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

Hang Hua*, **Yunlong Tang***, **Chenliang Xu, Jiebo Luo**

University of Rochester

{hhua2, jluo}@cs.rochester.edu, {yunlong.tang, chenliang.xu}@rochester.edu

Abstract

Video summarization aims to create short, accurate, and cohesive summaries of longer videos. Despite the existence of various video summarization datasets, a notable limitation is their limited amount of source videos, which hampers the effective training of advanced large vision-language models (VLMs). Additionally, most existing datasets are created for video-to-video summarization, overlooking the contemporary need for multimodal video content summarization. Recent efforts have been made to expand from unimodal to multimodal video summarization, categorizing the task into three sub-tasks based on the summary's modality: video-to-video (V2V), video-to-text (V2T), and a combination of video and text summarization (V2VT). However, the textual summaries in previous multimodal datasets are inadequate. To address these issues, we introduce 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%. Each video summary in Instruct-V2Xum is paired with a textual summary that references specific frame indexes, facilitating the generation of aligned video and textual summaries. In addition, we propose a new video summarization framework named V2Xum-LLM. V2Xum-LLM, specifically V2Xum-LLaMA in this study, is the first framework that unifies different video summarization tasks into one large language model's (LLM) text decoder and achieves task-controllable video summarization with temporal prompts and task instructions. Experiments show that V2Xum-LLaMA outperforms strong baseline models on multiple video summarization tasks. Furthermore, we propose an enhanced evaluation metric for V2V and V2VT summarization tasks.

Introduction

The interest in sharing life experiences has surged in recent years, making video the most informative and diverse visual medium on social media platforms. This trend has led to significant demands for a variety of video and language understanding tasks, such as video captioning, video question answering (Yu et al. 2019; Xiao et al. 2021), moment retrieval (Lei et al. 2021), and video summarization (Gygli et al. 2014; Song et al. 2015). Video summarization (V2V) provides an efficient way for humans to obtain key information from a long video. This process entails selecting the most significant information from a video and condensing it into a shorter form, while maintaining the essence of the original content. Beyond extensive research in V2V summarization, there are a few recent explorations of V2T and V2VT summarization. Most notably, VideoXum (Lin et al. 2023a) seeks to broaden the modality of video summaries to include text summaries. It utilizes the dense captions from ActivityNetCap (Krishna et al. 2017) videos as text summaries and annotates the corresponding video segments as summaries. Other datasets for video summarization include TVSum (Song et al. 2015), SumMe (Gygli et al. 2014), QFVS (Sharghi et al. 2017), MED Summaries (Potapov et al. 2014), and so on. While it is expected that powerful LLMs can help improve video summarization, the insufficient number of source videos may not be able to support the robust fine-tuning of LLMs to perform this task since finetuning a large model with a limited number of training examples is prone to overfitting (Hua et al. 2021, 2023). In addition, VideoXum data cannot be considered true summaries due to redundant information in ActivityNetCap's dense captions. More recently, Shot2Story20K (Han et al. 2023) collects 20k video-text data that enables

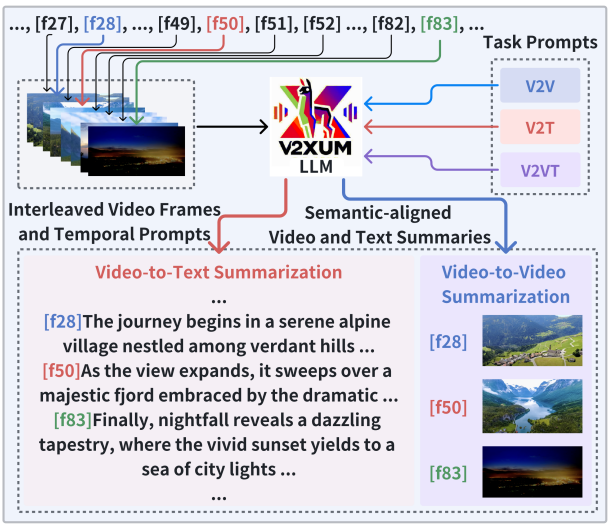

Figure 1: Illustration of cross-modal video summarization.

*These authors contributed equally.

the robust finetuning of LLMs, but it only supports video-to-text summarization. To address these issues, we propose Instruct-V2Xum, a new large-scale cross-model video summarization dataset that contains 30k open domain videos, partitioned as 25,000 in the training set, 1,000 in the validation set, and 4,000 in the test set. In Instruct-V2Xum, we obtain the source videos from YouTube using the video list provided by InternVid (Wang et al. 2023). The methodology of video summarization parallels that of extractive text summarization, where the objective is to isolate the pivotal frames or sentences from the source videos or documents, respectively. Extractive text summarization is a foundational task in the NLP field, with numerous LLMs setting the benchmark in this domain (Zhang et al. 2023a; Jia et al. 2020; He et al. 2023). Inspired by this, we first extract frames from the source videos and employ LLaVA-1.5-7B (Liu et al. 2023) to generate detailed captions for each frame. Then we take all the frame captions as a document to perform extractive document summarization using GPT-4 (Achiam et al. 2023). This approach enables us to obtain both video summaries and their corresponding textual summaries. Finally, the extracted text summaries are further refined by GPT-4.

