Fine-grained video paragraph captioning via exploring object-centered internal and external knowledge

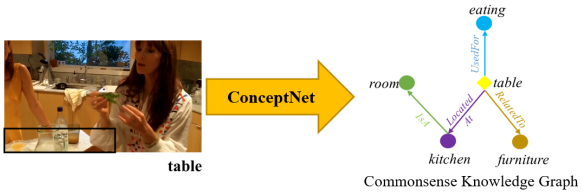
Anonymous ACL submission

Abstract

Video paragraph captioning task aims at generating a fine-grained, coherent and relevant paragraph for a video. Existing works often treat the objects (the potential main components in a sentence) isolated from the whole video content, and rarely explore the latent semantic relation between a certain object and the current video concepts, causing the generated description dull and even incorrect. Besides, different from images where objects are static, the temporal states of objects are changing in videos. The dynamic information could be contributed to better understand the whole video content. Towards generating a more detailed and stick-to-the-topic paragraph, we propose a novel framework that focuses on exploring the rich semantic and temporal meaning of objects, by constructing the concept graph from the external commonsense knowledge and the state graph from the internal video frames. Extensive experiments on ActivityNet captions and Youcook2 demonstrate the effectiveness of our method compared to the state-of-the-art works. We will release our code on GitHub community.

1 Introduction

In recent years, automatically generating a human-like paragraph to describe a video has gained a deal of interests in visual understanding domain. Unlike generating a single sentence from a short video, also known as video captioning (Gao et al., 2017), video paragraph captioning (Yu et al., 2016) aims at generating a coherent, accurate and informative description, which involves plentiful visual contents and activities. Towards this goal, many works (Zhang et al., 2020; Zhou et al., 2019; Shen et al., 2020) put their focus on utilizing the various objects that appear in the video, employing the off-the-shelf object detection techniques (Zou et al., 2019) on video frames. Park et al. (2019) designed an adversarial learning framework, and



GVDSup: Two girls are seen speaking to the camera and leads into a woman speaking to the camera.
Ours: A close up of a table is shown followed by a woman standing in a kitchen and speaking to the camera.
Ground Truth: A group of women are in a kitchen, eating lettuce leaves.

Figure 1: Illustration of exploring Commonsense Knowledge Graph in video paragraph captioning task. The high-level semantic entity could be inferred from the objects with the help of the external commonsense knowledge and suitable selection mechanism.

employed object features to enrich the video content. Zhou et al. (2019) came up with a new task named Grounded Video Description, by grounding the visual objects in generated sentences to avoid object hallucinations in descriptions. Zhang et al. (2020) further proposed a scene graph (Yang et al., 2018) based method for Grounded Video Description task.

Although the above-mentioned object-related methods proved the effectiveness of employing the object region features in video frames, there are still some dilemmas remained. On one hand, for example, in Figure 1, *table* can be related to *kitchen* or *eating* or some other semantic roles, commonsense concepts can be inferred from the *table* which might contain abundant semantic information. Current methods often fail to link such connection between the detected object with high level abstract concepts, and the object is isolated from the video concepts, or related to the video in a basic and limited level. However, in realistic world, much abstract external information could be inferred from a certain object. Equipped with such information, the model can understand the video concept better, gain the ability to generate

more fine-grained and more correct descriptions. Besides that, with reasonable selection mechanism and suitable multi-modal representation learning, the selected commonsense information can be the key component to form the concept for the entire video, resulting in more coherent multi-sentence descriptions. Lacking of key concepts often leads to the incorrect interpretation of the video contents, sabotaging the integrity and veracity of the generated results. On the other hand, unlike images, objects change through time in videos, the above-mentioned methods rarely put their focus on the crucial role of the temporal state change in a certain object. Just like the majority of the activities in the real world require detailed analysis, observing the dynamic temporal state of object can boost up the model understanding ability, and lead the model to generate the high quality sentences with a better view of the environment.

Towards filling this gap, we propose a novel framework to not only learn external meaningful information from the objects, but also explore internal temporal state change among them. Our model absorbs fine-grained semantic and temporal information to form the paragraph captions. Recently, taking the advantage of the external knowledge has been explored in many vision-language tasks such as Visual Question Answering (Marino et al., 2019), Visual Relation Detection (Wan et al., 2021). Commonsense knowledge graph like ConceptNet (Liu and Singh, 2004) provides sufficient external commonsense knowledge. The nodes in the graph can be nouns, adverbs, adjectives or terms, and connect with each other in a commonsense way, as shown in Figure 1. However, many concepts extracted from the commonsense knowledge graph for nodes are redundant for a specific scenario. For example, in Figure 1, the *furniture* is an irrelevant concept in this scenario. To make the descriptions both informative and accurate, we design an "imaginative-to-precise" network to encourage the suitable concepts to contribute more under the supervision of descriptions. We enrich the semantic meaning of objects by digging in their latent related concepts. It is worth noting that there is still abundant temporal information to be explored other than the external knowledge. To make the full use of the temporal and semantic information from objects, we further model the object temporal dynamic change by creating an object state tracking network, via which we endue the machine the ability of not only knowing what (semantically) but also knowing when (temporally).

