V2Xum-LLM: Cross-modal Video Summarization with Temporal Prompt Instruction Tuning

Hang Hua*, Yunlong Tang*, Chenliang Xu, Jiebo Luo
University of Rochester
*Equal Contribution

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

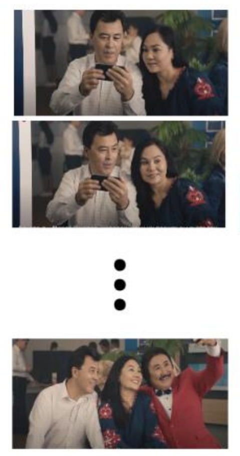
Contribution

- Propose **V2Xum-LLaMA**, a novel cross-modal video summarization framework that unifies different tasks into a single pre-trained language decoder, eliminating the need for task-specific heads used in prior methods.
- Introduce **Instruct-V2Xum**, a new instruction-following dataset for cross-modal video summarization.
- Present a comprehensive analysis of the limitations in current video summarization tasks, and propose new evaluation metrics F_{CLIP} and $Cross-F_{CLIP}$.

Data Curation

1. **Frame Captioning and Extractive Summarization.**
2. **Text Summarization Refinement.**
3. **Human Verification.**

Source Video



Video Summary

Text Summary

[3.0]A man and woman sit together, engrossed in a cell phone, sharing a moment of enjoyment. [5.0]In a room, a group, some in suits, sits on chairs and couches, engaged in a business meeting with a TV present, and cell phones in hand. [14.0]A man in a red jacket hands money to a woman at a desk. [23.0]A large white vault displays a sign for a 13,000 ruble deposit. [29.0]A man in a red jacket takes a selfie with a woman and another man.

Source Video

Video Summary

Text Summary

[2.0]A red wall features a \"Deposit\" sign. [9.0]A person interacts with a deposit machine, distinguished by its red background and a prominently placed green button on the right, to either deposit or withdraw cash. [23.0]The computer screen indicates that the transaction is currently processing. [32.0]A sign prompts users to enter the total deposit amount. [72.0]Another machine sign alerts users that their transaction is processing. [75.0]A gratitude sign thanks users for utilizing the cheque deposit service.

Instruct-V2Xum, a cross-modal video summarization dataset featuring 30,000 diverse videos sourced from YouTube, with lengths ranging from 40 to 940 seconds and an average summarization ratio of 16.39%.

