2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

CAVAN: Commonsense Knowledge Anchored Video Captioning

Huiliang Shao
School of Electrical,
Computer and Energy Engineering
Arizona State University
Email: hshao12@asu.edu

Zhiyuan Fang
School of Computing
and Augmented Intelligence
Arizona State University
Email: zfang29@asu.edu

Yezhou Yang
School of Computing
and Augmented Intelligence
Arizona State University
Email: yz.yang@asu.edu

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956241

***Abstract*—It is not merely an aggregation of static entities that a video clip carries, but also a variety of interactions and relations among these entities. Challenges still remain for a video captioning system to generate descriptions focusing on the prominent interest and aligning with the latent aspects beyond observations. In this work, we present a Commonsense knowledge Anchored Video cAptioNing(dubbed as CAVAN) approach. CAVAN exploits inferential commonsense knowledge to assist the training of video captioning model with a novel paradigm for sentence-level semantic alignment. Specifically, we acquire commonsense knowledge complementing per training caption by querying a generic knowledge atlas (ATOMIC [1]), and form the commonsense-caption entailment corpus. A BERT [2] based language entailment model trained from this corpus then serves as a commonsense discriminator for the training of video captioning model, and penalizes the model from generating semantically misaligned captions. Experimental results with ablations on MSR-VTT [3], V2C [4] and VATEX [5] datasets validate the effectiveness of CAVAN and reveal that the use of commonsense knowledge benefits video caption generation.**

I. Introduction

Human beings with extensive life experiences could describe observed daily events into narrative that semantically aligns with their contextual knowledge. For instance, given the video clips shown in Figure 1, one can identify the agent and the patient are "*people*" and "*food*" respectively by leveraging recognition, then supplement them with latent relations carrying interactions between the agent and the patient with multiple possibilities. The description could be as succinct as "*people are eating food*", or a verbose one, "*people are talking about food while eating*". Beyond straightforwardly narrating objects/entities of interest, an accumulation of good sense and sound judgement in practical matters connects them with latent relations, thus forming descriptions carrying prominent entities as well as suggesting probable causes, effects and attributes.

Motivated by the example in Figure 1, we argue that a video captioning system benefits from aligning descriptions semantically *w.r.t.* an inferable context (causes, effects and attributes). Advancements made in image/video sequence-to-sequence translation domain reveal the benefits of adopting the evaluation metrics (*e.g.*, CIDEr [6], BLEU [7] and SPICE [8] scores) as additional loss, together with a traditional word-level cross-entropy loss. More recently, reinforcement-based text generations, *e.g.*, policy gradient [9], [10], actor-critic [11], [12]

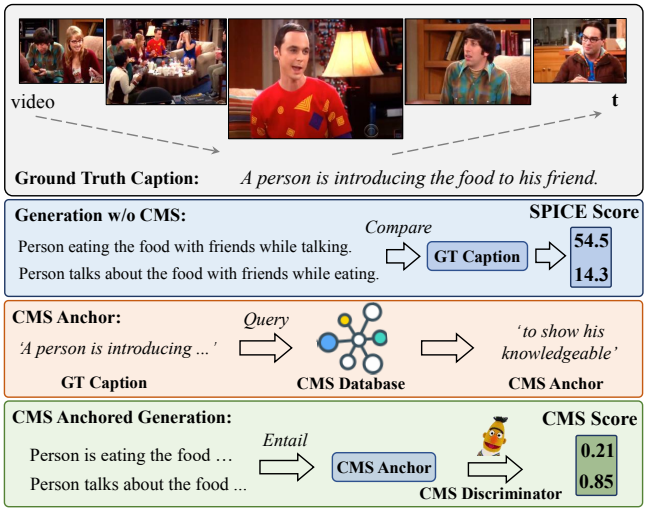

Fig. 1. We present CAVAN to address the semantic alignment for video captioning using commonsense alignment (CMS-A). Unlike traditional generation methods where the generations are only supervised by ground-truth annotations using token-level or phrase-matching metrics (*e.g.*, Cross-Entropy, CIDEr and BLEU). CMS-A leverages commonsense knowledge as anchors to constrain the overall semantics of the generated captions from deviating the current latent context.