Our contributions are summarized into three folds:

(1). We propose a novel object-centered semantic-temporal framework for video paragraph captioning, which can learn both internal and external knowledge to form a fine-grained and video-relevant multi-sentence description.

(2). We explore the external commonsense concept knowledge, and refine the concept knowledge through well-designed selection mechanism. Equipped with such commonsense knowledge, the model is able to generate stick-to-the-topic and coherent paragraphs.

(3). Extensive experiments on ActivityNet Captions and Youcook2 datasets demonstrate that our model outperforms the state-of-the-art methods.

2 Related Work

2.1 Video Captioning

Video captioning task has attracted widespread attention in recent years. With the remarkable progresses in Machine Learning, Transformer (Vaswani et al., 2017), Generative Adversarial network (Creswell et al., 2018) and Reinforcement Learning (Sutton and Barto, 2018) provide new solutions to this task.

The instinct thought of video captioning is "video to text". However, this task can be divided into two different sub-tasks based on whether generating a single brief sentence for a short video or generating a paragraph for a long video. The former task needs the model to be concise and accurate, while the video paragraph captioning task requires to generate more coherent and fine-grained descriptions. Generating multiple sentences to describe a video can lead to the cross-sentence redundancy. Yu et al. (2016) proposed a hierarchical LSTM-based caption decoder to pass on the cross-sentence context, and Lei et al. (2020) proposed a recurrent transformer to tackle this issue. Park et al. (2019) came up with a method using adversarial learning to train the model to generate coherent, relevant and less redundant descriptions. Reinforcement learning training was employed by Song et al. (2021) in the hope of generating more diverse descriptions.

2.2 Graph-based Neural Network

In recent years, the progresses made in visual relation detection (Zhang et al., 2017) have boosted many down-streaming tasks such as image captioning, video captioning and Visual Question Answering (Antol et al., 2015). In order to better understand the interaction between the visual objects, modeling the complicated visual relation between two objects has been widely explored. Relational-graph-based network is a reasonable solution to this issue. In our work, we model the concept relationship extracted from the commonsense knowledge graph by a relational graph bias network. Besides the widely adopted Graph Convolution Network, Graph Attention Network was proposed by Veličković et al. (2017) to learn the different importance weights among the neighbors of one node. In our work we employ a Graph Attention Network (GAT) to "track down" the temporal states of our objects.

3 Methodology

Firstly we introduce the video paragraph captioning task. Given a video V with annotated temporal video clips $[V_1, V_2, ..., V_T]$, the task aims at generating a relevant and coherent paragraph to describe the video events. We denote the generated sentences as $[S_1, S_2, ..., S_T]$.

Our model contains three major components. The visual and concept encoder will be introduced in section 3.1, object temporal encoder is in section 3.2, and the language decoder is in section 3.3.

3.1 Vision-guided Concept Selection Network

Commonsense Knowledge Graph Extraction. For a video clip V_t, we uniformly sample K frames and conduct a pre-trained Faster-RCNN (Ren et al., 2015) on them to gain the predicted object labels. After obtaining the top-n objects $\left[O_1^t, O_2^t, ..., O_n^t\right]$ for video clip V_t, we treat the object label O_i^t as the seed query to search in the ConceptNet commonsense knowledge database, and extract the nodes (embedded word vectors in the knowledge base) $\left[C_{i1}^t, C_{i2}^t, ..., C_{iQ}^t\right]$ connected to the seed node with the Q-th highest edge weights. We denote the seed node, its' j-th neighbor node and the relationship between them using a triplet $\left\{O_i^t, C_{ij}^t, R_{ij}^t\right\}$.

Relational Concept Encoder. In ConceptNet, concepts are defined as the graph nodes, and the edge between two nodes is composed by a weighted score and a relation term. The score shows how close the two nodes are related to each other and the relation term shows what kind of relation between them. For example, given a seed object *table*, the concepts related to it could be *furniture* and *kitchen*. However, the related term between *furniture* and *table* is *RelatedTo*, but the term between *kitchen* and *table* is *LocatedAt*. In order to understand the different semantic roles between different concepts for refining the representations of the concept nodes, we design a relational concept encoder. Inspired by relational graph network learning in many vision-language tasks (Yao et al., 2018; Johnson et al., 2018), we construct our relational graph as follows:

Node-to-node: We gather the concepts from the video clip and connect the related pairs, and each concept is a node in the graph. It is worth noting that each edge has direction. We denote the node vector as $\mathbf{c}$;

Edge-to-embedding: We further collect all the relations between the concepts, label them into different classes, and embed each class into a relational vector $\mathbf{b}$.