Numerous visual instruction tuning-based methods have been proposed for video-language understanding. These models can process the long videos for general video-language understanding and reasoning tasks such as video question answering, video captioning, and so on. Recently, there have been some attempts for fine-grained video moments understanding or video-language temporal grounding (Lin et al. 2023b) using large VLMs. However, these approaches, which typically process video frames as sequential images for a frozen visual encoder and train an LLM decoder for identifying video moment boundaries, are not well-suited for dense temporal prediction in video summarization tasks, particularly in the V2VT tasks. Furthermore, most existing models require large-scale data to train new parameters added to pre-trained VLMs (Lin et al. 2023b; Maaz et al. 2023; He et al. 2023). This requirement significantly limits their practicality in scenarios where only a limited number of training examples are available. To address these problems, we design a new temporal prompt instruction tuning framework – V2Xum-LLaMA. In V2Xum-LLaMA, we unify different modalities of video summary generation into one LLM decoder. This framework stands out by removing the dependence on task-specific layers that were required for video summarization in earlier VLM-based approaches. The main advantages of this method are that it enables the effective adaptation of the learned knowledge and the powerful capabilities of the pretrained language models to dense video temporal and content understanding. All the pretrained parameters of the VLMs are reused, and the model takes interleaved video frames and natural language temporal prompts as input to facilitate end-to-end model training. As video temporal prediction is performed by using language models, there is a challenge for calculating the correlation for V2V evaluation, since the language model-predicted video summaries are the discrete frame indexes. To overcome this challenge, we propose a solution to calculate the scores for language model-predicted video summaries. Furthermore, we also provide an analysis of the existing video summarization tasks and propose F_{CLIP} and $Cross-F_{CLIP}$, the enhanced evaluation metrics for V2V and V2VT summarization tasks.

In summary, our main contributions are as follows:

- We 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. By taking interleaved video frames and temporal prompts as input, our method enables end-to-end processing of long video sequences and outperforms all strong baseline models on mainstream V2V, V2T, and V2VT benchmarks.
- To address the lack of video-language data for fine-tuning large VLMs in video summarization tasks, we created Instruct-V2Xum, a new instruction-following dataset for cross-modal video summarization. It contains 30k diverse YouTube videos, ranging from 40 to 940 seconds, enabling VLMs to generate modality-controllable video summaries with task prompts. The experiments validate the rationality of our proposed dataset.
- We present a comprehensive analysis of the limitations in current video summarization tasks from the perspectives of data, methods, and evaluation. Based on this, we propose F_{CLIP} and $Cross-F_{CLIP}$, an enhanced evaluation metric for V2V and V2VT summarization tasks. Experimental results show that these metrics are highly consistent with the traditional evaluation metrics including F1, Spearman correlation, and Kendall correlation.

Related Work

Video Summarization

Traditional video summarization, also known as video-to-video summarization, typically generates a condensed version of the original video, comprising selected frames (Liu et al. 2020; Ghauri et al. 2021), shots (Ji et al. 2019; Feng et al. 2018; Zhang et al. 2018), or segments (Tang et al. 2022; Koutras et al. 2019). These models are commonly trained using supervised learning approaches, with reinforcement learning methods like policy gradient (Williams 1992), optimizing for diversity and representativeness in the summarized output (Zhou et al. 2018a), gaining popularity. The standard datasets for video summarization tasks include SumMe (Gygli et al. 2014) and TVSum (Song et al. 2015), which are widely used for benchmarking purposes. In recent years, cross-modal video summarization (Fu et al. 2020; Li et al. 2022; Huang et al. 2021) has emerged as an area of interest, incorporating additional modalities such as audio, speech, subtitles, and captions. These approaches leverage multimodal models to create more comprehensive summaries. Video-to-text summarization (Palaskar et al. 2019; Choi et al. 2018) is an evolving field that aims to generate descriptive paragraphs in natural language that encapsulate video content. VideoXum (Lin et al. 2023a) advances this field by repurposing the ActivityNetCap (Krishna et al. 2017) dataset and employing the BLIP-2 (Li et al. 2023) model for both V2V and V2T summarization. However, the data in (Lin et al. 2023a) may not be genuine summaries, as the dense captions from ActivityNetCap often include significant redundancy.