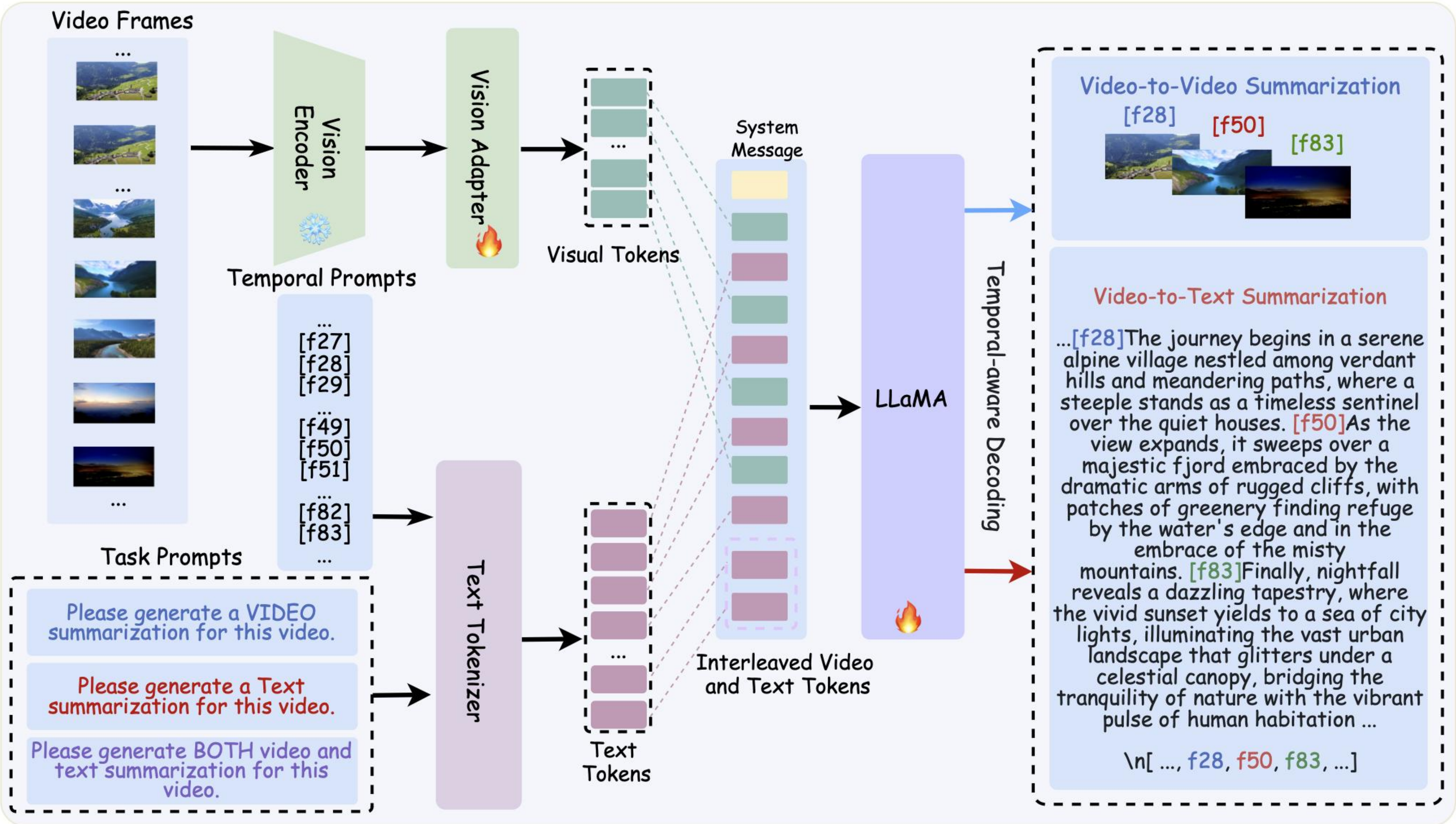
Proposed Evaluation Metrics

$$P_{CLIP}(v,\hat{v}) = \frac{1}{|\hat{v}|}\sum_{\hat{v}_i \in \hat{v}} \max_{v_i \in v} \mathbf{v}_i^\top \hat{\mathbf{v}}_j \qquad R_{CLIP}(v,\hat{v}) = \frac{1}{|v|}\sum_{v_i \in v} \max_{\hat{v}_j \in \hat{v}} \mathbf{v}_i^\top \hat{\mathbf{v}}_j$$

$$F_{CLIP}(v,\hat{v}) = 2\frac{P_{CLIP} \cdot R_{CLIP}}{P_{CLIP} + R_{CLIP}}$$

$$Cross - F_{CLIP}(v,\hat{v},t,\hat{t}) = \frac{F_{CLIP}(v,\hat{t}) + F_{CLIP}(\hat{v},t)}{2}$$

V2Xum-LLM


Video Frames
Vision Encoder
Vision Adapter
Visual Tokens
Temporal Prompts
[f27]
[f28]
[f29]
[f49]
[f50]
[f51]
[f82]
[f83]
Task Prompts
Please generate a VIDEO summarization for this video.
Please generate a Text summarization for this video.
Please generate BOTH video and text summarization for this video.
Text Tokenizer
Text Tokens
System Message
Interleaved Video and Text Tokens
LLaMA
Temporal-aware Decoding
Video-to-Video Summarization
[f28]
[f50]
[f83]
Video-to-Text Summarization