We employ graph bias convolution network on the created graph, treat each relational edge as a bias vector. Then the concept nodes can be learned by aggregating their neighbors and the relations as:

$$\tilde{\mathbf{c}}_i = \rho\left(\sum_{\mathbf{c}_j \in \mathcal{N}(\mathbf{c}_i)} \mathbf{W}_{\mathrm{dir}(\mathbf{c}_i,\mathbf{c}_j)}\mathbf{c}_j + \mathbf{b}_{\mathrm{rel}(\mathbf{c}_i,\mathbf{c}_j)}\right) \tag{1}$$

where $\mathbf{W}_{dir(\mathbf{c}_i,\mathbf{c}_j)}$ stands for the transformation matrix used for edges that connect $\mathbf{c}_i$ to $\mathbf{c}_j$, we choose different transformation matrix to differentiate the object and the subject (i.e., $\mathbf{W_1}$ for $\mathbf{c}_i$-to-$\mathbf{c}_i$, $\mathbf{W_2}$ for $\mathbf{c}_i$-to-$\mathbf{c}_j$, $\mathbf{W_3}$ for $\mathbf{c}_j$-to-$\mathbf{c}_i$). ρ denotes the relu function.

Vision-guided Concept-to-content Matching. Since we extract multiple concepts for objects, and a video clip often has nearly one hundred candidate concepts. The majority of them are irrelevant to the current description. Such redundant concepts would bring noises to the model thus hurt the model's captioning performance and lead to semantic hallucination. In order to convey important and accurate concepts as the semantic guidance for the sentence generation module, we conduct a cross-modal attention, by taking the segment-level visual features as the keys and the candidate concepts as

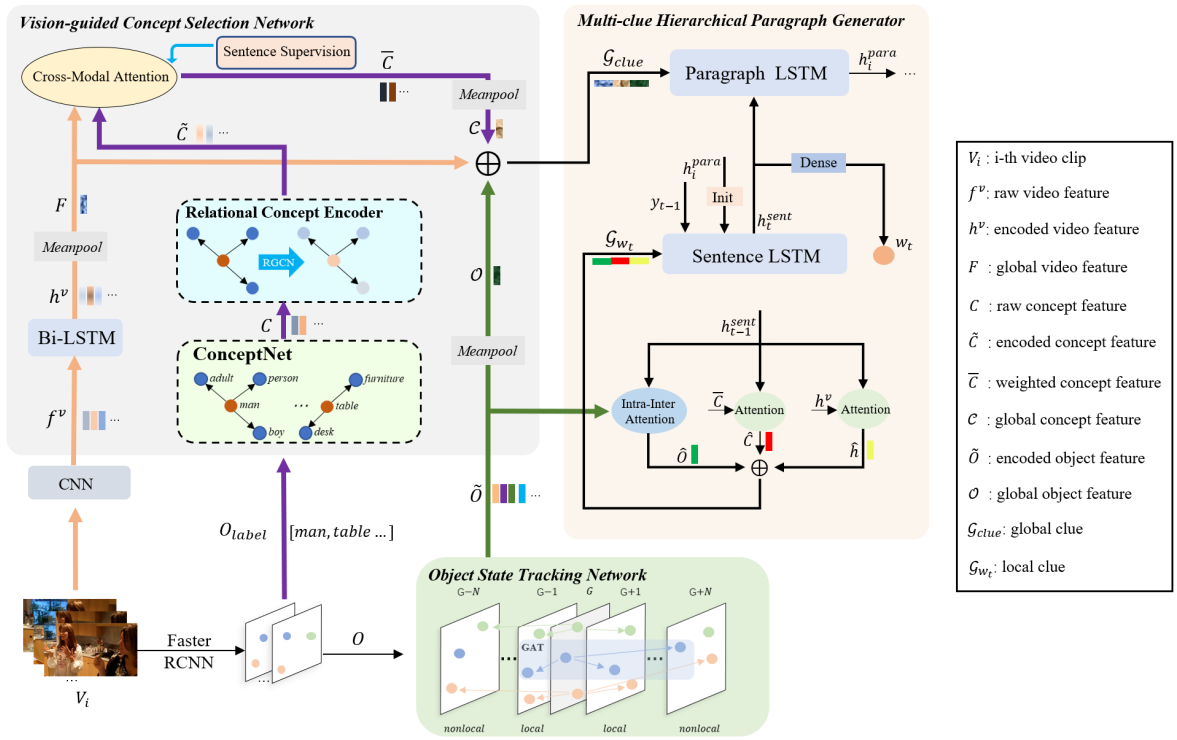


Figure 2: Overview of our framework. It mainly contains three sub-networks: Vision-guided Concept Selection Network (VGCSN), Object State Tracking Network (OSTN) and Multi-clue Hierarchical Paragraph Generator (MHPG).

