CVPR
JUNE 17-21, 2024
SEATTLE, WA

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

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

Snap Inc., University of California, Merced, University of Trento

Presenter: Sunghyun Ahn

sunghyun.ahn@pyler.tech

PYLER

Video Captioning

- Generates **natural language descriptions** from visual content
- Challenging due to temporal changes in scenes and events
- Applied in content retrieval, summarization, and multimodal learning

Chen, Tsai-Shien, et al. "Panda-70m: Captioning 70m videos with multiple cross-modality teachers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Introduction

Limitations of video-language dataset

- Large-scale video dataset but with ASR(automatic speech recognition) caption
- High-quality manual caption but with limited samples
- Lack of effective datasets for training video captioning models

Table 1. **Comparison of Panda-70M and other video-language datasets.** We split the datasets into two groups: the group at the top is annotated by ASR, and the group at the bottom is labeled with captions.

Dataset	Year	Text	Domain	#Videos	Avg/Total video len		Avg text len	Resolution
HowTo100M [52]	2019	ASR	Open	136M	3.6s	134.5Khr	4.0 words	240p
ACAV [32]	2021	ASR	Open	100M	10.0s	277.7Khr	-	-
YT-Temporal-180M [87]	2021	ASR	Open	180M	-	-	-	-
HD-VILA-100M [80]	2022	ASR	Open	103M	13.4s	371.5Khr	32.5 words	720p
MSVD [13]	2011	Manual caption	Open	1970	9.7s	5.3h	8.7 words	-
LSMDC [58]	2015	Manual caption	Movie	118K	4.8s	158h	7.0 words	1080p
MSR-VTT [79]	2016	Manual caption	Open	10K	15.0s	40h	9.3 words	240p
DiDeMo [3]	2017	Manual caption	Flickr	27K	6.9s	87h	8.0 words	-
ActivityNet [11]	2017	Manual caption	Action	100K	36.0s	849h	13.5 words	-
YouCook2 [93]	2018	Manual caption	Cooking	14K	19.6s	176h	8.8 words	-
VATEX [73]	2019	Manual caption	Open	41K	~10s	~115h	15.2 words	-
Panda-70M (Ours)	2024	Automatic caption	Open	70.8M	8.5s	166.8Khr	13.2 words	720p

Introduction

Why So Challenging?

- Labeling requires watching the entire video, which is time-consuming
- Frequent scene and content changes over time
- Meta information (e.g., subtitles, narration) often misaligned or inaccurate (e.g., ASR datasets)

