Cross-Modal Graph with Meta Concepts for Video Captioning

Hao Wang, Guosheng Lin, Steven C. H. Hoi, *Fellow, IEEE* and Chunyan Miao

***Abstract*—Video captioning targets interpreting the complex visual contents as text descriptions, which requires the model to fully understand video scenes including objects and their interactions. Prevailing methods adopt off-the-shelf object detection networks to give object proposals and use the attention mechanism to model the relations between objects. They often miss some undefined semantic concepts of the pretrained model and fail to identify exact predicate relationships between objects. In this paper, we investigate an open research task of generating text descriptions for the given videos, and propose Cross-Modal Graph (CMG) with meta concepts for video captioning. Specifically, to cover the useful semantic concepts in video captions, we weakly learn the corresponding visual regions for text descriptions, where the associated visual regions and textual words are named cross-modal meta concepts. We further build meta concept graphs dynamically with the learned cross-modal meta concepts. We also construct holistic video-level and local frame-level video graphs with the predicted predicates to model video sequence structures. We validate the efficacy of our proposed techniques with extensive experiments and achieve state-of-the-art results on two public datasets.**

***Index Terms*—Video Captioning, Vision-and-Language.**

I. Introduction

VIDEO captioning aims to give precise descriptions for input videos, which benefits many relevant applications such as human-computer interaction and video retrieval [1]. Although this task is trivial for humans, it can be very challenging for machine to achieve satisfying results. The main reasons are (i) videos contain complex spatial and temporal information within changing scenes, and (ii) captions have underlying syntax, including objects and their relationships. That means the captioning model is required to interpret input videos by identifying multiple objects as well as predicting their interactions.

Prior works [3]–[5] adopt 3D convolutional models to incorporate motion and temporal feature representations, which consider global visual information for videos but lack object-level representations. Recently, some works [6]–[8] focus on exploiting more fine-grained video representations by building graphs based on object proposals that are given by pretrained model. However, there are two main limitations of their work affecting model performance: (i) the pretrained object detector may fail to detect the undefined semantic concepts (Figure 1) in video frames and (ii) the predicate relationships between objects are not explicitly predicted.

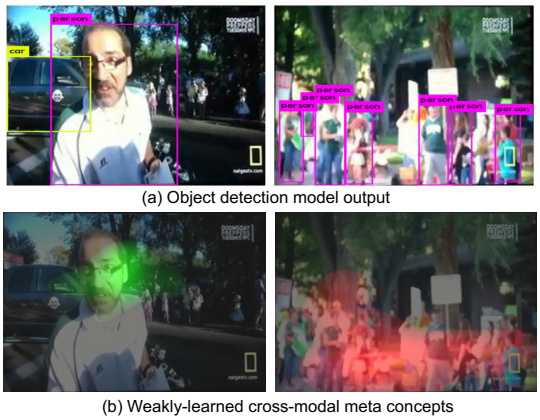


Fig. 1. **Two video frames extracted from MSR-VTT dataset as well as the caption.** We show the differences between object detection model outputs and our weakly-learned cross-modal meta concepts. In the top row, object proposals are produced with YOLOv3 model [2], which fails to detect the undefined semantic concept of pretrained model: *festival*. In the bottom row, we show the predicted visual regions for the given captioning words with our proposed model, where the green region and red region refer to *guy* and *festival* in the caption respectively.

In terms of the first limitation, as is shown in Figure 1, when we adopt a pretrained detection model to predict possible existing objects in the video, since the model is not trained on the video captioning datasets, it would miss some semantic concepts that are not defined during pretraining. For instance, in the top row of Figure 1, YOLOv3 model [2] fails to detect *festival*. Besides, there are many animation clips in given video captioning datasets [9], [10], the pretrained detection model may also fail at these scenarios. As a result, the constructed graphs can hardly provide enough fine-grained semantic information for the caption generation process. Regarding the second limitation above, previous works [6], [7] only construct soft connections between detected objects, which is realized by attention mechanism or similarity computation. This would result in the computed relationships between objects being implicit, in other words, the predicate information remains unclear. However, predicates play an important role in caption generation [8], [11], which can guide the model to be aware of the syntax and the exact interactions between objects.

To address the aforementioned limitations, we propose the Cross-Modal Graph (CMG) with meta concepts for video captioning. Specifically, to cover missed semantic concepts by

Hao Wang, Guosheng Lin and Chunyan Miao are with School of Computer Science and Engineering, Nanyang Technological University. E-mail: {hao005,gslin,ascymiao}@ntu.edu.sg.

Steven C. H. Hoi is with Singapore Management University and Salesforce Research Asia. E-mail: chhoi@smu.edu.sg.

Corresponding authors: Chunyan Miao and Guosheng Lin.

the pretrained detection model, we propose to learn the *cross-modal meta concepts*, consisting of the visual and semantic meta concepts, which are defined as the visual regions and their corresponding semantic words in captions. Since we do not have explicit pixel labeling, we adopt a weakly-supervised approach to uncover attended visual regions for words, which are learned through training the decoder to generate video captions. We further use the learned meta concepts as pseudo masks to train a localization model, which is incorporated into our captioning model and predicts the underlying meta concepts when processing video keyframes. After obtaining multiple predicted meta concepts, we construct meta concept graphs dynamically to output representations. In the bottom row of Figure 1, we show the learned cross-modal meta concepts, which can find the attended regions with their corresponding semantic classes. To give explicit predicate relationships between objects in videos, we also use a scene graph detection model [12] to predict object pairs along with their predicates. Based on the detected scene graphs, we then build holistic video-level and local frame-level graphs to give multi-scaled video structure representations.

Our contributions can be summarized as:

- We propose to use a weakly-supervised learning method to discover the corresponding visual regions of the given words of target captions, i.e. cross-modal meta concepts, which are used as the pseudo masks to train a localization model. This localization model is applied to predict meta concepts for the caption generation model.
- We build three types of cross-modal graph representations, i.e. the dynamically constructed meta concept graphs, holistic video-level and local frame-level video graphs from the detected scene graphs.
- We conduct extensive experiments to verify the usefulness of various modules of our model, and our model achieves state-of-the-art results on MSR-VTT [9] and MSVD [10] datasets.

II. Related Work

A. Video Understanding

Most video captioning works [3], [4], [13]–[20] are built with encoder-decoder architecture, where a CNN is adopted to extract video features and an RNN is used to recurrently generate the descriptions. Earlier work improves input video representations with different manners. Xu et al. [14] incorporate multimodal attention over LSTM to obtain video representations. Wang et al. [21] exploit a cycle-consistent idea to reproduce the visual contents after caption generation. In [22], Chen et al. introduce a frame picking module to select video keyframes. Recent video captioning works [6]–[8], [11] focus on improving correspondence between videos and captions. Both Zhang et al. [23] and Zheng et al. [8] propose to use Part-of-Speech (POS) information to guide video captioning process, where they generate objects first and consequently output actions of captions. Modeling interactions between objects in videos [6], [7] is an emerging way for video captioning. Specifically, Pan et al. [6] propose to construct a spatial-temporal graph and use knowledge distillation way to integrate the object features with video visual features. Zhang et al. [7] take an external large-scale language model to boost the original language decoder learning. Duan et al. [24] propose to use the weakly supervised learning method to address the video captioning problem. RCG [25] introduces to train an additional retrieval model for the video captioning task, Zhang et al. [25] firstly train an individual video-to-text retrieval model, where the corresponding descriptive sentences can be retrieved based on the given videos. During the caption generation phase, RCG learns a weighted sum between the retrieval words distribution and the captioning decoder output vocabulary distribution, based on which the final captions can be generated. Video reasoning [26]–[28] enables the model to build the relationships between visual concepts, which can infer objects and their interactions from videos and language. Using video reasoning can also benefit some down-steam tasks [26]–[28].