...[f28]The journey begins in a serene alpine village nestled among verdant hills and meandering paths, where a steeple stands as a timeless sentinel over the quiet houses. [f50]As the view expands, it sweeps over a majestic fjord embraced by the dramatic arms of rugged cliffs, with patches of greenery finding refuge by the water's edge and in the embrace of the misty mountains. [f83]Finally, nightfall reveals a dazzling tapestry, where the vivid sunset yields to a sea of city lights, illuminating the vast urban landscape that glitters under a celestial canopy, bridging the tranquility of nature with the vibrant pulse of human habitation ...

\n[..., f28, f50, f83, ...]

Performance on Cross-Modal Video Summarization Tasks

Method	Cross-Modal	LLM-Based	TSH-Free	V2T				V2V				V2VT
				B-4	M	R-L	C	F1	Spearman	Kendall	F_{CLIP}	Cross-F_{CLIP}
DENSE (Krishna et al. 2017)	✗	✗	✓	1.6	8.9	-	-	-	-	-	-	-
DVC-D-A (Li et al. 2018)	✗	✗	✓	1.7	9.3	-	-	-	-	-	-	-
Bi-LSTM+TempoAttn (Zhou et al. 2018)	✗	✗	✓	2.1	10.0	-	-	-	-	-	-	-
Masked Transformer (Zhou et al. 2018)	✗	✗	✓	2.8	11.1	-	-	-	-	-	-	-
Support-Set (Patrick et al. 2020)	✗	✗	✓	1.5	6.9	17.8	3.2	-	-	-	-	-
Frozen-BLIP (Li et al. 2023)	✓	✓	✗	0.0	0.4	1.4	0.0	16.1	0.011	0.008	-	-
Vid2Seq-HCY (Yang et al. 2023)	✓	✓	✓	2.3	8.2	19.0	7.6	24.2	-	-	0.888	0.214
Vid2Seq-HC (Yang et al. 2023)	✓	✓	✓	2.7	8.5	19.8	8.4	24.5	-	-	0.892	0.217
Vid2Seq-HCV (Yang et al. 2023)	✓	✓	✓	2.7	8.4	19.8	8.3	25.1	-	-	0.899	0.200
VSUM-BLIP (Lin et al. 2023b)	✗	✓	✗	-	-	-	-	21.7	0.207	0.131	-	-
TSUM-BLIP (Lin et al. 2023b)	✗	✓	✗	5.6	11.8	24.9	20.9	-	-	-	-	-
VTSUM-BLIP (Lin et al. 2023b)	✓	✓	✗	**5.8**	12.2	25.1	23.1	23.5	0.258	0.196	0.894	0.247
V2Xum-LLaMA-7B (ours)	✓	✓	✓	**5.8**	**12.3**	**26.3**	**26.9**	29.0	**0.298**	**0.204**	0.931	**0.253**
V2Xum-LLaMA-13B (ours)	✓	✓	✓	5.7	**12.3**	26.2	25.3	**31.6**	0.276	0.200	**0.957**	0.251
Human	✓	-	-	5.2	14.7	25.7	24.2	33.8	0.305	0.336	0.944	0.256

Performance on Video Summarization Tasks

Method	TVSum		SumMe	
	Spearman	Kendall	Spearman	Kendall
dppLSTM (Zhang et al. 2016)	0.055	0.042	-	-
DSN (Zhou, Qiao, and Xiang 2018)	0.020	0.026	-	-
Sumgraph (Park et al. 2020)	0.138	0.094	-	-
CLIP-it (Narasimhan et al. 2021)	0.147	0.108	0.120	0.109
TL;DW (Narasimhan et al. 2022)	0.167	0.143	0.128	0.111
iPTNet (Jiang and Mu 2022)	0.174	0.148	0.131	0.114
A2Summ (He et al. 2023a)	0.178	0.150	0.143	0.121
Standard ranker (Saquil et al. 2021)	0.230	0.176	0.014	0.011
VSUM-BLIP (Lin et al. 2023b)	0.261	0.200	0.365	0.268
V2Xum-LLaMA	**0.293**	**0.222**	**0.378**	**0.296**

