Concept Parser with Multi-modal Graph Learning for Video Captioning

Bofeng Wu, Buyu Liu, Peng Huang, Jun Bao, Xi Peng, Jun Yu*, *Senior Member, IEEE*

***Abstract*—Conventional video captioning methods are either stage-wise or simple end-to-end. While the former might introduce additional noise when exploiting off-the-shelf models to provide extra information, the latter suffers from lacking high-level cues. Therefore, a more desired framework should be able to capture multi-aspects of videos consistently. To this end, we present a concept-aware and task-specific model named CAT that accounts for both low-level visual and high-level concept cues, and incorporates them effectively in an end-to-end manner. Specifically, low-level visual and high-level concept features are obtained from the video transformer and concept parser of CAT. And a concept loss is further introduced to regularize the learning process of concept parser w.r.t. generated pseudo ground truth. To combine multi-level features, a caption transformer is later introduced in CAT, where visual and concept features are the inputs and caption is its output. In particular, we make critical design choices in the caption transformer to learn to exploit these cues with a multi-modal graph. This is achieved by a graph loss that enforces effective learning of intra and inter correlations between multi-level cues. Extensive experiments on three benchmark datasets demonstrate that CAT achieves 2.3 and 0.7 improvements in the CIDEr metric on MSVD and MSR-VTT compared to the state-of-the-art method SwinBERT [1], and also achieves a competitive result on VATEX.**

***Index Terms*—Video captioning, Transformer, Multi-modal Learning, Graph Learning.**

I. Introduction

As one of the most popular tasks in cross-modal learning, video captioning aims to take full advantage of vision and language information, then describe the content of the video with natural language. A family of existing approaches [2]–[14] to this problem learn to extract both low-level visual cues and extra scene [8], [15] or syntax [9], [10], [16], [17] cues to generate captions. Though achieving promising results, these methods are typically stage-wise [8], [13] and exploit off-the-shelf models that were originally designed for other tasks when performing feature extraction, such as from scene graph generation [18] and natural language translation [19], leading to interruption in gradient and addition noises when generating captions. Instead, another line of research [1] proposes an end-to-end framework where transformer is introduced to both perform feature extraction and bridge the gap between visual and caption spaces. Despite solving the above-mentioned problems of stage-wise models, this line of methods suffers from lacking high-level cues, *e.g.*, event or content. In summary, a unified framework that is capable of capturing multi-level cues in an end-to-end manner is lacking in video captioning.

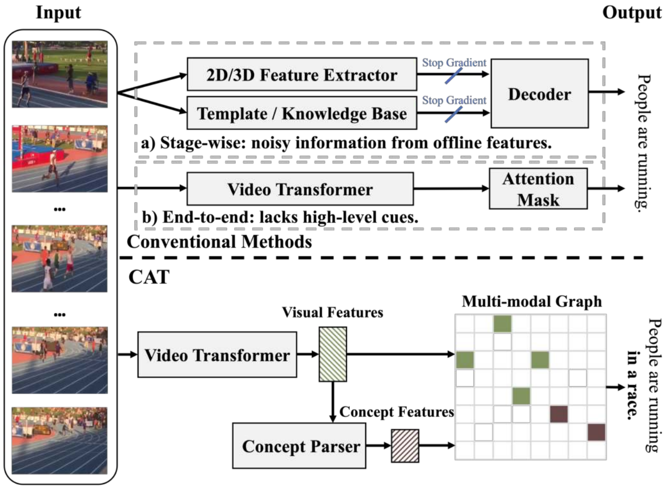


Fig. 1: Comparisons of different video captioning frameworks. Two types of conventional methods: a) A stage-wise pipeline that extracts low-level visual and extra scene or syntax cues with the off-the-shelf approaches to generate captions. b) An end-to-end framework where raw video frames are directly parsed for caption generation. Our CAT: a novel unified end-to-end framework that leverages high-level concepts and low-level visual cues extracted from video input, and further incorporates a multi-modal graph to effectively model the correlations between these two.

This unified end-to-end framework is challenging however, primarily due to a lack of supervisions at different levels and effective combinations of multi-level representations. Inspired by [20], we propose to exploit concept as high-level cues, *e.g.*, detailed events "race" in Figure 1, which is hard to produce by only visual information. On the one hand, concept lies in between visual and language spaces, which not only supplements the visual cues but also provides an efficient way of bridging the gaps between these two. On the other hand, the ground-truth concepts can be obtained from video captioning annotations effortlessly with the help of NLP tools [21], which addresses the supervision lacking problem. Moreover, we introduce a multi-modal graph to explicitly learn the relationship among three spaces, resulting in both state-of-the-

This work was supported by the National Natural Science Foundation of China under Grants 62125201, 62020106007.

*Jun Yu is the corresponding author.

B. Wu, P. Huang and J. Yu are with the Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China (email: wubofeng@hdu.edu.cn; hppp@hdu.edu.cn; yujun@hdu.edu.cn).

B. Liu is with NEC Laboratories of America (email: buyu@nec-labs.com)

J. Bao is with ZJU-Hangzhou Global Scientific and Technological Innovation Center (email: baojun@zju.edu.cn)

X. Peng is with College of Computer Science, Sichuan University (email: pengx.gm@gmail.com)

art performance and more interpretable predictions.

Specifically, we employ a transformer-based architecture for video captioning wherein both concept cues and multi-modal graph are incorporated. Our unified framework, or a **c**oncept-**a**ware and **t**ask-specific model named CAT, consists of three modules, a video transformer, a concept parser, and a caption transformer. The video transformer is in charge of extracting low-level visual features from input video sequence, which thereafter are the input of concept parser and caption transformer. Our concept parser leverages low-level features and extracts high-level concept cues, where deep-supervision of concept loss is introduced based on generated pseudo ground-truth as regularization. Later on, the caption transformer takes the output of both concept parser and video transformer, and generates captions on given video sequence. To effectively explore multi-level representations, a multi-modal graph is introduced in the caption transformer where the relationship between various spaces can be learned and modeled in an explicit manner. To the best of our knowledge, CAT is the first unified video captioning framework that is capable of capturing multi-level cues in an end-to-end fashion.

We validate our ideas on three publicly available datasets, including MSVD [22], MSR-VTT [23], and VATEX [24], and report our performances with four evaluation metrics designed for captioning tasks. We observe the SOTA performance on MSVD and MSR-VTT datasets, particularly 2.3 and 0.7 performance improvement of CIDEr [25] over the state-of-the-art methods. Extensive ablation studies also showcase the effectiveness and interpretability of our concept parser and multi-modal graph. The main contributions of this paper can be summarized as follows.

- A novel unified framework CAT to effectively learn and combine multi-level cues in the video captioning task in an end-to-end manner, which figures out the abundant information in videos and learns a common space for both visual and textual modality representations.
- A novel concept parser to extract concepts in videos, which denotes high-level cues for video understanding. Moreover, a multi-modal graph learns to fuse the visual and the concept representations during the training stage, bridging the gap between the multi-level representations.
- SOTA performances on three benchmark datasets, including MSVD, MSR-VTT, and VATEX.

II. Related Works

A. Video Captioning.

Recent researches mostly follow two types of frameworks, *i.e.*, stage-wise, and end-to-end, to solve the video captioning task. Most stage-wise methods [2]–[6], [11], [12], [26]–[28] have utilized a variety of 2D/3D feature extractors [29]–[32] or object detectors [33]–[35], to extract the low-level visual features, followed by a language decoder [36] to decode these features into captions. To supplement more information, a portion of efforts [7]–[10], [13], [14], [16], [17] proposes to extract syntax or scene cues through off-the-shelf models. In terms of syntax cues, [10] presents a syntax-aware model that generates syntax triplets from low-level visual features, and then fuses these syntax triplets and visual features for caption generation. As far as scene cues goes, [8] leverages the prior knowledge graph and a reasoning module to bridge the object and attributes that are detected offline to a commonsense, or scene graph, thus supplementing scene cues. Zhang *i.e.*, [37] propose a relational graph-based context-aware question understanding scheme, which designs a sparse graph attention network to enhance the user intention comprehension from local to global. Though the remarkable progress they achieved demonstrates these cues are beneficial to results, there still exist two challenges, *i.e.*, interruption in gradient caused by stage-wise manners, and additional noise introduced from off-the-shelf models. To deal with the first challenge stage-wise model occurs, an end-to-end framework named SwinBERT [1] is proposed and achieved promising results. SwinBERT is a transformer-based model that directly takes raw video frames as input to generate caption end-to-end, thus addressing the interruption of gradient propagation. The end-to-end approach did achieve good results, but there are no feasible methods yet in the video domain that extracts high-level cues within an end-to-end architecture. These challenges inspire us to propose a framework to extract high-level cues without the need for off-the-shelf models, and cues extraction can be merged into end-to-end training.

