Dual Video Summarization: From Frames to Captions

Zhenzhen Hu[1,2*], **Zhenshan Wang**[1], **Zijie Song**[1] and **Richang Hong**[1]
[1]Hefei University of Technology
[2] Institute of Artificial Intelligence Hefei Comprehensive National Science Center
{ huzhen.ice, wangzhenshan98, zjsonghfut, hongrc.hfut}@gmail.com

Abstract

Video summarization and video captioning both condense the video content from the perspective of visual and text modes, i.e., the keyframe selection and language description generation. Existing video-and-language learning models commonly sample multiple frames for training instead of observing all. These sampled deputies greatly improve computational efficiency, but do they represent the original video content enough with no more redundancy? In this work, we propose a dual video summarization framework and verify it in the context of video captioning. Given the video frames, we firstly extract the visual representation based on the ViT model fine-tuned on the video-text domain. Then we summarize the keyframes according to the frame-lever score. To compress the number of keyframes as much as possible while ensuring the quality of captioning, we learn a cross-modal video summarizer to select the most semantically consistent frames according to the pseudo score label. Top K frames (*K is no more than 3% of the entire video.)* are chosen to form the video representation. Moreover, to evaluate the static appearance and temporal information of video, we design the ranking scheme of video representation from two aspects: score-oriented and time-oriented. Finally, we generate the descriptions with a lightweight LSTM decoder. The experiment results on the MSR-VTT and MSVD dataset reveal that, for the generative task as video captioning, a small number of keyframes can convey the same semantic information to perform well on captioning, or even better than the original sampling.

1 Introduction

Video and language, as two kinds of sequence signals, provide a wealth of information for people's daily communication. A wild range task, such as video captioning [Chen and Jiang, 2019; Lin *et al.*, 2021], video question answering [Yang *et al.*, 2021], text-video retrieval [Gabeur *et al.*, 2020; Bain *et al.*, 2021] and video grounding [Anne Hendricks *et al.*, 2017; Escorcia *et al.*, 2019], has been designed due to the application potentials. Compared to the isolated image and abstract language description, videos are more engaging by conveying vivid and dynamic visual content. However, for video understanding and processing, it also contains quite a redundant.

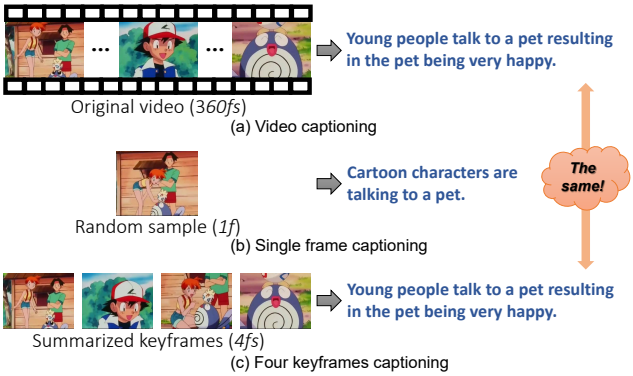

Figure 1: Video contains quite a redundancy towards video captioning task. Compared to the original video captioning in (a), one single frame sampled from the video can generate a nearly semantic expression (b). In our work, we learn to summarize the most compact keyframes to achieve consistency in the form of language description (c).

Video summarization and video captioning are tasks dedicating to extract a concise content of video from the perspective of visual and linguistic modes. Video summarization aims to select keyframes or shots while video captioning devotes to generate the natural language descriptions to express the video content. Video captioning is also known as a text form of video summarization [Sanabria *et al.*, 2018; Shang *et al.*, 2021]. Considering the redundancy in the video caused by the frame similarity, existing video-and-language learning model commonly chooses multiple frames as model inputs instead of observing all. A conventional way to process the input video is to randomly sample several frames [Baraldi *et al.*, 2017; Pei *et al.*, 2019; Lei *et al.*, 2021]. [Chen *et al.*, 2018] choose the informative frames according to the visual representation of them, like visual summarization. These sampled deputies greatly improve computational efficiency,

*Corresponding author.

but are they represent the original video content enough and with no more redundancy? The crucial issue of finding out the most informative frames for video captioning is: *In the video-language learning task, what is the essential visual content from the video?* [Buch *et al.*, 2022] and [Lei *et al.*, 2022] both revisit this problem by single frame training scheme and verify the pre-trained model on the discriminative downstream task, i.e., Video QA and cross-modal video retrieval. The result from [Lei *et al.*, 2022] reveals the existence of a strong "static appearance bias" in popular video-and-language datasets. Different from these discriminative tasks, video captioning is a generative task and requires a full coverage understanding of visual and temporal content. Single frame without any temporal clues is not satisfied to convey a comprehensive content for natural language expressing, as shown in Figure 1.