formulate reward functions incorporating the phrase-matching metrics. But a brutal integration of the phrase-level evaluation metric based reward function could trigger severe overall semantic misalignment. Figure 1 shows such a failure case, where comparing to the ground-truth annotation: "*person is introducing the food*", a caption like "*person eating while talking*" achieves a higher SPICE score than its semantic inverse: "*person talking while eating*", even though the latter one is semantically more correct according to the ground-truth. In essence, the captioning performance training and evaluation done by the aforementioned metrics are unanimously constricted by the ground-truth annotations from the datasets, neglecting the latent and probably inferable context that is not explicitly expressed by caption annotations, *i.e.*, cause and effect, entity and its attributes at sentence-level.

In this work, we propose a novel model supervised by a sentence-level metric ensuring semantic alignment exploiting commonsense knowledge. Specifically, we first design a fusion

module that reasons over multi-model visual features and dynamically aggregates them to obtain high-level semantic feature, which is conducive to infer the otherwise neglected sentence-level context. We then leverage a commonsense knowledge atlas to query semantic anchors carrying the inferable context, and adopt a sentence level entailment score comparing generated caption with the retrieved anchors as a semantic consistency measure. We present the **C**ommonsense knowledge **A**nchored **V**ideo c**A**ptio**N**ing (dubbed as CAVAN), where the commonsense entailment loss is introduced for the first time to supplement the existing caption generation supervisions. We compile a complementary set of commonsense knowledge by querying caption annotations from the ATOMIC dataset [1] and a human curate commonsense annotations of captions from the V2C dataset [4], then retrieving a set of probable causes, effects and attributes. With the augmented and paired (caption, commonsense knowledge) data, we train a generic natural language entailment model based on BERT [2] to serve as a discriminator during training by evaluating the entailment score of each generated caption (see Figure 1).

Empirically, we test and observe that our CAVAN model makes significant improvements over the baseline models and achieves competitive results with previous state-of-the-art video captioning methods. We summarize our contributions as:

- CAVAN is the first to leverage the complementary commonsense knowledge thus impose additional contextual constraints for video captioning training, and ultimately generates captions with better aligned sentence-level semantic.
- Our ablations on CAVAN comprehensively analyze the effect of incorporating different types of knowledge and modules, providing guiding insights for future research.

II. Related Work

Traditional captioning systems [13], [14], [15], [16] are trained typically with a teacher-forcing [17] manner and evaluated using discrete and non-differential metrics. However, such training schema suffers from exposure bias [18] and the inconsistency between the optimizing function and evaluation metrics. Recent work [19], [11], [20], [21] introduce Reinforcement-Learning (RL) techniques based on policy gradient to tackle these issues. Specifically, Ranzato *et al.* [18] adopt REINFORCE algorithm to sequence training with RNNs via treating the task metrics as optimization objectives. Later, Rennie *et al.* [22] directly optimize CIDEr metric with a self-critical sequence training (SCST) approach that harmonizes the model with respect to its test-time inference procedure. More relevant efforts are further discussed in the Appendix.

III. Commonsense Knowledge Anchored Video Captioning

CAVAN's backbone is an encoder-decoder architecture based on the transformer self-attention modules [23]. A two-branch encoder takes the input of global and object features respectively, and produces attentively aggregated visual representations. Notably, we develop a novel module that dynamically reasons over attended features and alternatively fusing them based on the high-level interactions across modalities. A transformer decoder then generates the caption taking the visual representations from the specifically designed fusion module, and is supervised by both traditional video captioning losses (*i.e.*, smoothed cross-entropy) and the newly introduced commonsense entailment reinforcement loss (see Figure 2).

A. Video Encoder

Given a sequence of video frames, a couple pre-trained networks are employed to extract both global (key frames or video snippets) and entity-level features (local regional features for objects) to form a holistic representations. Specifically, we obtain the per-frame features from pre-trained 2D recognition network by sampling one key frame from every 32 frames, $V^f = [v_1^f \dots v_T^f]$, with T denotes the temporal length of videos. For motion signals, we encode every non-overlapping 32 frames by a 3D activity recognition network [24], and yields $V^m = [v_1^m \dots v_T^m]$. Following recent work in video captioning [25], [26], we extract features of the class-agnostic object proposals sampled from keyframes of the input video. Then typical candidates proposals are obtained by clustering on the sampled candidates proposals and represented by the cluster centers. Let $V^o = [v_1^o \dots v_N^o]$ denote the features of typical object proposals, where N is the number of object proposals.