B. Vision-Language Models.

Previous works [38]–[42] propose pure-transformer [43] frameworks and derive superior performance in the image task. Inspired by them, recent researchers [1], [44]–[51] build video-language models and showcase great success in video tasks like video captioning [24], video question answers [52], video-textual retrieval [23]. They propose several new transformer architectures, including UniVL [50], ViViT [44], TimeSformer [45], HMTN [49] and VideoCoCa [51], that can leverage spatial-temporal attention for improving representation learning and also demonstrate the capability of the transformer in dealing with spatial-temporal sequences. In recent years, some efforts have focused on the computational efficiency of the transformer model. They propose to achieve a trade-off between speed and efficiency by variant the internal structure of the transformer. In particular, [53] prune the transformer architecture and show the close performance while model sparsity is maintained at 50%-70%. And the latest approach Video Swin Transformer [46] further variants the self-attention block of the transformer by introducing locality inductive bias into the self-attention algorithm, and achieves good performance on action recognition benchmark [32]. These works inspire us to inject an additional cues extractor into the transformer to achieve end-to-end cues extraction, and how to make the attention module enable dealing with our multi-level cues effectively and efficiently needs further consideration.

III. Methodology

In this section, we propose a novel concept-aware and task-specific framework named CAT for video captioning. We first introduce the model architecture in Sec. III-A and present the learning details in Sec. III-B.

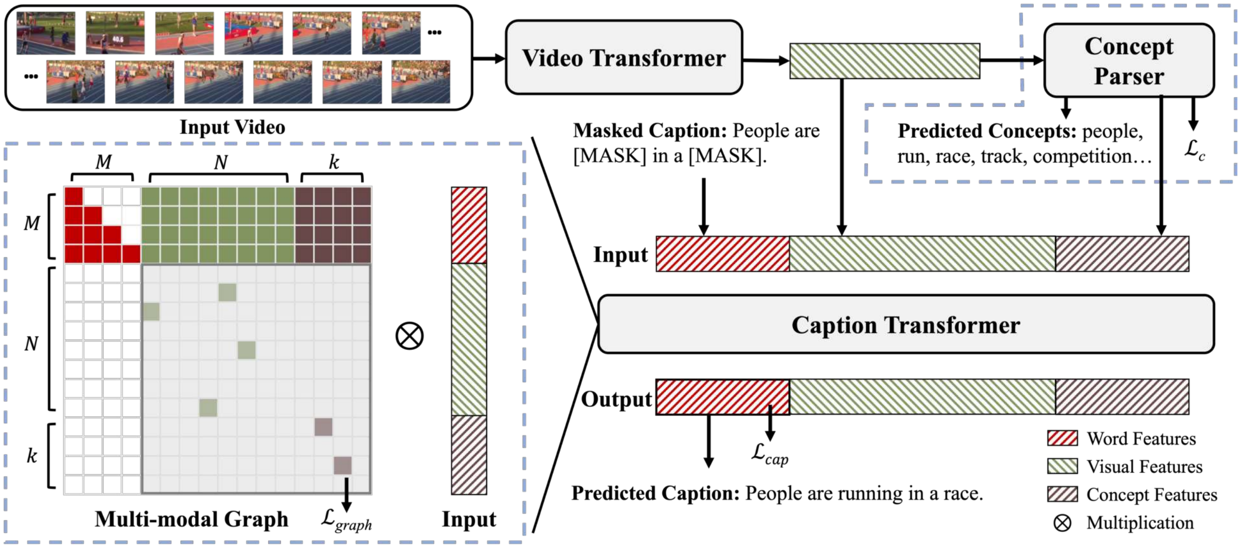


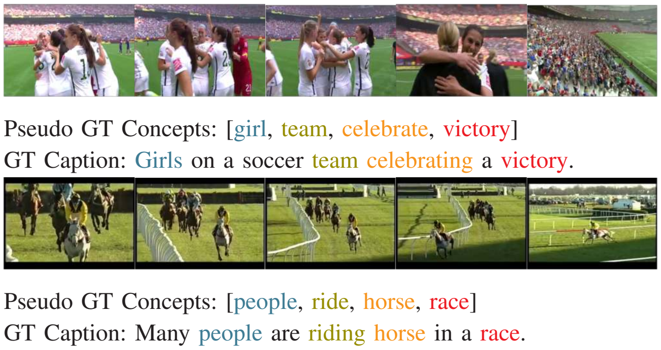
Fig. 2: Overview of our proposed framework: A sequence of raw video frames is fed into the video transformer and outputs visual features ($\boldsymbol{v}$). Then the concept parser takes $\boldsymbol{v}$ as inputs and produces concept features ($\boldsymbol{c}$). Lastly, visual and concept features are input into the caption transformer, generating a natural language sentence in a sequence-to-sequence way, with the help of multi-modal graph.

A. Model Architecture

Our proposed model consists of three modules, including a video transformer, a concept parser, and a caption transformer. The video transformer aims to encode the dense video frames by extracting features that capture low-level visual information. A concept parser later utilizes these features to generate concepts, which in the meantime also supplements high-level cues. Finally, features from the aforementioned modules are parsed to the caption transformer to produce natural language captions. We introduce each module in the following sections and provide the overall framework in Figure 2.

Video Transformer. Videos, compared to static images, provide more spatial-temporal cues. Therefore, modules that are capable of capturing these complex cues in an efficient manner are desired in video-related tasks. To this end, we introduce a Video Transformer, which inputs video frames and outputs visual features as our first module. On the one hand, the Video Transformer takes in dense frames where information loss is highly reduced. On the other hand, it modifies the attention from globality to the spatial-temporal locality to accelerate the computation. Specifically, we follow the design in [46] as they showcase good efficiency-accuracy tradedoffs in multiple video caption benchmarks. Please note that our Video Transformer is not restricted to a certain model architecture but other architectures can be also deployed. Our decision is made mainly for re-productivity and performance purposes.

We densely sample the video into $T \times H \times W \times 3$ size as the input of the video transformer, consisting of T frames and each has $H \times W \times 3$ pixels. The video transformer then outputs $N = \frac{T}{2} \times \frac{H}{32} \times \frac{W}{32}$ visual features, which we denote as $\boldsymbol{v} = \{v_i \in \mathbb{R}^D\}_i^N$, where v_i denotes i-th visual feature. N and D denote the total number and dimension of visual feature respectively. These output features are later utilized as input for subsequent concept parser and caption transformer to provide low-level visual information.


Pseudo GT Concepts: [girl, team, celebrate, victory]
GT Caption: Girls on a soccer team celebrating a victory.

Pseudo GT Concepts: [people, ride, horse, race]
GT Caption: Many people are riding horse in a race.

Fig. 3: Visualization of the two videos with their ground-truth captions and pseudo ground-truth concepts. The colored words are extracted by NLP toolkit as concepts. It is noted that the previous methods usually miss the "victory" and "race" events, resulting in rough consequences.

Concept Parser. Though the video transformer showcases good ability in terms of visual feature extraction, observations in recent work [20] demonstrate that concept is beneficial for captioning tasks. Intuitively, concepts reflect high-level information in videos, e.g. event in sports video, leading to more detailed descriptions that low-level information is less good at. Inspired by that, we introduce another module, or concept parser (CP), in video captioning. To effectively leverage caption-related concepts without requesting additional human annotations, we propose to obtain pseudo ground-truth of concept out of captioning ground-truth. Specifically, we deploy an NLP tools [21] to extract nouns and verbs of ground-truth captions as the pseudo ground-truth and we visualize some examples of obtained pseudo ground truth concepts in Figure 3.

Our concept parser consists of one pooling layer and two projection layers. The pooling layer first averages N visual features, resulting in a visual representation $\bar{\boldsymbol{v}} = \frac{1}{N}\sum_i^N v_i$,

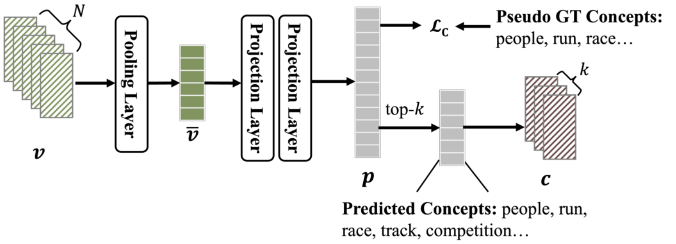


Fig. 4: Overview of our proposed concept parser. In this figure, we flatten out our visual features ($\boldsymbol{v}$) and concept features ($\boldsymbol{c}$) to help better understand the concept parser.

where $\bar{\boldsymbol{v}} \in \mathbb{R}^D$ denotes our visual-level representation. Then the projection layers receive $\bar{\boldsymbol{v}}$ as input and output concept probabilities $\boldsymbol{p} \in \mathbb{R}^K$, where K is the size of concept space. Specifically, we formulate the concept prediction problem as a multi-class classification task, where $p_{\boldsymbol{c}}$ denotes the probability of the existence of c-th concept in current video sequence. Later, we rank all concepts w.r.t. their probabilities and the top-k concepts are selected and tokenized to concept features $\boldsymbol{c} = \{c_j \in \mathbb{R}^D\}_j^k$. More details of our concept parser can be found in Figure 4.