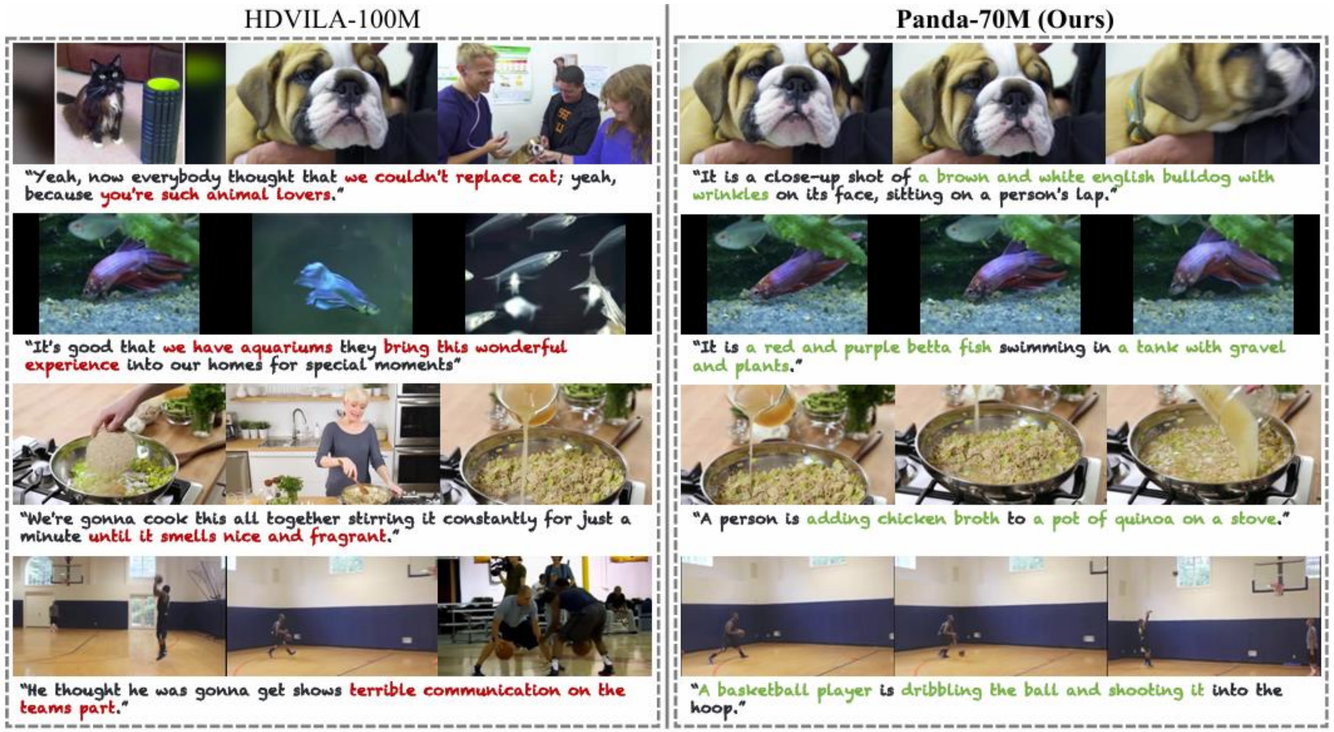


Figure 1. **Comparison of Panda-70M to the existing large-scale video-language datasets.** We introduce Panda-70M, a large-scale video dataset with captions that are annotated by multiple cross-modality vision-language models. Compared to text annotations in existing dataset [80], captions in Panda-70M more precisely describe the main object and action in videos (highlighted in green). Besides, videos in Panda-70M are semantically coherent, high-resolution, and free from watermarks. More samples can be found in Appendix E.

Contributions

- Turns long video into semantically consistent short clips using semantics-aware splitting
- Generate high-quality captions using diverse cross-modal teacher models
- Enables large-scale Panda-70M dataset construction with minimal human supervision

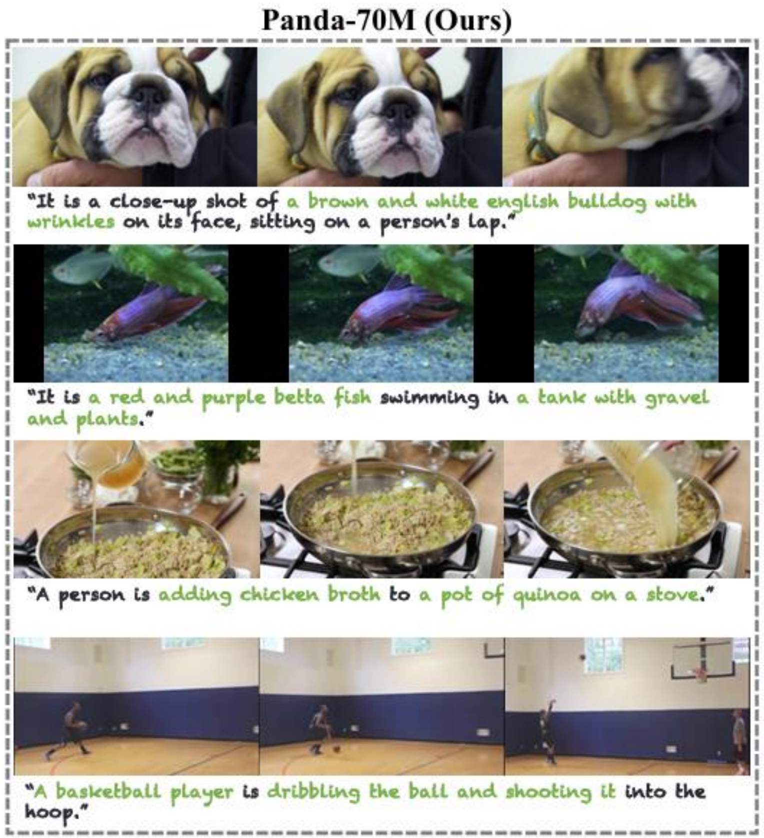


Table 1. **Comparison of Panda-70M and other video-language datasets.** We split the datasets into two groups: the group at the top is annotated by ASR, and the group at the bottom is labeled with captions.