We directly adopt the transformer-based visual encoder for encoding global and object features separately. Specifically, the object branch passes the features of candidate proposals V^o and generates enhanced local representations $L = [l_1 \dots l_N] \in \mathcal{R}^{N \times d}$ with interaction message between objects. The global branch takes the concatenation of appearance features V^f and motion features V^m as inputs to produce a global embedding $G = [g_1 \dots g_T] \in \mathcal{R}^{T \times d}$ of a temporal sequence, which provides additional global context that may be missing in the object branch.

B. Dynamic Fusion Module

Effective video captioning calls for a robust overall video encoding. It is critical for such encoding to incorporate representations with higher-order interactions. Existing research either apply simple concatenation [27], or a polynomial feature fusion [28]. Apart from that, **D**ynamic **M**emory **N**etworks (DMN) [29] has been applied in tasks across domains that require higher-order interactions among features, and is shown to be effective in VQA [30].

In CAVAN, we propose the **D**ynamic **F**usion **M**odule which builds on an attention module and a memory update module (dubbed as DFM). The attention module is responsible for producing global contextual representations from global features with relevance inferred by typical object features and previous memory status. Then the memory update module renews its internal episodic memory, based on the global contextual message, to retrieve new global context that were considered irrelevant during previous iteration.

Formally, given the refined global features $G = [g_1 \dots g_N]$ and object representations $L = [l_1 \dots l_N]$ from visual encoders,

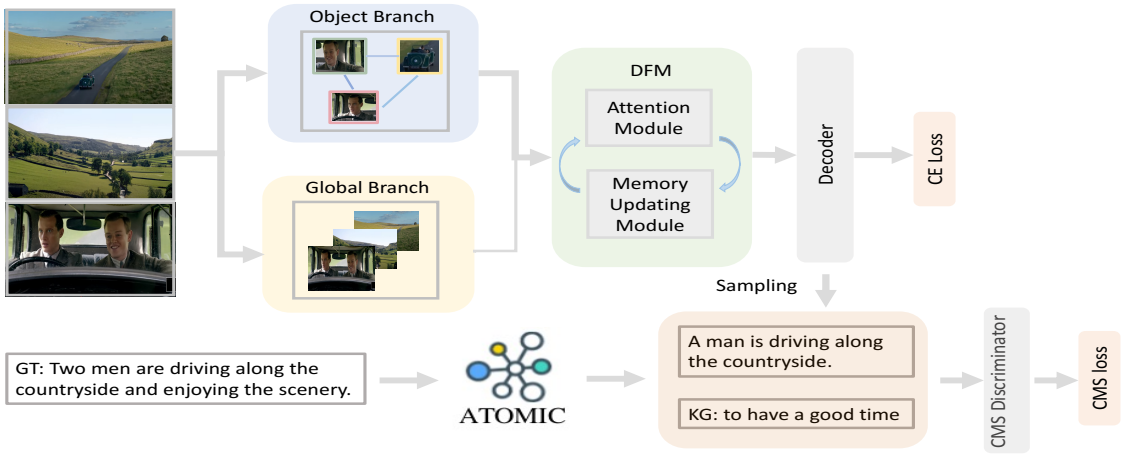

Fig. 2. Illustration of our proposed framework. CAVAN consists of transformer-based encoders and decoders, a dynamic fusion module, and a commonsense discriminator. Our model adopts a two-branch structure that generates attended object and global representations respectively. A fusion module is then adopted to fuse the outputs of two branches for decoding. The probability distribution is under the supervision of smooth cross-entropy loss. Meanwhile, a commonsense entailment loss is applied to guide the semantic alignment between current decoding captions and commonsense knowledge queried from ATOMIC [1].

an episodic memory $M = [m_1 \dots m_N]$ is initialized as $M^{(0)} = L$ and iteratively refined by an attention mechanism until the final step I is reached.