To construct better video representations, Wei et al. [29] explicitly separate the consistent features and the complementary features from the mixed information and harness their combination, aiming to improve the complementarity of different modality information in video. To reduce the computational overhead, Zheng et al. [30] design a dynamic sampling module to improve the discriminative power of learned clip-level classifiers and increase the inference efficiency during testing. SibNet [31] proposes a novel framework to extract video temporal features better, while it fails to consider the video-caption correspondence and did not use the audio features. Therefore, SibNet has inferior results than our proposed method. SwinBERT [32] adopts an end-to-end architecture, where the video features are extracted online. Technically, the video tokens and word tokens are simultaneously fed into the model, hence the attention masks can be learned, which reflect the visual semantic correspondence. However, end-to-end training requires intensive computation resources. By contrast, our proposed method is more lightweight.

B. Concept Prediction

Learning semantic concepts from visual input has been validated to be useful in the captioning task [33]–[35], where they mainly use a multi-label classification to predict the hidden high-level concepts. To be specific, You et al. [33] predict the semantic attributes from the given images before the captioning process, which are fused with the recurrent neural networks and the attention mechanism. Gan et al. [34] propose a semantic compositional network for the image captioning task, where they detect the semantic concepts from the input images and the probability of each of them is adopted to compose the parameters of the caption generator. Wu et al. [35] use the multi-label classification model to produce the predicted attributes, which can improve the performance of both the image captioning and VQA tasks on several benchmark datasets.

Recently, Zhou et al. [36] propose to use the grounding approach to find the attended visual regions for words in the given captions, which improves previous methods by introducing the localized visual representations. However, grounding

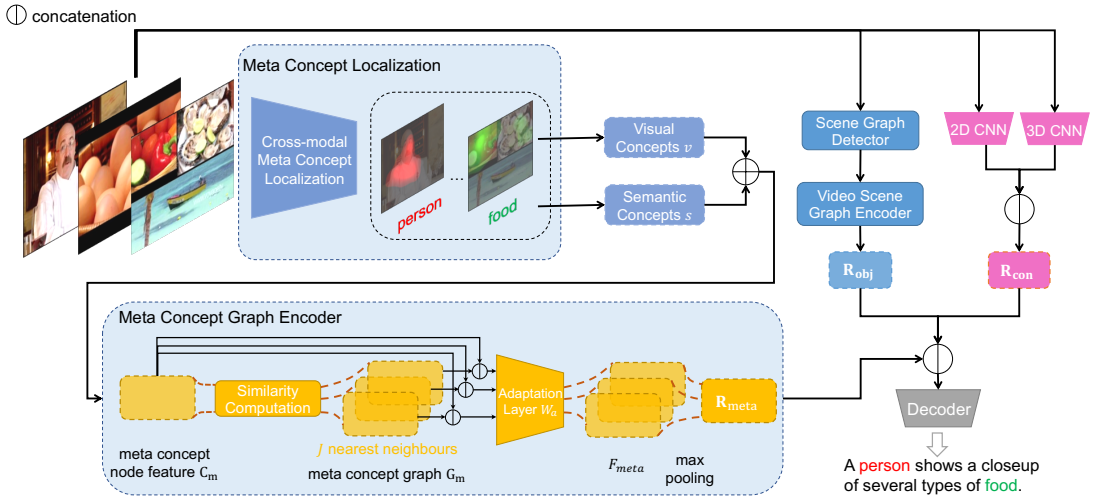


Fig. 2. **CMG: Cross-Modal Graph with meta concepts:** Our proposed framework to achieve effective video captioning, which consists of three modules: meta concept localization model, meta concept graph encoder and video scene graph encoder. In the first module, we first use a weakly-supervised method to learn cross-modal meta concepts for the given datasets, which are then used as the pseudo masks to train a localization model to output the visual regions and semantic information correlated with the target captions. In the meta concept graph encoder, to enable adaptive information flow across the video, we encode the predicted meta concepts $\mathrm{C_m}$ with a dynamic graph construction way $\mathrm{G_m}$ and give meta concept representation $\mathbf{R_{meta}}$. In the video scene graph encoder, we take a pretrained model to detect scene graphs for video frames, and then build local frame-level and holistic video-level graphs $\{\mathrm{G_f}, \mathrm{G_v}\}$ to give video object structural representation $\mathbf{R_{obj}}$. Finally, we concatenate video context features $\mathbf{R_{con}}$, $\mathbf{R_{meta}}$ and $\mathbf{R_{obj}}$ as decoder inputs to generate video captions.

methods [36], [37] are based on the object proposals that are produced by pretrained object detectors. That means these models can only find some object classes that are pre-defined. In contrast, our method is not constrained by the pretrained detector, we can predict a wider range of semantic concepts and the corresponded visual regions.

VinVL [38] and our proposed method are concurrent works. Specifically, VinVL introduced powerful detectors that can identify more than a thousand concepts including objects, relationships and attributes. However, their prediction ranges are still limited on the pretrained datasets. Given the MSR-VTT dataset contains a large portion of animations, the pre-trained models may fail in these scenarios, while our proposed method gives reasonable outputs. Moreover, VinVL requires large-scale training data with annotations, and needs many computation resources for training. By contrast, we use the weakly-supervised method for training, and the training time is much less than VinVL. We provide a solution to generate semantic concepts when ones do not have large-scale labeled training data or enough computation resources.

Specifically, instead of using object proposals detected by the pretrained models as the graph nodes [6], [7], we adopt a weakly-supervised learning approach to localize the visual regions, which are aligned with the corresponding caption semantic concepts, and use them as the nodes of our proposed meta concept graph. In this way, our method is not limited by the pretrained detector, we can predict a wider range of meta concepts.

C. Graph Models

To model non-Euclidean structures such as graphs and trees, Graph Convolution Networks (GCN) [39] are proposed to give graph structure representations. Later, graph attention networks (GAT) [40] introduce attention mechanism when encoding node features. In [41], Li et al. further explore the effect of stacking deeper layers for GCN. However, both GCN and GAT require the pre-defined edge information as the input to compute the embedded graph features. To alleviate this limitation, Wang et al. [42] compute the node relationships and aggregate neighbourhood features at each iteration, so that we can compute the graph representations without the pre-defined edges and build the relationships between nodes dynamically.

There have been some efforts to apply graph models on the captioning task [43]–[45]. To be specific, Yang et al. [43] use a pretrained detector to give predicted scene graphs of the images first, and then adopt a GCN to embed the scene graphs. In [45], semantic graph is built with directional edges on the detected object regions, where they also use GCN to embed and produce the contextual features. In terms of video related tasks, spatial and temporal information is of great significance. Xu et al. [46] perform graph convolution on segmented video snippets to leverage both spatial and temporal context for action localization. Liu et al. [47] take videos to be 3D point clouds in the spatial-temporal space. Wang et al. [48] build video graphs from the computed similarity and use GCN model to give representations for video action recognition. Zhang et al. [49] propose to tackle the problem of temporal language localization in videos with the multi-modal interaction graph convolutional network.

In this paper, we construct various types of graphs with different graph embeddings, where we use dynamic graph embedding way to model the cross-modal meta concepts across the whole video, and include the sequence information to the detected scene graphs in video frames. We also give a detailed analysis of their effects on video captioning tasks.