Dataset	Year	Text	Domain	#Videos	Avg video len	Total video len	Avg text len	Resolution
HowTo100M [52]	2019	ASR	Open	136M	3.6s	134.5Khr	4.0 words	240p
ACAV [32]	2021	ASR	Open	100M	10.0s	277.7Khr	-	-
YT-Temporal-180M [87]	2021	ASR	Open	180M	-	-	-	-
HD-VILA-100M [80]	2022	ASR	Open	103M	13.4s	371.5Khr	32.5 words	720p
MSVD [13]	2011	Manual caption	Open	1970	9.7s	5.3h	8.7 words	-
LSMDC [58]	2015	Manual caption	Movie	118K	4.8s	158h	7.0 words	1080p
MSR-VTT [79]	2016	Manual caption	Open	10K	15.0s	40h	9.3 words	240p
DiDeMo [3]	2017	Manual caption	Flickr	27K	6.9s	87h	8.0 words	-
ActivityNet [11]	2017	Manual caption	Action	100K	36.0s	849h	13.5 words	-
YouCook2 [93]	2018	Manual caption	Cooking	14K	19.6s	176h	8.8 words	-
VATEX [73]	2019	Manual caption	Open	41K	∼10s	∼115h	15.2 words	-
Panda-70M (Ours)	2024	Automatic caption	Open	70.8M	8.5s	166.8Khr	13.2 words	720p

Overview

- Semantics-aware Video Splitting
- Captioning with Cross-Modality Teachers
- Fine-grained Video-to-Text Retrieval
- Multimodal Student Captioning Model

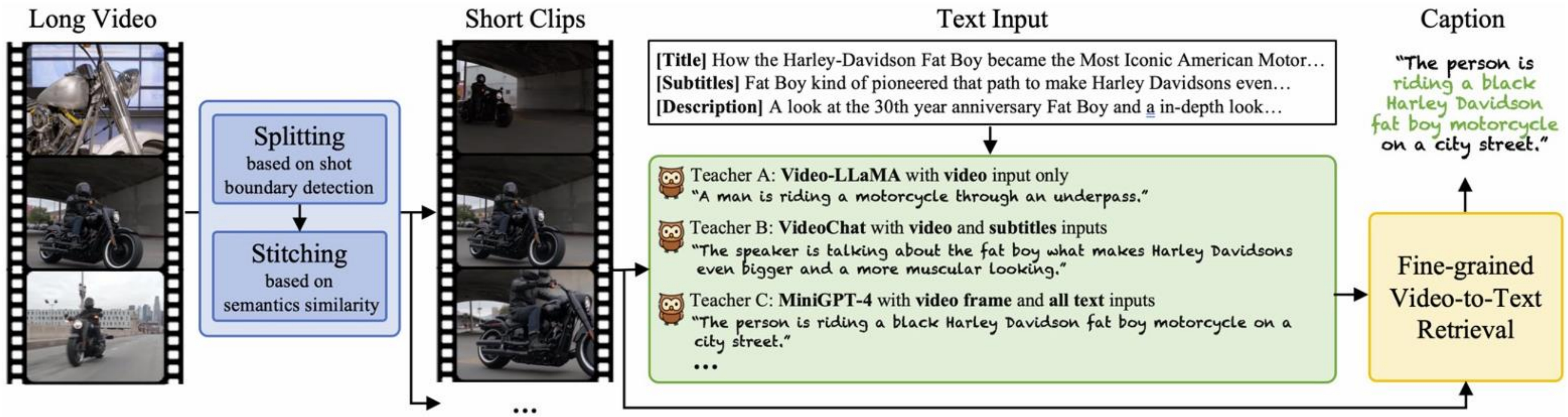


Figure 2. **Video captioning pipeline.** Given a long video, we first split it into several semantically coherent clips. Subsequently, we utilize a number of teacher models with different multimodal inputs to generate multiple captions for a video clip. Lastly, we finetune a fine-grained retrieval model to select the caption that best describes the video clip as the annotation.