In this paper, we consider that the caption generated from the keyframes should closely parallel that from the original video. Motivated by this, we propose a dual video summarization framework and verify it in the context of video captioning to find a moderate strategy and number of sampled frames with accuracy and efficiency. Given the sampling frames as candidates, we finetune the pre-trained ViT model [Radford *et al.*, 2021] based on the video-text retrieval task to extract the visual features. To select the most representative frames among them, we implement a cross-modal video summarization module as an auxiliary means to summarize the frames. We generate a pseudo score label of each frame as the reference to facilitate the summarizer. Then we sort the frames by their scores and select the Top K frames (K *is no more than* 3% *of the entire video.)* as the most compact and semantically consistent summary to represent the video. Moreover, to evaluate the static appearance and temporal information of video, we design the ranking scheme of video representation from two aspects: score-oriented and time-oriented. Finally, we generate the descriptions with a lightweight LSTM decoder. The experiment results on the MSR-VTT and MSVD dataset reveal that, for the generative task as video captioning, a small number of keyframes can convey the same semantic information to perform well on captioning, or even better than the original sampling.

In summary, our contributions are three-fold:

- We cooperate the video summarization and video captioning task with each other to investigate the video frame representation problem. We propose a dual video summarization framework that select accurate and compact keyframes without frame-level annotation.
- We design a semantic-consistency video summarization module to assist the video captioning. We utilize `clipscore` between the visual feature and text embedding as the pseudo label to facilitate the score learning module.
- We evaluate our model on the video captioning benckmarks MSR-VTT and MSVD. We find that, for the generative task as video captioning, a small number of keyframes can convey the same semantic information and is able to perform well in the captioning task, or even better than the original sampling.

2 Related Work

2.1 Video Captioning

Video captioning is a challenging task of yielding corresponding natural language description for a given video. In the past few years, the field of video captioning has been obtained great advancement with lots of newly proposed method. Existing works of video captioning mainly adapt an encoder-decoder framework. [Venugopalan *et al.*, 2015] firstly exploits LSTMs to learn the temporal structure of videos and then generate descriptions. [Pan *et al.*, 2016] proposes a hierarchical encoder which takes a series of visual feature sequences into a single vector as the main representation of the whole video. Following the same paradigm, [Baraldi *et al.*, 2017] encodes semantic content and video frames in a trainable encoding layer. [Chen and Jiang, 2019; Zhang and Peng, 2020] employ temporal and spatial attention for tackling video feature alignment and aggregation. [Song *et al.*, 2017] decides whether to depend on the visual information or the semantic context information especially when generating non-visual words(e.g."a", "the"). [Yan *et al.*, 2022] produces rich semantic vocabulary to obtain description of video contents from the proposed global-local representation granularity framework. [Gao *et al.*, 2022] tries to resolve the disconnection between offline extracted motion or appearance features and sentence generation by a dual-level transformer with image-text pre-training models. For improving model's generation efficiency and effectiveness, [Chen *et al.*, 2018] proposes frame selecting strategy to decrease input redundancy with little performance drop. In this paper, unlike previous works [Chen and Jiang, 2019; Zhang and Peng, 2020; Yan *et al.*, 2022; Gao *et al.*, 2022] use densely extracted offline features or complicated architecture, we rethink reducing video's content redundancy by learning to sample high informative frames from already sparsely sampled candidates as well as maintaining our model's lightweight.

2.2 Video Summarization

Video summarization aims to generate a subset of informative frames that can present the main contents from video sequences. Early works[Borji and Itti, 2012; Gygli *et al.*, 2013; Gygli *et al.*, 2014; Zhao and Xing, 2014] mainly concentrate on designing handed-craft video representation(e.g.visual attention, video interest) in unsupervised manner. [Gygli *et al.*, 2014] assigns frame scores according to multi level features and the summary would be selected as the optimal subset of them. [Zhou *et al.*, 2018] formulates video summarization as a sequential decision-making process and design reinforcement learning framework that doesn't rely annotated frame level labels. Some works take multimodal interacting into account. [Song *et al.*, 2015; Qiu *et al.*, 2022] bring in textual sources such as video's title or supplied articles to help key frame location while [Xiao *et al.*, 2020; Narasimhan *et al.*, 2021] consider users' preference and provide more fine-grained summarization through integrating query into visual features. In this work, we attempt to select frames on the unlabeled video captioning dataset by content-aware supervised learning.