Dataset	Domain	# Videos	Anno.	V2V	V2T	V2VT	Instruction
MSVD (Chen et al. 2011)	Open	1,970	M	×	✓	×	×
YouCook (Das et al. 2013)	Cooking	88	M	×	✓	×	×
UCF101 (Soomro et al. 2012)	Open	13,320	M	×	✓	×	×
ActivityNetCap (Krishna et al. 2017)	Activities	20,000	M	×	✓	×	×
Shot2Story20k (Han et al. 2023)	Open	20,000	M+S	×	✓	×	✓
SumMe (Gygli et al. 2014)	Events, holidays, sports	25	M	✓	×	×	×
TVSum (Song et al. 2015)	News, documentaries, vlogs	50	M	✓	×	×	×
VSUMM (De Avila et al. 2011)	Cartoons, news, commercials	50	M	✓	×	×	×
EDUVSUM (Ghauri et al. 2020)	Letures	98	M	✓	×	×	×
LoL (Fu et al. 2019)	Matches of League of Legends	218	M	✓	×	×	×
Ads-1K (Tang et al. 2022)	Commercials	1,041	M+S	✓	×	×	×
VideoXum (Lin et al. 2023a)	Activities	14,001	M	✓	✓	✓	×
Instruct-V2Xum	Open	**30,000**	M+S	✓	✓	✓	✓

Table 1: Comparison with existing video-to-video summarization and video-to-text summarization datasets. "V2V", "V2T", and "V2VT" indicate support for video-to-video, video-to-text, or both tasks. "Instruction" denotes whether the dataset supports video-text instruction tuning. M and S stand for manual and model synthesized, respectively.

Large Language Models

In recent years, Large Language Models (LLMs) have witnessed rapid advancements (Achiam et al. 2023; Touvron et al. 2023b,a). With pretraining on extensive corpora from the Internet, LLMs acquire substantial knowledge, enabling powerful zero-shot and in-context learning capabilities (Achiam et al. 2023; Wei et al. 2022; Hu et al. 2022b). Efforts have been increasingly directed toward leveraging LLMs for multimodal tasks (Lyu et al. 2023; Shu et al. 2023; Yu et al. 2024; Hua et al. 2024a). Techniques such as vision-language alignment and adapter fine-tuning are employed to integrate LLMs into the multimodal domain. These methods align the visual features extracted by visual encoders with the input token space of LLMs (Liu et al. 2023). Typically, the parameters of LLMs are frozen to retain their existing capabilities although LoRA (Hu et al. 2022a) fine-tuning is sometimes applied in low-resource settings. Based on this, several studies (Maaz et al. 2023; Zhang et al. 2023b) have successfully employed LLMs for video understanding tasks, referred to as Vid-LLMs (Tang et al. 2023). However, current research primarily concentrates on general video understanding tasks, such as video question-answering (QA) and video captioning, with less emphasis on temporal information. Recent works (Ren et al. 2023; Tang et al. 2024) explore LLMs' potential in temporal grounding and localization, emphasizing their untapped capability in temporal understanding.

The Instruct-V2Xum Dataset

Data Curation

Frame Captioning and Extractive Summarization. The source videos are sampled from InternVid (Wang et al. 2023). We filter the raw videos according to their duration and aesthetic scores. The filtered data is then used to generate both video and text summaries. We extract video frames at a rate of 1 FPS and then convert these frames into detailed textual descriptions using LLaVA-1.5-7B. After obtaining the textualized video frames, we utilize GPT-4V to perform extractive document summarization. Finally, the extracted summaries are converted into coherent video and text summaries.

Text Summarization Refinement. To further reduce the redundancy of the text summaries, we utilize BERT score (Zhang et al. 2019) to filter out the frame text representations that are similar to other summary frames. Here, we set the threshold to 0.93. The filtered video frame captions' indexes serve as the video summaries for the source videos. Then, we employ GPT-4 to further compress and rewrite the video summaries to be shorter and more grammar-fluent.

Human Verification. To enhance the quality of the collected data, we employed human annotators to filter the GPT-4-generated data, resulting in a final set of 30k data points. We also show examples of human-filtered data in the our technical appendices (Hua et al. 2024b).

Quantity Analysis and Quality Analysis

We provide statistical results and a comprehensive quantitative analysis in our technical appendices (Hua et al. 2024b). Additionally, we analyze text summaries for grammar fluency and commonsense plausibility, with a detailed quality analysis also included in the appendices (Hua et al. 2024b).

V2Xum-LLaMA

This section introduces our unified cross-modal video summarization framework, V2Xum-LLaMA, which employs interleaved video frames along with temporal and task prompts as input and converts videos to multimodal summaries, as shown in Figure 2.

Interleaved Video and Temporal Prompt Encoding

The temporal prompt mechanism binds visual tokens to their corresponding frame-level timestamps, injecting positional