Caption Transformer. Given visual and concept features, our next step is to integrate them in an effective manner so that information from various levels could contribute to each other. To this end, we introduce the last module of CAT, or caption transformer. Together with our multi-modal graph, the caption transformer can intellectually bridge the gap among the visual, concept, and caption spaces, resulting in natural language caption output.

The caption transformer f_{cap} takes three types of inputs, including the visual and concept features from video transformer and concept parser, as well as the sentence features $\boldsymbol{\omega} = \{\omega_i \in \mathbb{R}^D\}_i^M$ that is tokenized from masked sentence $\boldsymbol{s}_{mask}$ with an NLP tokenizer [19], where ω_i and M denote the feature vector of word in $\boldsymbol{s}_{mask}$ and the length of sentence, respectively. The goal of caption transformer f_{cap} is then to predict the masked words therefore to complete the full sentence $\boldsymbol{s}$. This is achieved by seq2seq [19]. Mathematically, we have:

$$\boldsymbol{s} = f_{cap}(\boldsymbol{s}_{mask}, \boldsymbol{v}, \boldsymbol{c}) \tag{1}$$

Multi-modal Graph. As observed in literature [46], long-range inputs of caption transformer and the activation function in attention module of transformer block [43] result in inefficiency in computation and inferior performance respectively. To alleviate these, we introduce a multi-modal graph (MG) in our caption transformer. Our multi-modal graph learns to model the relationship between multi-level features in an explicit manner, and further refines their importance. Specifically, our fully multi-modal graph $\boldsymbol{G} = \{n_p, e_{p,q}\}_{p,q}$ consists of $p \in [1, \ldots, (M{+}N{+}k)]$ nodes and $(M{+}N{+}k){\times}(M{+}N{+}k)$ edges. We denote n_p as the p-th node, which can be either word, visual, or concept feature. And $e_{p,q}$ defines the edge value between the p-th and q-th node. Our goal is to learn the $e_{p,q}$ such that our multi-modal graph captures and combines the multi-level features effectively.

We then formulate this learning process as a matrix learning problem. To this end, we represent $\boldsymbol{G}$ with a $(M+N+k) \times (M{+}N{+}k)$ matrix $\boldsymbol{A}$, where the value at position p, q is equal to $e_{p,q}$. Then we deploy this matrix at the attention module of caption transformer to refine the input of caption transformer, *i.e.*, $\boldsymbol{n} = \{n_p \in \mathbb{R}^D\}_p^{M+N+k}$, and produces refined nodes $\boldsymbol{n}'$. Mathematically, we have:

$$\begin{aligned}
&\boldsymbol{n} = [\boldsymbol{\omega}; \boldsymbol{c}; \boldsymbol{v}],\\
&\mathbf{q}, \mathbf{k}, \mathbf{v} = \boldsymbol{n}\mathbf{W_q}, \boldsymbol{n}\mathbf{W_k}, \boldsymbol{n}\mathbf{W_v},\\
&\boldsymbol{A} = \lambda(1 - \boldsymbol{A}),\\
&\boldsymbol{n}' = softmax(\boldsymbol{A} + \mathbf{q}\mathbf{k}^{\mathrm{T}}/\sqrt{d_k})\mathbf{v},
\end{aligned} \tag{2}$$

where $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$ denote query, key and value features, respectively, $\mathbf{W_q}$, $\mathbf{W_k}$, and $\mathbf{W_v}$ are three learnable projections that share same size. Furthermore, λ is a hyper-parameter, usually set to a large negative value, d_k is the dimension of $\mathbf{k}$, and $[;]$ denotes concatenation.

B. Model Learning

Our overall loss function $\mathcal{L}$ consists of three terms and our model is optimised in an end-to-end manner. Specifically, we have:

$$\mathcal{L} = \mu\mathcal{L}_{\boldsymbol{c}} + \gamma\mathcal{L}_{graph} + \mathcal{L}_{cap}, \tag{3}$$

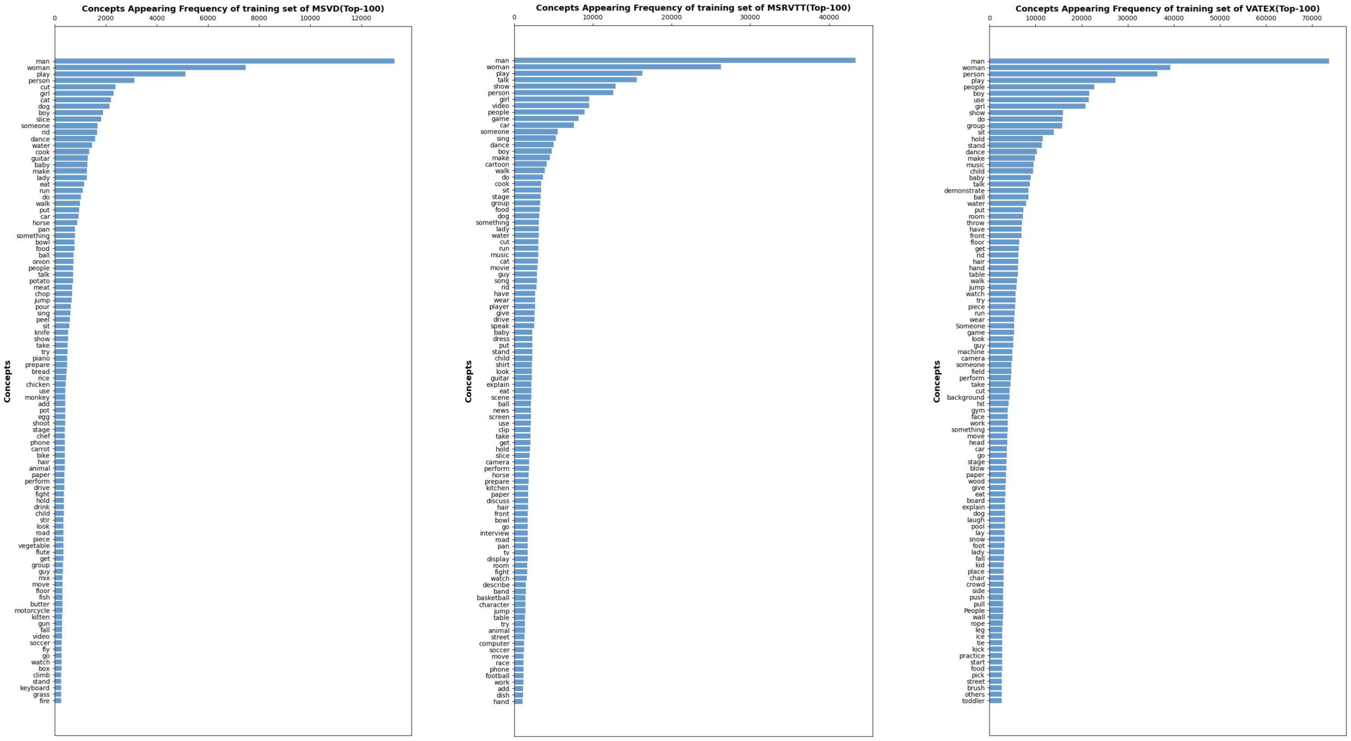
where μ and the γ are hyper-parameters. And our training object is to minimize the overall loss.

Concept Loss $\mathcal{L}_{\boldsymbol{c}}$. To ensure that features from concept parser are concept-related, we introduce a concept loss w.r.t. generated pseudo ground-truth (See Sec. III-A). Assuming that we have obtained pseudo ground-truth for concept, we adopt the cross-entropy loss as below:

$$\mathcal{L}_{\boldsymbol{c}} = -\frac{1}{K}\sum_{c=1}^{K}((1 - p_c^g)\log(1 - p_c) + p_c^g \log p_c), \tag{4}$$

where $p_c^g = \{0, 1\}$ is the pseudo label of c-th concept. And $p_c \in [0, 1]$ denotes the predicted probability of c-th concept.