Performance on Video Summarization Tasks

Method	V2T				V2V		V2TV
	B-4	M	R-L	C	F1	F_{CLIP}	Cross-F_{CLIP}
Vid2Seq-HC (Yang et al. 2023)	3.8	6.1	22.6	0.4	23.0	80.5	16.1
Vid2Seq-HCY (Yang et al. 2023)	3.7	6.2	22.4	0.5	24.7	81.3	16.0
Vid2Seq-HCV (Yang et al. 2023)	3.6	6.2	22.5	0.4	25.1	81.5	16.3
V2Xum-LLaMA-7B	**6.8**	**15.8**	26.9	**0.9**	**31.7**	**95.5**	**23.1**
V2Xum-LLaMA-13B	6.7	**15.8**	**27.0**	0.8	31.3	95.3	23.0

Ablation Study

Method	V2T				V2V				V2VT
	BLEU-4	METEOR	ROUGE-L	CIDEr	F1-Score	Spearman	Kendall	F_{CLIP}	Cross-F_{CLIP}
V2Xum-LLaMA	**5.8**	**12.3**	**26.3**	**26.9**	**29.0**	**0.298**	**0.204**	**0.931**	**0.253**
w/o simultaneous VT-Sum	5.6	12.2	25.6	25.9	25.1	0.249	0.203	0.926	0.251
w/o Instruct-V2Xum	4.9	12.0	24.3	21.6	23.1	0.260	0.191	0.921	0.252
w/o fully fine-tuning	4.5	11.7	24.7	22.8	23.4	0.222	0.175	0.915	0.250
w/o temporal prompts	4.4	11.7	24.4	21.2	23.9	0.258	0.192	0.910	0.249
w/o pretrained adapter	3.1	11.1	21.9	9.5	3.7	-	-	-	-

Case Study

[09]A breathtaking mountain range is set against a clear blue sky. The scene is marked by a green sign reading \"GREAT VALUE STEWART & LLOYD\". [31]A man in a black shirt is seen cleaning the floor with a red and black vacuum cleaner. [86]In a warehouse, another individual is diligently sweeping the floor.

[00]A man and a woman, both in suits with the woman in a pink jacket, are seated at a desk with a laptop, possibly discussing news. [19]A diverse crowd, some holding signs, stands in front of a building, engaging in a protest or public gathering. [27]A police officer in a blue uniform is positioned before a crowd. [34]A police officer in a blue uniform holds a baton, facing a crowd. [41]A police officer in a blue uniform, a woman, draws her service weapon in response to a protest. [58]A police officer in a blue uniform and a badge stands in front of a car. [65]A police officer in a blue uniform holds a gun while conversing with a man in a black shirt. [72]A police officer in a blue uniform stands by a white car, likely responding to a call. [80]A man in a black shirt is handcuffed by a police officer in a blue uniform. [87]A man in a black shirt is placed in a police car by an officer in a blue uniform.

[00]A woman in a black suit stands before a large airplane in a disaster area, showing the crashed airplane, a truck, and several people. [33]The frame shows a CNN reporter speaking on camera, likely reporting on an event related to ISIS-K, with the background featuring illuminated city lights and structures at night. [48]The frame shows the aftermath of a drone strike in a residential area, with damaged buildings and debris scattered around. [52]The personnel is carrying a flag-draped casket into a large aircraft, with another aircraft visible in the background. [67]A news broadcast showing two individuals walking near barbed wire fences, with news headlines reporting the death of a notable sports figure and a stock market update. [76]A busy street in a city with cars and pedestrians are visible, and the news ticker reports a rocket attack near an airport.

Visit our GitHub repo for code and data

https://github.com/hanghuacs/V2Xum-LLM