Exploring Group Video Captioning with Efficient Relational Approximation

Wang Lin*, Tao Jin*, Ye Wang*, Wenwen Pan, Linjun Li, Xize Cheng, and Zhou Zhao†
Zhejiang University, Hangzhou, China
{linwanglw,jint_zju,yew,wenwenpan,lilinjun21,chengxize,zhaozhou}@zju.edu.cn

Abstract

Current video captioning efforts most focus on describing a single video while the need for captioning videos in groups has increased considerably. In this study, we propose a new task, group video captioning, which aims to infer the desired content among a group of target videos and describe it with another group of related reference videos. This task requires the model to effectively summarize the target videos and accurately describe the distinguishing content compared to the reference videos, and it becomes more difficult as the video length increases. To solve this problem, 1) First, we propose an efficient relational approximation (ERA) to identify the shared content among videos while the complexity is linearly related to the number of videos. 2) Then, we introduce a contextual feature refinery with intra-group self-supervision to capture the contextual information and further refine the common properties. 3) In addition, we construct two group video captioning datasets derived from the YouCook2 and the ActivityNet Captions. The experimental results demonstrate the effectiveness of our method on this new task.

1. Introduction

Video captioning aimed to understand the scene and describe it in words has recently attracted extensive research attention. Currently, mainstream video captioning works mostly focus on describing individual videos [27, 8, 17, 34]. However, since the amount of online videos has been growing at an exponential rate, the need for captioning the video groups has increased considerably like titling a categorized video folder and query suggestions for text-based video retrieval. Although [16] has studied group image captioning which boosts many real-world applications, there is no existing work in the literature that addresses the task of group-based video captioning.

Thus, we are inspired by the group image captioning[16] and propose the novel problem of **group video caption**: given a group of target videos and a group of reference videos, to generate a description that simultaneously identifies both important generalities arising in target videos, as well as, particularities captured from reference videos. A promising application scenario is shown in Figure 1, the search engine returns a group of topically close videos with the query of *the onions* and the user indicates his/her interest in some of the videos (*i.e.* the target group shown in green area). With the remaining videos (*i.e.* the reference group shown in yellow area), we can infer the user's hidden preferences among multiple events in the target videos and suggest a refined search query *chop the onions* accordingly.

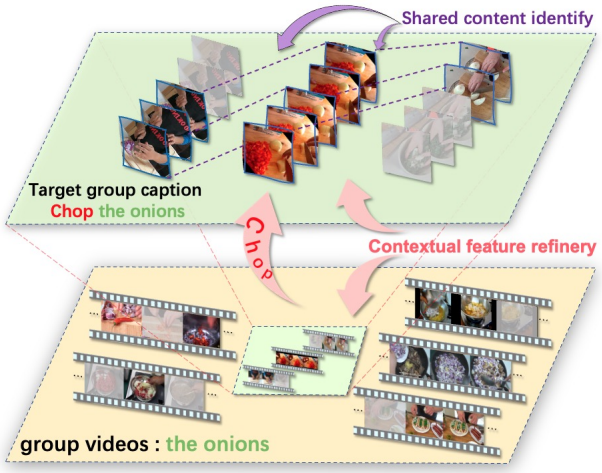

Figure 1. An example of context-aware group video captioning. We aim to generate a description *chop the onion* that best describes the target group (shown in the green area) with the contextual information from the reference group (shown in the yellow area).

Compared to the conventional setting of single-based video captioning, the challenges of our group-based video captioning are two-fold: 1) identifying which temporal features correspond to the shared content for videos in the group, and 2) distinguishing the shared content of target videos from all videos in the reference group, *i.e.* group-level distinctiveness.

For identifying the shared content, the method of group image captioning is not suitable for video since the dispar-

* Equal contribution.
† Corresponding author

ity of solution space size. We argue that the fundamental issue of group-based captioning among group videos lies in modeling the long sequence relevance from the cross-video perspective in an efficient manner. Following this premise, we first investigate the traversal method whose computation complexity is $O(m^n)$ for n videos with m frames. Then, we further introduce an Efficient Relational Approximation (ERA) to summarize the shared content in video groups. In particular, we find a new random feature mapping that can be equivalent to the softmax-kernel method and the complexity scales linearly $O(n)$ with the number of videos.

To achieve group-level distinctiveness, we propose the contextual feature refinery which can learn to capture the salient feature difference between target and reference videos precisely. Specifically, the contextual feature refinery enables cross-group interactions between target and reference video groups through a multi-layer co-attention. To avoid the model mainly extracting information from certain unrelated/noisy content like background, we further introduce Intra-Group Contrastive (IGC) learning into the refinery. The key idea of IGC is leveraging the intra-group self-supervision to learn desirable representations that keep alignment between semantically-related the contextual feature and the shared target group feature.

As the first step in this type of problem, we constructed two new datasets for our task by using the existing dense video captioning dataset YouCook2[37] and ActivityNet Captions[15]. The reason is that dense video captioning dataset annotations map sentences describing events to the segments in the videos which is easier to be grouped. Specially, we parse the single-segment caption and use the shared verb phrases as the groups' ground-truth captions.