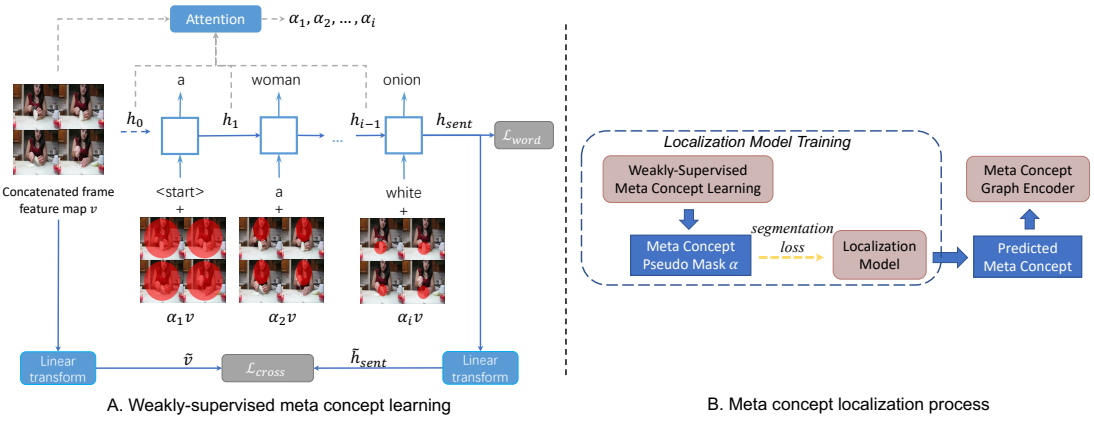


Fig. 3. **The demonstration of our proposed weakly-supervised meta concept learning process. Step A:** We use an LSTM model to localize the visual meta concepts, where the input is the concatenated feature maps v from ResNet-101. In each training step, we compute the attention map α_i from h_{i-1} and v, then use $\alpha_i v$ to generate each word, which is supervised by cross-entropy loss $\mathcal{L}_{word}$. We also give cross-modal alignments over visual and sentence features by $\mathcal{L}_{cross}$ to improve the correspondence between vision and language. **Step B:** The learned attention maps α of each caption word from step A are used as the pseudo masks, to train a localization model. The localization model takes the video frames as the inputs, and predicts the visual regions (visual meta concepts) along with the corresponding words (semantic meta concepts). The predicted cross-modal meta concepts are fed into the meta concept graph encoder to give the graph features.

III. Method

The proposed Cross-Modal Graph (CMG) with meta concepts for video captioning is presented in Figure 2. Here we define *cross-modal meta concepts* as: the visual regions and the corresponding caption words, which cover both visual and semantic information. Note that for computational efficiency, we first find the absolute difference between frames, then we sort the computed difference and pick the top N frames to be keyframes, which are inputs for our proposed model. Our goals have two folds: (i) learn informative cross-modal meta concepts and bridge the gap between videos and their descriptions and (ii) capture the complex relationships between various objects in the video. To this end, we design a novel model that consists of three modules: meta concept localization model, meta concept graph encoder and video scene graph encoder.

In the meta concept localization model, we use a weakly-supervised manner to discover the cross-modal meta concepts with caption guidance, which is shown in Figure 3. It is observed that there are multiple text descriptions for each video, we use a sentence scene graph parsing tool[1] to extract object tokens $\mathrm{T} = \{t_1, \ldots, t_k\}$ from all captions, then we group T based on synonym rules and sort them by frequency order, we take the top K groups of synonyms $\mathrm{T}' \in \mathrm{T}$ to be semantic classes. We adopt the attended visual regions for T' to be pseudo masks to train a semantic segmentation network, which is used to localize objects from T' during caption generation. Note that the localization model is trained individually to ease training difficulty, otherwise we need to jointly train the whole model and the training speed would be slow.

With the trained localization model, we are able to localize caption word $t_i \in \mathrm{T}'$ in video frames and get visual representations v_i for the attended regions. We encode the one-hot vectors of t_i as semantic information, denoting as s_i. Our predicted meta concepts can be formulated as: $\mathrm{C_m} = \{c_1, \ldots, c_L | c_i = [v_i, s_i]\}_{i=1}^{L}$, where L is the number of predicted meta concepts across the video frames. The cross-modal meta concept representation c_i is the sum of v_i and s_i. Since $\mathrm{C_m}$ have implicit interactions between each other, we propose to use a dynamic graph construction way to give output representation $\mathbf{R_{meta}}$ of meta concept graph encoder.

In the video scene graph encoder module, we adopt the method of [50] to generate the scene graphs for video frames. Technically, Tang et al. [50] propose to build a causal graph for scene graph generation. They firstly perform traditional scene graph training. In the second step, the counterfactual causality can be drawn from the trained graph, where they further infer the effect from the bad bias and attempt to remove it. With the generated scene graphs on video frames, we take two ways to build the object graph $\mathrm{G_{obj}} = \{\mathrm{G_f}, \mathrm{G_v}\}$, i.e. local frame-level graphs $\mathrm{G_f}$ and holistic video-level graphs $\mathrm{G_v}$, as shown in Figure 4. In $\mathrm{G_f}$, we first encode scene graphs through GAT [40] and obtain graph feature for each frame. We then adopt a transformer [51] to embed frame scene graph features to model sequence dependency between frames, giving $\mathbf{R}_{\mathrm{G_f}}$. We follow [6], [7] to construct $\mathrm{G_v}$, which is based on the cosine similarity and interaction over union (IoU) between adjacent frame objects. We also use GAT to encode $\mathrm{G_v}$ and output $\mathbf{R}_{\mathrm{G}_v}$. We concatenate $\mathbf{R}_{\mathrm{G_f}}$ and $\mathbf{R}_{\mathrm{G_v}}$, and get the video scene graph representations as $\mathbf{R_{obj}} = [\mathbf{R}_{\mathrm{G_f}}, \mathbf{R}_{\mathrm{G_v}}]$.

Apart from the aforementioned modules, we also follow [8], [52] to do prepossessing and extract video context representations $\mathbf{R_{con}}$ given input videos. We concatenate all obtained features together and get $\mathbf{R_{dec}} = [\mathbf{R_{con}}, \mathbf{R_{meta}}, \mathbf{R_{obj}}]$ for the decoder to recurrently generate captions, where the decoder is a one-layer plain LSTM. We train our model by two methods: minimizing the cross-entropy loss or maximizing a reinforcement learning (RL) [53] based reward.

[1]https://github.com/vacancy/SceneGraphParser

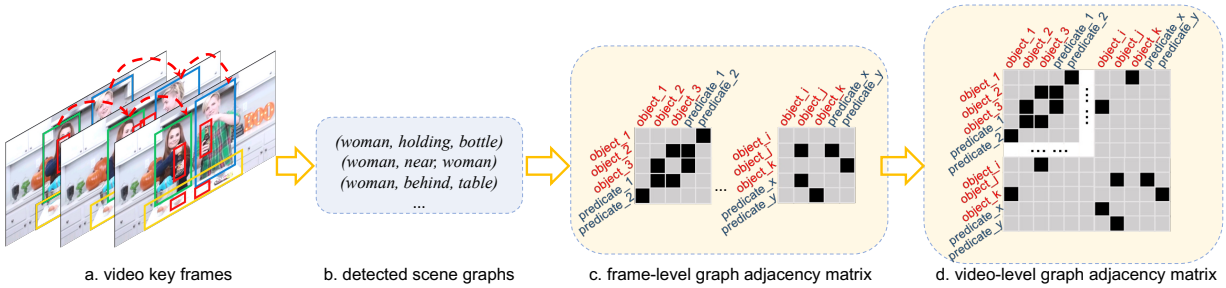


Fig. 4. **The demonstration of our constructed frame-level and video-level graphs.** We use the pretrained scene graph detector to generate the detected scene graphs for the given video keyframes. For each scene graph triplet, we construct two edges between the objects and the predicates. There are multiple frame-level scene graphs in videos, while we only plot two of them for simplicity. The video-level graph is built by grouping frame-level graphs, where we connect nodes from the adjacent frames with high similarity. The black blocks denote connections between nodes.

A. Weakly-Supervised Meta Concept Learning

The learning process is shown in Figure 3. We first learn the cross-modal meta concepts in a weakly-supervised approach, which are adopted as pseudo masks to train the localization model. Then we use the trained localization model to predict the meta concepts, which are embedded with meta concept graph encoder and further used to generate video captions. In other words, the meta concept model is trained to cover semantic concepts in captions.