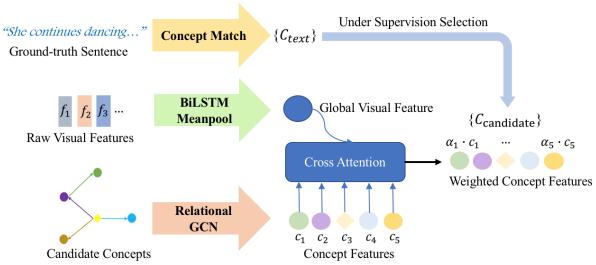


Figure 3: An illustration of Vision-guided Concept Selection network (VGCSN).

the queries. We match the relevant concepts with the visual clue, thus change participant degree of the concept features based on their attention scores. Given the frame features $[f_1^t, f_2^t, ..., f_s^t]$ in video clip V_t , we employ a Bi-LSTM (GRA, 2005) to encode the frame features for capturing the temporal relations among frames. Then we mean pool the encoded frame-wise features to get the global video clip feature $\mathbf{F}^t$, and conduct the cross-modal attention:

$$\beta_i^t = (\mathbf{W}_v \mathbf{F}^t)^T \mathbf{W}_c \tilde{\mathbf{c}}_i^t \quad (2)$$

$$\alpha_i^t = \sigma(\beta_i^t) \quad (3)$$

$$\bar{\mathbf{c}}_i^t = \alpha_i^t \cdot \tilde{\mathbf{c}}_i^t \quad (4)$$

We compute the attention scores between each concept and the global visual feature as the different semantic importance of concepts.

Sentence-supervised Concept Alignment. After the vision-guided cross-modal attention, we gather the content-aware candidate concepts $\bar{\mathbf{c}}_\mathbf{i}$. However, visual information can only ground the concepts in a limited level, because visual information feature is global information, while the descriptions selectively focus on the salient parts. As a common fact, people hardly describe everything in a scene. The semantic entities , i.e., objects or abstract concepts inside the sentences often value most. We hope our concepts can be more correctly grounded to guide the sentence generation, and enable the model to generate stick-to-the-topic sentences. We design a sentence-supervised manner to suppress the attention weights of the text-irrelevant concepts and meanwhile encourage the concepts mentioned in the text to gain more attention. We denote the ground truth words set as $\mathcal{C}_{text}$ and the candidate concept set as $\mathcal{C}_{candidate}$. The candidate concepts inside the intersection of two sets $\mathcal{C}_{text} \cap \mathcal{C}_{candidate}$

are treated as the positive examples, while others are the negative ones. Our vision-guided concept weights are trained under the sentence-supervision using cross-entropy loss.

3.2 Object State Tracking Network

Object State Tracking Graph Construction. Unlike on images where contents are static, the objects are changing dynamically in videos. To catch the dynamic change of an object, we construct a temporal state tracking graph. This graph crosses multi-frames to aggregate long-range information, because some movements cannot be grounded in a few seconds but requires more time duration. For G-th frame of video clip V_t, we define the neighbor frame set as $\{G \pm i\}_{i=1}^{N}$, and the nearest frames as their local frames, while other frames are the non-local frames. Given j-th object in G-th frame, we first compute the cosine similarity between O_j^G and all object vectors $O^{G\pm1}$ in local frames, and select the most related objects in both frames as neighbors. Then, we treat the largest similarity score as the threshold values γ_{thresh} to select the objects in non-local frames. This above-threshold selection can help to prevent noises being selected to a certain extent. The objects connected with O_j^G are denoted as set $\mathcal{O}_{neighbor}^{j}$.

Object Graph Attention Network. The information from local and non-local neighbor frames contributes differently. Thus we employ a graph attention network to pay more attention to significant ones:

$$\tilde{O}_j = \mathrm{GAT}(O_j, O_i^j \in \mathcal{O}_{neighbor}^{j}) \tag{5}$$

The specific formulations of GAT layer are:

$$\mu_{ji} = \rho(\mathbf{W}_a[\mathbf{W}_p h_i; \mathbf{W}_k h_j]) \tag{6}$$

$$\eta_{ji} = \frac{\exp(\mu_{ji})}{\sum_{l\in\mathcal{N}_j}\exp(\mu_{jl})} \tag{7}$$

$$\tilde{\boldsymbol{h}}_j = \sigma\left(\sum_{i\in\mathcal{N}_j}\eta_{ji}\mathbf{W}_t\boldsymbol{h}_i\right) \tag{8}$$

where h_i and h_j in the GAT layers denote the hidden states of O_i and O_j.

3.3 Multi-clue Hierarchical Paragraph Generator

For our sentence decoder, we employ a hierarchical architecture to gather clue information for more coherent sentence generation. The hierarchical paragraph generator contains a paragraph LSTM and a sentence LSTM. The paragraph LSTM processes the global context information. The sentence LSTM generates the word in each time step guided by the visual, textual information and the contextual memories provided by the paragraph LSTM.

Global Clue Gathering for Paragraph LSTM. With the purpose of forming an informative guidance vector for sentence generator, we gather the global clues (global visual features F , global concept features $\mathcal{C}$ and global state-aware object features $\mathcal{O}$) from the encoder by respectively mean-pooling the clip-level features. For video clip V_i's sentence generation, we concatenate the multi-modal global clues together with the previous V_{i-1} sentence's last hidden state from sentence generator as the concept and memory guidance for video clip:

$$h_i^{para} = \mathrm{LSTM}_{para}(h_{i-1}^{para}, [\mathcal{C}; \mathcal{O}; F; h_{i-1}^{sent}]) \tag{9}$$