Our main contributions can be summarized as:

- We propose a novel task of group video captioning that can boost many real-world applications like video retrieval and classification.
- We introduce a new model for group video captioning with an efficient relational approximation that summarizes relevant shared information in the groups. Also, our model proposes a contextual feature refinery to capture discriminative information.
- We constructed two new video datasets specifically for the group captioning problem. Experiments on the two datasets demonstrate that our model outperforms various baselines on the group video captioning task.

2. Related work

2.1. Video Captioning

Deep neural networks have achieved remarkable success in various fields of computer vision[9, 12, 31, 30, 32]. Classical approaches usually extract visual representations with an encoder, then feed them to a language decoder and output sequences of words [27, 10, 11, 3]. Recent advances mainly focus on improving visual representation. Existing solutions can be coarsely divided into two categories: Object-based and Frame-based methods. Object-based models mainly exploit the spatio-temporal object interaction.[35] aggregated salient objects according to their spatial coordinate to capture the dynamic information in the temporal domain. [20, 36] constructed a spatio-temporal graph according to the position as well as representation of objects to enhance object-level representation. In contrast, Frame-based methods focus on the relationship between contents. [7] used boundary-aware pooling to select the features of different scenarios and reduce redundancy. Another related work is dense video captioning[15, 38] which studies the event-based visual representation. [29] focused on the holistic scene and event-level features to generate a more comprehensive description. However, our work focuses on the novel setting of group-based video captioning which aims to extract cross-video visual representations, not objects or frames.

2.2. Multi-input Captioning

There are several captioning settings that need multiple inputs in image captioning. The existing multi-input captioning setting has two main tasks: change captioning and distinctive captioning. Change captioning [25, 22, 33, 23] take before and after images as input and describe the changes between them(i.e., $1-1$). Thus, the two images in their settings always have strong correlations. Distinctive captioning initially uses an image as input and generates recognizable captions for each image [18, 19]. Recent work[28] proposes to study the task based on a target image and a group of semantic-similar reference images(i.e., $1-N$). There are also some $N-N$ works that use a group of images as references to investigate certain properties of the target images. [2] firstly model both relevance and diversity among image contents in group-based image captioning. [16] summarized common information in the target group while capturing discriminative information between them. However, the methods of multi-input image captioning cannot be simply applied to video since 1) the complexity of identifying the relationship among videos increases with the video length, and 2) there is a large amount of unrelated content makes it more difficult to focus the discriminative information.

3. Dataset

To train our model, we need a large-scale dataset where each data sample consists of a group of target videos with its shared content description and a larger group of reference videos. The reference videos need to be relevant to target videos while containing a larger variety of visual content and thus providing context for describing target videos. The

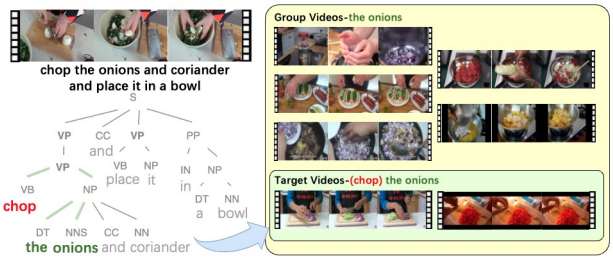

Figure 2. Dataset construction method. A constituency parse tree is used to extract the common VP. Then the videos with shared VP are grouped to form the target group, the videos with same verb or noun that partially match the targets form the reference group.

description should be simultaneously specific to the target group and conditioned on the reference group.

3.1. Dataset Construction

We construct two datasets with pairs of a video group and its shared content description on top of two existing dense video captioning datasets: ActivityNet Captions[15], which is the largest existing public dense video captioning dataset, and YouCook2[37], which contains videos that are visually different but semantically similar that suitable for our experiment.

First, for the target videos, what is the shared content? Different from images, events are important for video understanding and applications. So, we extracted verb phrases (VP) from sentences attached to segments in all videos by using a constituency parse tree[14]. More specifically, if a common VP appeared in different videos, we grouped them. As shown in Figure 2, videos with the verb phrases *chop the onions* are selected to form the target group. Note that there are multiple events in a video, which means that more than one common content can be identified in the video group. Without any reference videos, people won't know what should be emphasized in the target videos. In contrast, they will focus on the unique events of *chop*, and predict *chop the onions* when they use the videos in the yellow region as references.

Thus, after getting the shared verb phrases among target videos, the videos containing *the onions* are selected to form the reference group paired with the target group. In this way, the reference group contains a larger variety of contents (onion in any places or conditions) which distinct the desired shared content. Finally, the shared verb phrases serve as the ground truth group description. Details about the construction of the two datasets are provided as follows. For simplicity, in this paper, we call our newly constructed group captioning datasets by the same name as their parent datasets: ActivityNet Captions and YouCook2.

Datasets	Groups	Train	Val	Test	Video_len	Vocab
ActivityNet	4035	2421	403	1211	40.27	1466
YouCook2	1763	1058	176	529	16.09	633

Table 1. Statistics of ActivityNet Captions and YouCook2.

3.2. Statistical Information

ActivityNet Captions This dataset is based on different human activity divided into 200 classes and consists of about 100,000 sentences to describe all 20,000 videos, and on average each video has 3.65 events annotated. The high diversity of visual content and a large number of videos makes ActivityNet Captions a suitable choice.