Different from using the detected bounding boxes of the pretrained model, we are able to customize the classes of cross-modal meta concepts for various datasets. To this end, we adapt the model of [54] to video captioning datasets. Specifically, we randomly sample 4 frames from picked keyframes for each video and use ResNet-101 [55] to embed sampled frames to get the output feature maps from the last convolutional layer. We then concatenate 4 extracted feature maps to be the video representation v for model training. We use a LSTM to weakly learn the corresponding regions for words in captions, where we impose word-level constraints and cross-modal alignment. In the word-level training, we feed in the initialized hidden states h_0 and previous word t_{i-1} for LSTM model:

$$\alpha_i = softmax(W_f ReLU(W_v v + W_h h_{i-1})), \tag{1}$$

$$h_i = \textbf{LSTM}([W_e t_{i-1}, \alpha_i v], h_{i-1}), \tag{2}$$

where $[\cdot,\cdot]$ indicates concatenation operation, W_v, W_h, W_f and W_e denote different learnable embedding matrix and α_i corresponds to the attention weights on visual feature map v for the ith word. That means we use the frame features with attention weights to be the context vector for LSTM. The cross-entropy loss is used to decode caption sequence for word-level training:

$$p_i = softmax(W_p h_i), \tag{3}$$

$$\mathcal{L}_{word} = -\sum_i \log\, p(\hat{t}_i = t_i), \tag{4}$$

where p_i denotes the predicted probability over the vocabulary and $\hat{t}_i$ is the predicted word in training phase.

To improve the matching between video frame visual representations and caption text representations, we propose to use cross-modal alignment. Technically, we extract the final LSTM output to be sentence embedding h_{sent}, and apply linear transforms on v and h_{sent} to map them to the same dimension as $\widetilde{v}$ and $\widetilde{h}_{sent}$. We aim to impose alignment constraints over $\widetilde{v}$ and $\widetilde{h}_{sent}$, hence in the mini-batch, we sample M video-caption pairs $\{(\widetilde{v}^i, \widetilde{h}^i_{sent})\}_{i=1}^M$ from different identities, where M is the batch size. We define only $\widetilde{v}^i$ matches $\widetilde{h}^i_{sent}$ while the rest $M-1$ $\widetilde{h}_{sent}$ are all mismatched with $\widetilde{v}^i$. We adopt triplet loss to achieve the alignment, the objective is given as:

$$\begin{aligned}\mathcal{L}_{cross} = &\sum_v \Big[s(\widetilde{v}^a, \widetilde{h}^p_{sent}) - s(\widetilde{v}^a, \widetilde{h}^n_{sent}) + m\Big]_+ \\ &+ \sum_{h_{sent}} \Big[s(\widetilde{h}^a_{sent}, \widetilde{v}^p) - s(\widetilde{h}^a_{sent}, \widetilde{v}^n) + m\Big]_+,\end{aligned} \tag{5}$$

where $s(\cdot,\cdot)$ denotes the Euclidean distance measurement, superscripts a, p and n refer to anchor, positive and negative instances respectively, and m is the margin of error.

The whole training objective for weakly-supervised meta concept learning is given as:

$$\mathcal{L}_{meta} = \mathcal{L}_{word} + \lambda \mathcal{L}_{cross}, \tag{6}$$

where λ is the trade-off parameter.

To localize the corresponding meta concepts, we take α_i as pseudo masks to train another semantic segmentation model [56] to infer meta concepts when generating video captions. We first extract all semantic concepts T from given captions and group them based on synonym rules, then the top K classes of cross-modal meta concepts $\mathrm{T}' \in \mathrm{T}$ are taken to be the pseudo labels for segmentation model training. Since each video may have various meta concepts, we train the segmentation model with multi-label loss, i.e. the probability of each class is computed separately and optimized with a binary cross-entropy loss.

B. Meta Concept Graph Encoder

With the trained localization model, we can obtain a set of cross-modal meta concepts for given video: $\mathrm{C_m} = \{c_1, \ldots, c_L | c_i = [v_i, s_i]\}_{i=1}^L$, where L is the number of predicted meta concepts, v_i and s_i denote visual and semantic features from localization model output. We use the sum of v_i and s_i to be the representation of c_i. Note the span of $\mathrm{C_m}$ is

not restricted in a single frame, but covers all detected meta concepts from the picked keyframes. We propose to construct a dynamic graph $\mathrm{G_m} = \{\mathrm{C_m}, \mathrm{E_m}\}$ to integrate features of $\mathrm{C_m}$, where $\mathrm{C_m}$ is regraded as the node set and $\mathrm{E_m}$ denotes edge set.

Our target of constructing this dynamic graph is to capture interactions between semantically similar meta concepts, which can be defined by feature distance [46]. Hence we use the k-nearest neighbour algorithm to build edges between nodes as follows:

$$\mathrm{E_m} = \{(c_i, n_j^{(c_i)}) | j \in \{1, \ldots, J\}\}_{i=1}^{L}, \tag{7}$$

where $n_j^{(c_i)}$ denotes the jth neighbour of c_i. After we build edges for $\mathrm{C_m}$, we can get the adjacency matrix $\mathrm{A_m}$ for the constructed graph. Since we connect nodes dynamically during training phase, the model can keep updating node features and aggregating both intra- and inter-frame information.

Based on the obtained adjacency matrix $\mathrm{A_m}$ and node features $\mathrm{C_m}$, we perform graph convolution as follows:

$$F_{meta}(\mathrm{C_m}, \mathrm{A_m}) = ([\mathrm{C_m^T}, \mathrm{A_m C_m^T}] W_a)^T, \tag{8}$$

where $[\cdot, \cdot]$ denotes concatenation operation, W_a is a learnable adaption layer.

$$\mathbf{R_{meta}} = \mathrm{maxpool}(F_{meta}(\mathrm{C_m}, \mathrm{A_m})), \tag{9}$$

We apply a max-pooling operation on the graph convolution output $F_{meta}(\mathrm{C_m}, \mathrm{A_m})$ and obtain the cross-modal meta concept graph representation $\mathbf{R_{meta}}$.

C. Video Scene Graph Encoder

In this module, we use an off-the-shelf scene graph model [12], [50], to give video frame-level graph results $\mathrm{G_f} = \{\mathrm{G_f^1}, \ldots, \mathrm{G_f^N} | \mathrm{G_f^i} = (o_x, r_{xy}, o_y)^P\}_{i=1}^{N}$, where P denotes the number of predicted relationship triplets, N is the frame number and $o \rightarrow \mathbb{O}^{150}$, $r \rightarrow \mathbb{R}^{50}$, meaning we have 150 classes of objects and 50 types of predicate relationships.

We build edges for (o_x, r_{xy}) and (r_{xy}, o_y) respectively, and then we can obtain the adjacency matrix $\mathrm{A_f}$ for $\mathrm{G_f}$. We further extract node features for $\mathrm{G_f}$, where we take the one-hot vectors for each node and use a linear layer to encode them, then we are able to get the node features $\mathrm{F_n}$. $\mathrm{A_f}$ and $\mathrm{F_n}$ are fed into GAT [40] to obtain frame-level graph features $\mathrm{F_{G_f}} = \{\mathrm{F_{G_f^1}}, \ldots, \mathrm{F_{G_f^N}}\}$. We introduce to apply transformer [51] on $\mathrm{F_{G_f}}$ to further include the temporal dependency between frame-level graph features and give the final representation $\mathbf{R}_{\mathrm{G_f}}$ for frame-level graph $\mathrm{G_f}$.

We also construct a holistic video-level graph $\mathrm{G_v}$ to capture fine-grained temporal node connections, which builds edges between nodes is not only one single frame but also adjacent frames. Specifically, we compute the cosine similarity and Interaction over Union (IoU) between node pairs from adjacent frames. If the computed similarity and IoU are greater than the pre-defined thresholds, we connect them together. By this way, we group all $\mathrm{G_f}$ together and build $\mathrm{G_v}$. We also adopt GAT [40] to encode $\mathrm{G_v}$ and give $\mathbf{R}_{\mathrm{G_v}}$. The output representation of this module $\mathbf{R_{obj}}$ is the concatenation of $\mathbf{R}_{\mathrm{G_f}}$ and $\mathbf{R}_{\mathrm{G_v}}$.

