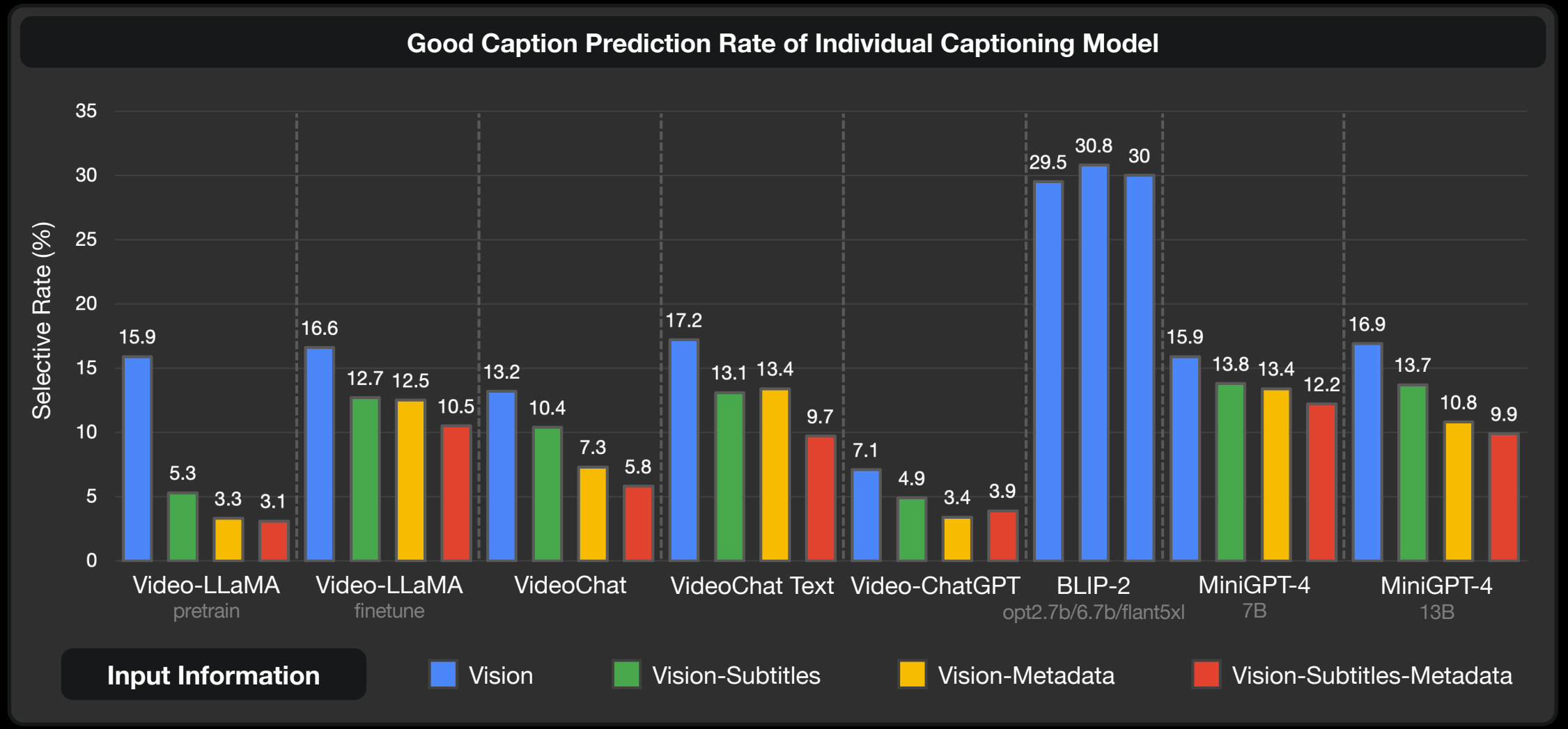


Dataset Collection Pipeline: Splitting

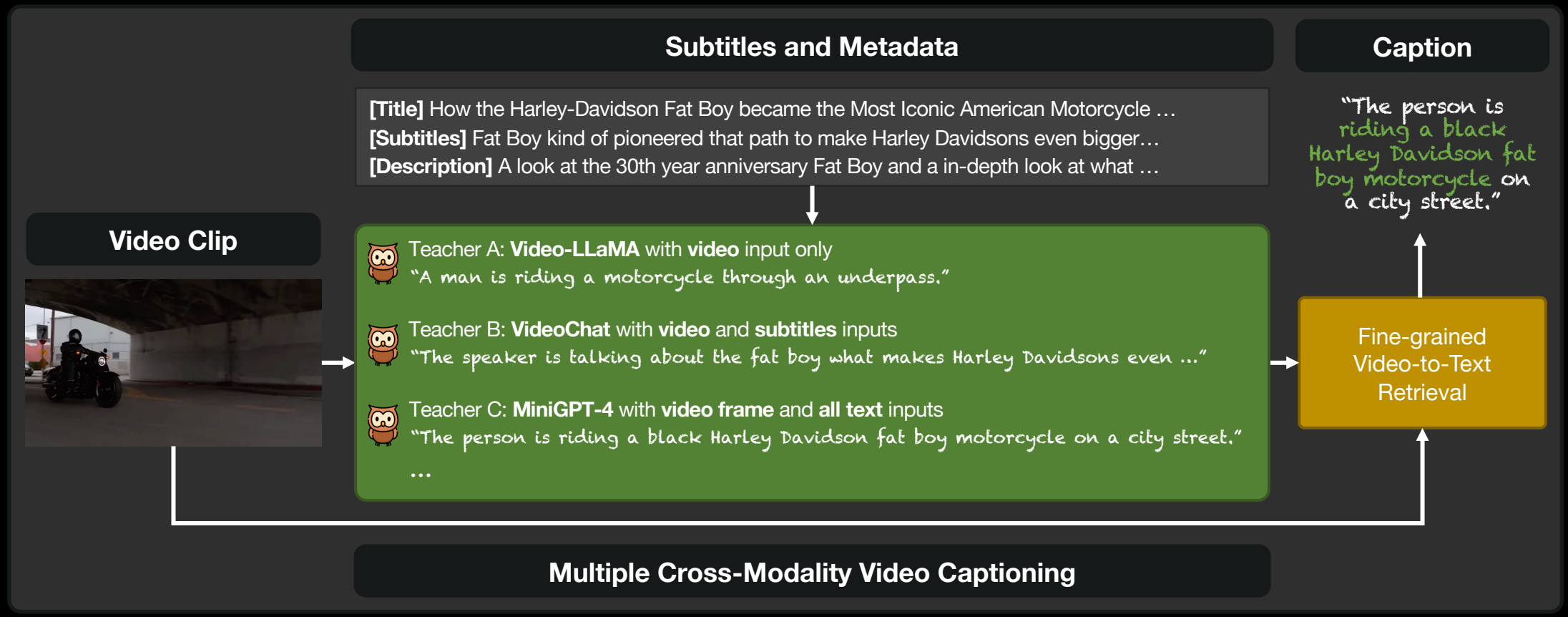
Comparison of Splitting Algorithms

	Max Running LPIPS↓	Avg Video Length↑
Subtitles Align	0.408	11.8 sec
PySceneDetect	0.247	4.1 sec
Our Splitting	0.256	7.9 sec