After sampling from 20,000 clips from YouCook2, we obtain around 4,035 samples with 12,947 videos included. Each sample contains 3 target videos and 5 reference videos, where target and reference videos share the same video pool. The videos with rare verb phrases that cannot be made into groups are not used. We manually clean the sampled data to remove samples that are not meaningful. We also clean the vocabulary to remove rare words. The 4,035 samples are split into test, validation, and train splits, where these three splits share the same video pool.

YouCook2 While the ActivityNet Captions dataset excels in video diversity, we found that its captions are often long and sometimes noisy. Motivated by the query suggestion application where the suggested search queries are usually short and compact, we propose to construct the dataset on another dataset named YouCook2 Captions which contains fine-grained action annotation but is visually different. YouCook2 has 2,000 videos and the average number of segments per video is 7.70. Many of the captions in this dataset are more verb-like short events titles.

After grouping and filtering the 15,400 clips, we get 5613 videos, grouped into 1,736 data samples for the YouCook2 Captions dataset as shown in Table 1. The dataset sampling and split details are similar to ActivityNet Captions.

4. Method

Formally, given a group of n_{tar} target videos and a group of n_{ref} reference videos, we aim to generate a description $D = \langle w_1, ..., w_T \rangle$ to describe the target video group in the context of the reference group. In our setting, $n_{tar} = 3$, $n_{ref} = 5$. The input of our model is the extracted target video features set $V_{tar} = \left[\{v_1^m\}, ..., \{v_{n_{tar}}^m\}\right]$ and reference video features set $V_{ref} = \left[\{v_1^m\}, ..., \left\{v_{n_{ref}}^m\right\}\right]$, where m is number of frames.

Different from the conventional video captioning task, the generated description D should describe both the com-

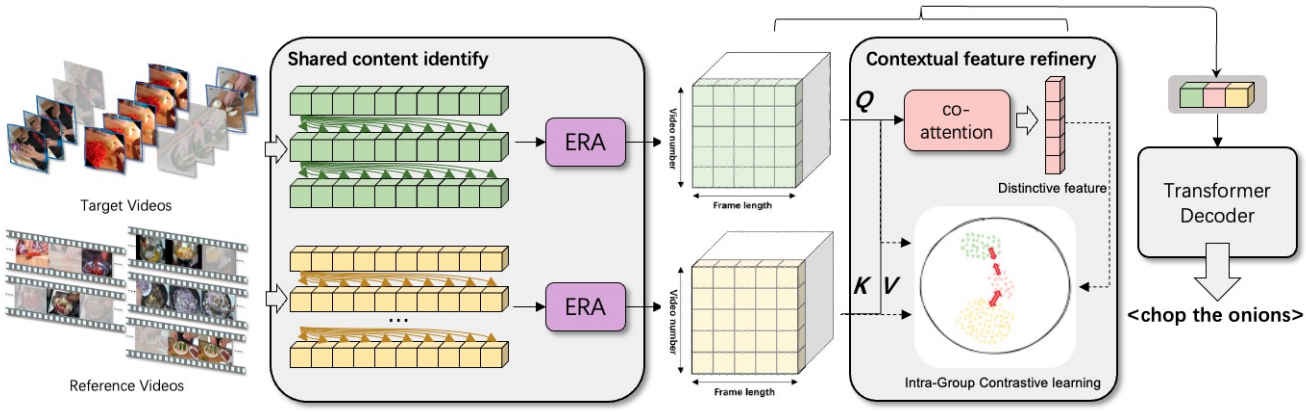

Figure 3. An overview of our framework. For the video group input, we apply the efficient relation approximation (ERA) to obtain the joint representations that summarize the shared content among the videos. Then, we leverage the intra-group supervision to distill the desired contextual feature. Finally, the two shared representations are concatenated with contextual features to compose the input to the Transformer decoder for description generation.

mon content in the target videos and also highlight the uniqueness of the target videos compared to other reference videos in the same group. To perform this task, we explore methods to address the two main challenges in our proposed problem: **a)** how to identify the shared content for all videos simultaneously by considering the relationships among videos and **b)** how to figure out the difference between two video groups.

In the subsequent subsections, we describe our method explorations path starting with an intuitive baseline. We then gradually introduce more computationally specialized modules. For each module, we describe our intuition and back them up with quantitative results and visual illustrations.

4.1. Baseline: video relation traversing

The intuitive approach would be to summarize the target and reference features by traversing. We refer to this method as **Traversal** which identifies the temporal frames corresponding to the common content in each video by traversing all videos.

Suppose the extracted feature lists of videos are $\left\{v_i \in \mathbb{R}^{d\times m} | i \in [1,n]\right\}$ where n is the number of videos and m is the frame length of v_i, we could obtain m^n groups of joint representations. To combine these information, we introduce the attention map $\mathrm{A} \in \mathbb{R}^{m^n}$ as follows:

$$\mathrm{A}_{a_1,a_2,\ldots,a_n} = \sum_{i=1}^{n}\sum_{j<i}^{n}\left(v_i^{a_i} v_j^{a_j}\right) \tag{1}$$

where $v_i^{a_i} \in \mathbb{R}^d$ denotes the a_i-th frame of video v_i and $\mathrm{A}_{a_1,a_2,\ldots,a_n}$ denotes an element in tensor A with indices $[a_1, a_2, ..., a_n]$. We consider the score $\mathrm{A}_{a_1,a_2,\ldots,a_n}$ indicates the similarity of frame $(v_1^{a_1}, v_2^{a_2}, ..., v_n^{a_n})$.