Multi-clue Attention Sentence Generator. At the beginning of generating V_i's sentence in the paragraph, we utilize the last hidden state from the paragraph LSTM to initialize the hidden state of the sentence LSTM. For every time step t, we design a multi-clue attention method to stimulate the model to choose wisely from the various inputs of the visual, semantic and textual clues. For object clues, we conduct inter-intra frame attention with h_{t-1}^{sent}. For concept and frame features, we attend their features by cross-modal attention with the previous hidden state h_{t-1}^{sent}. The hidden state of the sentence LSTM is generated by:

$$h_t^{sent} = \mathrm{LSTM}_{sent}([w_{t-1}; \hat{O}; \hat{C}; \hat{h}], h_{t-1}^{sent}) \tag{10}$$

where w_t is generated by the hidden state h_t^{sent}, and $\hat{O}$, $\hat{C}$, $\hat{h}$ are the attended clues at each time step.

3.4 Training

Unlikelihood Training. In order to reduce sentence repeated n-grams, we conduct unlikelihood training (Welleck et al., 2019) for our generation:

$$\mathcal{L}_{caption} = -\frac{1}{T}\sum_{t=1}^{T}\left(\log p\left(w_t \mid w_{<t}, v\right)\right) + \sum_{e\in E^t}\log\left(1 - p\left(c \mid w_{<t}, v\right)\right) \tag{11}$$

Methods	Det.	ae-val				ae-test			
		B@4	M	C	R@4↓	B@4	M	C	R@4↓
Transformer based									
VTransformer (CVPR) (2018b)	✗	9.75	15.64	22.16	7.79	9.31	15.54	21.33	7.45
Transformer-XL (ACL) (2019)	✗	10.39	15.09	21.67	8.54	10.25	14.91	21.71	8.79
Transformer-XLRG	✗	10.17	14.77	20.40	8.85	10.07	14.58	20.34	9.37
MART (ACL) (2020)	✗	10.33	15.68	23.42	5.18	9.78	15.57	22.16	5.44
TowardsDiv* (CVPR) (2021)	✗	-	-	-	-	**12.20**	16.10	27.36	2.63
ParallelDecoding (ICCV) (2021)	✗	11.80	15.93	27.27	-	-	-	-	-
VPCSum (ACL) (2021)	✗	-	-	-	-	10.89	15.84	24.33	**1.54**
LSTM based with Detection Features									
GVD (CVPR) (2019)	✓	11.04	15.71	21.95	8.76	10.50	15.60	21.60	-
GVDSup (CVPR) (2019)	✓	11.30	16.41	22.94	7.04	10.70	16.10	22.20	-
AdvInf (CVPR) (2019)	✓	10.04	**16.60**	20.97	5.76	-	-	-	-
Ours	✓	**12.30**	16.52	**29.56**	**4.64**	11.60	**16.30**	**29.19**	4.33

Table 1: Paragraph-level automatic evaluation results on ActivityNet Cations. The * in row 5 means that the method uses an additional RGB feature and is under the reinforcement training setting. Det. means whether object region features are used.

where E^t denotes the previously generated word set. Under such setting, the sentence redundancy can be reduced by maximizing w_t probability and meanwhile suppressing the previously generated word probability.

Total loss. The total loss contains sentence generation loss and concept supervision loss:

$$\mathcal{L}_{total} = \mathcal{L}_{caption} + \lambda \mathcal{L}_{concept} \quad (12)$$

where λ is a hyper parameter. We conduct our training in an end-to-end manner.

4 Experiments

4.1 Datasets

We conduct our experiments on two benchmark datasets ActivityNet Captions (Krishna et al., 2017) and Youcook2 (Zhou et al., 2018a). ActivityNet Captions is a large-scale dataset of indoor and outdoor activities, which includes 10,009 videos in training set and 4,917 videos in validation set. For better comparing with other baselines, we use the commonly used splits in Zhou et al. (2019), where the original validation set is split into two subsets, i.e., ae-val with 2,460 videos for validation and ae-test with 2,457 videos for testing. Youcook2 is a task-specific dataset composed by indoor cooking videos, which has 1,333 training videos and 457 validation videos.

4.2 Data Preprocessing

We use appearance and optical flow features provided by Zhou et al. (2018b). For object region features, we employ a pre-trained Faster-RCNN model to extract top-K object region features from frames every two seconds for a video clip. For creating the vocabularies of these two datasets, we add the word into our dictionary if the word frequency is larger than 4 in ActivityNet captions and 2 in youcook2.

Methods with Detection features	B@4	M	C
GVD (CVPR) (2019)	2.16	10.8	44.9
GVDSup (CVPR) (2019)	2.35	11.0	45.5
RelGraph (ACMMM) (2020)	2.59	11.0	47.2
HieAtt (IJCAI) (2020)	2.65	11.2	49.3
Ours	**2.88**	**11.3**	**52.1**

Table 2: Sentence-level automatic evaluation results on ActivityNet ae-test.