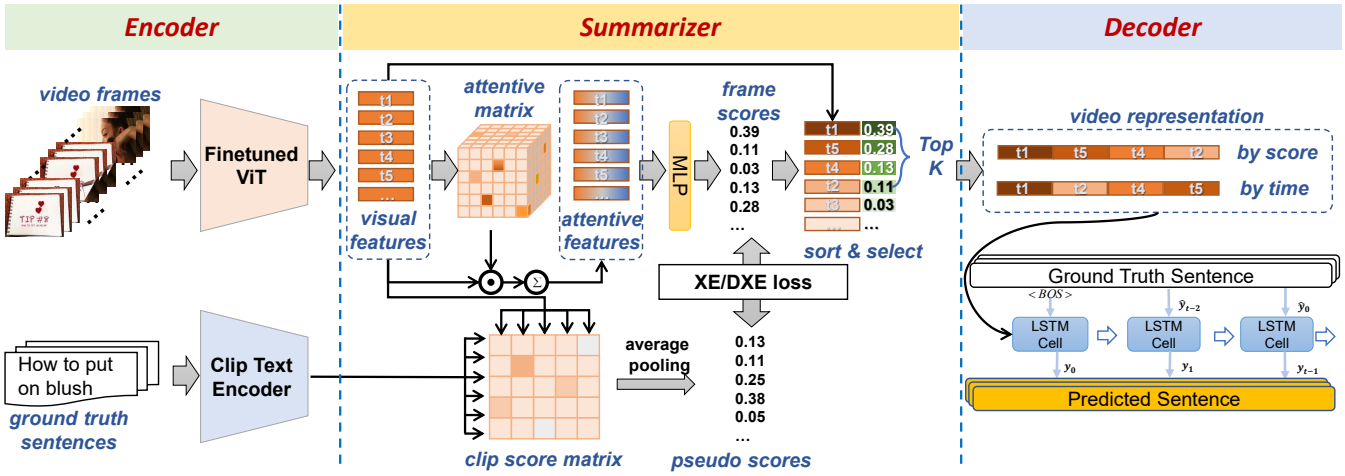

Figure 2: The illustration of our framework. Given the video frames, we first encode the visual feature by the ViT model finetuned with the video-text retrieval task. After attention augmentation, we summarize the TOP K frames by learning scores. We utilize `clipscore` between the visual feature and text embedding as the pseudo label to facilitate the score learning module. The decoder is a light weight LSTM to generate caption according to the video summarization.

3 Framework

The proposed dual video summarization framework follows the pipeline with three components: an encoder, a summarizer and a decoder, as shown in Figure 2. The encoder extracts visual features from candidate frame samples. The summarizer selects the most representative and compact frames according to the frame scores, while the decoder generates natural language descriptions based on the selected keyframes.

3.1 Problem Formulation

Given an input video V with totally T frames, we follow the common measures to sample N frames and represent them as visual feature $F = \{f_1, f_2, ..., f_N\}, f_i \in \mathbb{R}^{d_f}$ with equal time spacing. The target of our framework to select a subset F_{key} from F with K frames ($K < N \& K \ll T$) as keyframes' features, while the caption generated from the keyframes should closely parallel that from the original video.

For each video in the training set, there are M captions represented by the text embedding $C = \{c_1, c_2, ..., c_M\}, c_i \in \mathbb{R}^{d_c}$. Each caption is a sentence (i.e., word sequence) $Y = \{y_1, y_2, ..., y_W\}$ to express the video's content.

3.2 Encoder

Considering the redundancy of video caused by the frame similarity, existing video-and-language learning model commonly chooses multiple frames as model inputs instead of observing all. This multi-frame training strategy has been the norm and is shown to work well [Lei *et al.*, 2022]. We follow this procedure and sample the candidate frames time equally. The visual feature of each frame is extracted by a fine-tuned ViT model [Dosovitskiy *et al.*, 2020]. The outputs of encoder are defined as $f_i, i \in [1, ..., N]$, where N denotes the length of input frames.

We adapt pre-trained CLIP [Radford *et al.*, 2021] with 12 layers' ViT-B/32 as our visual encoder. Although CLIP has been pre-trained on 400M image-text pairs, we tend to narrow the gap between videos and images by performing video-text retrieval task to fine-tune CLIP's parameters. Many previous works, such as [Chen and Jiang, 2019; Zhang *et al.*, 2021; Yan *et al.*, 2022], take off-line CNNs pre-trained on other datasets to extract temporal representation or object features. Since the differences in data distribution, these operations might suffer a disconnection between the target task and the pre-trained domain [Gao *et al.*, 2022]. Video-text retrieval is a coarse-grained multi-modal task compared with video captioning. Thus we use it to fine-tune the pre-trained CLIP(ViT-B/32) on the mainstream captioning datasets and seek to weaken the disconnection while not hurting the visual representation ability of the model.

For the fine-tuning, ViT first reshapes images into flattened patches. Then the 2D patches would be further flattened and mapped to 1D vectors through trainable linear projection for adjusting standard Transformer [Vaswani *et al.*, 2017]. The specific prepend token `[CLS]` interacting with each input patch is regarded as the image representation. Following [Luo *et al.*, 2021], we average the generated features along the temporal dimension and get the video representation $\hat{f}$ by average pooling. We directly apply CLIP's text encoder to output caption embeddings $c_j, j \in [1, ..., M]$ corresponding to the given video. The visual-language similarity function $s(\hat{f}, c)$ can be defined as

$$s(\hat{f}, c) = \frac{c^{tr} \hat{f}}{||c||\ ||\hat{f}||}. \quad (1)$$

where tr denotes vector transposition .

3.3 Summarizer

For the traditional video summarization task, the evaluation metrics are the precision, recall and F-score. These metrics all require supervised frame-level annotations as ground truth, which limits to summarize the video without manually labels.