Method

Semantics-aware Video Splitting

- Good clips are **semantically coherent** and **long enough to contain meaningful event**
- (1) Split videos using shot boundary detection, which may break a single semantic clip into multiple short clips
- (2) Merge adjacent clips if embeddings (e.g., ImageBind) are similar

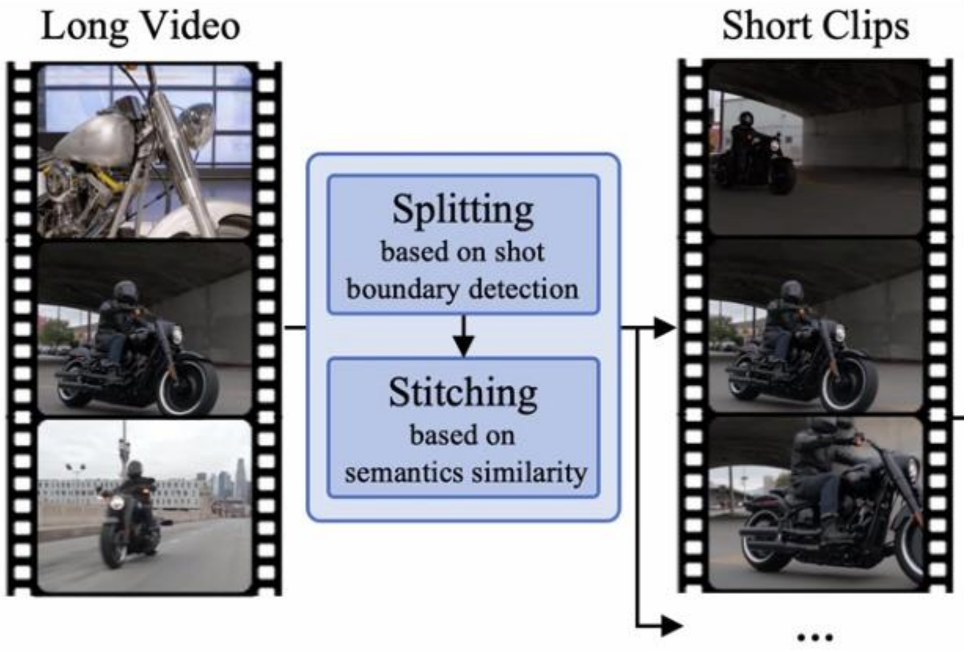


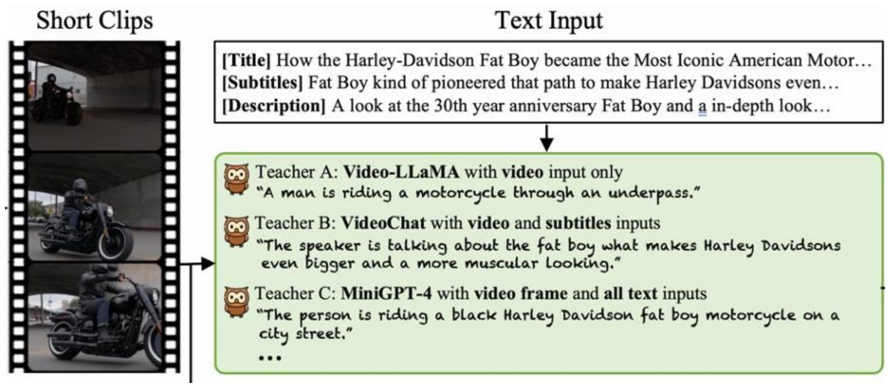
Table 2. **Comparison of splitting algorithms.** We split 1K long videos by three algorithms and test the semantics consistency of the output clips by the proposed Max Running LPIPS. Our splitting strikes a better balance for the trade-off between semantics consistency and clip length.