After we traversed all video frames and obtain m^n groups of joint representations A. The softmax function is applied element-wisely along all dimensions. And, we consider the score w_i^j which indicates whether the content of j-th frame in video v_i is common to all videos.

$$\begin{aligned} \mathbb{A} &:= \text{softmax}\left(\mathrm{A}\right), \\ w_i^j &= \sum_{q=1,\neq i}^{n}\sum_{t=1}^{m}\mathbb{A}_{v_i^j, v_q^t} \end{aligned} \tag{2}$$

Then, we compute the target common feature ψ_{tar} and the reference common feature ψ_{ref} as follows:

$$\psi_{tar} = \sum_{i=1}^{n_{tar}}\sum_{j=1}^{m} w_i^j v_i^j \qquad \psi_{ref} = \sum_{i=1}^{n_{ref}}\sum_{j=1}^{m} w_i^j v_i^j \tag{3}$$

Finally, we follow the standard captioning pipeline and take the concatenation of the two group features as the decoder input to predict the descriptions. We use Transformer as the sentence decoder. At time step t, we have the equations for decoding:

$$\begin{aligned} x &= [\psi_{tar}, \psi_{ref}] \\ z_t &= \text{Transformer}(x, \hat{w}_{t-1}) \\ \hat{w}_t &\sim \text{Softmax}(z_t) \end{aligned} \tag{4}$$

The entire system is trained by minimizing the negative log-likelihood as follows.

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} log\left(\hat{w}_t = w_t | w_{1:T-1}\right), \tag{5}$$

This decoding architecture is used in all of our subsequent model variants.

4.2. Shared feature identification with efficient relation approximation

The traversal is the intuitive and theoretically optimal method, it calculates all possible combinations of all frames in a video group. Obviously, the computation complexity is the limitation of the traversal method. For a video group containing n videos, and each video extracts m frame features, the computation complexity is $O\left(m^n\right)$. In real video applications, the complexity of the traversal method is unacceptable as shown in Figure 6.

We find that the complexity of the traversal method lies in the computation of attention map A. However, such a high-order tensor A cannot be decomposed directly due to the non-linearity caused by the softmax function. Thus, it is necessary to find a substitute approximate solution.

The softmax-kernel function is given as : $\sigma\left(x, y\right) = \exp\left(x^T y\right)$. In the traversal method, we obtain the cross-video joint representations as follows:

$$\sigma\left(v_i | i \in [1, n]\right) = \exp\left(\sum_{i=1}^{n}\sum_{j<i}^{n}\left(v_i^T v_j\right)\right) \tag{6}$$

Inspired by the Performer [4], we propose the Efficient Relational Approximation (ERA) mechanism that can reduce the computation complexity to linear. σ could be rewritten as follows:

$$\sigma\left(v_i^j\right) = \mathbb{E}_{\omega\sim\mathcal{N}(0,\mathbf{I}_{\mathbf{K}'})} exp\left(\prod_{i=1}^{n}\omega^T v_i^j - \frac{\left\|v_i^j\right\|^2}{2}\right) \tag{7}$$

where $\mathbb{E}$ and $\mathcal{N}\left(0, \mathbf{I}_{\mathbf{K}'}\right)$ denote expectation and sampling distribution. With Eq 7, we can projecting the features of i-th video and obtain $\vartheta_i \in \mathbb{R}^{m\times d'}$. The detailed derivations and theoretical errors can be found in Appendix **??**.

Note that the softmax function consists of exponential operation and sum normalization, we need the value of the normalized denominator. And the score w_i can be calculate as follows:

$$w_i = \frac{\vartheta_i(\prod_{t=1,\neq i}^{n}\mathbf{1}\cdot\vartheta_t)}{\sum(\prod_{t=1}^{n}\mathbf{1}\cdot\vartheta_t)} \tag{8}$$

where $\mathbf{1} \in \mathbb{R}^m$, $\sum$ denotes the sum of all elements in the tensor.

In this way, the complexity scales linearly $O\left(n\right)$ with the number of videos. In the subsequent analysis, we show that the proposed efficient relational approximation accelerates the process of identifying shared features while maintaining the desired level of accuracy.

4.3. Contextual feature refinery with intra-group self-supervision

We propose the contextual feature refinery with contrastive learning to obtain group-level distinctiveness from the reference group. We first set the target shared feature as a query and the reference shared feature as the key and value in the co-attention mechanism.

$$\phi \quad = \text{MHA}\left(\psi_{tar}, \psi_{ref}, \psi_{ref}\right) \tag{9}$$

where MHA denotes the multi-head attention. By the residual connection in self-attention blocks, feature ψ_{tar} gradually distills useful information from the reference videos.