Attention Component: For the n_{th} object proposal, the attention is implemented by allowing the interaction between object feature vector $l_n \in L$ and both the global features $G = [g_1 \dots g_N]$ and previous memory states $m_n^{(i-1)} \in M^{(i-1)}$. The context $c_n^{(i)}$ is obtained by applying soft attention procedure as:

$$\begin{aligned} z_n^{(i)} &= \Big[\, G \odot l_n \;;\; G \odot m_n^{(i-1)} \;;\; |G - l_n| \;; \\ &\qquad |G - m_n^{(i-1)}| \,\Big]; \\ \alpha^{(i)} &= \textsc{Softmax}(W_2(\textsc{tanh}(W_1 z_n^{(i)} + b_1)) + b_2); \\ c_n^{(i)} &= \sum_{t=1}^{T} \alpha_t^{(i)} \cdot g_t, \end{aligned} \tag{1}$$

where $\odot$ denotes element-wise multiplication; $|\cdot|$ is the the element-wise absolute value; $[;]$ represents concatenation operation. $\alpha_t^{(i)}$ is the t_{th} element of $\alpha^{(i)}$ which denotes the normalized attention weight for g_t at i_{th} iteration. W_1, W_2, b_1 and b_2 are the parameters in the linear operation.

Memory Updating Component: The memory vector is updated as

$$m_n^{(i)} = \textsc{ReLU}(W_3[m_n^{(t-1)}; c_n^{(i)}; l_n] + b_3), \tag{2}$$

where W_3, b_3 are the parameters for the linear layer. m_n^i is the memory vector for n_{th} object proposal at the t_{th} iteration.

By the I_{th} iteration, the memory vector $m_n^{(I)}$ that memorizes the most relevant context from global features for n_{th} object proposal, is fused with the object vector l_n to generate globally contextualized object representations $\tilde{l}_n$ for decoding.

$$\tilde{l}_n = \textsc{ReLU}(W_4[l_n; m_n^{(I)}] + b_4), \tag{3}$$

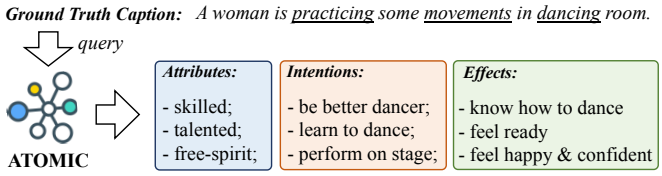

Fig. 3. Inferential commonsense knowledge retrieved from ATOMIC includes several types, *e.g.*, intentions, effects and attributes of the agents.

where W_4, b_4 are the linear parameters.

C. Language Decoder

We design the language decoder by compiling a stack of transformer attention blocks using self-attention module. During training, it takes as input of the encoded word embedding and their corresponding positional encoding [23] and attend to visual representations from the fusion module. The training criterion is based on cross-entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{T} \phi(w_t^*)' \cdot \log(P(w_t)), \tag{4}$$

where T denotes the total training step of the ground-truth captions; $\mathbf{P}(w_t)$ represents the probability distribution across the vocabulary at time t; $\phi(w_t^*)$ is the one-hot vector of ground-truth word at time t.

D. Commonsense Entailment Loss

Supervising captioning model learning with existing short textual annotations largely limits training efficacy. The semantic carried by caption only is often with weak expressive power without latent inferable context. Instead, we leverage inferable commonsense knowledge to complement each video caption and treat them as additional constrains to regularize the generating process. Figure 3 depicts an example of three types

of complementing commonsense knowledge (intention, effect, and attribute) paired with the ground-truth caption. In practice, we acquire the commonsense knowledge description k, *w.r.t.* the video caption by either retrieving from knowledge base (MSR-VTT + ATOMIC) or directly from human annotations (V2C). We discuss the commonsense knowledge retrieval procedure in Section IV-A.

Given a textual sequence $w^s = \{w_1^s \dots w_T^s\}$ sampled from language decoder, we regularize the generation by an entailment reward leveraging the commonsense description k. Intuitively, we encourage the model to caption by entailing the commonsense knowledge. To enable optimizing over non-differentiable metrics, previous efforts adopt the policy gradient approach [18], [22] and treat the task as a reinforcement learning one, with the testing metrics as the reward function. Particularly, Pasunuru *et al.* [31] implement an entailment-enhanced score from a pre-trained model as the reward. Formally, with the policy gradient strategy, model like an active agent which generates words (as action) and the learning process is supervised by minimizing the negative expected reward function:

$$\mathcal{L}(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s)], \tag{5}$$

where p_θ is the policy and θ is the model parameters.