In this work, we evaluate the video summarization result from a high-level semantic aspect. The language description generated from the image or video is the abstractive summary of the visual content. If the captions generated from the video and keyframes are consistent, then they convey the same semantic information for human cognitive understanding.

To learn the summarizer, we leverage the local self-attention module to capture semantic relation among all the frames as well as outputting predicted scores for these frames following [Xiao *et al.*, 2020]. Given the encoder output $F = \{f_1, f_2, ..., f_N\}$, we compute the relation score map as:

$$r(f_i, f_j) = \mathbf{P} \tanh(\mathbf{W_1} f_i + \mathbf{W_2} f_j + b), \tag{2}$$

where $\mathbf{P}, \mathbf{W_1}, \mathbf{W_2} \in \mathbb{R}^{d_f \times d_c}$ are parameter matrices and b is bias vector, d_f and d_c are the dimension of input features and outputs. Note that the score map with shape $N \times N \times d_f$ means the features not only interact with each other along the temporal perspective but also their inner dimension. Then we can get the local attentive visual feature f_i^{att} by

$$a_{ij} = \frac{\exp(r(f_i, f_j))}{\sum_{k=0}^{N} \exp(r(f_i, f_k))}, \tag{3}$$

$$f_i^{att} = \sum_{j=0}^{N} a_{ij} \odot f_i, \tag{4}$$

where $\odot$ denotes element-wise multiplication.

Finally, we get the predicted score $scr_i^p, i \in [1, ..., N]$ of each frame using a trainable MLP after processing the attentive visual feature employing a residual connection and GeLU activation

$$scr_i^p = MLP(f_i^{att} + GeLU(f_i^{att})), \tag{5}$$

We choose the Top K frames according to the score ranking as the video summarization. In the practices, $K = 4$ achieves the best performance, which only accounts for $2 \sim 3\%$ for the original video ($T > 200$).

To train the summarizer, we generate a pseudo score label of each frame to facilitate video summarization. Concretely, we compute cosine similarity score for each sampled frame in the video with all the video-related texts, then average them as the current frame's visual correlation. Here we choose the `clipscore` [Hessel *et al.*, 2021], a reference-free metric reaching the highest correlation with human judgements, to assess the image-caption compatibility. This metric is outperforming existing reference-based metrics like CIDEr and SPICE. We slightly modify the definition of `clipscore` and formulate it as:

$$scr_i^c = \frac{1}{M} \sum_{j=1}^{M} s(f_i, c_j), \tag{6}$$

where M denotes the number of references and m_j is the j_{th} references. This pseudo score scr_i^c is regarded as the ground truth label when training the summarizer based on the binary cross-entropy loss function

$$\mathcal{L} = \frac{1}{N} \sum_{i}^{N} scr_i^c \log(scr_i^p) + (1 - scr_i^c) \log(1 - scr_i^p), \tag{7}$$

3.4 Decoder

The summarizer selects the Top K frames as the video summarization and the features are $F_{key} = \{f_1, f_2, ..., f_K\}, f_i \in \mathbb{R}^{d_c}$. As a feature sequence, we concatenate the key frame features by two strategies, i.e., score-oriented and time-oriented. We use two ranking mechanisms to decide the elements' position depending on their temporal position or predicted scores. The vector $V_s = [f_1, f_2, ..., f_K], V_s \in \mathbb{R}^{1 \times (K \times d_c)}$ is the final representation of the input video .

The captioning decoder is a light weight LSTM which produces a hidden state h_i and a cell state $cell_i$ at the i_{th} step,

$$h_i, c_i = LSTM([h_{i-1}; \Phi(y_{i-1}, \hat{y}_{i-1}); cell_{i-1}]), \tag{8}$$

where $h_{i-1}, y_{i-1}, \hat{y}_{i-1}$ and $cell_{i-1}$ are the previous hidden state, the predicted word, the ground truth and the cell state respectively. $[.;.]$ denotes the concatenation. We introduce scheduled sampling method [Bengio *et al.*, 2015] to solve the inconsistent distribution of input representation during training. Concretely, $\Phi(\cdot)$ can randomly choose y_{i-1} or $\hat{y}_{i-1}$ as the i_{th} input token. As the training epoch increasing, $\Phi(\cdot)$ tend to choose y_{i-1} and the initial input of LSTM is the reshaped representation V_s when i=0.

Our objective function for the decoder with trainable parameters θ is formulated as:

$$\mathcal{L}_{XE}(\theta) = -\sum_{t=1}^{W} \log p_\theta(\hat{y}_t | \Phi(\cdot)_{1:t-1}), \tag{9}$$

where $\Phi(\cdot)_{1:t-1}$ denotes the above scheduled sampling sequences.