Then the contextual feature ϕ is added to target common feature ψ_{tar} to generate the comprehensive feature ψ'_{tar}.

$$\psi'_{tar} \quad = \psi_{tar} + \phi \tag{10}$$

The comprehensive target feature ψ'_{tar} and reference common feature ψ_{ref} are concatenated and fed into Transformer to generate captions.

However, co-attention ignores the self-supervision within each group, thus failing to guarantee the desirable precision of learned features. The reason is that a) text usually captures most of the salient events in the paired visual and overlooks background features; and b) videos are inherently noisy, which makes the problem in i) even worse. Therefore, obtaining ϕ without any constraints will result in degraded representations.

To mitigate the limitations, we propose to further make use of intra-group self-supervision by introducing an Intra-Group Contrastive objective. In other words, IGC aims to maximize the Mutual information (MI) between the contextual feature and the shared target group feature. Specifically, we pair contextual feature ϕ with target videos feature V_{tar} as positive examples V_t^+ while the feature V_r from reference videos in the same group are used to build up negative examples $\tilde{V}_r$.

$$\pounds_{igc}\left(\psi_{tar}, V_t^+, \tilde{V}_r\right) = -log\left(\frac{e^{\left(s\left(\psi_{tar}, v_t^+\right)/\tau\right)}}{\sum_{i=1}^{n_{ref}} e^{\left(s\left(\psi_{tar}, \tilde{v}_r^i\right)/\tau\right)}}\right) \tag{11}$$

where $s\left(p, q\right) = p^T q / \left\|p\right\| \left\|q\right\|$ denotes the dot product between l_2 normalized p and q; τ is the temperature parameter.

We finally introduce the hyper-parameter λ to seek a trade-off between the two learning losses (details about λ can be found in the appendix), and combine the two losses:

$$\mathcal{L} = \mathcal{L}_{cap} + \lambda\mathcal{L}_{igc}, \tag{12}$$

Intuitively, by minimizing $\mathcal{L}_{igc}$, we encourage the contrasting feature refinery to condense the contextual feature and in turn ease the feature fusion.

As an example shown in Figure 4, our model can learn to focus on the refined features. When predicting unique objects, for *dishes* in V_{tar}, because all reference videos for it do not contain the same concept, the refinery learns that *dishes* is a discriminative object in V_{tar} and overlooks background features like kitchen and bathroom by leveraging the intra-group supervision.

Methods	Contextual feature			BLEU1	BLEU2	BLEU3	BLEU4	WAC	METEOR	ROUGE	CIDEr	WER↓
	Subtract[16]	Refinery	IGC									
Average (13.2s)	✘	✘	✘	39.36	24.77	15.47	9.98	28.75	16.25	37.15	155.8	89.53
	✔	✘	✘	41.17	26.48	16.56	11.37	30.89	19.35	40.29	159.7	86.78
	✘	✔	✘	44.98	28.52	17.97	12.03	34.33	20.95	43.46	169.3	86.03
	✘	✔	✔	44.93	28.77	19.03	12.69	34.13	20.72	43.86	170.6	85.34
Traversal (3884.4s)	✘	✘	✘	42.22	27.74	17.76	13.65	32.14	18.63	41.37	162.6	81.84
	✔	✘	✘	44.76	27.33	19.54	13.86	36.81	20.59	43.25	169.3	77.53
	✘	✔	✘	45.77	28.17	20.84	14.12	38.22	21.91	44.87	176.6	76.16
	✘	✔	✔	46.15	29.52	**21.22**	14.04	38.80	21.84	**45.97**	179.4	**75.61**
ERA (17.6s)	✘	✘	✘	42.34	27.87	18.61	13.55	32.58	19.34	41.08	161.7	80.37
	✔	✘	✘	45.07	28.96	18.39	13.09	36.49	20.77	43.53	168.3	77.74
	✘	✔	✘	46.37	30.22	20.55	14.37	38.61	21.63	45.06	178.7	75.42
	✘	✔	✔	**47.32**	**30.74**	20.78	**14.74**	**39.55**	**22.18**	45.77	**180.8**	75.77

Table 2. Comparison and ablation study with other methods on the YouCook2 dataset. The time below the methods is the average time required to train an epoch. For each method, the first row indicates no contextual features, the second row indicates the use of subtraction to obtain contextual features, the third row indicates the use of refinery instead of subtraction, and the shaded fourth row indicates the complete refinery of contextual features using IGC.

Methods	Contextual feature			BLEU1	BLEU2	BLEU3	BLEU4	WAC	METEOR	ROUGE	CIDEr	WER↓
	Subtract[16]	Refinery	IGC									
Average (41.7s)	✘	✘	✘	39.48	23.06	16.40	12.33	28.42	17.89	37.88	149.4	89.21
	✔	✘	✘	40.95	27.64	17.43	14.48	31.16	19.13	39.27	157.7	87.84
	✘	✔	✘	41.21	28.73	19.81	15.49	33.27	19.66	41.06	171.6	87.10
	✘	✔	✔	41.87	28.46	19.53	16.17	33.62	20.32	40.17	172.8	86.01
Traversal (11456.4s)	✘	✘	✘	39.94	26.07	18.92	16.24	31.67	19.10	39.23	159.4	79.28
	✔	✘	✘	41.76	27.11	20.13	16.64	35.35	20.08	40.88	166.2	77.36
	✘	✔	✘	42.52	29.59	21.05	17.65	37.09	21.40	41.51	177.0	75.11
	✘	✔	✔	**44.57**	29.70	21.28	16.53	37.85	**21.64**	**42.78**	180.7	74.94
ERA (49.3s)	✘	✘	✘	39.45	25.58	18.66	15.37	31.42	19.19	38.79	158.2	79.18
	✔	✘	✘	41.17	27.04	19.87	16.56	34.64	20.24	40.36	165.5	76.55
	✘	✔	✘	42.68	28.44	20.72	16.73	36.93	21.04	41.80	177.9	74.85
	✘	✔	✔	44.26	**29.75**	**21.61**	**16.97**	**37.93**	21.51	42.42	**181.3**	**74.41**