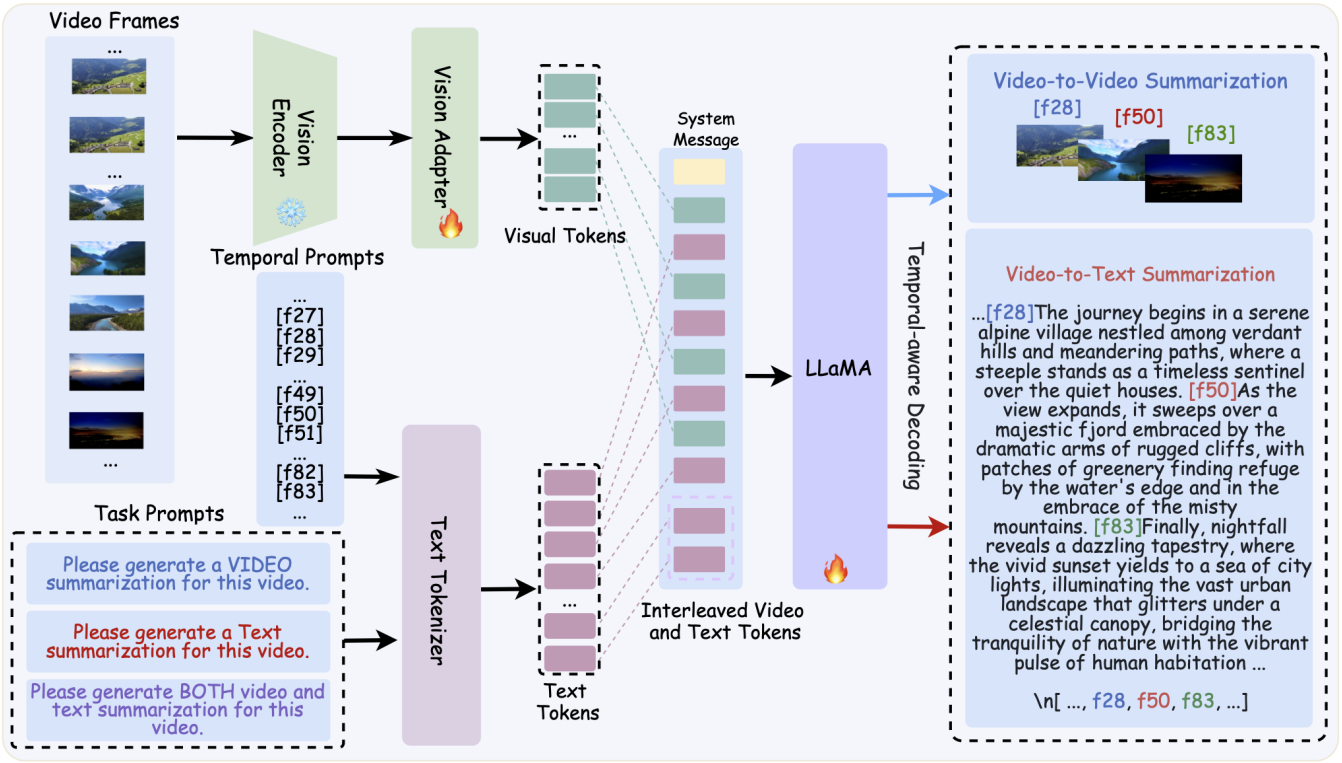

Figure 2: The architecture of the proposed V2Xum-LLaMA.

information for each frame. To this end, we first encode each video frame f_i using pretrained CLIP (Radford et al. 2021) encoder E_v. The frames' encoding $v_i = E_v(f_i)$ are involved in the visual tokens sequence $V = \{v_1, v_2, ..., v_L\}$, where L is the number of sampled frames.

We then bind temporal prompts with each visual token. The temporal prompts are tokenized zero-padded numbers in natural language format like "[f00]", "[f06]", "[f12]", "[f99]", etc., indicated by $T = \{t_1, t_2, ..., t_L\}$. They are inserted into the visual token sequence V to form a new sequence with interleaved visual tokens and temporal prompts:

$$S = \{t_1, v_1, t_2, v_2, ..., t_L, v_L\} \tag{1}$$

Compared to the original visual token sequence, the temporal prompted sequence can better capture the relations between the timestamps and the visual semantics. Then the sequence S is projected into the word embedding space by the vision adapter, which is represented as $\overline{S}$.

Temporal-Aware Decoding

We use LLaMA-2 (Touvron et al. 2023a) as the decoder to generate V2V summarization A^v, V2T summarization A^t, or V2VT summarization A^b. The I^v, I^t, and I^b represent the task instructions for V2V, V2T, and V2VT summarization. With the instruction I and temporal prompted sequence $\overline{S}$, the output is given by the temporal-aware decoding:

$$A^x = LLM(\overline{S}, I^x),\ x \in \{v, t, b\} \tag{2}$$

The V2V summarization A^v is defined as a sequence of temporal tokens sequence, i.e., $A^v = \{v_i\}_{i=1}^{M}$, which is a subset of temporal prompts sequence T and $M \leq L$. The V2T summarization is a summarized caption in a natural language format. In our implementation shown as Figure 2, the V2T summarization also contains frame referring temporal tokens:

$$A^t = \{..., w_{i-1}, t_j, w_{i+1}, w_{i+2}...\} \tag{3}$$

where the temporal token t_j is bind with the words or sentences that consist of the summarization of visual content and can be further extracted from A^t to get the V2V summarization A^v. This is called temporal-aware decoding. An example is shown in Fig. 2. When decoding, the temporal tokens can be decoded together with the text summaries and represent the temporal position where the described visual content occurred in the input video.