4.3 Evaluation Metrics

We conduct our evaluation on paragraph level captioning performance as Xiong et al. (2018), reporting the standard metrics including BLEU@4 (B@4) (Papineni et al., 2002), METEOR (M) (Denkowski and Lavie, 2014), CIDEr-D (C) (Vedantam et al., 2015). For our paragraph repetition evaluation, we follow Xiong et al. (2018), and use the R@4 metric. Besides the automatic evaluation, we also conduct the human evaluation to evaluate the coherence, relevance and expressiveness.

4.4 Baselines

State-of-the-art Methods. We compare our model with multiple methods and separate them by their main architectures. As for LSTM-based architecture, we compare with AdvInf (Park et al.,

Method	B@4	M	C
Transformer-XL (2019)	6.6	14.8	26.35
VPCSum (2021)	6.1	**15.1**	23.92
Vanilla PE	6.5	14.3	23.23
Vanilla+Concept	6.3	14.5	23.82
Vanilla+VGCSN	**6.8**	14.5	**27.21**

Table 3: Evaluation results on Youcook2 val.

	Ours(%)	GVDSup(%)	Vanilla+Object(%)	tie(%)
Relevance	**33.65**	26.92	20.19	19.23
Coherence	**35.71**	26.79	26.79	10.71
Expressiveness	**33.91**	29.57	22.61	13.91

Table 4: Human evaluation results between ours and the baselines. "tie" means the caption quality between the three models are close.

Method	Det.	B@4	M	C	R@4↓
Vanilla PE	✗	11.0	16.1	25.56	5.97
Vanilla+Object	✓	11.3	16.0	26.63	4.85
Vanilla+OSTN	✓	11.8	16.0	27.41	4.76
Vanilla+Concept	✗	11.2	16.1	27.13	4.45
Vanilla+VGCSN	✗	11.6	16.3	28.25	5.14
OSTN+VGCSN	✓	**12.0**	16.3	29.12	4.98
OSTN+VGCSN (PUlk.)	✓	11.4	**16.5**	28.71	**3.88**
OSTN+VGCSN (SUlk.)	✓	11.6	16.3	**29.20**	4.33

Table 5: Ablation studies on ActivityNet ae-test. Det. indicates object region features.

2019), GVDSup (Zhou et al., 2019), RelGraph (Zhang et al., 2020) and HieAtt (Shen et al., 2020). For transformer-based methods, we compare with VTranformer (Zhou et al., 2018b), Transformer-XL (Dai et al., 2019), MART (Lei et al., 2020), TowardsDiv (Song et al., 2021), ParallelDecoding (Wang et al., 2021) and VPCSum (Liu and Wan, 2021).

Vanilla PE + Object. This model is our basic architecture without Vision-guided Concept Selection Network (VGCSN) and Object State Tracking Network (OSTN), and the decoder remains to be the hierarchical network. We also employ our intra-inter frame region attention when generating words. In addition, we embed the relatively temporal location of each clip in the whole video as Mun et al. (2020). Here, we denote this model as Vanilla PE (Position-Enriched) + Object.

4.5 Experimental Results

Automatic Evaluation. Table 1 shows the results on both ActivityNet ae-val and ae-test, and several methods only provided their results on one of these two splits. For fair comparisons, we mark the methods with their architectures and the features they use. It can be observed that stronger or comparable results on B@4, M, C and redundancy metric R@4 (the lower the better) are achieved. Comparing to all the methods employing the same visual features as ours, our model outperforms them on most of the metrics. Table 2 shows the sentence-level automatic evaluation results between several methods and ours, because some of them were only evaluated under such different settings. RelGraph and HieAtt did not test the redundancy R@4. It is worth noting that the related methods in Table 2 use region detection features and grounding supervision, we outperform them on all three major caption metrics by a large margin even without the grounding supervision. Table 3 shows the results of our VGCSN on youcook2 val compared to Transformer based methods. It can be viewed that with only our concept network (without both object state tracking network and object region features) provides fine-grained information to boost the captioning performance, even though youcook2 is a specific indoor-cooking dataset that contains less kinds of activities and environments than ActivityNet captions.

Human Evaluation. We compare our method with the baselines that also employ object region features, i.e., Vanilla PE + Object and GVDSup, because RelGraph and HieAtt have not released their source code. We randomly sample 186 video segments from ActivityNet ae-val. We extract the GVDSup generation results from their pre-trained model. Then we ask 20 volunteers to evaluate the generated results from three aspects, i.e., coherence, relevance and expressiveness. The evaluation is anonymous, and each video is judged by 3 volunteers. From the results we can see that our model significantly outperforms GVDSup and Vanilla PE + Object on both three aspects.

Model Ablation Study. Table 4 shows the ablation

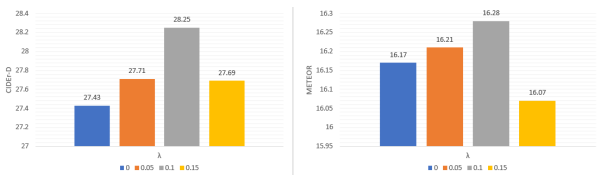


Figure 4: Experimental results of different hyper-parameter λ settings for concept loss on VGCSN.

Ground Truth: A large group of people are seen standing around when one walks into the center of the gym floors and begins a routine, she performs a dance routine while spinning her batons up and down and ends with her running off to meet others.