Graph Learning Loss $\mathcal{L}_{graph}$. We introduce a graph learning loss $\mathcal{L}_{graph}$ to regularize our caption transformer, as we have described in Sec. III-A. Our matrix $\boldsymbol{A}$ is firstly initialized with prior knowledge and then learned with $\mathcal{L}_{graph}$. As for initialization, we induce task-specific priors to define the $e_{p,q}$. Specifically, we set the value at position p, q, where $p \in [1, ..., M]$ and $q \in [(1, ..., (M+N+k)]$, to one, which indicates that caption generation will resort to visual and concept features. Then a sequence mask [19] is deployed on the caption region, *i.e.*, the first M rows and M columns of $\boldsymbol{A}$ to achieve sequence-to-sequence captioning. Afterward, we propose to update only the visual and concept part of the matrix, or $\boldsymbol{A}' \in \mathbb{R}^{(N+k)\times(N+k)}$ consisting of only visual and concept nodes and edges between these nodes. The main intuition behind this design is that while concept and visual features can be mutually informative, they might also contribute some unrelated semantics, and our multi-modal graph should be in a position to eliminate edges that bridge these semantics as possible and reserve the significant information. To validate



MSVD MSR-VTT VATEX

Fig. 5: Stats of three benchmark datasets. We summarize the top-100 most frequently appeared concepts extracted by NLTK toolkit and demonstrate the concept distribution over datasets with three histograms.

this, we conduct ablations studies on graph initialization and report our results in Sec. IV-C.

Our graph loss is defined as follows:

$$\mathcal{L}_{graph} = \frac{1}{(N+k)\times(N+k)} \sum_{q'=1}^{N+k} \sum_{p'=1}^{N+k} A'_{p',q'}, \tag{5}$$

where $A'_{p',q'}$ equals to $e_{p'+M,q'+M}$.

Captioning Loss $\mathcal{L}_{cap}$. As described in Equ. 1 in Sec. III-A, we introduce a masked sentence as input to our caption transformer. Specifically, we follow the design in BERT [19] that masks a portion of words by replacing them with a special placeholder [MASK]. And the caption transformer predicts the true value of masked ones, which has demonstrated itself superior in captioning tasks. We implement Masked Language Modeling on the last layer of the caption transformer and produce the captioning loss $\mathcal{L}_{cap}$ [19].

IV. Experiments

In this section, we first introduce our experimental settings, and then evaluate our model on three video captioning datasets, MSVD [22], MSR-VTT [23], and VATEX [24], via four metrics including BLEU@4 [54], METEOR [55], ROUGE-L [56], and CIDEr [25]. Comprehensive ablation studies on the effectiveness of each proposed module are also conducted and reported.

A. Experimental Settings

Datasets. We mainly work on the following three datasets of various sizes and difficulties. MSVD and MSR-VTT are the most popular and most difficult one respectively, and VATEX is the largest one with long and high-quality annotations.

- **MSVD** contains 1970 YouTube short video clips. Each video is annotated with roughly 40 captions in English. We separate the dataset into 1,200 train, 100 validation, and 670 test videos, the same as previous works [1], [13], [14].
- **MSR-VTT** consists of 10,000 open-domain videos and each video is annotated with 20 English captions. We follow the official split which separates the dataset into 6,513 training, 497 validation, and 2,990 test videos [23].
- **VATEX** is a large-scale dataset that contains 41,269 videos. Each video is annotated with 10 longer and higher-quality captions in English. We follow the official split: 25,991 videos for training and 6,000 public test videos for testing [24].
- **Stats of Concepts** Figure 5 demonstrates the concept distributions over three benchmark datasets. Specifically, it summarizes the top-100 most frequently appeared concepts extracted by NLTK toolkit [21]. As can be found in this figure, there exists strong bias where these distributions are highly long-tailed. Since the videos in three datasets are generally human-dominated, nouns like “man” and “woman” appear frequently and occupy the top of the distribution. Moreover, ball sports and instrument usage also appear frequently as events, so “play” is the most frequent verb. In addition, those concepts located at the tail mean that they appear roughly the same times in the dataset.

Datasets	Video		Avg. Captions	Avg. Concepts
	Training	Testing		
MSVD	1200	670	40	27.9
MSR-VTT	6513	497	20	33.8
VATEX	25991	6000	10	27.6

TABLE I: Properties of three datasets.

Pseudo Ground-truth Concepts. In order to generate pseudo ground-truth labels for concept decoder from the above-mentioned datasets, we utilize open-sourced NLP tools [21] to extract the verbs and nouns in the ground-truth captions as the concepts. We present the scale of the training/testing set, number of captions, and the average number of concepts ground-truth in each video in Table I.

Implementation Details. We implement our model mainly with the PyTorch [58] and huggingface libraries [59], and deep-speed [60] is used for automatic mixing precision training. In our experiments, we choose Video Swin Transformer [46] and BERT [19] as video transformer and caption transformer, both the video and caption transformer consist of 12 layers, where the sizes of hidden layers of these two modules are 512 and 768 respectively, the video transformer is initialized with Kinetics-600 pre-trained weights [46] and caption transformer is initialized randomly. The sizes of two projection layers in concept parser are 512 and 30522 (*i.e.*, size of word space). In experiment, dropout rate is set to 0.1 to mitigate overfitting. To achieve a trade-off between efficiency and effectiveness, we densely sample 64 frames with the size of 224×224 of each video on all datasets as the input of the video transformer. Based on statistics reported in Table I, we set k to 25, meaning that concept decoder predicts 25 concepts for each video. Through multiple sets of experiments, we finally set M and N to 50 and 1568. In terms of loss weights, we set μ and γ both to 0.5. Similar to CTN [20], the concept shares the same vocabulary as the caption in experiments, which reduces parameter size by avoiding introducing an extra tokenization module. We exploit the AdamW optimizer [61] with a warm-up scheduler to tune the learning rate.

During training, we follow the hyper-parameter design in [19], [46] and set the number of epoch to 15 without any early stop mechanism. In the Masked Language Modeling phase, we mask half of the words in each sentence. All experiments are conducted on 8 Nvidia A100 GPUs (40GB). And it takes about 0.4h, 22h, and 42h to train the full model and 1min, 3min, and 7min to perform dataset-wise inference on MSVD, MSR-VTT, and VATEX respectively.

B. Main Results

To validate our ideas, we conduct extensive experiments on publicly available datasets and report our performance on four evaluation metrics, including BLEU@4, METEOR, ROUGE-L, and CIDEr. Our results are then compared with several state-of-the-art works [1]–[7], [9]–[14] in Table II. Specifically, results are obtained on MSVD and MSR-VTT datasets, and we also showcase the diverse architectures of the listed methods to prove the improvement carried from the end-to-end architecture. Our method outperforms existing methods on both datasets and yields significant improvements on mostly metrics. In particular, the proposed method improves the CIDEr, which is specially designed for captioning and is considered more consistent with humans, by 2.3 and 0.7 points on MSVD and MSR-VTT datasets, respectively. We believe these improvements are due to the way our model extracts multi-level cues and dynamically integrates them through a learnable multi-modal graph. Compared with previous methods that only focused on the implicit representations in the video or used kinds of fixed graphs to integrate presentations, our method can obtain more explicit video content representations, i.e., concept, and leverage the learnable graph to bridge the gap between multi-level representations, which is more in line with the video captioning task. The superior of these two parts is detailed analyses in the Ablation Studies.

We further report our performance on VATEX in Table III. Our method achieves competitive results, where most metrics, including BLEU@4, METEOR, and ROUGE-L are on par with the state-of-the-art method. The performance reveals that our method also performs well on videos that have richer and more specific content.

C. Ablation Study

In this section, we will investigate the impact of each proposed module, followed by a graph learning study to demonstrate the reasonable design of our framework. For efficient training and inference, we further lighten our model, wherein the number of input frames is reduced to 16 and the layers of the caption transformer are halved. For a fair comparison, we reproduce SwinBERT, an end-to-end and transformer-based approach, and perform the same lightweight to conduct experiments, while we also obtained the lightened results from their official GitHub. We show the results of these lightened models in Table IV.

Impact of Proposed Modules. In order to investigate the impact of the proposed two modules, *i.e.*, concept parser and multi-modal graph, we showcase the ablation study in Table V. At first, we present a baseline that consists of video and caption transformer without concept and multi-modal graph, demonstrated in the first row of Table V. Then we impose each module on the baseline independently to prove the validity of each individual one. The results in the second and third rows prove that each module is helpful to the baseline. Particularly, we observe the results in the fourth row can further prove that our two modules are compatible and complementary when impose simultaneously.