CAVAN exploits commonsense-caption entailment score as a reward for training. We adopt a BERT model as the commonsense (CMS) discriminator $\mathcal{D}_{cms}$, which returns the entailment score for the caption and commonsense description pair. Following [4], we pre-train the BERT model on ATOMIC dataset using the next sentence prediction task, whose input is an event description sentence and its associated commonsense description. Then, this BERT model is frozen and applied to our entailment score computation as offline. Further details for BERT pre-training are given in Section IV-B. $\mathcal{D}_{cms}$ computes a probability (as SE score) for whether the sampled caption (w^s) entails the commonsense anchor:

$$r_{cms}(w^s) = \mathcal{D}_{cms}(w^s, k). \tag{6}$$

Here, the commonsense-caption entailment score essentially encodes whether the generated caption semantically aligns with the caption w.r.t. the sentence-level meaning. Applying $r_{cms}(w^s)$ to *e.q.* (5) yields a commonsense entailment loss $\mathcal{L}_{cms}$. The gradient is estimated as follows:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{cms}(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[(r_{cms}(w^s) - r_{cms}(\hat{w}))\\ \nabla_\theta \log p_\theta(w^s)],\end{aligned} \tag{7}$$

where $\hat{w}$ is the generated sequence obtained by the current model using greedy decoding. The corresponding entailment reward $r_{cms}(\hat{w})$ is seen as a baseline to reduce the variance of the gradient estimate without changing the expected gradient.

For our experiments, we also adopt the commonsense-caption entailment score as an extra evaluating metric on the testing split. Note that, the queried commonsense knowledge and $\mathcal{D}_{cms}$ are only needed to form supervision signal $\mathcal{L}_{cms}$, but not required during inference.

E. Training

Putting all the loss terms together for an end-to-end training yields an overall optimization target:

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{cms}, \tag{8}$$

where β is a trade-off hyper-parameter weighting each loss term. During the training process, we freeze the CMS discriminator and compute the $r_{rms}(w^s)$ with an inference mode.

IV. Experiments

We conduct experiments and ablation studies on two benchmarks, MSR-VTT [3] and V2C [4] dataset. We evaluate the performance on CAVAN with standard caption evaluation metrics: BLEU@4 [7], METEOR [32], ROUGE-L [33], CIDEr [6], and our newly proposed commonsense-caption entailment score (SE) (see Section III-D). We compare CAVAN with other state-of-the-art methods with systematic ablations and different experimental settings, and exhibit a few captioning examples by CAVAN (see Figure 4). To further validate the effectiveness of CAVAN, we conduct experiments on VATEX [5] dataset, and observe similar performance improvements on it. Detailed conclusions and results can be found in Appendix.

A. Dataset and Augmentation

MSR-VTT as a large-scale video description dataset, contains 10,000 video clip with 200,000 clip-sentence pairs in total. Each video is annotated with 20 English descriptions. Following the official split, we use 6,513 videos for training, 457 videos for validation and 2,990 videos for testing.

We augment each caption in MSR-VTT dataset with 3 types of complementary commonsense descriptions (intention/attribute/effect) retrieved from **ATOMIC** dataset as commonsense anchors for training. More concretely, ATOMIC is an atlas of everyday commonsense knowledge. It consists of 880k triplets of annotations that contain causes and effects of human activities/events as an *if-then* relations. We then extract the most related events from ATOMIC for each ground-truth caption by encoding the key nouns and verbs of captions and events into word vectors [40] and computing their cosine similarities. Following [4], we then select the top-3 most plausible commonsense descriptions for each type of knowledge associated with the events using the BERT discriminator to produce the ranking score. As the queried knowledge from ATOMIC unavoidably comes with noises and incorrect annotations, we further move to V2C dataset with more reliable commonsense knowledge for CAVAN.