Task-Controllable Video Summarization Training

As mentioned before, the types of summarization can be controlled by the task instructions I^x. Specifically, we use task prompts like "Please generate a BOTH/VIDEO/TEXT summarization for this video." to instruct models to generate the corresponding video summaries. We then train the model end-to-end using negative log-likelihood loss:

$$\mathcal{L} = -\sum_{\mathcal{D}} \sum_{i=1}^{N} \log p(A_i^x \mid \overline{S}, A_{\leq i-1}^x). \tag{4}$$

where $\mathcal{D}$ denotes the training samples in the dataset, and N is the length of the generated video and text summaries. During training, all the parameters of the vision encoder and update the vision adapter, and the language decoder are freezed.

Method	Cross-Modal	LLM-Based	TSH-Free	V2T				V2V				V2VT
				B-4	M	R-L	C	F1	S	K	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	×	×	✓	2.1	10.0	-	-	-	-	-	-	-
Masked Transformer	×	×	✓	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. 2023a)	×	✓	×	-	-	-	-	21.7	0.207	0.131	-	-
TSUM-BLIP (Lin et al. 2023a)	×	✓	×	5.6	11.8	24.9	20.9	-	-	-	-	-
VTSUM-BLIP (Lin et al. 2023a)	✓	✓	×	**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

Table 2: Comparison Results on the VideoXum dataset. "TSH-Free" indicates the model is task-specific-head-free; "B-4" denotes BLEU-4; "M" denotes METEOR; "R-L" refers to ROUGE-L metric; "C" represents CIDEr. For all metrics, higher scores indicate better performance. "S" and "K" are Spearman and Kendall correlation metrics, respectively.

Experiments

Baseline Models

We evaluate our V2Xum-LLaMA model, both 7B and 13B versions, against various models on V2V, V2T, and V2VT summarization tasks using the VideoXum dataset. Baseline models include LLM-based approaches such as Frozen-BLIP (Li et al. 2023), VSUM-BLIP (Lin et al. 2023a), TSUM-BLIP (Lin et al. 2023a), and VTSUM-BLIP (Lin et al. 2023a). We compare with task-specific-head-free (TSH-Free) models like DENSE (Krishna et al. 2017), DVC-D-A (Li et al. 2018), Bi-LSTM+TempoAttn (Zhou et al. 2018b), Masked Transformer (Zhou et al. 2018b), and Support-Set (Patrick et al. 2020), which do not rely on regression-based timestamp prediction with extra task-specific heads. Additionally, on the classical TVSum (Song et al. 2015) and SumMe (Gygli et al. 2014) datasets, we compare our 7B version V2Xum-LLaMA with the following V2V summarization methods: dppLSTM (Zhang et al. 2016), DSN (Zhou et al. 2018a), Sumgraph (Park et al. 2020), CLIP-it (Narasimhan et al. 2021), TL;DW (Narasimhan et al. 2022), iPTNet (Jiang et al. 2022), A2Summ (He et al. 2023), Standard ranker (Saquil et al. 2021), and VSUM-BLIP (Lin et al. 2023a). We also evaluated Vid2Seq (Yang et al. 2023), initially pre-trained on the HowTo100M (Miech et al. 2019), VidChapter-7M (Yang et al. 2024), YouCook2 (Zhou et al. 2018c), and ViTT (Huang et al. 2020), and fine-tuned on VideoXum for fair comparison.

Evaluation Metrics

We introduce new CLIP-based evaluation metrics for V2V and V2VT summarization evaluation. While the F1 score is a common metric for V2V summarization tasks, the process of video summarization annotation is highly subjective, leading to considerable variance among human annotators (Otani et al. 2019). The traditional F1 score, which compares predicted video frames directly with the ground truth, fails to account for semantically similar frames that are close in time but not exactly matching, thus potentially undervaluing accurate summaries. To mitigate this, we introduce the F_{CLIP} metric for V2V summarization evaluations, which is designed to recognize and reward semantic similarities between predicted and ground truth frames even when they are not identical. And we also propose the $Cross - F_{CLIP}$ metric for the V2VT summarization tasks. Unlike the VT-CLIPScore metric used in VideoXum—which calculates the average cross-modal CLIP score as an indicator of semantic alignment between predicted video and text summaries—our $Cross - F_{CLIP}$ calculates the average F_{CLIP} scores between the predicted video summaries and the ground truth text summaries, as well as between the predicted text summaries and the ground truth video summaries. This approach aims to provide a more nuanced evaluation of summarization tasks by acknowledging the importance of semantic content alignment across modalities. For a reference video summary v and predicted video summary $\hat{v}$, the recall, precision, and F1 scores are:

$$R_{CLIP}(v, \hat{v}) = \frac{1}{|v|} \sum_{v_i \in v} \max_{\hat{v}_j \in \hat{v}} \mathbf{v}_i^\top \hat{\mathbf{v}}_j \tag{5}$$

$$P_{CLIP}(v, \hat{v}) = \frac{1}{|\hat{v}|} \sum_{\hat{v}_j \in \hat{v}} \max_{v_i \in v} \mathbf{v}_i^\top \hat{\mathbf{v}}_j \tag{6}$$

$$F_{CLIP}(v, \hat{v}) = 2\frac{P_{CLIP} \cdot R_{CLIP}}{P_{CLIP} + R_{CLIP}} \tag{7}$$