Method	Max running LPIPS↓	Avg Video Len
Sub. Align [52, 80]	0.408	11.8s
PySceneDetect [1]	0.247	4.1s
Our Splitting	0.256	7.9s

Method

Captioning with Cross-Modality Teachers

- HD-VILA-100M provides rich multimodal data (video, subtitles, descriptions); 3.8M high-res videos used
- Various teacher models (e.g., Image Captioning, Video-VQA, Image-VQA) improve caption accuracy across video types
- BLIP-2 works well for static scenes, Video-LLaMA for dynamic ones, and cross-modal models for complex content
- **Selected 8 teachers based on high selective rate, video understanding, and multimodal capability from 31 models**



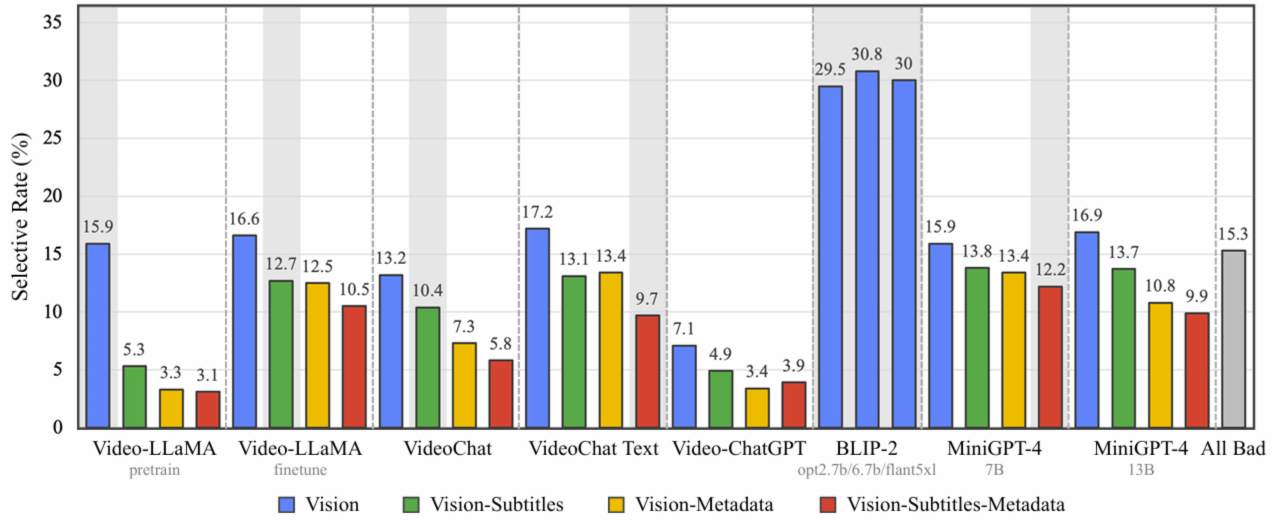
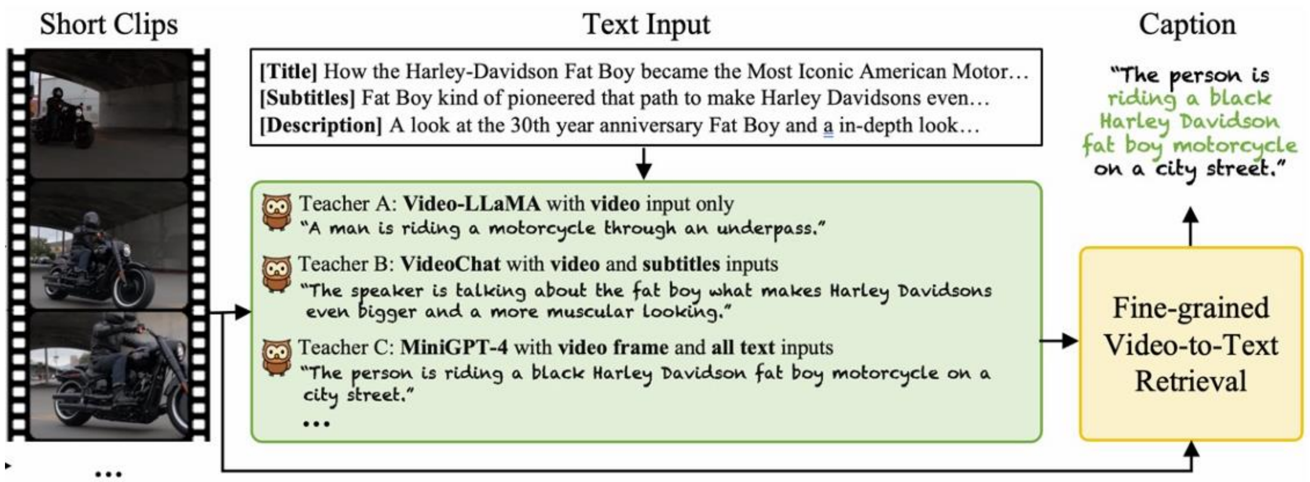