Table 3. Comparison and ablation study with other methods on ActivityNet dataset. Same experimental setup as YouCook2.

5. Experiment

5.1. Experimental settings

Evaluation Metrics: We consider the standard metrics widely used in image captioning literature. BLEU[21] for a sanity check, METEOR[1] based on unigram precision and recall, ROUGE-L[24] based on the longest common subsequence cooccurrence, and CIDEr[26] based on human-like consensus. In addition, following [16], we also consider two additional metrics, Word-by-word accuracy (WAC) and word error rate (WER), that specifically assess word-based accuracy due to the group descriptions being often compact.

Implementation Details: For videos, we inspect 10 frames for each video. And the ResNet-101[6] model pretrained on ImageNet[5] is used to extract the representation of video. For descriptions, we set the max length as 15. When training our system, we used the Adam[13] optimizer with a $5 * 10^{-5}$ learning rate with a batch size of 32 and $\lambda = 1 * 10^{-3}$. In the Transformer decoder, we set the dimensionality d of each layer to 512, and the number of heads to 8. For inference, we used a beam search with 5 and train our models on an NVIDIA GeForce RTX 2080.

5.2. Group Captioning Performance

To the best of our knowledge, no methods are doing precisely the same task as ours before. The most relevant to our research is [16], which studied the semantic understanding of groups of visual information in the spatial direction, not the temporal direction. Therefore, we followed the path of exploration to thoroughly analyze the behavior of our model

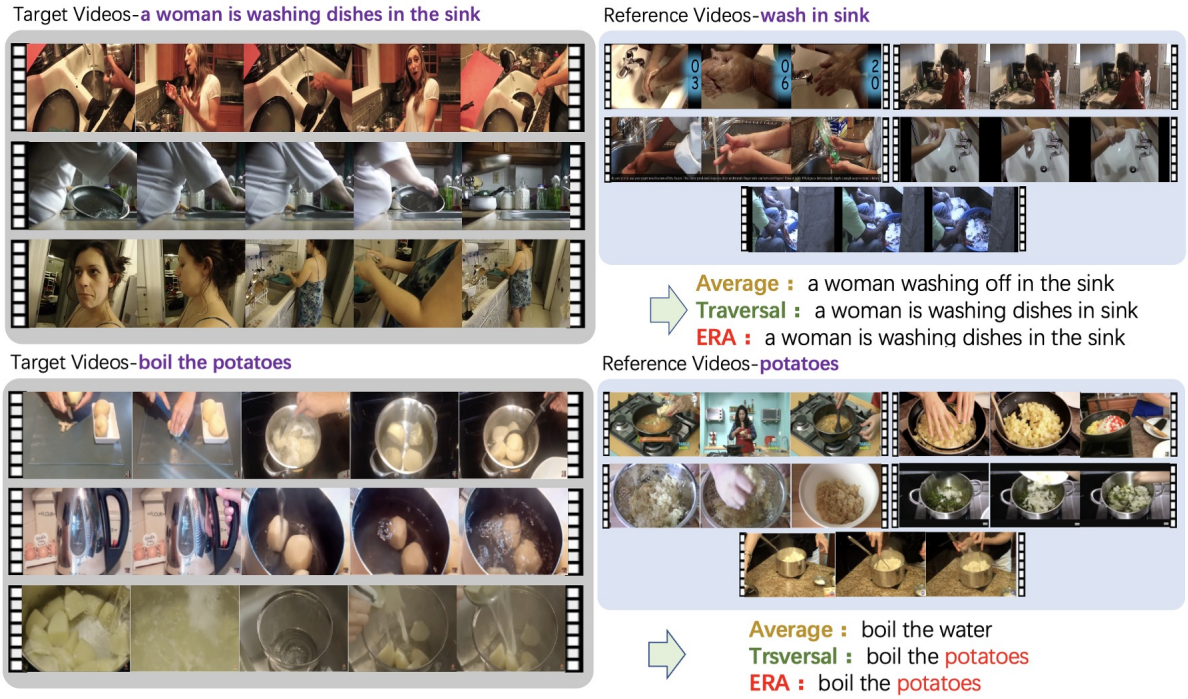

Figure 4. Qualitative prediction examples on YouCook2 Captions (up) and ActivityNet Captions (bottom) datasets. Our model can effectively summarize the shared information and takes contextual information between the target and reference group into account during captioning to predict accurate group captioning results.