IV. Experiments

We evaluate the efficacy of our proposed framework in two public datasets: MSR-Video To Text (MSR-VTT) dataset [9] and Microsoft Video Description (MSVD) dataset [10]. The results are obtained from four captioning metrics: BLEU, METEOR, ROUGE-L and CIDEr. "-" means number is not available. The reported results are evaluated with the Microsoft COCO evaluation server [57]. We compare our results with the previous state-of-the-art models and report extensive ablation studies to show the effectiveness of each module of our model.

A. Datasets

MSR-VTT [9]. MSR-VTT is a commonly used benchmark dataset for video captioning task. It is composed of $10,000$ video clips, where each video clip is annotated with 20 English text. These video clips are categorized into 20 classes, such as music, cooking and etc. We follow the standard splits [6]–[8], [11], i.e. there are $6,513$, 497 and $2,990$ for training, validation and testing respectively.

MSVD [10]. MSVD is a relatively small-scale dataset compared with MSR-VTT, as it in total contains $1,970$ video clips. MSVD has multilingual captions, while we only consider the English annotations. There are roughly 40 English sentences for each video clip. Similar with prior work [6]–[8], [11], the dataset is separated into $1,200$ training clips, 100 validation clips and 670 test clips.

B. Implementation Details

1) Feature Extraction: We follow [8], [60] to extract video context features for MSR-VTT dataset, and use four types of features. Specifically, we extract 2D features from the last avg-pooling layer of pretrained InceptionResnetV2 (IRV2) [61]. We adopt a C3D [62] model pretrained on Sports-1M [63] dataset to capture short-term motion features. In terms of audio features, they are extracted from audio segments within frame steps from MFCC [64]. Since MSR-VTT provides category information, we also use GloVe [65] to encode the semantic labels for each video.

For MSVD, we take 2D and 3D visual features as the video context features following the previous practice [8], [66]. Since it has limited number of training video clips, we only take two features to avoid over-fitting in this dataset, i.e. ResNeXt [67] pretrained on the ImageNet dataset is adopted to extract visual features, an ECO [68] pretrained on the Kinetics400 dataset is used to give video temporal features. Specifically, we use 32 evenly extracted video frames as the input, which are fed into ResNeXt and ECO respectively. We take the averaged ResNeXt conv5/block3 output as 2D visual features and the global pool results of ECO as 3D features.

2) Model Setting: We pick $N = 10$ keyframes to be the input based on the difference between frames, and we take top $K = 60$ synonym categories out of T to be semantic classes of cross-modal meta concepts.

In the weakly-supervised meta concept learning, we train the model with batch size of 60 and learning rate of 4×10^{-4}, and set the parameter m and λ as 0.3 and 0.5 respectively. Specifically, we use the output feature maps from the ResNet-101 last

TABLE I
MAIN RESULTS. EVALUATION OF PERFORMANCE COMPARED AGAINST VARIOUS BASELINE MODELS ON THE MSR-VTT DATASET, WE EVALUATE THE RESULTS WITH BLEU@1∼4, METEOR, ROUGE-L AND CIDER SCORES (%). WE ALSO STATE THE VIDEO CONTEXT FEATURES USED BY THE LISTED METHODS, WHERE V, G, R-*N*, A, CA, IRV2 AND ROI DENOTE VGG19, GOOGLENET, *N*-LAYER RESNET, AUDIO, CATEGORY, INCEPTIONRESNETV2 AND REGION OF INTEREST (ROI) FEATURES RESPECTIVELY. BERT AND H-LSTMS DENOTE BERT PRETRAINED MODEL AND THE HIERARCHICAL-LSTMS RESPECTIVELY. XE AND RL DENOTE TRAINING WITH CROSS-ENTROPY LOSS AND REINFORCEMENT LEARNING RESPECTIVELY.

Model	Backbone	BLEU@1	BLEU@2	BLEU@3	BLEU@4	Meteor	Rouge-L	CIDEr	Training
SA [13]	V+C3D	72.2	58.9	46.8	35.9	24.9	-	-	XE
M3 [3]	V+C3D	73.6	59.3	48.26	38.1	26.6	-	-	XE
MA-LSTM [14]	G+C3D+A	-	-	-	36.5	26.5	59.8	41.0	XE
VideoLab [5]	R-152+C3D+A+Ca	-	-	-	39.1	27.7	60.6	44.1	XE
v2t_navigator [4]	C3D+A+Ca	-	-	-	42.6	28.8	61.7	46.7	XE
RecNet [15]	InceptionV4	-	-	-	39.1	26.6	59.3	42.7	XE
OA-BTG [23]	R-200+RoI	-	-	-	41.4	28.2	-	46.9	XE
MARN [58]	R-101+C3D+Ca	-	-	-	40.4	28.1	60.7	47.1	XE
MGSA [59]	IRV2+C3D	-	-	-	42.4	27.6	-	47.5	XE
STG [6]	IRV2+I3D+RoI	-	-	-	40.5	28.3	60.9	47.1	XE
ORG-TRL [7]	IRV2+C3D+RoI+BERT	-	-	-	43.6	28.8	62.1	50.9	XE
POS-CG [11]	IRV2+I3D+Ca	79.1	66.0	53.3	42.0	28.1	61.1	49.0	XE
SAAT [8]	IRV2+C3D+Ca+RoI	80.2	66.2	52.6	40.5	28.2	60.9	49.1	XE
SibNet [31]	Temporal networks	-	-	-	41.2	27.8	60.8	48.6	XE
RCG [25]	IRV2+C3D+h-LSTMs	-	-	-	43.1	29.0	61.9	52.3	XE
CMG (ours)	IRV2+C3D	79.2	65.7	52.6	40.9	28.7	61.3	49.2	XE
CMG (ours)	IRV2+C3D+Ca	81.6	67.0	54.4	43.1	29.2	61.8	51.5	XE
CMG (ours)	IRV2+C3D+A+Ca	**83.5**	**70.7**	**57.4**	**44.9**	**29.6**	**62.9**	53.0	XE
PickNet [22]	R-152+Ca	-	-	-	41.3	27.7	59.8	44.1	RL
SAAT [8]	IRV2+C3D+Ca+RoI	79.6	66.2	52.1	39.9	27.7	61.2	51.0	RL
POS-CG [11]	IRV2+I3D+Ca	81.2	67.9	53.8	41.3	28.7	62.1	53.4	RL
CMG (ours)	IRV2+C3D+A+Ca	83.4	70.1	56.3	43.7	29.4	62.8	**55.9**	RL

convolutional layer as the video frame feature representations, the final spatial dimension is 14×14 and the feature dimension is 2048. Then we randomly sample 4 frame features and input these into the weakly-supervised meta concept learning model. The model aims to generate captions and produce the attended regions α_i with the size of $(4, 14, 14)$ for each caption token. We set a threshold to alleviate the noise in α heatmap. To specific, the threshold is set as 80, the heatmap values smaller than 80 are set as background. When we use α as the pseudo masks for training, we resize them to the same size as the input video frames. We train semantic segmentation model PSPNet [56] to localize the learned meta concepts. The segmentation model is trained with batch size of 8 and learning rate of 0.05.

In the dynamic meta concept graph construction, we allow each node to connect its $J = 3$ nearest neighbour and output 256-dimensional features. The adopted scene graph detector [12], [50] is pretrained on Visual Genome (VG) dataset [69] with Faster R-CNN [70] backbone. We use a two-layer graph attention networks (GAT) [40] to encode the constructed video graphs, where we set the hidden dimension as 8, head number as 8 and output dimension as 256. Then we use a one-layer transformer [51] with 4 heads to give temporal representations of frame-level graphs.