Dataset Collection Pipeline: Captioning



Dataset Collection Pipeline: Captioning

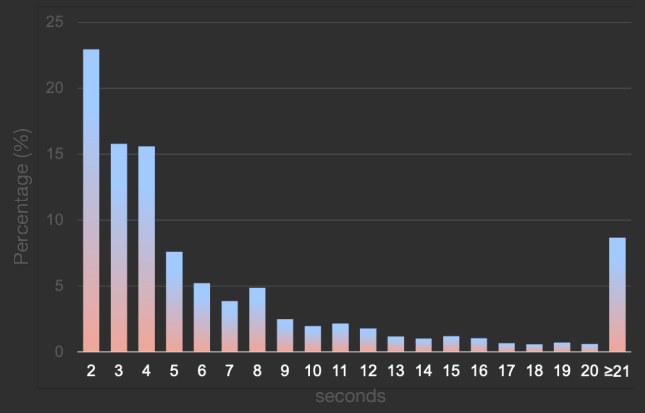


Dataset Statistics

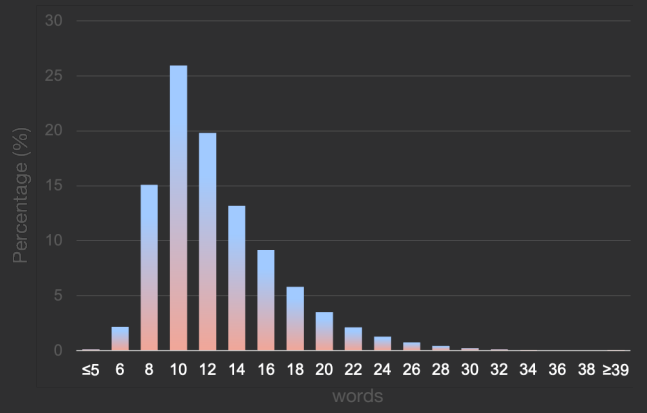
General Information

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 ↑

Distribution of Video Length



Distribution of Caption Length



Word Cloud

Pre-training 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 T2V 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**

Other 2.5M vid **Panda-2M**

[Prompt] "Cut tusuncub walking in the snow, blurry, looking at viewer, [...]"