V2C is a video description dataset adapted from a subset of MSR-VTT [3]. It contains 9,725 videos, 121,651 captions with each surrounded by 3 types of commonsense descriptions, *i.e.*, intention, attribute and effect. We use the standard splits with 6,819 videos for training, and 2,906 videos for testing.

B. Implement Details

To obtain global visual representations, We use I3D [24] network pre-trained on Kinetics dataset [41] for motion feature extraction, and InceptionResNetV2 [42] pretrained

Method	Motion	Appearence	Object	External Knowledge	B@4	M	R	C	SE
RecNet [34]	-	Inception-V4	-	-	39.1	26.6	59.3	42.7	-
PickNet [35]		ResNet152	-	-	41.3	27.7	59.8	44.1	-
MARN [36]	C3D	ResNet-101	Faster-RCNN	-	40.4	28.1	60.7	47.1	-
OA-BTG [37]	-	ResNet-200	MASK-RCNN	-	41.4	28.2	-	46.9	-
GRU-EVE [38]	C3D	InceptionResnetV2	YOLO	-	38.3	28.4	60.7	48.1	-
MGSA [39]	C3D	InceptionResnetV2	Faster-RCNN	-	42.4	27.6	-	47.5	-
ORG-TRL [25]	C3D	InceptionResnetV2	Faster-RCNN	TBC&WiKi	**43.6**	**28.8**	**62.1**	50.9	-
STG-KD [26]	I3D	ResNet-100	Faster-RCNN	-	40.5	28.3	60.9	47.1	
Baseline(ours)	I3D	InceptionResNetV2	-	-	40.9	27.6	60.5	47.3	47.8
CAVAN	I3D	InceptionResNetV2	Faster-RCNN	ATOMIC	43.0	**28.8**	61.6	**51.0**	**49.2**

TABLE I
WE COMPARE CAVAN WITH PREVIOUS MODELS ON MSR-VTT PUBLIC TESTING SPLIT. "EXTERNAL K." REPRESENTS THE SOURCE OF EXTERNAL KNOWLEDGE. "SE" DENOTES THE AVERAGE ENTAILMENT SCORES OF GENERATED CAPTIONS WITH THEIR CORRESPONDED COMMONSENSE KNOWLEDGE USING A GENERIC COMMONSENSE DISCRIMINATOR MODEL (BERT) PRE-TRAINED ON ATOMIC DATASET.

Method	K. Type	B@4	M	R	C	SE
V2C [4]	-	34.2	-	-	-	-
V2C [4]	INT.	34.6	-	-	-	-
CAVAN	-	38.0	26.6	59.1	57.3	48.3
	INT.	**38.6**	**26.8**	**59.4**	**58.7**	**49.6**

TABLE II
VIDEO CAPTIONING RESULTS ON V2C TESTING SPLIT USING CAVAN AND INTENTION-TYPE OF KNOWLEDGE.

on ImageNet [43] for appearance features of frames. As for object features, we utilize a ResNet152 backbone based Faster-RCNN [44] pretrained on VisualGenome [45]. For caption pre-processing, all captions are truncated to a maximum of 24 words. We replace all words with less than 2 word counts into $\langle$UNK$\rangle$ token in the vocabulary.

Our BERT based CMS discriminator consists of 12 transformer blocks, 12 attention heads, and is with 768 hidden dimensions. For the entailment pre-training, we choose the event sentence and its corresponded commonsense description as a positive pair, and another random commonsense sentence from the ATOMIC as a negative pair. In total, we have 230,624 event-commonsense pairs constructed, with 70% for training, and 30% for testing. Our discriminator achieves 85% accuracy on the testing split.

We use 3 transformer blocks in the visual encoders and decoders, with 768 hidden dimensions and 8 attention heads. We find optimum result by setting the weighting loss term β=0.5. We use the warm-up strategy for the first 5 epochs. We set the batch size as 32 and train the model for 50 epochs. The reinforcement loss $\mathcal{L}_{cms}$ is not applied until 15 epochs. During testing, we use greedy decoding to generate sentences.

C. Experimental Results

We show performances of CAVAN in Table I and compare them with state-of-the-art methods. To translate content-rich videos into human language, current methods not only extract multi-modal visual features *i.e.*motion, appearance and object features, but also bridge the semantic gap to generate accurate captions by introducing external knowledge. For comparison, we list the feature extractors and the sources of external knowledge in Table I.