For the $Cross - F_{CLIP}$, given a reference video and text summary v and t and the predicted video summary $\hat{v}$ and text

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

Table 3: Comparison results on the Instruct-V2Xum test set.

summary $\hat{t}$:

$$Cross - F_{CLIP}(v, \hat{v}, t, \hat{t}) = \frac{F_{CLIP}(v, \hat{t}) + F_{CLIP}(\hat{v}, t)}{2} \tag{8}$$

Given that the cosine similarity values range from -1 to 1, we adjust the similarity scores by applying the operation $\max(\cos(\mathbf{v}, \hat{\mathbf{v}}), 0)$. This ensures that only non-negative similarity scores are considered in our analysis.

Implementation Details

We use CLIP ViT-L/14@336 as the vision encoder and Vicuna-v1.5-7B/13B as the text decoder. Other implementation details are included in our technical appendices (Hua et al. 2024b).

Quantitative Results

For cross-modal video summarization, adopt the VideoXum dataset (Lin et al. 2023a) and our proposed V2Xum dataset. For V2V summarization, we used the TVSum (Song et al. 2015) and SumMe (Gygli et al. 2014) benchmarks.

Method	TVSum		SumMe	
	S	K	S	K
dppLSTM (Zhang et al. 2016)	.055	.042	-	-
DSN (Zhou et al. 2018a)	.020	.026	-	-
Sumgraph (Park et al. 2020)	.138	.094	-	-
CLIP-it (Narasimhan et al. 2021)	.147	.108	.120	.109
TL;DW (Narasimhan et al. 2022)	.167	.143	.128	.111
iPTNet (Jiang et al. 2022)	.174	.148	.131	.114
A2Summ (He et al. 2023)	.178	.150	.143	.121
VSUM-BLIP (Lin et al. 2023a)	.261	.200	.365	.268
V2Xum-LLaMA	**.293**	**.222**	**.378**	**.296**

Table 4: Comparison results on the TVSum and SumMe datasets.

Cross-Modal Video Summarization. We use the VideoXum dataset to evaluate our model's capability to cross-model video summarization. The experimental results are shown in Table 2. It can be summarized that our proposed method outperforms all the baseline models. Specifically, in V2V summarization, V2Xum-LLaMA achieves a **8.1%** higher F1-Score than VTSUM-BLIP, alongside significant improvements in both Spearman and Kendall correlation metrics. In addition, V2Xum-LLaMA-13B achieves higher evaluation scores on the V2V summarization task than V2Xum-LLaMA-7B. On the contrary, V2Xum-LLaMA-7B performs better on V2T summarization. We attribute this result to the increased complexity of the V2T summarization compared to V2V summarization. Additionally, the VideoXum training set comprises only 8,000 examples, a quantity insufficient for effectively training models with large language decoders.

We also evaluate various models on our newly proposed Instruct-V2Xum dataset, as detailed in Table 3. The results indicate that the models are well-adapted to the dataset, exhibiting sound performance. Moreover, V2Xum-LLaMA-13B outperforms V2Xum-LLaMA-7B in V2T summarization, which we believe is due to the larger language models benefiting from the increased volume of training data.

Video-to-Video Summarization. We evaluate V2Xum-LLaMA on VideoXum, TVSum, SumMe, and Instruct-V2Xum datasets. The results are shown in Table 2 and Table 4. It can be concluded that the unified video summarization using the language decoders in V2Xum-LLaMA can effectively perform traditional V2V summarization tasks. V2Xum-LLaMA outperforms all the previous methods that relied on task-specific regression heads for generating video summaries. This result indicates that LLMs with temporal prompts are capable of performing fine-grained video temporal understanding. The results presented in Table 3 affirm the validity of our proposed V2Xum dataset. It demonstrates that the model can properly fit the data.

Ablation Study

To better evaluate the effectiveness of our proposed V2Xum-LLaMA framework and to underscore the importance of augmenting the training dataset, we conduct an ablation study on V2V and V2VT summarization tasks, detailed in Table 5. A comprehensive ablation analysis is included in our technical appendices (Hua et al. 2024b).

Limitations of Current Video Summarization: Data, Methods, and Evaluation

In this section, we discuss the limitations of existing vision summarization tasks from the perspective of data, method, and evaluation. As mentioned before, current existing video summarization datasets, including V2V and V2VT video

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	-	-	-	-

Table 5: Ablation Study of our V2Xum-LLaMA (7B) on the VideoXum dataset.

summarization datasets, contain few training examples and cannot support training large-scale deep neural networks. The TVSum dataset comprises merely 50 YouTube videos; the SumMe dataset includes only 25 personal videos sourced from YouTube, while the QFVS dataset provides 135 video-query training samples. Additionally, VideoXum, the first cross-modal video summarization dataset, provides only 8k training examples. Our experimental results reveal that it is insufficient for training 13B models for the V2T summarization task. This issue is one of the significant drawbacks of current existing datasets. To address this problem, we collect more videos from YouTube and generate the corresponding cross-modal summaries for the videos using GPT-4 and propose a large-scale cross-modal video summarization dataset.