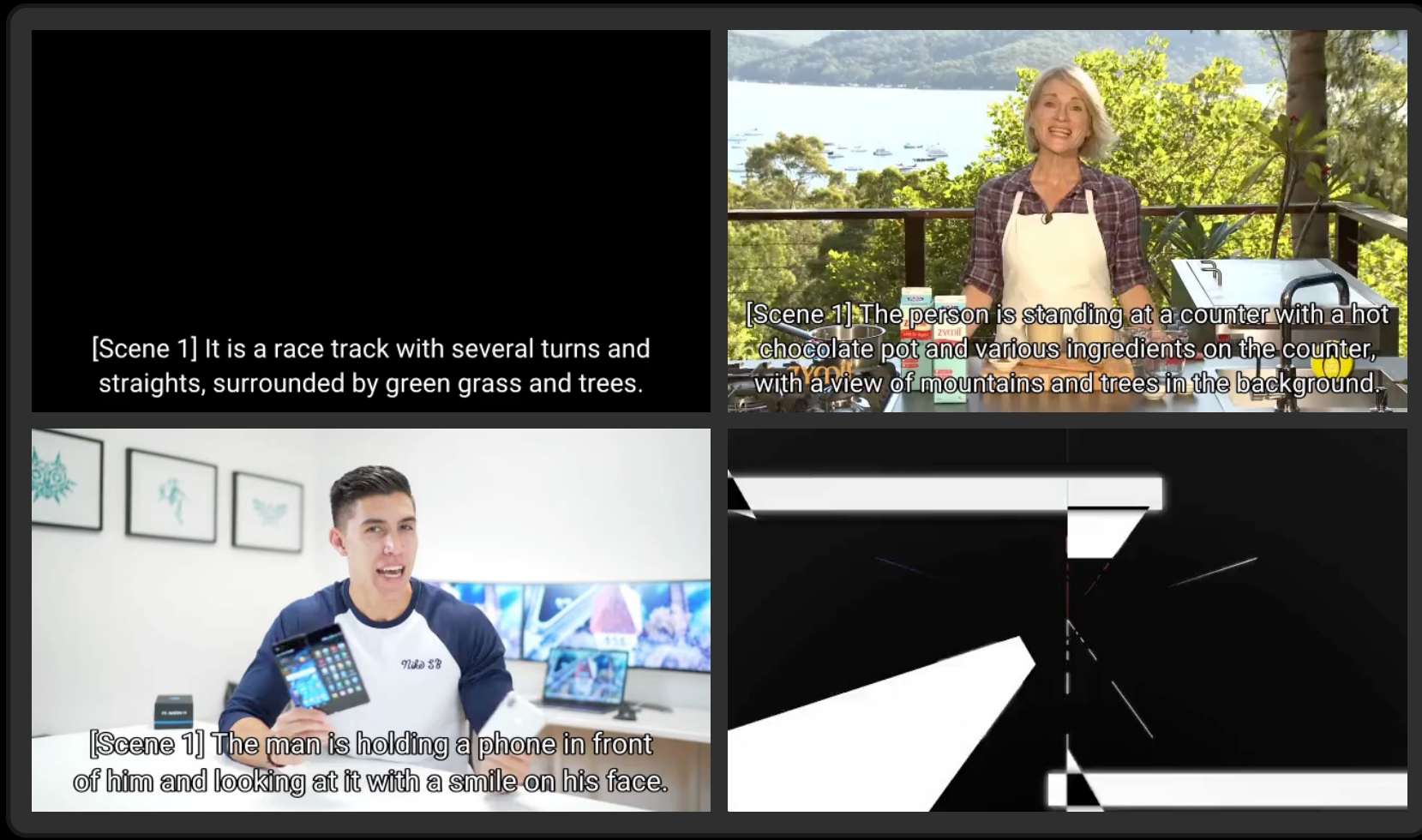
Zero-shot T2V Retrieval

Pretraining Data	R@1↑ on MSR-VTT	R@1↑ on DiDeMo	B4↑ on MSVD
Other 5M vid+img	30.2%	33.6%	66.3%
Panda-5M (Ours)	**37.2%**	**34.2%**	**71.2%**