Figure 10. **Ratio of an individual captioning model to predict a good caption.** Each bar represents an individual model and is colored by its input information. We highlight the 8 selected teacher models with gray. Note that we also report the ratio of "All Bad" at rightmost.

Method

Fine-grained Video-to-Text Retrieval

- Simple approach: Select the best caption from teacher outputs using a generic retrieval model
- However, retrieval results often differ from human judgment and may be inaccurate
- **Sampled 100K videos and collected human-verified best captions as ground truth**
- Fine-tuned the UMT retrieval model with contrastive learning and implicit hard-negative mining to improve accuracy



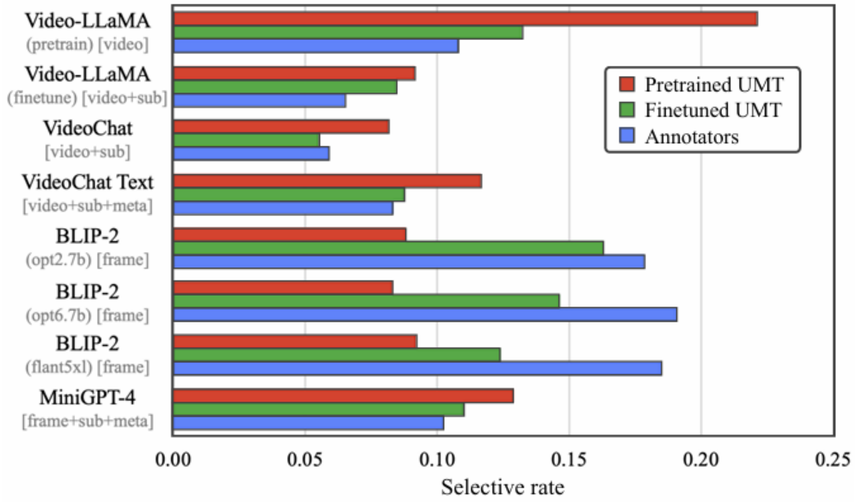


Figure 3. **Distributions of the selective rate of teacher models.** We plot the distributions of the selective rate of eight teachers on 1,805 testing videos. The results are based on the selection of the pretrained (red) or finetuned (green) Unmasked Teacher [39] and human annotators (blue).

Multimodal Student Captioning Model

- Using 8 teacher models and a retrieval model is inefficient → **Train a student model on Panda-70M**
- Adopt visual prompt tuning with Q-former to extract fixed-length text embeddings for long inputs
- Achieves over 2× higher preference ratio than single student models; multimodal training yields better performance