A conventional approach to V2V summarization involves training a regression head to assign an importance score to each frame, ranking frames based on these scores, and selecting the top K% frames to evaluate performance using metrics like the F1 score or Kendall/Spearman correlation against the ground truth. However, as demand for cross-modal video summarization grows, this method is inadequate for multi-modal video summarization. Recent studies have begun to explore the use of language models for generating temporal and spatial references in videos (Li et al. 2024; Ren et al. 2023), demonstrating the viability of using language models to generate the video interval indexes. In addition, leveraging large language models' text decoders for both V2V and V2T summarization tasks is a logical step. Therefore, in this study, we investigate how to prompt large VLMs to understand the fine-grained video content along with temporal information, how to maximally leverage the powerful capability of content understanding and reasoning of large language models, and how to achieve task controllability via natural language instructions. To that end, we first propose the temporal prompt mechanism, and then design the visual encoder that takes the interleaved video frames, temporal, and language as input. Instead of using the regression head to perform V2V summarization, we unify different video summarization tasks into one language decoder.

Evaluating video summaries poses a significant challenge, primarily because the criteria for quality are inherently subjective, vary across different viewers and even fluctuate over time. This subjectivity, coupled with the limited availability of evaluation videos and annotations, further exacerbates the ambiguity in assessing video summary quality (Otani et al. 2019). The traditional F1 score, designed to directly compare predicted video frames with ground truth, ignores the nuances of semantically similar frames that, while temporally proximate, do not exactly match. This oversight can lead to the undervaluation of otherwise accurate summaries. To address this, we design new CLIP-based F scores for evaluating V2V and V2VT summarization tasks. Unlike the traditional F1 score, which relies on exact matches for its precision and recall calculations, F_{CLIP} evaluates the V2V summarization from the perspective of semantic similarity. This approach allows for a more comprehensive yet meaningful evaluation that recognizes the importance of semantic accuracy over mere frame-by-frame accuracy. To evaluate the alignment of the generated video and text summaries, VideoXum employs an approach that calculates the average CLIP score across video frames and text summaries. This method involves computing vectors for all frames and sentences, applying mean pooling to these vectors to represent the video and text summaries, and then calculating the cosine similarity between them. However, this approach evaluates the video and text summaries as unified entities and thus neglects the detailed alignment at the sentence level with specific video frames. In contrast, our proposed $Cross - F_{CLIP}$ adopts a greedy matching strategy to optimize the similarity score between individual video frames and corresponding sentences, ensuring a more granular and accurate alignment.

Conclusion

In this study, we address the deficiencies in current video summarization datasets including the insufficient number of training examples and the insufficient evaluation of video summarization by building a new large-scale cross-model video summarization dataset Instruct-V2Xum and designing the improved video summarization evaluation metrics. We also propose V2Xum-LLM, a novel temporal prompt instruction tuning method that unifies the generation of various video summary modalities within the text decoder of VLMs, eliminating the need for task-specific heads. This approach supports interleaved long video and language input sequences and allows modality-controllable summary generation through language instructions. Experimental results demonstrate the effectiveness of our method.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; et al. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Chen, D. L.; et al. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Annual Meeting of the Association for Computational Linguistics*.

Choi, J.; et al. 2018. Contextually customized video summaries via natural language. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1718–1726. IEEE.

Das, P.; et al. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2634–2641.

De Avila, S.; et al. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern recognition letters*, 32(1): 56–68.

Feng, L.; et al. 2018. Extractive video summarizer with memory augmented neural networks. In *Proceedings of the 26th ACM international conference on Multimedia*, 976–983.

Fu, T.-J.; et al. 2019. Attentive and Adversarial Learning for Video Summarization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1579–1587.

Fu, X.; et al. 2020. Multi-modal summarization for video-containing documents. arXiv:2009.08018.

Ghauri, J. A.; et al. 2020. Classification of Important Segments in Educational Videos using Multimodal Features. *International Workshop on Investigating Learning During Web Search (IWILDS 2020) co-located with CIKM*.

Ghauri, J. A.; et al. 2021. Supervised video summarization via multiple feature sets with parallel attention. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6s. IEEE.

Gygli, M.; et al. 2014. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, 505–520. Springer.

Han, M.; et al. 2023. Shot2Story20K: A New Benchmark for Comprehensive Understanding of Multi-shot Videos. arXiv:2312.10300.

He, B.; et al. 2023. Align and Attend: Multimodal Summarization with Dual Contrastive Losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, E. J.; et al. 2022a. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, Y.; et al. 2022b. PromptCap: Prompt-guided task-aware image captioning. arXiv:2211.09699.