Additionally, to visualize the practical role that two modules play in our proposed method, we pick one video from MSVD and masked the key content (*i.e.*, dog) with a black box, then make CAT generate captions with the input of three groups of concepts, including pseudo ground-truth (pseudo GT), predicted, and fake, respectively. The results in the right column in Figure 6 show that the additional concept group with the "dog" or "cat" can help produce the missing subject when "dog" is masked in the video, thus demonstrating the effectiveness of concepts. While the video is unmasked, CAT generates the proper caption even when misleading concepts

Models	Features			MSVD				MSR-VTT			
	Appearance	Motion	Region	B@4	M	R	C	B@4	M	R	C
PickNet [2]	ResNet152	-	-	52.3	33.3	69.6	76.5	41.3	27.7	59.8	44.1
SibNet [3]	GoogleNet	-	-	54.2	34.8	71.7	88.2	40.9	27.5	60.2	47.5
OA-BTG [4]	ResNet200	-	MaskRCNN	56.9	36.2	-	90.6	41.4	28.2	-	46.9
GRU-EVE [5]	IncepResnetV2	C3D	YOLO	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1
MGSA [6]	IncepResnetV2	C3D	-	53.4	35.0	-	86.7	42.4	27.6	-	47.5
POS+CG [7]	IncepResnetV2	OpticalFlow	-	52.5	34.1	71.3	88.7	42.0	28.2	61.6	48.7
POS+VCT [9]	IncepResnetV2	C3D	-	52.8	36.1	71.8	87.8	42.3	29.7	62.8	49.1
SAAT [10]	IncepResnetV2	C3D	-	46.5	33.5	69.4	81.0	39.9	27.7	61.2	51.0
STG-KD [11]	ResNet101	I3D	FasterRCNN	52.2	36.9	73.9	93.0	40.5	28.3	60.9	47.1
OpenBook [12]	IncepResnetV2	C3D	-	-	-	-	-	33.9	23.7	50.2	52.9
ORG-TRL [13]	InceptionResnetV2	C3D	FasterRCNN	54.3	36.4	73.9	95.2	**43.6**	28.8	62.1	50.9
SGN [14]	ResNet101	ResNext101	-	52.8	35.5	72.9	94.3	40.8	28.3	60.8	49.5
SwinBERT [1]	Transformer			58.2	41.3	77.5	120.6	41.9	29.9	62.1	53.8
CAT	Transformer			**59.9**	**41.7**	**78.4**	**122.9**	42.1	**30.2**	**62.5**	**54.5**

TABLE II: This table reveals the performance comparison with existing methods on MSVD and MSR-VTT datasets in terms of BLEU@4(B@4), METEOR(M), ROUGE-L(R), and CIDEr(C) scores. In which the "Features" column denotes the features used by this method which are extracted by 2D-CNN, 3D-CNN, and object detector respectively.

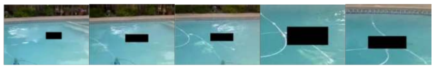
	Raw Video	Masked Video
		
Pseudo GT Concepts: [dog, pool, swim ...]	A dog is swimming in a pool.	A dog is swimming in a pool.
Predicted Concepts: [dog, pool, swim ...]	A dog is swimming in a pool.	-
Predicted Concepts: [man, pool, swim ...]	-	A man is swimming in a pool.
Fake Concepts: [cat, pool, swim ...]	A dog is swimming in a pool.	A cat is swimming in a pool.

Fig. 6: Visualization of the role that concepts play in our CAT. We generate the caption with CAT by taking in different video and concept pairs. The results present that our method is able to generate the caption associated with the input concept when the visual subject is missing.

	Raw Video	Masked Video
Pseudo GT Concepts: [cat, lick, watermelon...]	A cat is licking a piece of watermelon.	A cat is playing with a watermelon.
Predicted Concepts: [cat, lick, watermelon...]	A cat is licking a piece of watermelon.	-
Predicted Concepts: [cat, lick, clean...]	-	A cat is licking its lips.
Fake Concepts: [cat, lick, ball...]	A cat is licking a piece of watermelon.	A cat is playing with a ball.
	Raw Video	Masked Video
Pseudo GT Concepts: [girl, play, phone...]	A little girl is playing with a phone.	A little girl is playing.
Predicted Concepts: [girl, play, phone...]	A little girl is playing with a phone.	-
Predicted Concepts: [girl, play, use...]	-	A little girl is playing.
Fake Concepts: [girl, play, toy...]	A little girl is playing with a phone.	A little girl is playing with a toy.

Fig. 7: Additional visualization of the role that concepts play in our CAT. With the original and masked videos, as well as three categories of concept sets (*i.e.*, pseudo GT, predicted, and fake), CAT takes different pairs of video and concept set as inputs to predict captions, and the results further demonstrate the interpretability of our CAT.

are introduced. We argue that this finding mostly benefits from our multi-modal graph, proving that when dealing with the unmasked video, the multi-modal graph consciously reduces the importance of the concept information that mismatches the video content.

Figure 7 showcases two pairs of examples where each pair

Models	Modules		MSVD			
	CP	MG	B@4	M	R	C
Baseline	×	×	53.4	38.8	75.6	108.9
	✓	×	54.8	39.5	76.1	111.3
	×	✓	55.2	39.4	76.1	110.3
CAT	✓	✓	**56.6**	**39.6**	**76.4**	**112.5**

TABLE V: Ablation studies of concept decoder and multi-modal graph on MSVD benchmark.

Region		MSVD			
B	C	B@4	M	R	C
×	×	54.4	38.9	75.5	109.2
✓	×	53.4	38.5	74.9	107.2
×	✓	54.8	38.9	75.8	110.1
✓	✓	**56.6**	**39.6**	**76.4**	**112.5**

TABLE VI: Ablation studies of the graph initialization on MSVD benchmark.

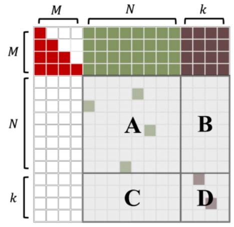


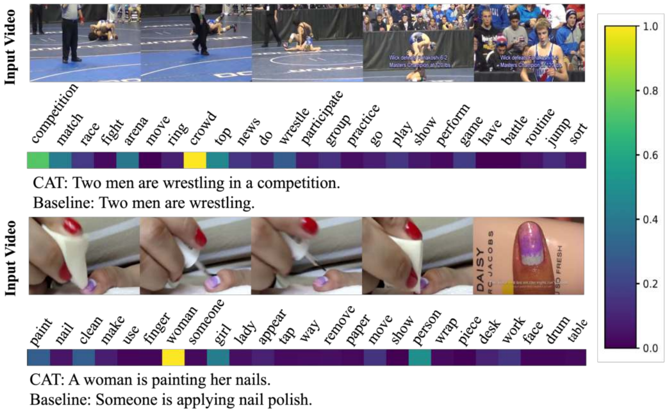
Fig. 8: Rough visualization of multi-modal graph. Four of these regions, A, B, C, and D, consist the learnable part A'.

Models	VATEX			
	B@4	M	R	C
Shared Enc [24]	28.4	21.7	47.0	45.1
Shared Enc-Dec [24]	27.9	21.6	46.8	44.2
Support-set [57]	32.8	24.4	49.1	51.2
Open-Book [12]	33.9	23.7	50.2	57.5
ORG-TRL [13]	32.1	22.2	48.9	49.7
SwinBERT [1]	**38.7**	**26.2**	53.2	**73.0**
CAT	**38.7**	**26.2**	**53.5**	72.4

TABLE III: Performance comparison on public test set of VATEX.

Models	Settings		MSVD			
	Frames	Layers	B@4	M	R	C
SwinBERT (official)	32	12	55.7	39.7	75.7	109.4
SwinBERT (repo)	16	6	55.8	39.5	76.0	110.8
CAT	16	6	**56.6**	**39.6**	**76.4**	**112.5**

TABLE IV: Results of the Lightened model.



CAT: Two men are wrestling in a competition.
Baseline: Two men are wrestling.

CAT: A woman is painting her nails.
Baseline: Someone is applying nail polish.

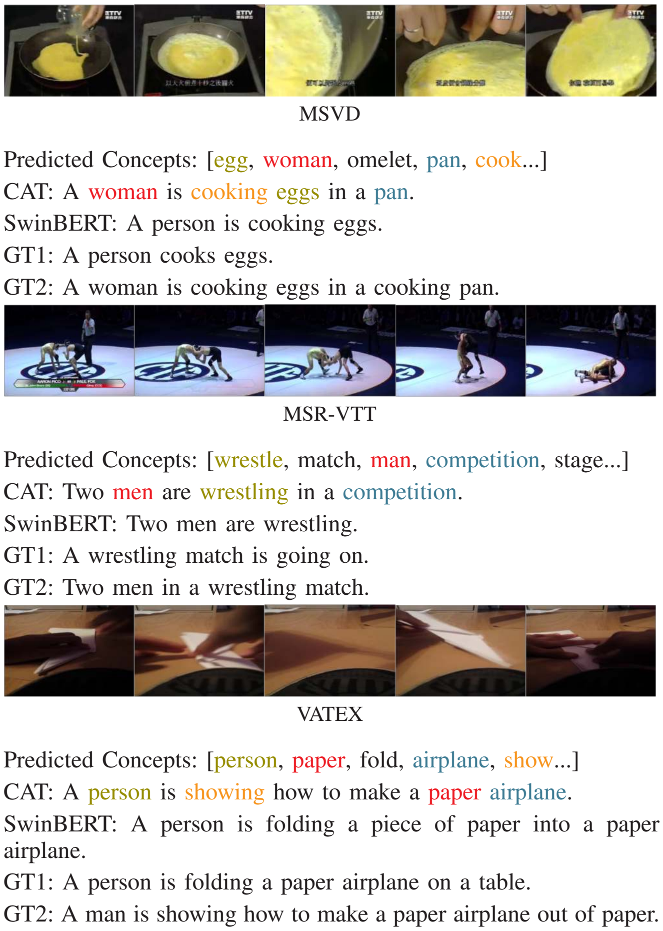
Fig. 9: Heatmap of region D in multi-modal graph.