In the caption decoder, we use word embedding layer to give word representations, whose dimension is 512. We also map all the used visual context features onto the space of 512-dimensional space and then concatenate them together to be decoder input. We take a one-layer plain LSTM as the decoder. We train the decoder with batch size of 32 and learning rate of 8×10^{-5}. Our implemented reinforcement learning (RL) strategy is based on SCST [53]. We use beam search for evaluation, and set the beam size as 5.

TABLE II
PERFORMANCE COMPARISONS WITH DIFFERENT BASELINE METHODS ON THE TESTING SET OF THE MSVD DATASET. THE RESULTS ARE EVALUATED WITH BLEU@4, METEOR, ROUGE-L AND CIDER SCORES (%).

Model	B@4	M	R	C
MA-LSTM [14]	52.3	33.6	-	70.4
MGSA [59]	53.4	35.0	-	86.7
OA-BTG [23]	56.9	36.2	-	90.6
POS-CG [11]	52.5	34.1	71.3	88.7
SAAT [8]	46.5	33.5	69.4	81.0
STG [6]	52.2	36.9	73.9	93.0
ORG-TRL [7]	54.3	36.4	73.9	95.2
SibNet [31]	55.7	35.5	72.6	88.8
CMG (ours)	**59.5**	**38.8**	**76.2**	**107.3**

3) Model Efficiency: We run all the experiments on a single V100 GPU. We need to train three components in our proposed framework. For the weakly-supervised meta concept learning module, we set the training epoch number as 120. The total training process costs about 3 hours. For the meta concept localization (segmentation) model, we set the training epoch number as 80. The total training process costs about 1 day. For the captioning model training, we run 50 epochs for the non-RL and RL settings respectively. In the non-RL training, the total training process costs about 2 hours. In the RL training, the total training process costs about 8 hours. During inference phase, each sample costs about 0.1 s. To summarize, our proposed method is efficient for training and inference.

C. Experimental Results

1) Performance Comparison: In Table I we compare our results against earlier models under different training strategies, i.e. cross-entropy loss and reinforcement learning, on MSR-VTT dataset [9]. In table II, we show model performance in MSVD dataset [10]. It can be observed that our results gain remarkable improvement across various metrics on both MSR-VTT and MSVD datasets. We also conduct experiments with different video backbone features, to indicate the importance of multi-sourced information.

STG [6] and ORG-TRL [7] models utilize similar methodologies, which construct graphs based on object proposals. STG uses a transformer [51] as the decoder, while ORG-TRL uses an extra pretrained BERT [71] model, it gets higher results than STG. Another reason for STG having relatively low performance is that, MSR-VTT contains a large portion of animations, making pretrained object detection models often fail in these scenarios. When we use similar video context features (*IRV2+C3D*) as STG, our model outperforms STG by around 5% in CIDEr score, indicating our proposed cross-modal meta concepts can be adapted to different datasets and help alleviate such issues even without external language model.

To enable the generation model to keep aware of the syntax information, Zheng et al. [8] adopt predicate and object information to guide the language decoder in generation. In SAAT [8], instead of construing graph representations, they directly use the attention mechanism to encode the predicted predicates and objects as the input vectors for decoder. In contrast, when we use similar context features (*IRV2+C3D+Ca*) as SAAT, our model can outperform SAAT by a margin, where we propose to build holistic and local video graphs for the predicted syntax, indicating the effectiveness of our model. RCG [25] incorporates the retrieval learning into the captioning process, where they use *IRV2+C3D* as the video context features and adopt the hierarchical-LSTMs to generate captions. While here we only use the plain LSTM for the caption generation, in order to make fair comparisons with most of the previous works [8], [11], [58], [59]. It may indicate better decoder can give better generation results.

When we shift to the RL training strategy that directly optimizes our model with CIDEr scores, we achieve the highest CIDEr score. In general, the performance of our proposed model is shown to be very promising, having improvements in all metrics consistently.

2) Ablation Studies: We conduct extensive ablation studies as shown in Table III and IV and V.

Effectiveness of the meta concept graph. To observe the impact of the number of neighbours J in the dynamic meta concept (MC) graph embedding process, we change J to different values. The results show there is minor difference between different settings for J, one possible reason is: with adaptive updating on node embeddings and edge connections, nodes can aggregate around semantically similar node features. To validate the efficacy of our dynamic graph construction method, we follow [36] to build an attention-LSTM to encode the learned cross-modal meta concepts, denoting as *-attention-LSTM* in Table III, for comparison. It can be seen that the dynamic graph encoding method of our framework gives better performance than the attention-LSTM embedding method, indicating the graph embedding gives better feature aggregation. Besides, we also evaluate the usefulness of our learned visual and semantic concepts separately, the results suggest that visual features can give better performance than pure semantic features.

TABLE III
ABLATION STUDIES. EVALUATION OF THE BENEFITS OF DIFFERENT MODULES OF THE PROPOSED MODEL, WHERE J, MC, FG AND VG DENOTE THE NUMBER OF CONNECTED NEIGHBOURS, META CONCEPT GRAPHS, FRAME-LEVEL AND VIDEO-LEVEL GRAPHS RESPECTIVELY. WE ALSO SHOW THE RESULTS OF MC WITH AND WITHOUT THE DYNAMIC GRAPH ENCODING OR VISUAL/SEMANTIC FEATURES. THE RESULTS ARE EVALUATED WITH BLEU@4, METEOR, ROUGE-L AND CIDER SCORES (%) ON MSR-VTT DATASET.

Method	B@4	M	R	C
Baseline (BL)	43.0	28.1	61.5	50.2
BL + MC ($J = 20$)	44.0	29.5	62.5	51.9
BL + MC ($J = 10$)	44.1	29.5	62.6	52.0
BL + MC ($J = 3$)	44.7	29.4	62.9	52.2
- attention-LSTM	43.6	29.1	62.4	51.3
- Semantic Only	43.1	28.9	62.0	51.0
- Visual Only	43.7	29.1	62.4	51.6
BL + FG	43.8	29.1	62.1	51.4
BL + VG (no rel)	43.3	29.0	61.9	51.2
BL + VG	44.0	29.1	62.2	51.7
BL + FG + VG	44.3	29.1	62.4	51.8
All	**44.9**	**29.6**	**62.9**	**53.0**

TABLE IV
CAPTION GENERATION PERFORMANCE OF WEAKLY-SUPERVISED META CONCEPT LEARNING EVALUATED WITH BLEU@4 AND TOP-5 ACCURACY (%) AT MSR-VTT AND MSVD DATASETS, WHERE CA DENOTES CROSS-MODAL ALIGNMENT.

	Methods	B@4	Top-5 Acc
MSR-VTT	without CA	15.7	61.5
	with CA	**16.3**	**63.3**
MSVD	without CA	24.8	68.6
	with CA	**26.1**	**70.7**

Effectiveness of scene graph. We compare the performance of video-level graphs (VG) without and with predicate information, which corresponds to VG (no rel) and VG respectively. In VG (no rel), which is similar with the setting in [6], [7], we only connect object nodes together without using predicate relationships. We observe that with predicates, our model can gain better results. We incrementally add frame-level graphs (FG) and VG on the baseline (BL) respectively, which can be seen both types of graphs boost baseline performance. VG give better scores regarding BLEU@4 and CIDEr, showing that graphs with more informative connections can output better representations. On the whole, we can see that each proposed module gives positive effects to our model by improving captioning performance, and they can work collaboratively with other modules to output overall boosted results.