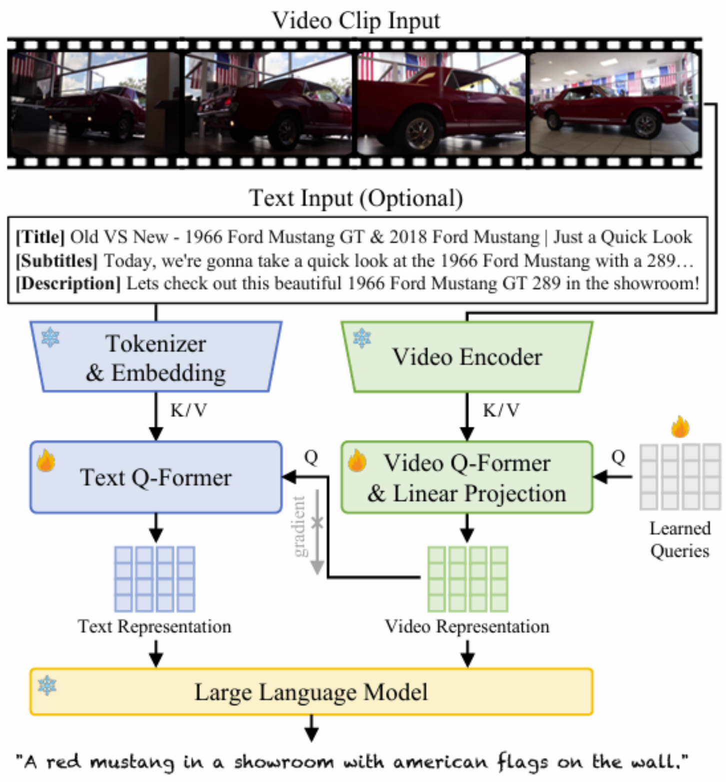


Figure 4. **Architecture of student captioning model.**

Table 4. **Comparison of the teacher(s) and student captioning models (%).** We conduct a user study to compare single teacher, all teacher, and two student models (with and without text).

Model	Preference Ratio↑
Video-LLaMA [88] (pretrain)	9.4
Video-LLaMA [88] (finetune)	7.0
VideoChat [38]	7.7
VideoChat Text [38]	3.3
BLIP-2 [37] (opt2.7b)	10.7
BLIP-2 [37] (opt6.7b)	9.0
BLIP-2 [37] (flant5xl)	9.9
MiniGPT-4 [94]	3.1
Student (video input) (Ours)	18.4
Student (video+text inputs) (Ours)	<u>21.4</u>
All Teachers (Ours)	**23.3**

Experiments

Qualitative comparison

- Student model returns captions closely aligned with ground-truth
- Enables accurate category-wise grouping of video clips

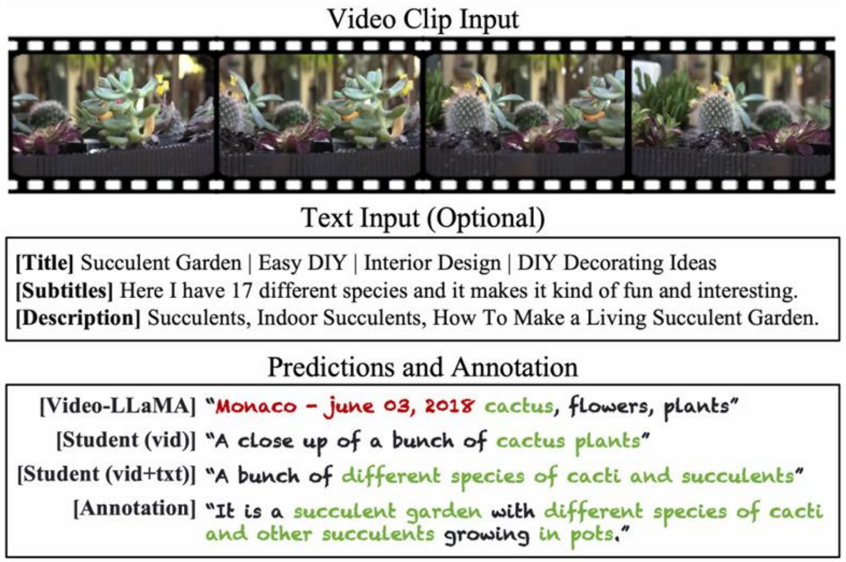
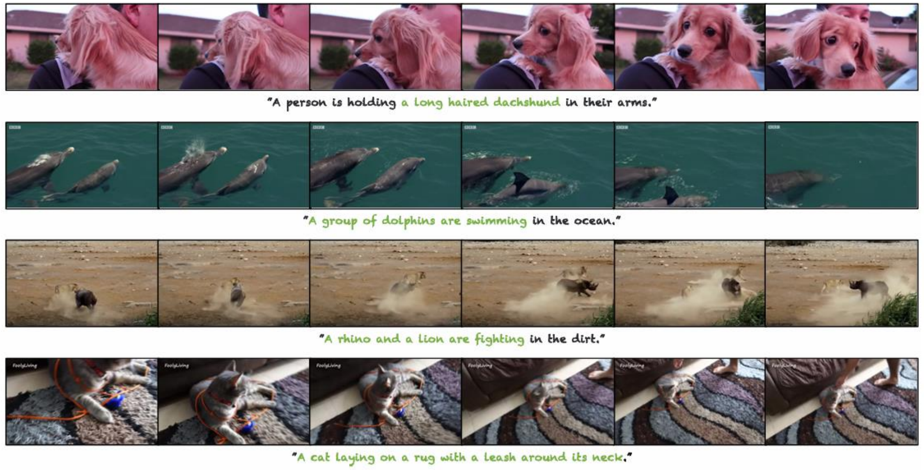