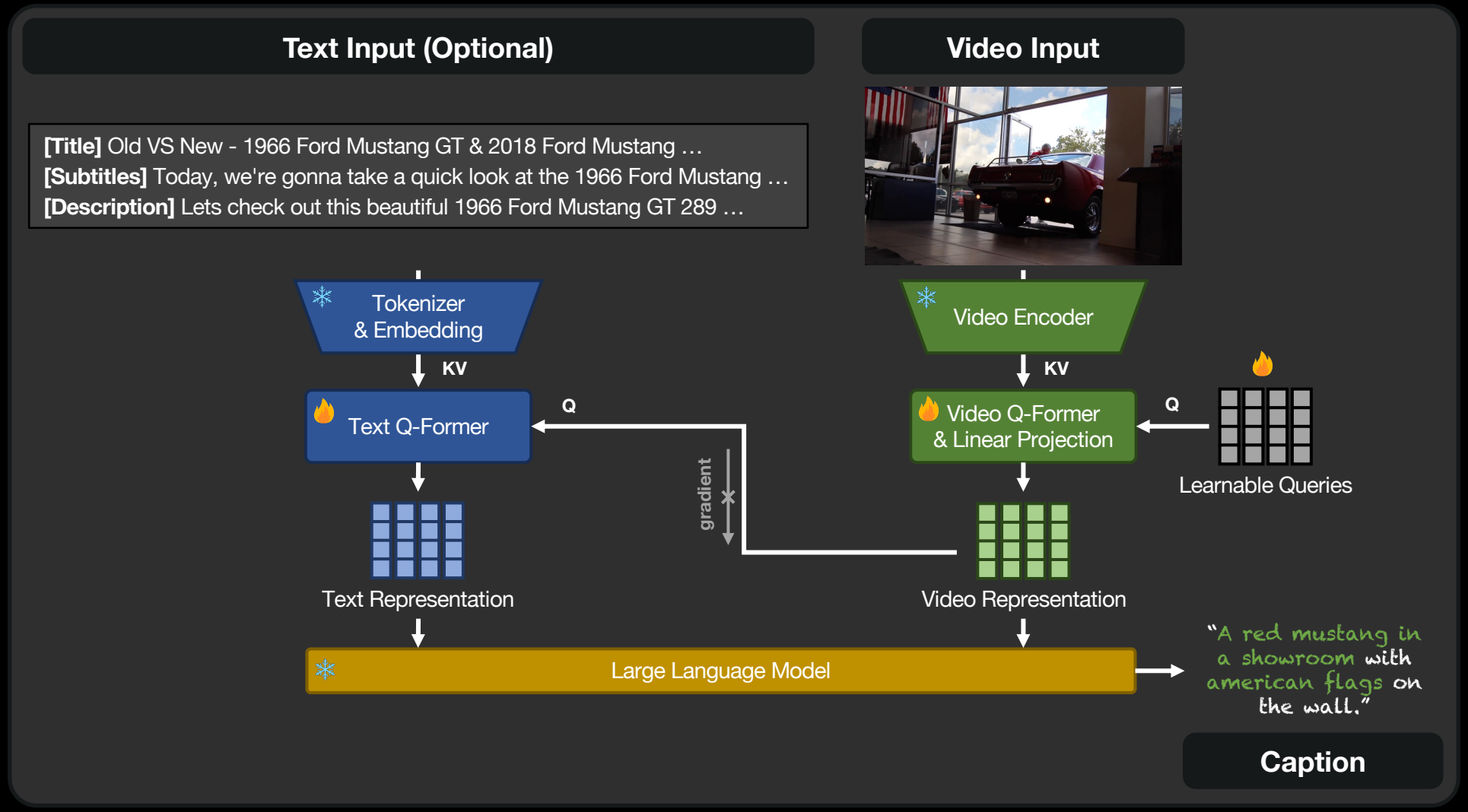
Zero-shot V2T Retrieval

Pretraining Data	R@1↑ on MSR-VTT	R@1↑ on DiDeMo	B4↑ on MSVD
Other 5M vid+img	33.3%	32.1%	**44.4%**
Panda-5M (Ours)	**36.3%**	**33.4%**	37.5%