We also adapt discriminative cross-entropy(DXE) [Yan *et al.*, 2022] as the learning objective. Each video is attached with M captions $\hat{Y} = \{Y_1, Y_2, ..., Y_M\}$, the qualities of the captions is not equivalent which can be evaluated by metric scores $m(\hat{Y})$ pre-computed using BLEU@4 or CIDEr, $m(\hat{Y})$ serves as the discriminative weight in cross-entropy loss to promote the model to more concentrate on high-quality captions. The DXE loss function is formulated as:

$$\mathcal{L}_{DXE}(\theta) = -\frac{1}{M} \sum_{j=1}^{M} m(Y_j) \log p(Y_j | V_s; \theta), \tag{10}$$

3.5 Training

The training procedure is split into two stages. In the first stage, we train the summarizer module with our automatic content-aware scores to choose a subset of frames containing more visual diversity and less noise. In the second stage, we freeze the summarizer's parameters while update the captioning decoder. The two-stage training strategy makes the both module more stable. We consider not every frame is possible for a video especially when captioning, our summarizer module pre-excludes most of video's inherent redundancy and the following captioning would suffer from less interference.

4 Experiments

4.1 Implement Details

Dataset. We evaluate our model on MSR-VTT [Xu *et al.*, 2016] and MSVD [Chen and Dolan, 2011] datasets.

Training	Method	Feature	MSR-VTT				MSVD			
			B@4	M	R	C	B@4	M	R	C
XE	PMI-CAP [Chen *et al.*, 2020]	IRV2+C3D	44.0	29.6	-	50.7	54.7	36.4	-	95.2
	SAAT [Zheng *et al.*, 2020]	IRV2+C3D+Ca	40.5	27.9	61.2	51.0	46.5	33.5	69.4	81.0
	STGraph [Pan *et al.*, 2020]	RN+I3D+F	40.5	28.3	60.9	47.1	52.2	36.9	73.9	93.0
	SGN [Ryu *et al.*, 2021]	RN+3D-RN	40.8	28.3	60.8	49.5	52.8	35.5	72.9	94.3
	O2NA [Liu *et al.*, 2021]	RN+3D-RX	41.6	28.5	62.4	51.1	55.4	37.4	74.5	96.4
	RCG [Zhang *et al.*, 2021]	IRV2+C3D	42.8	29.3	61.7	**52.9**	-	-	-	-
	ORG-TRL [Zhang *et al.*, 2020]	IRV2+C3D+F	43.6	28.8	62.1	50.9	54.3	36.4	73.9	95.2
	GL-RG [Yan *et al.*, 2022]	RN+3D-RN+RX	**45.5**	**30.1**	**62.9**	51.2	**55.5**	**37.8**	**74.7**	**94.3**
	Ours	ViT	**45.5**	**30.5**	**63.6**	**55.0**	**64.2**	**41.4**	**79.1**	**118.7**
RL	PickNet [Chen *et al.*, 2018]	RN	38.9	27.2	59.5	42.1	46.1	33.1	69.2	76.0
	SAAT [Zheng *et al.*, 2020]	IRV2+C3D+Ca	39.9	27.7	61.2	51.0	46.5	33.5	69.4	81.0
	POS [Wang *et al.*, 2019]	IRV2+Motion I3D	41.3	28.7	62.1	53.4	53.9	34.9	72.1	91.0
	$\mathcal{D}^2$ [Gao *et al.*, 2022]	ViT	44.5	30.0	63.3	**56.3**	56.9	38.4	**75.1**	**99.2**
DXE	GL-RG [Yan *et al.*, 2022]	RN+3D-RN+RX	**46.9**	**30.4**	**63.9**	55.0	**57.7**	**38.6**	74.9	95.9
	Ours	ViT	**45.9**	**30.5**	**64.2**	**57.8**	**60.1**	**40.7**	**77.4**	**109.6**

Table 1: Performance Comparisons with state-of-the-art methods on the testing set of the MSR-VTT and MSVD datasets in terms of BLEU@4, METHOR, ROUGE-L and CIDEr scores. The **best** and the **second-best** methods are highlighted. In the first column, "XE" is cross-entropy; "DXE" is discriminative cross-entropy which is compared with "RL"(reinforcement learning). "IRV2", "Ca", "F", "RN", "RX" denote Inception ResNet-v2, Category features, Faster RCNN, ResNet and ResNeXt repectively.

- **MSR-VTT** is a large-scale open domain dataset. It contains 10K videos crossing a wide range categories including music, game, sports and movie. Each video is annotated with 20 references. The duration of each video in MSR-VTT is between 10 and 30 seconds. We split the data into a 6,513 training set, 497 validation set and 2,990 testing set.
- **MSVD** has 1,970 Youtube videos. This dataset mainly contains short video clips with a single action, and the average duration is about 9 seconds. We follow the data split of 1,200 videos for training, 100 videos for validation and the rest for testing. The number of references of each video in MSVD dataset is not fixed and we set the the number to 17 following [Yan *et al.*, 2022].

Evaluation Metrics. We use four universal metrics for evaluation: BLEU@4, ROUGE-L, METHOR and CIDEr [Vedantam *et al.*, 2015], which are denoted as B@4, M, R, and C respectively. We mainly compare CIDEr as the previous video captioning works.