consists of raw and masked videos, different input concepts and their outputs. Our goal here is to demonstrate the impact of concepts together with various input videos. Taking the first pair of videos as examples, we mask the watermelon in the raw video and thus produce the masked video. In addition, pseudo GT concepts are extracted by NLTK toolkit [21] from ground-truth captions, predicted concepts are generated by our concept parser based on the video content, and fake concepts are created by manually replacing the concept in pseudo GT concepts that are associated with the masked content (*i.e.*, "watermelon") with other concepts (*e.g.*, "ball"). As can be found in this example, CAT is able to generate correct captions for raw video regardless of the concepts input. When we mask the watermelon in the video and input predicted concepts, CAT assumes that the cat is licking its own lips because there is no object to lick, and therefore predicts the event as "cat-lick-lips". In addition, when we input the pseudo GT or fake concepts, CAT successfully fills the object but predicts an inexact action "play". We empirically believe this is due to a lack of visual interaction between cat and object, thus CAT tends to produce a verb that often appears together with "cat" in the dataset, *i.e.*, "play".

Such observation is not always valid across all video sequences. And we provide a corner case in our second pair. For instance, "phone" would not be occurred in output even being provided as input through pseudo GT concepts, together with masked video. In contrast, "toy" in fake concepts benefits the caption output when video is masked. We believe this is due to the fact that the event of "girl-play-phone" rarely appears in the dataset while "girl-play-toy" are more frequently occurs. Therefore, CAT takes "toy" as an important cue for caption generation rather than "phone".

Impact of Graph Initialization. One another important matter is why we reduce the interplay between irrelevant information by graph learning instead of directly blocking the interactions between multi-level features. As shown in Figure 8, the learnable part of our multi-modal can further divide into four regions, where A and D are responsible for the self-integration of visual and concept features, respectively, while B and C are in charge of cross-integration. Specifically, we can set the values of any region to zero, to enforce information not to interact. To demonstrate the validity of graph learning, we conduct experiments by deploying the zero mask on different regions in the following three combinations:1) Mask the region C. 2) Mask the region B. 3) Mask region B and region C simultaneously.

Table VI presents the results of three settings on MSVD datasets, and we also put the results of our CAT in the last row. We observe that the results after masking the regions are all worse than our CAT, proving that our learning strategy has the advantage to reduce irrelevant edges. Moreover, noting the fact that if we mask region C will result in the worst consequence, we argue that the reason is top-k concepts still have some wrong information. And without help from visual features through graph learning, the wrong information still exists and harms either the visual feature refining or caption generation. We also find that when visual features are involved in graph learning, indeed achieve better results. Referring to



MSVD

Predicted Concepts: [egg, woman, omelet, pan, cook...]
CAT: A woman is cooking eggs in a pan.
SwinBERT: A person is cooking eggs.
GT1: A person cooks eggs.
GT2: A woman is cooking eggs in a cooking pan.

MSR-VTT

Predicted Concepts: [wrestle, match, man, competition, stage...]
CAT: Two men are wrestling in a competition.
SwinBERT: Two men are wrestling.
GT1: A wrestling match is going on.
GT2: Two men in a wrestling match.

VATEX

Predicted Concepts: [person, paper, fold, airplane, show...]
CAT: A person is showing how to make a paper airplane.
SwinBERT: A person is folding a piece of paper into a paper airplane.
GT1: A person is folding a paper airplane on a table.
GT2: A man is showing how to make a paper airplane out of paper.

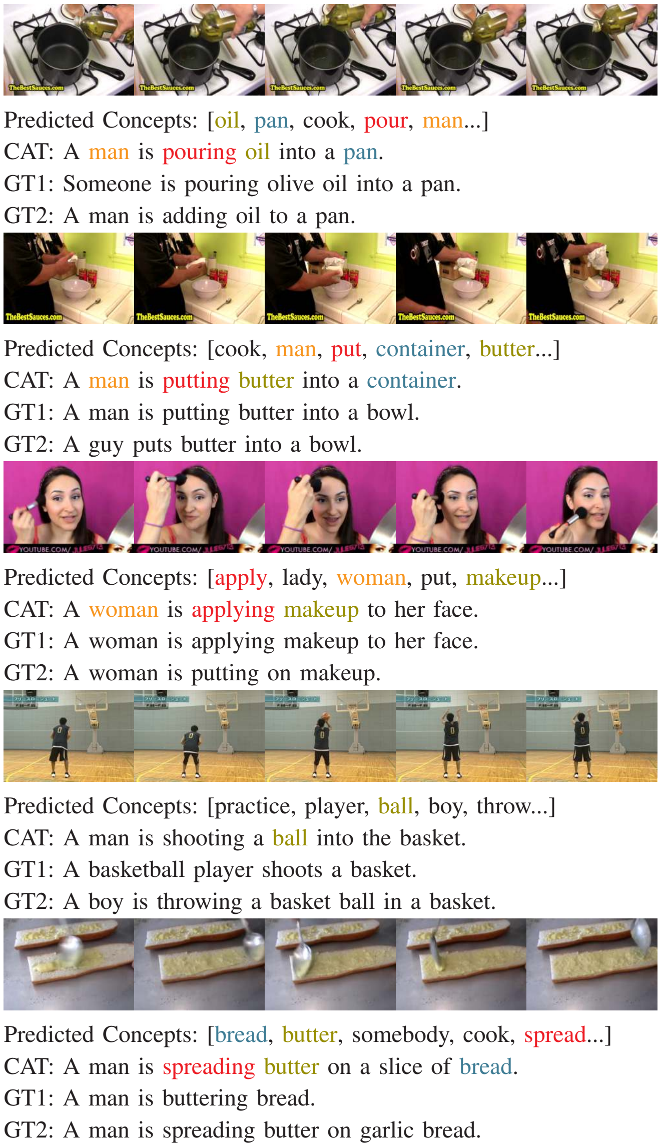
Fig. 10: Examples of predicted concepts and captions on three benchmark datasets with our proposed CAT. Furthermore, the predicted captions of SwinBERT have also been illustrated for comparison. Although both two methods produce correct captions at a coarse-grained level, the predicted captions of CAT are more detailed and diverse captions.

the predicted captions of two videos in Figure 6, we believe that thanks to our multi-modal graph, concept and visual features can flexibly contribute caption generation.

To explicitly demonstrate how the multi-modal graph influences self-integration of concepts, We further show in the heatmap of region D generated by CAT in Figure 9. The heatmap is obtained by summing the vertical axes and then scaling summed values to [0,1]. Take the second video for instance, we can find that concepts like "nail" and "finger", which can be covered in visual features, will not be highlighted in concept groups. On the contrary, those concepts denote high-level information such as "woman" and "girl", which do not appear in video, contribute more to self-integration, further producing the correct subject of the video. This fact also demonstrates the effectiveness of our multi-modal graph.

D. Qualitative Results

Figure 10 visualizes the qualitative examples of CAT together with generated concepts. We further include the results predicted by the remarkable model SwinBERT, thus intuitively showing the improvement that our approach brings. It is noted that the head predicted concepts almost hit the content in the video, thus allowing CAT to predict more detailed events than the previous method. For example, in the case video of VATEX, the "showing" usually can not be predicted by the video captioning method based on vision detection or action recognition however, relying on our concept parser, we have covered the "show" in the concept group, and further generate a caption that is close to GT2.



Predicted Concepts: [oil, pan, cook, pour, man...]
CAT: A man is pouring oil into a pan.
GT1: Someone is pouring olive oil into a pan.
GT2: A man is adding oil to a pan.

Predicted Concepts: [cook, man, put, container, butter...]
CAT: A man is putting butter into a container.
GT1: A man is putting butter into a bowl.
GT2: A guy puts butter into a bowl.