Long Video Splitting and Captioning



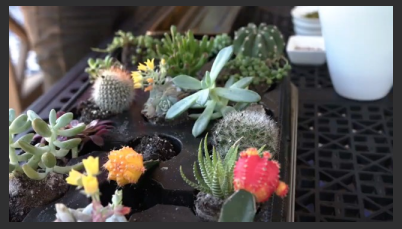
Multimodal Student Captioning Model



Performance of Student Captioning Model

Qualitative Results

Video Input



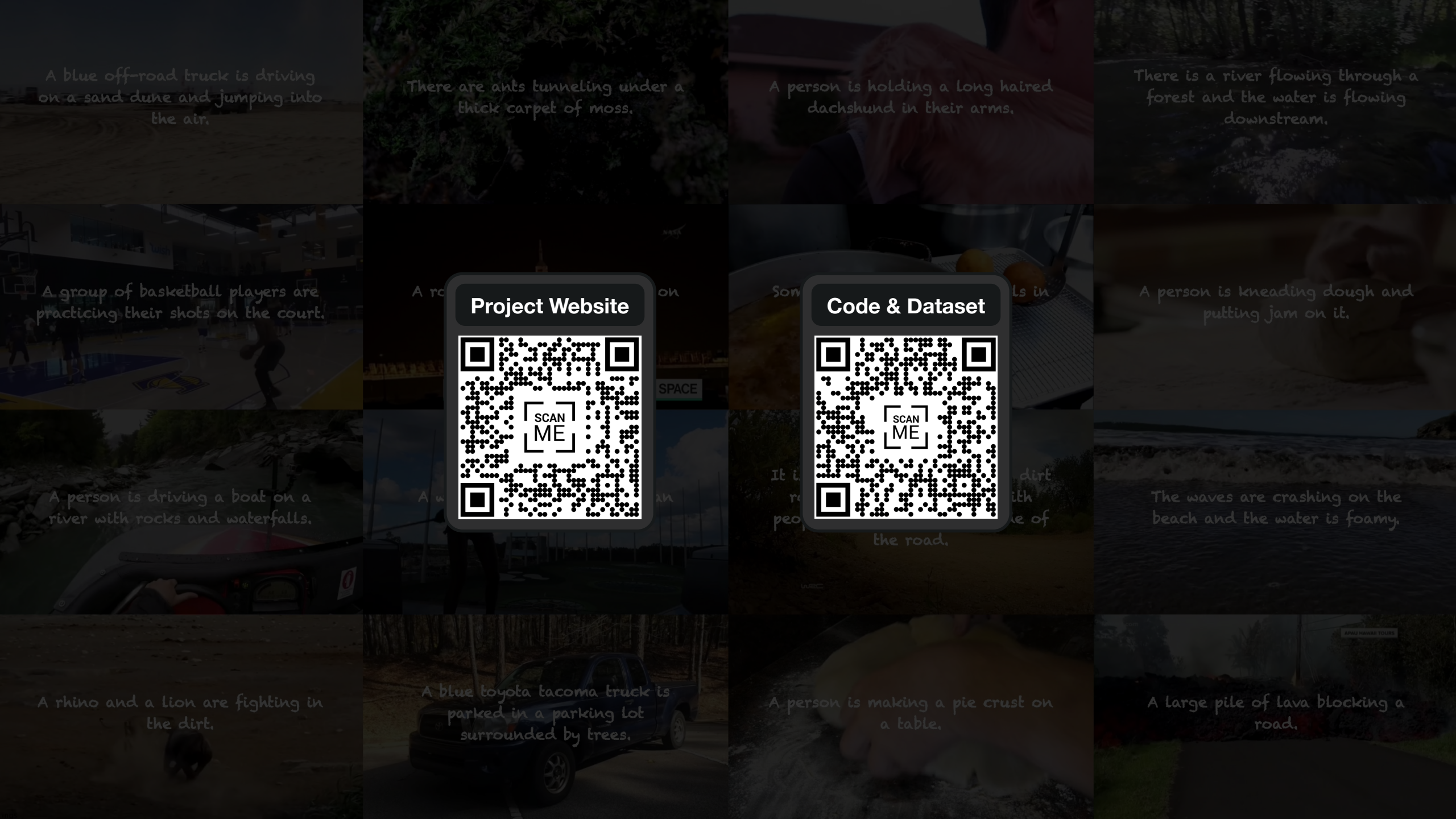
Text Input

[Title] Succulent Garden | Easy DIY | Interior Design | DIY Decorating Ideas
[Subtitles] Here I have 17 different species and it makes it kind of fun and interesting.
[Description] Succulents, Indoor Succulents, How To Make a Living Succulent Garden.

Video-LLaMA [Zhang et al. 2023]	"Monaco - June 03, 2018 cactus, flowers, plants"
Student (Ours) Video input only	"A close up of a bunch of cactus plants"
Student (Ours) Video and text inputs	"A bunch of different species of cacti and succulents"
Panda-70M Annotation	"It is a succulent garden with different species of cacti and other succulents growing in pots."

User Study

Model	Preference Ratio↑
VideoLLaMA [Zhang et al. 2023]	9.4%
BLIP-2 [Li et al. 2023]	10.7%
Student (Ours) Video input only	18.4%
Student (Ours) Video and text inputs	21.4%
All Teachers (Ours)	**23.3%**


A blue off-road truck is driving on a sand dune and jumping into the air.
There are ants tunneling under a thick carpet of moss.
A person is holding a long haired dachshund in their arms.
There is a river flowing through a forest and the water is flowing downstream.
A group of basketball players are practicing their shots on the court.
A person is kneading dough and putting jam on it.
A person is driving a boat on a river with rocks and waterfalls.
The waves are crashing on the beach and the water is foamy.
A rhino and a lion are fighting in the dirt.
A blue toyota tacoma truck is parked in a parking lot surrounded by trees.
A person is making a pie crust on a table.
A large pile of lava blocking a road.
Project Website
SCAN ME
Code & Dataset
SCAN ME