Methods	ActivityNet↓	YouCook2↓
Average	13.77	11.67
Traversal	10.13	9.24
ERA	**10.05**	**9.16**

Table 4. Results of the variance of the extracted feature ψ within each video group are shown.

Methods	B1	B4	WAC	M	R	C	WER↓
T1+R5	41.94	14.83	31.63	19.86	40.86	163.1	87.43
T2+R5	45.64	14.54	37.36	21.68	43.84	174.3	77.23
T3+R0	19.87	8.32	13.92	8.63	19.15	69.81	97.62
T3+R5	**47.32**	**14.74**	**39.55**	**22.18**	**45.77**	**180.8**	**75.77**

Table 5. Performance with varying the number of target and reference videos. (evaluated on YouCook2 Captions dataset)

and provide insights into this new problem. We compare our model with two baselines: **Average** which calculates ψ by averaging the video features in the temporal direction and **Traversal** which serves as a theoretical upper bound performance. The detail about the Average can be found in the appendix. To analyze the effect of different components, we perform ablation studies with the proposed contextual feature module on all three models.

The captioning performance on the YouCook2 Captions and ActivityNet Captions datasets are reported in Table 2 and Table 3, and we make the following observations. First, when compared to the Average model, our model achieves impressive improvement for all metrics. For example, ERA brings +10.2% CIDEr boost on YouCook2 and +8.5% CIDEr boost on ActivityNet. Second, when compared to the Traversal model which serves as the upper bound performance of group video captioning theoretically, our model also achieves competitive results, 179.4 vs 180.8 on YouCook2. It is worth mentioning that ERA achieved those competitive results with only 0.4% training time of Traversal. This observation suggests that our ERA is an efficient substitute approximate solution. Note that Traversal would perform worse than ERA in some cases while Traversal searches the space of all possible solutions but lacks a better way of integration.

Specifically for the influence of each component, models trained without contextual features tend to perform worse. We also compare refinery with the subtraction proposed by [16], which removes the similarity portion in two group features without constraints and deduces noisy representations. When we instead use our contextual feature refinery, which emphasizes the difference between the target video group in the context of the reference group, we observe a further performance improvement (*e.g.* 168.3 vs 178.7 CIDEr on YouCooK2). And on top of Refinery, applying IGC to either method leads to a performance boost.

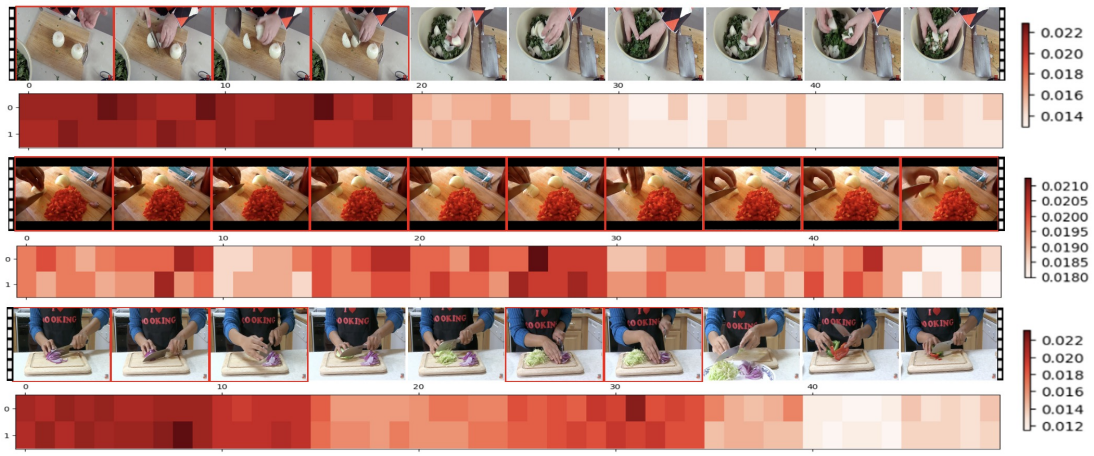

Figure 5. Visualization of the shared content identification. The frame with the red box is the shred content. The first row of each weight matrix is calculated by Traversal and the second row is calculated by ERA. Darker color means more relevant.

5.3. Discussion

Quantitative and qualitative analysis of shared group features identification. We first design quantitative experiments to evaluate the shared group feature. Specially, we evaluated the variance of the extracted features, which is calculated as $var = \frac{1}{n}\sum_{i=1}^{n}(w_i v_i - \psi)^T(w_i v_i - \psi)$. A smaller var value means that the extracted shared feature ψ among all the videos is more convergent. Table 4 shows the results of the YouCook2 and ActivityNet Captions datasets.

To better understand the effectiveness of EAR in shared content identification, we visualize the score w calculated by Traversal and EAR. As shown in Figure 5, EAR performs similarly to Traversal in identifying shared content. Qualitative and quantitative experiments further demonstrate that our proposed EAR can correctly identify the shared content and achieve competitive results with Traversal by linear complexity.

Importance of multiple target and reference videos. To investigate the effectiveness of giving multiple videos in each group, we vary the number of target and reference videos. As the results are shown in Table 5, fewer target or reference videos results in a performance decline. The results of T3+R0 indicate that when not given a reference group the predictions tend to be more generic and less discriminative, which indicates that the contextual information contained in reference videos is necessary.