Figure 5. **Qualitative comparison of video captioning.** We visualize a sample from the testing set of Panda-70M and show its annotation (bottommost). We also show the captions predicted from three models, including Video-LLaMA [88] with official weight and the student models with video-only or video and text inputs.

E.1. Category: Animal



E.3. Category: Food

Experiments

Video and text retrieval

- Evaluated the proposed dataset on downstream tasks (T2V, V2T)
- Models pretrained on Panda-5M outperform those using other data
- **semantically consistent clips and high-quality captions benefit video pretraining**

Table 5. **Video and text retrieval (%).** We compare the Unmasked Teacher [39] with the official checkpoint (pretrained on 2.5M videos and 3M images) and our Panda-5M pretraining. We evaluate their performance on zero-shot and finetune text-to-video (T2V) and video-to-text (V2T) retrieval. We report R@1, R@5, and R@10 accuracy on three benchmarks: MSR-VTT [79], DiDeMo [3], and MSVD [13].

Method	Pretraining Data	MSR-VTT			DiDeMo			MSVD		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Zero-shot T2V / V2T Retrieval										
AlignPrompt [34]	2.5M vid + 3M img	24.1 / -	44.7 / -	55.4 / -	23.8 / -	47.3 / -	57.9 / -	- / -	- / -	- / -
BridgeFormer [24]	2.5M vid + 3M img	26.0 / -	46.4 / -	56.4 / -	25.6 / -	50.6 / -	61.6 / -	43.6 / -	74.9 / -	84.9 / -
UMT [39]	2.5M vid + 3M img	30.2 / 33.3	51.3 / 58.1	61.6 / 66.7	33.6 / 32.1	58.1 / 57.3	65.5 / **66.7**	66.3 / **44.4**	85.5 / **73.3**	89.3 / **82.4**
UMT [39]	**Panda-5M (Ours)**	**37.2 / 36.3**	**58.1 / 61.0**	**69.5 / 69.7**	**34.2 / 33.4**	**58.4 / 57.9**	**66.5** / 65.8	**71.2** / 37.2	**88.4** / 65.1	**92.7** / 75.6
Finetune T2V / V2T Retrieval										
CLIP4Clip [49]	400M img	44.5 / 40.6	71.4 / 69.5	81.6 / 79.5	43.4 / 42.5	70.2 / 70.6	80.6 / 80.2	46.2 / 62.0	76.1 / 87.3	84.6 / 92.6
X-CLIP [50]	400M img	49.3 / 48.9	75.8 / 76.8	84.8 / 84.5	50.4 / **66.8**	80.6 / **90.4**	- / -	47.8 / 47.8	79.3 / 76.8	- / -
InternVideo [74]	146M vid + 100M img	55.2 / 57.9	- / -	- / -	57.9 / 59.1	- / -	- / -	**58.4** / 76.3	- / -	- / -
UMT [39]	2.5M vid + 3M img	53.3 / 51.4	76.6 / 76.3	83.9 / 82.8	59.7 / 59.5	84.9 / 84.5	90.8 / **90.7**	53.7 / 77.2	80.5 / 91.6	86.8 / 94.8
UMT [39]	**Panda-5M (Ours)**	**58.4 / 58.5**	**80.9 / 81.0**	**86.9 / 87.0**	**60.6** / 58.9	**86.0** / 84.6	**92.4** / 90.4	57.5 / **81.3**	**83.6 / 93.7**	**89.5 / 96.6**

Conclusion

Contribution and Limitation

- Proposed Panda-70M: 70M video clips with high-quality captions
- Automatically constructed via semantic-aware splitting and fine-grained video-to-text retrieval

- Built from HD-VILA-100M, leading to category imbalance in some domains
- Videos with rapid motion changes (e.g., sports) have low semantic consistency, resulting in less accurate captions

Thank you