Training setup. Our encoder is adapted ViT model fine-tuned on video-text retrieval task. Our summarizer module is trained with 10 epochs on the above datasets' with learning rate 1e-3 and dropout 0.2. Our captioning module is trained with learning rate 1e-4 and 40 epochs, and we set the batch size to 32. Both the summarizer and captioning decoder employ Adam optimizer [Kingma and Ba, 2014] to minimize the loss. The candidate frames number is set to 12 with time-equally sampling and we set the maximum sequence length to 30 following [Luo *et al.*, 2021] and [Yan *et al.*, 2022] respectively. The dimension of visual embeddings and text embeddings is 512.

4.2 Comparison to State-of-the-art

The performance of our proposed framework and other top-performing baselines are presented in Table 1. In the practice, we only summarize four keyframes to present each video. We compare the XE training result with other XE-based methods. For the DXE training results, we compare them with DXE-based method and reinforcement learning methods.

As it can be observed, our model achieves best performances on all metrics over two benchmarks under the XE training. The CIDEr score of our model reaches 55.0, which achieves increments of 4.0% and 6.9% to strong models RCG [Zhang *et al.*, 2021] and GL-RG [Yan *et al.*, 2022]. Moreover, our model achieves improvements of 1.1% and 1.3% on R and M respectively while bringing into correspondence on B@4 with previous best results. Under DXE training, our margins over [Yan *et al.*, 2022] are 0.3% on M, 0.5% on R and 5.1% on C. It's worth mentioning that we only use ViT feature, while other methods employ various features from visual to temporal. During all the training stages, we do **not** employ any temporal information such as 3D temporal visual feature. This reveals that, for the video understanding, there is quite a redundancy and the visual appearance is much more essential than the temporal information.

Compared to the MSR-VTT, our model achieves a more superior performance on the MSVD over all metrics. Under the XE training, our advancements over the second best results are 15.7% on B@4, 9.5% on M, 5.9% on R, and especially 23.1% on C. DXE training decreases the model's performance on MSVD unexpectedly although still surpasses the other methods in a large margin, e.g. 10.4% improvements on C over [Gao *et al.*, 2022].

Number	MSR-VTT		MSVD	
	B@4	C	B@4	C
1	40.9	49.5	58.0	97.8
2	43.9	52.0	60.0	109.6
4	**45.5**	55.0	**64.2**	**118.7**
8	45.4	**55.4**	61.8	109.0
12	44.5	53.6	60.1	104.4

Table 2: Comparison of the different K frames on MSR-VTT and MSVD.

Number	B@4	M	R	C
1→4	58.2	37.9	75.1	101.8
2→4	61.7	39.9	77.5	111.0
4	**64.2**	**41.4**	**79.1**	**118.7**
1→8	58.1	37.9	75.4	103.9
2→8	60.1	40.1	76.7	107.9
4→8	61.5	40.1	77.2	**111.3**
8	**61.8**	**40.7**	**77.6**	109.0

Table 3: Evaluation the effects of feature dimension duplicaiton on MSVD dataset. → denotes the expansion process.

U	F	S	T	B@4	M	R	C
×	✓	✓	×	**45.5**	**30.5**	**63.6**	**55.0**
×	✓	×	✓	44.5	30.3	62.9	53.9
✓	×	✓	×	44.5	30.0	62.8	53.3
✓	×	×	✓	43.9	29.9	62.3	52.7
×	✓	✓	×	**64.2**	**41.4**	**79.1**	**118.7**
×	✓	×	✓	62.2	40.6	77.9	111.5
✓	×	✓	×	59.4	38.5	75.5	100.6
✓	×	×	✓	59.6	38.6	76.1	104.0

Table 4: Performance comparison of different sources and rank order of the selected frames on MSR-VTT(the upper) and MSVD(the lower). "U", "F", "S", "T" denote unfinetuned ViT features, finetuned ViT features, scoring order and temporal order respectively. The number of selected frames here is 4.

4.3 Ablation Study

We conduct several ablation studies to quantify the influences of different configuration of our model.

Summarized frames. We measure different K numbers of summarized frames. We list the results of K from single to 12 in Table 2. It is shown that, $K = 4$ achieves the best results. We notice that the model trained with only two frames can exceed many previous state-of-the-art results. Even the single frame selected by our model produces considerable captioning performance. The metric scores maintain growth as the input frames increasing on MSR-VTT dataset, but the rating is gradually declined, using more frames(e.g. 8) can not bring corresponding improvements(55.4 CIDEr vs. 55.0 CIDEr) or even produce negative effect(45.4 B@4 vs. 45.5 B@4) when compared with 4 frames input. As for MSVD dataset, 4 frames is the optimal input, which produce the most obvious advancement on B@4 and C. One possible reason is that the noise frames existed in videos affect the training of the model. We set the number of input frames to 4 as the final configuration. In case the improvement is benefited by the feature dimension extension from the frame increasing, we verify the performance by duplicating the features into the same dimension. The range of dimension is $\{512, 1024, 2048, 4096\}$ corresponding to $\{1, 2, 4, 8\}$ frames. The results is in Table 3. 'X→Y' means X frames are expanded to Y by simple temporal duplication to obtain identical dimension and not provide additional information. It can be observed '1→Y' could improve the performance when compared with single frame input while still dropped behind selected Y (line 1vs line 3 and line 5 vs. line 7). It's apparent that the latter contains more visual information and the content in single frame is quite weak. We also confirm the above conclusion that 4 frames input is ultimate (line 2 vs. line 5 and line 3 vs. line 6).