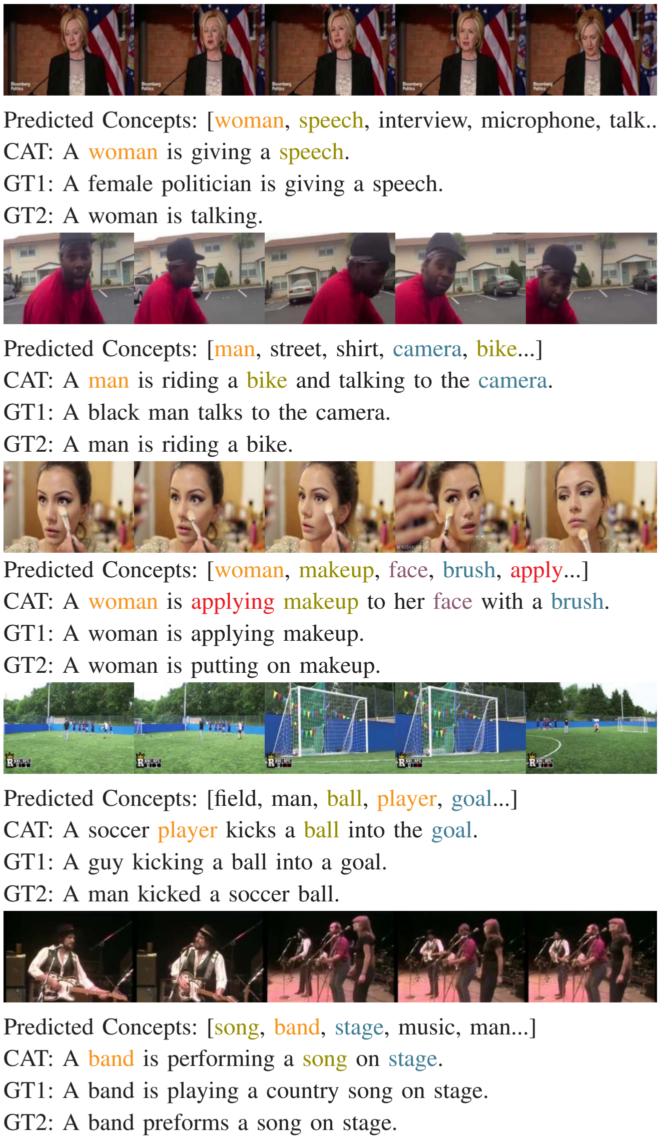
Predicted Concepts: [apply, lady, woman, put, makeup...]
CAT: A woman is applying makeup to her face.
GT1: A woman is applying makeup to her face.
GT2: A woman is putting on makeup.

Predicted Concepts: [practice, player, ball, boy, throw...]
CAT: A man is shooting a ball into the basket.
GT1: A basketball player shoots a basket.
GT2: A boy is throwing a basket ball in a basket.

Predicted Concepts: [bread, butter, somebody, cook, spread...]
CAT: A man is spreading butter on a slice of bread.
GT1: A man is buttering bread.
GT2: A man is spreading butter on garlic bread.

Fig. 11: Additional qualitative results of MSVD.

We supplement more qualitative results in Figure 11, 12, and 13. We present the predicted concepts, predicted captions and two ground-truth captions for each video. In Figure 11, we notice that the video content and captions of videos in MSVD [22] are straightforward, thus it is relatively easy to predict accurate captions on these videos. Examples in Figure 12 are from MSR-VTT [23]. Some events in example videos cannot be directly reflected visually, but can be reasoned based on visual cues, such as "speech" in the first video. Again, CAT is capable of generating highly accurate captions with most of events covered. VATEX [24] provides more descriptive captions (See Figure 13), resulting in a more


Predicted Concepts: [woman, speech, interview, microphone, talk...]
CAT: A woman is giving a speech.
GT1: A female politician is giving a speech.
GT2: A woman is talking.

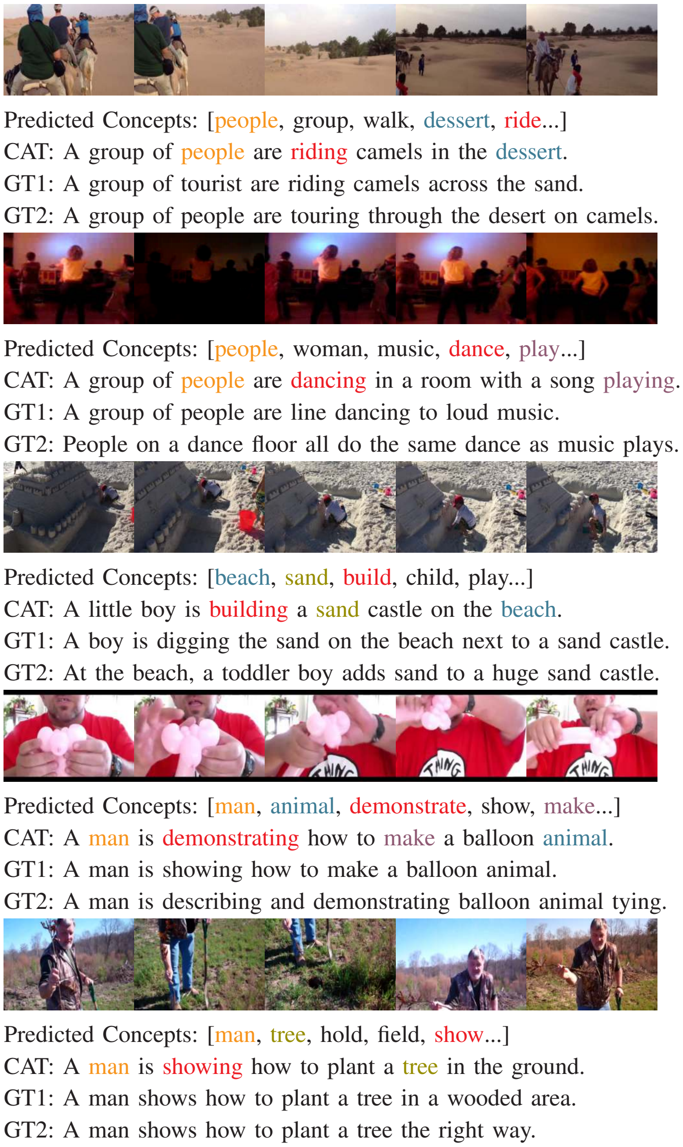
Predicted Concepts: [man, street, shirt, camera, bike...]
CAT: A man is riding a bike and talking to the camera.
GT1: A black man talks to the camera.
GT2: A man is riding a bike.

Predicted Concepts: [woman, makeup, face, brush, apply...]
CAT: A woman is applying makeup to her face with a brush.
GT1: A woman is applying makeup.
GT2: A woman is putting on makeup.

Predicted Concepts: [field, man, ball, player, goal...]
CAT: A soccer player kicks a ball into the goal.
GT1: A guy kicking a ball into a goal.
GT2: A man kicked a soccer ball.

Predicted Concepts: [song, band, stage, music, man...]
CAT: A band is performing a song on stage.
GT1: A band is playing a country song on stage.
GT2: A band preforms a song on stage.

Fig. 12: Additional qualitative results of MSR-VTT.


Predicted Concepts: [people, group, walk, dessert, ride...]
CAT: A group of people are riding camels in the dessert.
GT1: A group of tourist are riding camels across the sand.
GT2: A group of people are touring through the desert on camels.

Predicted Concepts: [people, woman, music, dance, play...]
CAT: A group of people are dancing in a room with a song playing.
GT1: A group of people are line dancing to loud music.
GT2: People on a dance floor all do the same dance as music plays.

Predicted Concepts: [beach, sand, build, child, play...]
CAT: A little boy is building a sand castle on the beach.
GT1: A boy is digging the sand on the beach next to a sand castle.
GT2: At the beach, a toddler boy adds sand to a huge sand castle.

Predicted Concepts: [man, animal, demonstrate, show, make...]
CAT: A man is demonstrating how to make a balloon animal.
GT1: A man is showing how to make a balloon animal.
GT2: A man is describing and demonstrating balloon animal tying.

Predicted Concepts: [man, tree, hold, field, show...]
CAT: A man is showing how to plant a tree in the ground.
GT1: A man shows how to plant a tree in a wooded area.
GT2: A man shows how to plant a tree the right way.

Fig. 13: Additional qualitative results of VATEX.

challenging problem. In contrast, captions generated by CAT in VATEX are less accurate compared to that in MSVD/MSR-VTT. Nevertheless, they are semantically correct w.r.t given videos.

E. Limitations and Future Works

In this work, we leverage an NLP tool to extract pseudo ground-truth concepts. Our concepts have a few of noise due to grammatical errors in the ground-truth caption. Moreover, the strategy of selecting top-k concepts often introduces inexact ones such as "do", "someone", *etc.*, in the tail of concept groups, that occur frequently in the ground-truth captions, which affects the caption generation. In our future works, we aim to develop a concept parser that can produce a variable number of concepts, which can further reduce the impact of concepts from the tail of the ordering.