Hua, H.; et al. 2021. Noise stability regularization for improving BERT fine-tuning. arXiv:2107.04835.

Hua, H.; et al. 2023. Improving Pretrained Language Model Fine-Tuning With Noise Stability Regularization. *IEEE Transactions on Neural Networks and Learning Systems*.

Hua, H.; et al. 2024a. FINEMATCH: Aspect-based Fine-grained Image and Text Mismatch Detection and Correction. arXiv:2404.14715.

Hua, H.; et al. 2024b. V2Xum-LLM: Cross-modal video summarization with temporal prompt instruction tuning. arXiv:2404.12353.

Huang, G.; et al. 2020. Multimodal pretraining for dense video captioning. arXiv:2011.11760.

Huang, J.-H.; et al. 2021. Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 580–589.

Ji, Z.; et al. 2019. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6): 1709–1717.

Jia, R.; et al. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 3622–3631.

Jiang, H.; et al. 2022. Joint Video Summarization and Moment Localization by Cross-Task Sample Transfer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16367–16377.

Koutras, P.; et al. 2019. Susinet: See, understand and summarize it. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Krishna, R.; et al. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lei, J.; et al. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Li, H.; et al. 2022. Progressive Video Summarization via Multimodal Self-supervised Learning. arXiv:2201.02494.

Li, J.; et al. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; et al. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7492–7500.

Li, Z.; et al. 2024. LEGO: Language Enhanced Multi-modal Grounding Model. arXiv:2401.06071.

Lin, J.; et al. 2023a. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Transactions on Multimedia*.

Lin, K. Q.; et al. 2023b. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2794–2804.

Liu, H.; et al. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Y.-T.; et al. 2020. Transforming multi-concept attention into video summarization. In *Proceedings of the Asian Conference on Computer Vision*.

Lyu, C.; et al. 2023. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. arXiv:2306.09093.

Maaz, M.; et al. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv:2306.05424.

Miech, A.; et al. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2630–2640.

Narasimhan, M.; et al. 2021. Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34: 13988–14000.

Narasimhan, M.; et al. 2022. Tl; dw? summarizing instructional videos with task relevance and cross-modal saliency. In *European Conference on Computer Vision*, 540–557. Springer.

Otani, M.; et al. 2019. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7596–7604.

Palaskar, S.; et al. 2019. Multimodal abstractive summarization for How2 videos. arXiv:1906.07901.

Park, J.; et al. 2020. Sumgraph: Video summarization via recursive graph modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 647–663. Springer.

Patrick, M.; et al. 2020. Support-set bottlenecks for video-text representation learning. arXiv:2010.02824.

Potapov, D.; et al. 2014. Category-specific video summarization. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 540–555. Springer.

Radford, A.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, S.; et al. 2023. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. arXiv:2312.02051.

Saquil, Y.; et al. 2021. Multiple Pairwise Ranking Networks for Personalized Video Summarization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1698–1707.

Sharghi, A.; et al. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4788–4797.

Shu, F.; et al. 2023. Audio-Visual LLM for Video Understanding. arXiv:2312.06720.

Song, Y.; et al. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5179–5187.

Soomro, K.; et al. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402.

Tang, Y.; et al. 2022. Multi-modal segment assemblage network for ad video editing with importance-coherence reward. In *Proceedings of the Asian Conference on Computer Vision*, 3519–3535.

Tang, Y.; et al. 2023. Video understanding with large language models: A survey. arXiv:2312.17432.

Tang, Y.; et al. 2024. Empowering LLMs with Pseudo-Untrimmed Videos for Audio-Visual Temporal Understanding. arXiv:2403.16276.

Touvron, H.; et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Touvron, H.; et al. 2023b. Llama: Open and efficient foundation language models. arXiv:2302.13971.

Wang, Y.; et al. 2023. InternVid: A large-scale video-text dataset for multimodal understanding and generation. arXiv:2307.06942.

Wei, J.; et al. 2022. Emergent abilities of large language models. arXiv:2206.07682.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8: 229–256.

Xiao, J.; et al. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Yang, A.; et al. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10714–10726.

Yang, A.; et al. 2024. Vidchapters-7m: Video chapters at scale. *Advances in Neural Information Processing Systems*, 36.

Yu, Y.; et al. 2024. PromptFix: You Prompt and We Fix the Photo. arXiv:2405.16785.

Yu, Z.; et al. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9127–9134.

Zhang, H.; et al. 2023a. Diffusum: Generation enhanced extractive summarization with diffusion. arXiv:2305.01735.

Zhang, H.; et al. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv:2306.02858.

Zhang, K.; et al. 2016. Video summarization with long short-term memory. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, 766–782. Springer.

Zhang, K.; et al. 2018. Retrospective encoders for video summarization. In *Proceedings of the European conference on computer vision (ECCV)*, 383–399.

Zhang, T.; et al. 2019. BERTScore: Evaluating text generation with BERT. arXiv:1904.09675.

Zhou, K.; et al. 2018a. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhou, L.; et al. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8739–8748.

Zhou, L.; et al. 2018c. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.