Video representation. We elaborate the influence of ranking order and fineturned ViT in Table 4. As we can see(line 1 vs. line 2), score-oriented video representation, which completely overlooks temporal information, is better than the time-oriented, increasing B@4 by 2.2% and C by 2%. This finding reaches an agreement with "static appearance bias" [Lei *et al.*, 2022] existing in the popular video captioning datasets. Finetuned ViT (line 1 vs. line 3 and line 6 vs. line 8) helps reduce the gap between images and videos and achieves improvement of 3.2% and 7.2% on C respectively. Notice that the finetuning process has more obvious effect on MSVD dataset, we assume the reason is that the video in MSVD is attached with average 36 captions while 20 captions in MSR-VTT during finetuning. The larger scale video-caption pairs lead to the better visual representation ability for ViT.

Sampling method. We investigate the influence of using different frame sampling method in Table 5. We introduce Clip [Radford *et al.*, 2021] embedding relation score to indicate correlation among frames, we calculate the cosine similarity of extracted embeddings then sum them along the temporal dimension, that is to say, the value report the degree of correlation between the frame and the others. Intuitively, people prefer these frames appeared more frequently in the video when labeling the importance of them, here we consider the frames with high relation score coincide this conclusion. We choose the input subset based on maximal or mini-

Method	B@4	M	R	C
Random	43.4	30.0	62.0	52.4
Uniform	43.4	30.1	62.2	51.7
Clip_Max	44.0	29.8	62.6	51.9
Clip_Min	40.6	28.4	60.1	45.9
Ours	**45.5**	**30.5**	**63.6**	**55.0**

Table 5: Comparison of the influence of different sampling methods on MSR-VTT dataset. The selected input frames here is set to 4. **Ours** is trained on XE.

Figure 3: Qualitative examples on the MSR-VTT testing set. Compared to the previous method GL-RG[Yan *et al.*, 2022], our model can generate more accurate and more diverse captions. **TO** and **SO** are time-oriented and score-oriented of the video representation.

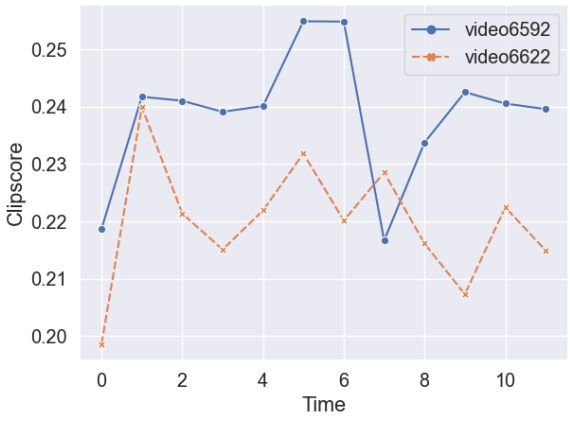

Figure 4: Example results of frame-level score on MSR-VTT dataset.

mal **K**(e.g.4) values which are denoted as Clip-Max and Clip-Min respectively. Randomly and uniformly sampling produce almost the same results, Clip-Max get higher scores on B@4 and R. The fourth line(Clip-Min) shows that the frames appeared rarer can not cover the main content compared with Clip-Max. The last line shows the importance of frame is not only determined by their frequency and our model using content aware labeling in consideration of visual and lingual perspective can find approximately optimal subset from a series of frames.

4.4 Qualitative Results

Figure 3 shows the qualitative examples of our method. As indicated by the examples, with only 4 selected frames input, our method can generate more accurate captions like the lower sub-figure, while GL-RG[Yan *et al.*, 2022] produces wrong description like 'a woman'. Figure 4 shows the

(a) Video 6592 (b) Video 6622

Figure 5: (a) and (b) show the video contents and the videos come from Figure 4, the blue and orange boxes indicate picked frames. Frames are organized from left to right, then top to bottom in temporal order.

`clipscore` labeling scores of two random sampled videos on MSR-VTT dataset. And the frames of (a) and (b) in Figure 5 are from the TOP 4 of labeling scores in Figure 4. We notice higher scores normally indicate the frames are dominant in the whole video which report the main content.

5 Conclusion

In this paper, we cooperate the video summarization and video captioning task with each other to investigate the video frame representation problem. We propose a dual video summarization framework composed of an encoder, a summarizer and a decoder. By verifying it in the context of video captioning, for the generative task as video captioning, a small number of keyframes can convey the same semantic information. This reveals the redundancy and frame bias of video captioning benchmarks.