V. Conclusion

In this work, we propose a unified framework, which consists of a video transformer, a concept parser, and a caption transformer. With the supervision of concept loss based on generated pseudo ground-truth, we can produce the high-level concept features within an end-to-end fashion. Furthermore, a multi-modal graph is particularly learned to better integrate the multi-level features. Extensive experimental results on three benchmark datasets verify the effectiveness of CAT.

References

[1] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "Swinbert: End-to-end transformers with sparse attention for video captioning," arXiv preprint arXiv:2111.13196, 2021.

[2] Y. Chen, S. Wang, W. Zhang, and Q. Huang, "Less is more: Picking informative frames for video captioning," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 358–373.

[3] S. Liu, Z. Ren, and J. Yuan, "Sibnet: Sibling convolutional encoder for video captioning," in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1425–1434.

[4] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8327–8336.

[5] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 487–12 496.

[6] S. Chen and Y.-G. Jiang, "Motion guided spatial attention for video captioning," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 8191–8198.

[7] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, "Controllable video captioning with pos sequence guidance based on gated fusion network," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 2641–2650.

[8] J. Hou, X. Wu, X. Zhang, Y. Qi, Y. Jia, and J. Luo, "Joint commonsense and relation reasoning for image and video captioning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 10 973–10 980.

[9] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, "Joint syntax representation learning and visual cue translation for video captioning," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8918–8927.

[10] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13 096–13 105.

[11] B. Pan, H. Cai, D.-A. Huang, K.-H. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, "Spatio-temporal graph for video captioning with knowledge distillation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10 870–10 879.

[12] Z. Zhang, Z. Qi, C. Yuan, Y. Shan, B. Li, Y. Deng, and W. Hu, "Open-book video captioning with retrieve-copy-generate network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9837–9846.

[13] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang, W. Hu, and Z.-J. Zha, "Object relational graph with teacher-recommended learning for video captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13 278–13 288.

[14] H. Ryu, S. Kang, H. Kang, and C. D. Yoo, "Semantic grouping network for video captioning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 3, 2021, pp. 2514–2522.

[15] M. Qi, Y. Wang, A. Li, and J. Luo, "Sports video captioning via attentive motion representation and group relationship modeling," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 8, pp. 2617–2633, 2019.

[16] L. Wang, H. Li, H. Qiu, Q. Wu, F. Meng, and K. N. Ngan, "Pos-trends dynamic-aware model for video caption," IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[17] J. Deng, L. Li, B. Zhang, S. Wang, Z. Zha, and Q. Huang, "Syntax-guided hierarchical attention network for video captioning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 2, pp. 880–892, 2021.

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International journal of computer vision, vol. 123, no. 1, pp. 32–73, 2017.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[20] Z. Fang, J. Wang, X. Hu, L. Liang, Z. Gan, L. Wang, Y. Yang, and Z. Liu, "Injecting semantic concepts into end-to-end image captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18 009–18 019.

[21] E. Loper and S. Bird, "Nltk: The natural language toolkit," arXiv preprint cs/0205028, 2002.

[22] D. Chen and W. 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, 2011, pp. 190–200.

[23] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5288–5296.

[24] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4581–4591.

[25] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[26] S. Chen, T. Yao, and Y.-G. Jiang, "Deep learning for video captioning: A review." in IJCAI, vol. 1, 2019, p. 2.

[27] A. Wu, Y. Han, Y. Yang, Q. Hu, and F. Wu, "Convolutional reconstruction-to-sequence for video captioning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 11, pp. 4299–4308, 2019.

[28] B. Wu, G. Niu, J. Yu, X. Xiao, J. Zhang, and H. Wu, "Towards knowledge-aware video captioning via transitive visual relationship detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 10, pp. 6753–6765, 2022.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[30] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Thirty-first AAAI conference on artificial intelligence, 2017.

[31] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.

[32] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] H. Zhang, M. Liu, Z. Gao, X. Lei, Y. Wang, and L. Nie, "Multimodal dialog system: Relational graph-based context-aware question understanding," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 695–703.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10 012–10 022.

[40] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao et al., "mplug: Effective and efficient vision-language learning by cross-modal skip-connections," arXiv preprint arXiv:2205.12005, 2022.

[41] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv preprint arXiv:2301.12597, 2023.

[42] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," arXiv preprint arXiv:2205.14100, 2022.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[44] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.

[45] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in ICML, vol. 2, no. 3, 2021, p. 4.

[46] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3202–3211.

[47] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7331–7341.

[48] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, “Frozen in time: A joint video and image encoder for end-to-end retrieval,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1728–1738.

[49] Y. Zhao, W. Nie, Z. Gao, and A.-a. Liu, “Hmtn: Hierarchical multi-scale transformer network for 3d shape recognition,” in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 316–324.

[50] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, “Univl: A unified video and language pre-training model for multimodal understanding and generation,” arXiv preprint arXiv:2002.06353, 2020.

[51] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu, “Video-text modeling with zero-shot transfer from contrastive captioners,” arXiv preprint arXiv:2212.04979, 2022.

[52] J. Lei, L. Yu, M. Bansal, and T. Berg, “TVQA: Localized, compositional video question answering,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1369–1379. [Online]. Available: https://aclanthology.org/D18-1167

[53] Z. Gan, Y.-C. Chen, L. Li, T. Chen, Y. Cheng, S. Wang, J. Liu, L. Wang, and Z. Liu, “Playing lottery tickets with vision and language,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, 2022, pp. 652–660.

[54] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[55] S. Banerjee and A. 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, 2005, pp. 65–72.

[56] C.-Y. Lin and F. J. Och, “Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics,” in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004, pp. 605–612.

[57] M. Patrick, P.-Y. Huang, Y. Asano, F. Metze, A. Hauptmann, J. Henriques, and A. Vedaldi, “Support-set bottlenecks for video-text representation learning,” arXiv preprint arXiv:2010.02824, 2020.

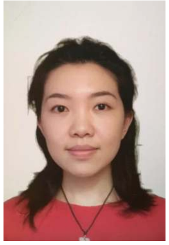
[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., “Pytorch: An imperative style, high-performance deep learning library,” Advances in neural information processing systems, vol. 32, 2019.

[59] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz et al., “Transformers: State-of-the-art natural language processing,” in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[60] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, “Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters,” in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3505–3506.

[61] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” arXiv preprint arXiv:1711.05101, 2017.

Bofeng Wu received the B.Eng. from Zhejiang Chinese Medical University, Zhejiang, China. He is currently pursuing the Ph.D. degree with the Key Laboratory of Complex Systems Modelling and Simulation in the School of Computer Science, Hangzhou Dianzi University, Zhejiang, China. His research interests include machine learning and multi-modal learning.



Buyu Liu (Member, IEEE) is currently a senior researcher at NEC Laboratory America. She was a Post-doc in the School of Informatics, at the University of Edinburgh, collaborating with Prof. Vittorio Ferrari. She received her Ph.D. degree from the ECE Department, Australian National University, under the supervision of Xuming He and Steven Gould. Buyu is broadly interested in the fields of computer vision, machine learning, and their interdisciplinary applications. Her latest interests focus on Multi-modality learning and 3D scene understanding.

Peng Huang received a bachelor's degree from Beijing Wuzi University, Beijing, China. He is currently pursuing the master's degree at Hangzhou Dianzi University, Zhejiang. His research interests include multimodal machine learning and video understanding.

Jun Bao (Member, IEEE) is currently a researcher at Zhejiang University - Hangzhou Global Scientific and Technological Innovation Center. He received his Ph.D. degree from the School of Informatics, the University of Edinburgh, under the supervision of Richard Shillcock. His research interest includes eye tracking and video analysis. His latest interests focus on remote gaze estimation and its applications.

Xi Peng is currently a full professor at College of Computer Science, Sichuan University. His current interests mainly focus on machine learning and multi-media analysis. On these areas, he has authored more than 80 articles published in JMLR, TPAMI, IJCV, ICML, NeurIPS, and so on. Dr. Peng has served as an Associate Editor for five journals such as " IEEE Trans on SMC: Systems", a Guest Editor for four journals such as "IEEE Trans. on Neural Network and Learning Systems".

Jun Yu (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from Zhejiang University, Zhejiang, China. He was an Associate Professor with the School of Information Science and Technology, Xiamen University, Xiamen, China. From 2009 to 2011, he was with Nanyang Technological University, Singapore. From 2012 to 2013, he was a Visiting Researcher with Microsoft Research Asia (MSRA). He is currently a Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. He has authored or coauthored more than 100 scientific articles. Over the past years, his research interests have included multimedia analysis, machine learning, and image processing. In 2017, he received the IEEE SPS Best Paper Award. He has (co-)chaired several special sessions, invited sessions, and workshops. He has served as a program committee member for top conferences including CVPR, ACM MM, AAAI, IJCAI, and has served as associate editors for prestigious journals including IEEE Trans. CSVT and Pattern Recognition.