GVDSup: A girl is standing on a court, she begins to **dance** around the gym floor.

Vanilla+Object: A girl is seen standing in a large **gymnasium** while others watch on the side, the girls **perform a routine** while the girl continues to perform the routine.

Ours: A girl in a **pink shirt** is standing in front of a large **crowd**, the girl **performs a routine** while **spinning her baton** and **twirling her arms** and twirling.

Ground Truth: Two reporters talk in a TV set, girls talk in a beach, then girls spray sunscreen on the back of women, also a woman wearing a white coat talks and shows products, a man sprays sunscreen on his back, and then he sprays sunscreen to the back of a woman, after, a man sprays sunscreen on his arm and back, other people spray sunscreen on their legs, the reporters in the TV set continue talking.

GVDSup: A man and a woman are talking to the **camera**, a woman is seen speaking to the camera and leads into her putting **sunscreen** on a bottle, the woman is talking to the camera, the woman is then shown talking to the camera and showing off her tattoo and then shows off her tattoo, the man is talking to the camera.

Vanilla+Object: A **news anchor** news anchor is talking about a news segment, they are shown **putting sunscreen** on their sunscreen and talking, they are shown talking to the camera and showing the pictures of people, the woman is now talking to the camera, the **news reporter** is talking about the **news**.

Ours: **A news anchor and a woman** are sitting in a **chair**, they are then shown putting **sunscreen** on their heads, they then begin to talk about their **experience**, the woman is now putting on her **legs** and she is **doing a tutorial** on how to use the **sunscreen**, the **news reporter** talks to the **camera**.

Ground Truth: A musician plays saxophone on front people sitting in a room, a person in the audience flips the pages of a book, people applaud while the musician plays the saxophone, then, the musician end her performance and left the room while people applaud.

GVDSup: A woman is seen sitting on a chair and playing a **saxophone** while a man watches her behind her, the man plays the saxophone and the man is playing the saxophone, a man in a black shirt is playing the saxophone, the man finishes and finishes the song.

Vanilla+Object: A woman in a **dress** is playing a **saxophone**, she plays the saxophone in front of the camera, he stops playing the saxophone, the man finishes and walks away.

Ours: A woman is **standing on a stage** playing a **saxophone**, she is playing a saxophone on the stage, she stops playing and **smiles** at the camera, she finishes playing and the **audience claps.**

Figure 5: Qualitative results of ours and the baselines. The blue and bold words stand for the fine-grained or video-relevant descriptions. The underlined words mean irrelevant or incorrect descriptions.

study results. We observe that with the help of our OSTN, the performance boosts on all four metrics, when compared with Vanilla PE + Object. The extracted concepts alone (Vanilla + Concept) can help the model to gain a better performance, proving the external knowledge can bring about more semantic information. The under-supervision manner Vision-guided Concept Selection Network (VGCSN + Vanilla) greatly increases captioning ability of the model. We also notice that our model performs excellently even without object region features. To generate less redundant sentences, we train our model under unlikelihood training setting. PUlk. sign indicates paragraph-level unlikelihood training, and SUlk. sign indicates sentence-level unlikelihood training, while other experiments shown in this table are trained under MLE setting. We trade off captioning and redundancy performance by employing SUlk. to reduce the intra-sentence redundancy. The final results exhibit a good combination of the knowledge-enriched and temporal-enriched modules.

Parameter Experiment. Figure 4 and Figure 6 show how sentence supervision mechanism and the concept loss affect our captioning module. Without sentence supervision (λ=0), the captioning performance decreases due to the noises from the concepts. We also conduct experiments under different settings of non-local frame number, and the detailed results can be found in Appendix A.

Qualitative Analysis. Figure 5 shows the qualitative results of ours, GVDSup and Vanilla+Object.

Concept Set : {car, **city**, vehicle…**person**, **people**…sidewalk, walkway, **roller blade**…sky, blue…}
True concepts: {city, person, people, roller blade}
Ground Truth: a woman is seen roller blading down an alley as well as several clips of other people riding around the city.
w/o sentence supervision: a person is seen riding down a skateboard on a skateboard while the camera follows him from behind.
w/ sentence supervision: a person is seen riding around on **roller blades** while a camera follows him in the **city**.

Figure 6: An illustration for the importance of sentence supervision.

The results clearly demonstrate that the descriptions generated by our method can be fine-grained, video-relevant and coherent at the same time. Moreover, our model has the tendency to describe various events with highly abstract concepts involved in the videos, showing a deep understanding of the video contents.

5 Conclusion

In this paper, we propose a novel framework addressing on the importance of the potential relations between video concepts and objects. With the external commonsense knowledge and the internal temporal knowledge being engaged and well-designed multi-modal representation network, our model achieves high paragraph captioning performance. In the future, we will extend our work to different datasets, and employ more efficient and powerful decoder or pre-trained models.

References

2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. 2017. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228.