Evaluation of weakly-supervised meta concept learning. In

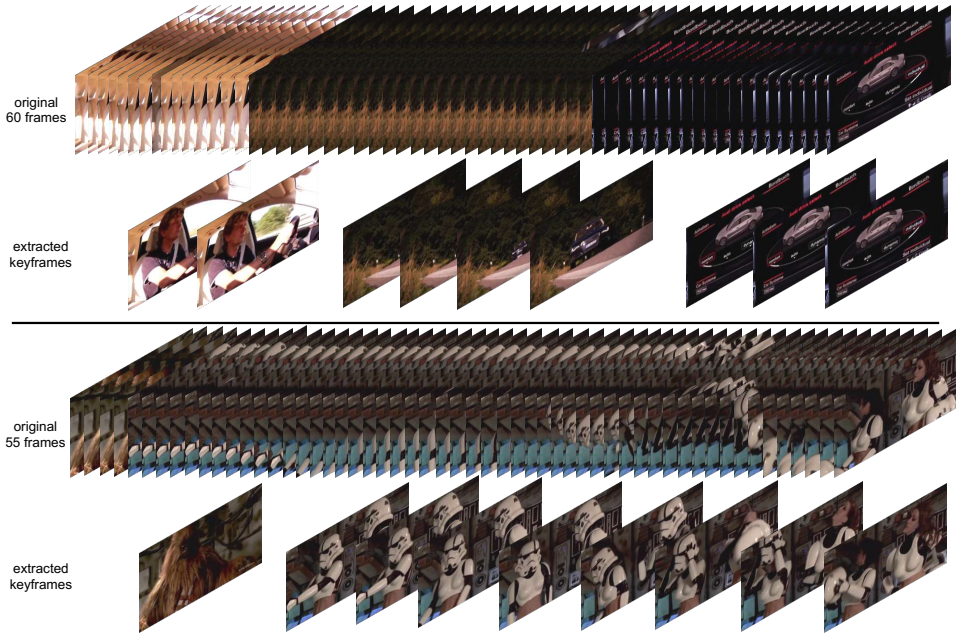


Fig. 5. **Qualitative analysis of the selected keyframes.** The upper rows show the original full video frames, the bottom rows show our generated keyframes.

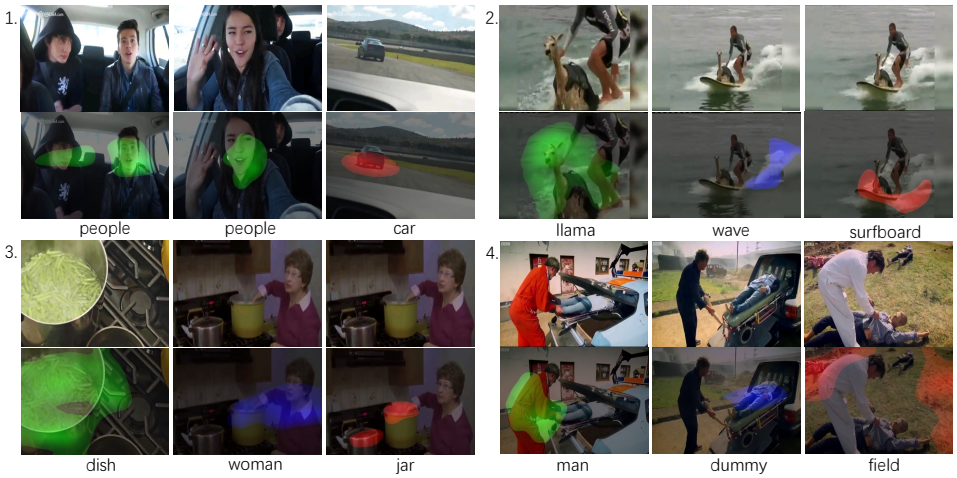


Fig. 6. **Qualitative analysis of the segmented regions of our proposed weakly-supervised learning method.** We present 4 groups of video frames and the corresponding segmented visual regions of the given semantic meta concepts. The first row are the picked video keyframes and the second row stated the learned visual meta concepts, which are the visual segmented areas. The segmented areas are obtained through the weakly-supervised learning method, which are used as the meta concepts in our proposed CMG video captioning framework.

TABLE V
ABLATION STUDY ON λ OF WEAKLY-SUPERVISED META CONCEPT LEARNING MODEL. THE PERFORMANCE IS EVALUATED WITH BLEU@4 AND TOP-5 ACCURACY (%) AT THE MSR-VTT DATASET.

λ	B@4	Top-5 Acc
0.1	15.7	61.9
0.5	**16.3**	**63.3**
1.0	16.0	62.4

Table IV, we show the efficacy of our proposed cross-modal alignment (CA) for weakly-supervised meta concept learning. In Table V, we show the ablation study on λ. It is hard to evaluate the quality for learned cross-modal meta concepts directly, as we do not have any ground truth. Hence we choose to evaluate the model caption generation performance, since the sequence is produced based on localized meta concepts, such that the generation results can reflect the quality of learned cross-modal meta concepts to some extent. It can be observed that CA can help improve captioning performance on both MSR-VTT and MSVD datasets, and setting $\lambda = 0.5$ gives the best results.

3) Qualitative Results: We show the qualitative results in Figure 5, 6, 7, 8 and 9.

The demonstration of our generated keyframes. In Figure 5, we present qualitative results to demonstrate the efficacy of our keyframe extraction method. Specifically, we decoded videos with 5 FPS. According to our observations, the decoded

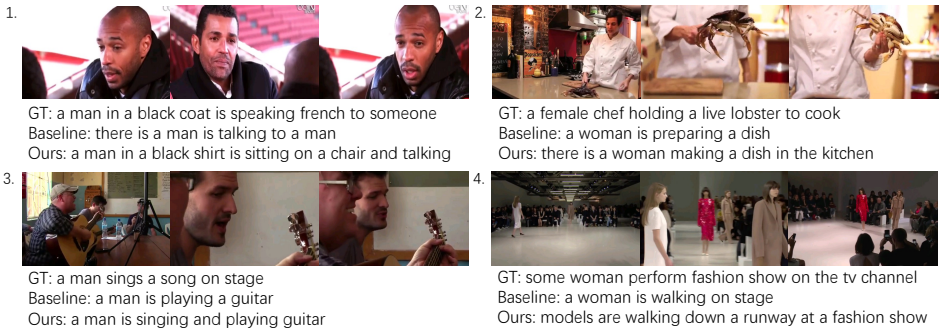


Fig. 7. **Qualitative analysis of our generated video captions.** We present 4 groups of video frames and their ground truth (GT), baseline and our model generated video captions. For the baseline model, we do not apply the proposed cross-modal graph framework on it, and it produces relatively short captions, which may lose some semantic attributes.

Fig. 8. **Qualitative analysis of failure cases.** We indicate the cross-modal meta concepts with the same colour. It is observed our model can hardly give reasonable attended visual regions for items that are not shown in the video frames.

video frames have repeated information. Our difference-based keyframe extraction method appears to successfully identify the change of actions and scenes. For example, in the upper row of Figure 5, the extracted keyframes show three different scenes of the given video, i.e., 1) a man is driving a car, 2) a car is running on the road, and 3) a demonstration of the car.

The visualization of the weakly-learned visual meta concepts. In Figure 6, we present the learned visual meta concepts with the weakly-supervised learning method. Since the only supervision for the attended semantic visual regions is the given video captions, we do not expect precise pixel labelling for the regions. However, we can still observe the proposed framework outputs mostly reasonable results. To be specific, we show the most activated regions across the video frames of the given semantic meta concepts. For example, in the third sample, the model gives the coarse region of the given semantic meta concept *dish*, which is not defined by the prevailing object detectors. Besides, based on the given captions, the model can also localize the visual regions of *jar*. These predicted visual meta concepts help give more fine-grained information compared to the traditional object detectors, which allows the model to learn different classes of meta concepts for different datasets. This property enables our proposed framework to produce some semantic information that is missed by the existing object detection methods.

The visualization of the generated captions from the videos. In Figure 7, we present the visualization examples of our generated video captions. For each video, we show the picked keyframes, ground truth (GT) captions and the captions generated by the baseline model and our proposed CMG model respectively. To be specific, in the first example, our generated caption also gives the information about the person's clothes, which is the *black shirt*, while the baseline results lose such fine-grained attributes. In the last example, our generated captions produce more descriptions on the video context: *a fashion show*. We observe our generated video captions generally contain more useful semantic information than the baseline results. The proposed cross-modal graph with the learned meta concepts guides the video captioning model to focus on the fine-grained semantic attributes of the video frames, hence it allows the model to produce textual descriptions with more sufficient desired information than the baseline model.