Effectiveness of EAR. To investigate the effect of EAR on reducing model complexity, we measured the average time to train an epoch on YouCook2 with the same batch size of 32 for each model. Figure 6 shows that the time consumed by our method and average is basically the same during the increase of the number of videos from 1 to 5. While the time consumed by Traversal starts to rise sharply as the number of videos increases (when the number of videos is 5, the time has reached 3884s) which is unacceptable.

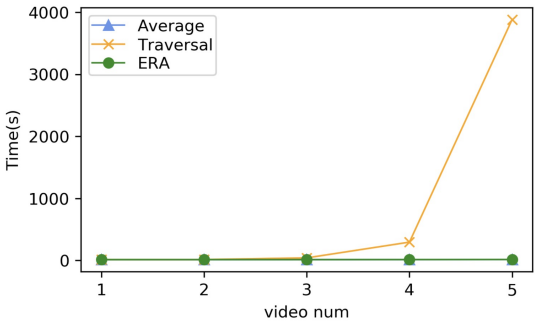

Figure 6. The average training time for one epoch with different video nums on YouCook2.

6. Conclusion

In this paper, we introduce the novel group-based video captioning, which requires a semantic understanding of the relationships among videos to identify the desired content. To solve this problem, firstly, we present an efficient relation approximation to capture the shared content among the videos while the complexity scales linearly $O(n)$ with the number of videos. Then, we propose the contextual feature refinery and further consider intra-group supervision to guarantee that the learned representations are meaningful for target videos. To evaluate our system, we construct two new datasets using the YouCook2 and ActivityNet Caption datasets. Qualitative and quantitative experimental results demonstrate the effectiveness of our method.

Acknowledge: This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No. 62222211, National Natural Science Foundation of China under Grant No.61836002 and National Natural Science Foundation of China under Grant No.62072397.

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[2] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1345–1353, 2018. 2

[3] Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan, and Zhou Zhao. Opensr: Open-modality speech recognition via maintaining multi-modality alignment. *arXiv preprint arXiv:2306.06410*, 2023. 2

[4] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 5

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[7] Tao Jin, Siyu Huang, Ming Chen, Yingming Li, and Zhongfei Zhang. Sbat: Video captioning with sparse boundary-aware transformer. *arXiv preprint arXiv:2007.11888*, 2020. 2

[8] Tao Jin, Siyu Huang, Yingming Li, and Zhongfei Zhang. Low-rank hoca: Efficient high-order cross-modal attention for video captioning. *arXiv preprint arXiv:1911.00212*, 2019. 1

[9] Tao Jin, Yingming Li, and Zhongfei Zhang. Recurrent convolutional video captioning with global and local attention. *Neurocomputing*, 370:118–127, 2019. 2

[10] Tao Jin and Zhou Zhao. Contrastive disentangled meta-learning for signer-independent sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5065–5073, 2021. 2

[11] Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. Mc-slt: Towards low-resource signer-adaptive sign language translation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4939–4947, 2022. 2

[12] Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. Prior knowledge and memory enriched transformer for sign language translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3766–3775, 2022. 2

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*, 2018. 3

[15] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 2, 3

[16] Zhuowan Li, Quan Tran, Long Mai, Zhe Lin, and Alan L Yuille. Context-aware group captioning via self-attention and contrastive features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3440–3450, 2020. 1, 2, 6, 7

[17] Wang Lin, Tao Jin, Wenwen Pan, Linjun Li, Xize Cheng, Ye Wang, and Zhou Zhao. Tavt: Towards transferable audio-visual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14983–14999, 2023. 1

[18] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 338–354, 2018. 2

[19] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2018. 2

[20] 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. 2

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[22] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633, 2019. 2

[23] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1971–1980, 2021. 2

[24] Lin CY ROUGE. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004. 6

[25] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019. 2

[26] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[27] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 1, 2

[28] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Group-based distinctive image captioning with memory attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5020–5028, 2021. 2

[29] Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1890–1900, 2020. 2

[30] Ye Wang, Tao Jin, Wang Lin, Xize Cheng, Linjun Li, and Zhou Zhao. Semantic-conditioned dual adaptation for cross-domain query-based visual segmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9797–9815, 2023. 2

[31] Ye Wang, Wang Lin, Shengyu Zhang, Tao Jin, Linjun Li, Xize Cheng, and Zhou Zhao. Weakly-supervised spoken video grounding via semantic interaction learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10914–10932, 2023. 2

[32] Zehan Wang, Yang Zhao, Haifeng Huang, Yan Xia, and Zhou Zhao. Scene-robust natural language video localization via learning domain-invariant representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 144–160, 2023. 2

[33] An Yan, Xin Eric Wang, Tsu-Jui Fu, and William Yang Wang. L2c: Describing visual differences needs semantic understanding of individuals. *arXiv preprint arXiv:2102.01860*, 2021. 2

[34] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. Hierarchical modular network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17939–17948, 2022. 1

[35] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8327–8336, 2019. 2

[36] 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 computer vision and pattern recognition*, pages 13278–13288, 2020. 2

[37] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 3

[38] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748, 2018. 2