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

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Hui Liu and Xiaojun Wan. 2021. Video paragraph captioning as a text summarization task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han. 2020. Streamlined dense video captioning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. 2019. Adversarial inference for multi-sentence video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Kai Shen, Lingfei Wu, Fangli Xu, Siliang Tang, Jun Xiao, and Yueting Zhuang. 2020. Hierarchical attention based spatial-temporal graph-to-sequence learning for grounded video description. In *IJCAI*, pages 941–947.

Yuqing Song, Shizhe Chen, and Qin Jin. 2021. Towards diverse paragraph captioning for untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11245–11254.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Hai Wan, Jinrui Liang, Jianfeng Du, Yanan Liu, Jialing Ou, Baoyi Wang, Jeff Z Pan, and Juan Zeng. 2021. Iterative visual relationship detection via commonsense knowledge graph. *Big Data Research*, 23:100175.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-end dense

video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Yilei Xiong, Bo Dai, and Dahua Lin. 2018. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 468–483.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540.

Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang. 2020. Relational graph learning for grounded video description generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3807–3828.

Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. 2019. Grounded video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6578–6587.

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

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.

Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2019. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*.

A Appendices

A.1 Experiments on Non-local frame Number Setting

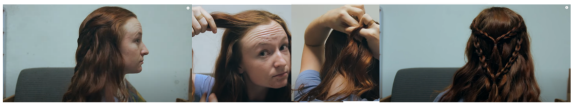
From Table 6, we can observe that attending non-local frames leads to better captioning performance. However, involving too many non-local frames may lead to noises being aggregated and lower the model captioning ability.

Non-Local Frame Counts	B@4	M	C
0	**11.8**	16.2	28.22
2	11.7	16.3	28.61
4	11.6	**16.3**	**29.19**
6	11.8	16.0	28.69

Table 6: Experimental results on different settings of attended non-local numbers setting with SUlk. training.

A.2 Additional Qualitative Examples

The additional qualitative examples demonstrate that our model can achieve high accuracy and content-rich paragraph captions in complicated and various scenarios, while the baselines tend to generate dull and video-irrelevant descriptions.



Ground Truth: A woman is showing her hair to the camera, she starts braiding her hair, she turns around and shows her hair to the camera.

GVDSup: A woman is seen sitting in a **chair** and leads into her brushing her hair and looking back to the, the girl then begins **braiding her hair** and then proceeds to braid her hair into the hair, she finishes her hair and smiles.

Vanilla+Object: A woman is seen speaking to the camera and leads into her hair, she begins **braiding her hair** and begins braiding her hair, she finishes by the camera and shows off the finished product.

Ours: A woman is seen speaking to the camera and leads into her **holding a hair and putting it into a ponytail**, she is shown **doing a tutorial** on **how to braid**, she finishes her hair and **shows off her hair**.

Ground Truth: A man is on the diving board preparing to dive standing on his head, he dives in and his score is displayed on the score board above, the audience cheers and claps at the results of the dive, then another man in a speedo facing backwards dives in to the pool.

GVDSup: A person is seen **standing on a diving board** and leads into a **person jumping off into the water**, A man jumps off the diving board and jumps into the pool, the athlete is shown again in slow motion , a man jumps off the diving board and jumps into the pool.

Vanilla+Object: A close up of a board is shown followed by a man jumping off and diving board, the girl **jumps off of the diving board** and **lands in the water**, another man is seen speaking to the camera, the diver jumps off the diving board and then jumps off of the diving board.

Ours: A camera pans around a **large indoor pool** and leads into a person walking into a pool, a man is seen **jumping off the diving board** and **jumps into the pool**, the **crowd cheers and cheers** as they jump, the crowd cheers and cheers as they **dive in the water**.

Ground Truth: A person is seen sitting behind a set of bongo drums and speaking to people off in the distance, the men then plays on the drums while stopping to speak and continuing to play.

GVDSup: A man is seen sitting in front of a drum set and begins **playing the drums**, he continues playing the drums and ends by speaking to the camera.

Vanilla+Object: A man is seen **sitting behind** a drum set playing drums and **playing the drums**, he continues playing the drums on the drums and ends by walking away.

Ours: A man is seen **sitting behind** a set of **bongo drums** while **looking to the camera**, the man **continues playing the drums** while the camera **captures his movements** and ends with him **hitting the drums**.

Ground Truth: A person is seen riding in on a horse in front of a large group of people, the person chases and calf and ropes him up while walking away, several more shots are shown of people chasing cattle in after riding on a horse.

GVDSup: The man runs and runs around the field and the man **jumps off the horse** and the man runs off, the man runs and runs around the field and the **bull** runs and runs around the field, the man rides the **calf** and runs around the field.

Vanilla+Object: A man is seen **riding a horse** and leads into a man running around a field and leads into a man running around a field, the man runs around the horse and begins to run around the field, the man continues **riding around the arena** and ends by walking away.

Ours: A man is seen **riding on a horse** and a man **chases a calf**, the man throws the calf and **runs back to the horse**, the man **throws the rope around and ties it up**.

Figure 7: More qualitative results of ours and the baselines. The blue and bold words stand for the fine-grained or video-relevant descriptions. The underlined words mean irrelevant or incorrect descriptions.