The visualization of the correspondence between the generated visual and semantic meta concepts. In Figure 9, we show the qualitative analysis for our learned visual and semantic meta concepts, where we visualize the video frames from different scenarios. Specifically, in the second example that is an animation clip, our learned cross-modal meta concepts can localize the visual regions of cartoon characters, while some pretrained object detection model may fail [6]. The learned meta concepts also allow the generation model to keep aware of the visual context information, such as *tournament* in the first video and *race* in the third one, which generate precise captioning words on the video context. Generally, the learned cross-modal meta concepts show promising results, where the CMG model gives more useful textual descriptions than the baseline model, thus boosts the model captioning performance.

Failure cases. Figure 8 and the last example of Figure 9 are failure cases. We observe the proposed model can hardly give reasonable attended visual regions for items that are not shown in the video frames, for instance, the *table* in the upper row of Figure 8. However, there are only small amounts of failure cases, the learned meta concepts are demonstrated to improve the baseline model performance by around 4% at CIDEr score.

Fig. 9. **Qualitative analysis of the learned cross-modal meta concepts and our generated captions.** We present 4 groups of video frames and their ground truth (GT), baseline and our model generated captions, where the first row are the picked keyframes and the second row stated the learned visual meta concepts. We indicate the cross-modal meta concepts with the same color in the GT captions, and the underline generated words denote correspondence with the localized meta concepts.

V. Conclusion

In this paper, we propose CMG with meta concepts for video captioning. Specifically, we use a weakly-supervised learning approach to localize the attended visual regions and their semantic classes for objects shown in captions, in an attempt to cover some undefined classes of pretrained models. We then use dynamic graph embeddings to aggregate semantically similar nodes and give meta concept representations. To include predicate relationships between objects, we adopt detected scene graphs in frames to build video- and frame-level graphs and give structure representations. We conduct extensive experiments and ablation studies, and achieve state-of-the-art results on MSR-VTT and MSVD datasets for video captioning.

Acknowledgments

This research is supported, in part, by the National Research Foundation (NRF), Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-003) and under its NRF Investigatorship Programme (NRFI Award No. NRF-NRFI05-2019-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This research is supported, in part, by the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/HAIG03/2017). This research is also supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2018-003), and the MOE AcRF Tier-1 research grant: RG95/20.

References

[1] Y. Yu, H. Ko, J. Choi, and G. Kim, "End-to-end concept word detection for video captioning, retrieval, and question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3165–3173.
[2] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
[3] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, "M3: Multimodal memory modelling for video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7512–7520.
[4] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. Hauptmann, "Describing videos using multi-modal fusion," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1087–1091.
[5] V. Ramanishka, A. Das, D. H. Park, S. Venugopalan, L. A. Hendricks, M. Rohrbach, and K. Saenko, "Multimodal video description," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1092–1096.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] 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 International Conference on Computer Vision*, 2019, pp. 2641–2650.
[12] K. Tang, "A scene graph generation codebase in pytorch," 2020, https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch.

[13] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4507–4515.

[14] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention lstm networks for video captioning," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 537–545.

[15] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7622–7631.

[16] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6504–6512.

[17] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "Stylenet: Generating attractive visual captions with styles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3137–3146.

[18] X. Long, C. Gan, and G. De Melo, "Video captioning with multi-faceted attention," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 173–184, 2018.

[19] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," *arXiv preprint arXiv:1910.01442*, 2019.

[20] Z. Chen, J. Mao, J. Wu, K.-Y. K. Wong, J. B. Tenenbaum, and C. Gan, "Grounding physical concepts of objects and events through dynamic visual reasoning," *arXiv preprint arXiv:2103.16564*, 2021.

[21] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S. C. Hoi, "Learning cross-modal embeddings with adversarial networks for cooking recipes and food images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 572–11 581.

[22] 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.

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

[24] X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, and J. Huang, "Weakly supervised dense event captioning in videos," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[25] 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.

[26] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan, "Star: A benchmark for situated reasoning in real-world videos," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[27] M. Ding, Z. Chen, T. Du, P. Luo, J. Tenenbaum, and C. Gan, "Dynamic visual reasoning by learning differentiable physics models from video and language," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[28] Z. Chen, K. Yi, Y. Li, M. Ding, A. Torralba, J. B. Tenenbaum, and C. Gan, "Comphy: Compositional physical reasoning of objects and events from videos," *arXiv preprint arXiv:2205.01089*, 2022.

[29] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Transactions on Image Processing*, vol. 29, pp. 1–14, 2019.

[30] Y.-D. Zheng, Z. Liu, T. Lu, and L. Wang, "Dynamic sampling networks for efficient action recognition in videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 7970–7983, 2020.

[31] S. Liu, Z. Ren, and J. Yuan, "Sibnet: Sibling convolutional encoder for video captioning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 9, pp. 3259–3272, 2020.

[32] 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.

[33] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[34] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5630–5639.

[35] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 203–212.

[36] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, "Grounded video description," in *CVPR*, 2019.

[37] C.-Y. Ma, Y. Kalantidis, G. AlRegib, P. Vajda, M. Rohrbach, and Z. Kira, "Learning to generate grounded visual captions without localization supervision," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 2. Springer, 2020.

[38] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[40] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[41] G. Li, M. Muller, A. Thabet, and B. Ghanem, "Deepgcns: Can gcns go as deep as cnns?" in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9267–9276.

[42] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[43] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

[44] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 2117–2130, 2019.

[45] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[46] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-tad: Sub-graph localization for temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 156–10 165.

[47] X. Liu, J.-Y. Lee, and H. Jin, "Learning video representations from correspondence proposals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4273–4281.

[48] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 399–417.

[49] Z. Zhang, X. Han, X. Song, Y. Yan, and L. Nie, "Multi-modal interaction graph convolutional network for temporal language localization in videos," *IEEE Transactions on Image Processing*, 2021.

[50] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3716–3725.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[52] S. Phan, G. E. Henter, Y. Miyao, and S. Satoh, "Consensus-based sequence training for video captioning," *ArXiv e-prints*, 2017.

[53] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich, "Discriminability objective for training descriptive captions," *arXiv preprint arXiv:1803.04376*, 2018.

[54] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[55] 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.

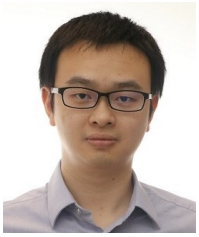
[56] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[57] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[58] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, "Memory-attended recurrent network for video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8347–8356.

[59] S. Chen and Y.-G. Jiang, "Motion guided spatial attention for video captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8191–8198.
[60] S. Phan, G. E. Henter, Y. Miyao, and S. Satoh, "Consensus-based sequence training for video captioning," *arXiv preprint arXiv:1712.09532*, 2017.
[61] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.
[62] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
[63] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.
[64] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.
[65] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
[66] H. Chen, K. Lin, A. Maye, J. Li, and X. Hu, "A semantics-assisted video captioning model trained with scheduled sampling," *arXiv preprint arXiv:1909.00121*, 2019.
[67] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.
[68] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 695–712.
[69] 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.
[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.
[71] 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.

Hao Wang is a PhD candidate with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include cross-modal generation and computer vision.



Guosheng Lin is currently an Assistant Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include computer vision and machine learning.

Steven C. H. Hoi is currently the Managing Director of Salesforce Research Asia, and a Professor of Information Systems at Singapore Management University, Singapore. He has served as the Editor-in-Chief for Neurocomputing Journal, guest editor for ACM Transactions on Intelligent Systems and Technology. He is an IEEE Fellow and ACM Distinguished Member.

Chunyan Miao is the chair of School of Computer Science and Engineering in Nanyang Technological University (NTU), Singapore. Dr. Miao is Director of the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University (NTU), Singapore. She is the Editor-in-Chief of the International Journal of Information Technology published by the Singapore Computer Society.