Acknowledgements

This work was supported by the NSFC NO. 62172138 and 61932009. This work was also partially supported by The University Synergy Innovation Program of Anhui Province NO. GXXT-2021-007.

References

[Anne Hendricks *et al.*, 2017] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE ICCV*, pages 5803–5812, 2017.

[Bain *et al.*, 2021] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF ICCV*, pages 1728–1738, 2021.

[Baraldi *et al.*, 2017] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the IEEE Conference on CVPR*, pages 1657–1666, 2017.

[Bengio *et al.*, 2015] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

[Borji and Itti, 2012] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on PAMI*, 35(1):185–207, 2012.

[Buch *et al.*, 2022] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 2917–2927, 2022.

[Chen and Dolan, 2011] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[Chen and Jiang, 2019] Shaoxiang Chen and Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8191–8198, 2019.

[Chen *et al.*, 2018] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In *Proceedings of the ECCV*, pages 358–373, 2018.

[Chen *et al.*, 2020] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *European Conference on Computer Vision*, pages 333–351, 2020.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Escorcia *et al.*, 2019] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019.

[Gabeur *et al.*, 2020] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229, 2020.

[Gao *et al.*, 2022] Yiqi Gao, Xinglin Hou, Wei Suo, Mengyang Sun, Tiezheng Ge, Yuning Jiang, and Peng Wang. Dual-level decoupled transformer for video captioning. *arXiv preprint arXiv:2205.03039*, 2022.

[Gygli *et al.*, 2013] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *Proceedings of the IEEE ICCV*, pages 1633–1640, 2013.

[Gygli *et al.*, 2014] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520, 2014.

[Hessel *et al.*, 2021] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lei *et al.*, 2021] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 7331–7341, 2021.

[Lei *et al.*, 2022] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022.

[Lin *et al.*, 2021] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 7005–7015, 2021.

[Liu *et al.*, 2021] Fenglin Liu, Xuancheng Ren, Xian Wu, Bang Yang, Shen Ge, and Xu Sun. O2na: An object-oriented non-autoregressive approach for controllable video captioning. *arXiv preprint arXiv:2108.02359*, 2021.

[Luo *et al.*, 2021] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

[Narasimhan *et al.*, 2021] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34:13988–14000, 2021.

[Pan *et al.*, 2016] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on CVPR*, pages 1029–1038, 2016.

[Pan *et al.*, 2020] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020.

[Pei *et al.*, 2019] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 8347–8356, 2019.

[Qiu *et al.*, 2022] Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. Mhms: Multimodal hierarchical multimedia summarization. *arXiv preprint arXiv:2204.03734*, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[Ryu *et al.*, 2021] Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D Yoo. Semantic grouping network for video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2514–2522, 2021.

[Sanabria *et al.*, 2018] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding. In *NeurIPS*, 2018.

[Shang *et al.*, 2021] Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. Multimodal video summarization via time-aware transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1756–1765, 2021.

[Song *et al.*, 2015] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on CVPR*, pages 5179–5187, 2015.

[Song *et al.*, 2017] Jingkuan Song, Zhao Guo, Lianli Gao, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. Hierarchical lstm with adjusted temporal attention for video captioning. *arXiv preprint arXiv:1706.01231*, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on CVPR*, pages 4566–4575, 2015.

[Venugopalan *et al.*, 2015] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE ICCV*, pages 4534–4542, 2015.

[Wang *et al.*, 2019] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable video captioning with pos sequence guidance based on gated fusion network. In *Proceedings of the IEEE/CVF ICCV*, pages 2641–2650, 2019.

[Xiao *et al.*, 2020] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. Convolutional hierarchical attention network for query-focused video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12426–12433, 2020.

[Xu *et al.*, 2016] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on CVPR*, pages 5288–5296, 2016.

[Yan *et al.*, 2022] Liqi Yan, Qifan Wang, Yiming Cui, Fuli Feng, Xiaojun Quan, Xiangyu Zhang, and Dongfang Liu. Gl-rg: Global-local representation granularity for video captioning. *arXiv preprint arXiv:2205.10706*, 2022.

[Yang *et al.*, 2021] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF ICCV*, pages 1686–1697, 2021.

[Zhang and Peng, 2020] Junchao Zhang and Yuxin Peng. Video captioning with object-aware spatio-temporal correlation and aggregation. *IEEE Transactions on Image Processing*, 29:6209–6222, 2020.

[Zhang *et al.*, 2020] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 13278–13288, 2020.

[Zhang *et al.*, 2021] Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 9837–9846, 2021.

[Zhao and Xing, 2014] Bin Zhao and Eric P Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE Conference on CVPR*, pages 2513–2520, 2014.

[Zheng *et al.*, 2020] Qi Zheng, Chaoyue Wang, and Dacheng Tao. Syntax-aware action targeting for video captioning. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 13096–13105, 2020.

[Zhou *et al.*, 2018] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.