CAVAN outperforms all of the earlier methods on four metrics except ORG-TRL. We summarize the following reasons for this: (1) ORG-TRL carefully designs a relation graph to encode the cross-object interactions later aggregated with global features via a temporal-spatial attention module. Since the main focus lies on the novel commonsense supervision, CAVAN puts less effort on visual encoding. (2) Different video pre-processing and feature extraction methods make it harder to get a completely fair comparison and have a great impact on the results: the baseline models only using appearance and motion features for CAVAN and ORG-TRL achieve 40.9 and 41.9 on BLEU@4 metric respectively. Despite the performance gap for baseline results, CAVAN still gets as competitive improvement as ORG-TRL in comparison with their own baseline models.

It is worth noting that CAVAN gets outstanding result on CIDEr metric because semantically aligning with commonsense knowledge encourages accurate and informative details to be involved in the output descriptions, which coincides with the mechanism of CIDEr. In addition, we propose to evaluate the generated captions using the BERT produced semantic score on testing split, which heuristically measures the caption quality and its semantic alignment to commonsense knowledge.

We report video captioning results on V2C dataset in Table II using CAVAN. Comparing with V2C which also uses intention-type knowledge, CAVAN shows a great improvement on B@4 score: 4.0 higher on B@4 metric. The consistent improvements on both MSR-VTT and V2C datasets corroborate that CAVAN is not dataset specific, it is applicable for video captioning task as a generic and novel training schema.

D. Effectiveness of Components

To demonstrate the effectiveness of the proposed DFM module and commonsense entailment loss, we design control experiments. First, baseline model applies appearance and motion features with only global branch. After fusing the object-level features with global representations using DFM module (see baseline+DFM), performances of our model are dramatically improved, which clearly indicates that the enhanced object-level features aggregated by DFM module help boost the results. Also, we combine the commonsense entailment loss to the baselines(see baseline+CMS) to verify the

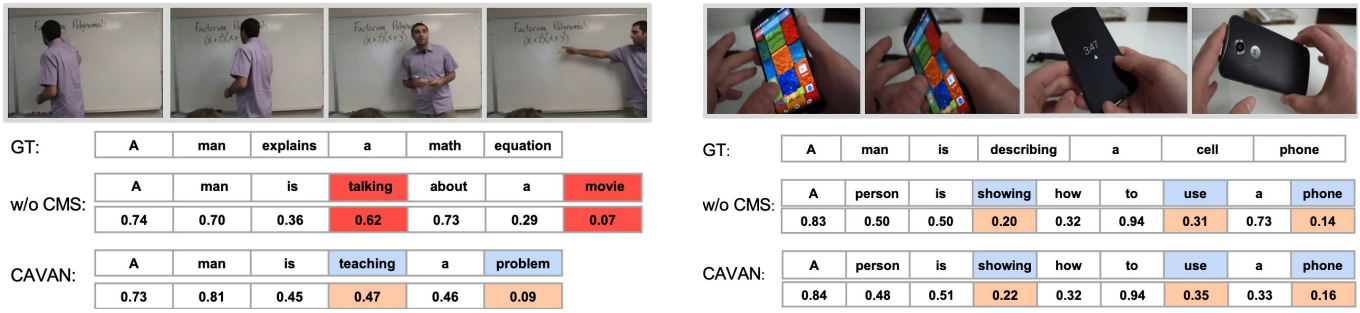

Fig. 4. Caption generation examples using CAVAN on V2C dataset with intention-type knowledge (KG). GT represents the ground-truth captions. w/o CMS denotes the model (baseline+DFM) without CMS constrains.

Model	B@4	M	R	C	SE
Baseline	40.9	27.6	60.5	47.3	47.8
Baseline + DFM	42.5	28.6	61.2	49.6	48.2
Baseline + CMS	41.5	27.9	60.9	48.4	48.8
Baseline + DFM + CMS	**43.0**	**28.8**	**61.6**	**51.0**	**49.2**

TABLE III
EFFECT OF EACH COMPONENT ON MSR-VTT DATASET.

K. Type	B@4	M	R	C	SE
-	38.0	26.6	59.1	57.3	48.3
ATT.	38.3	26.7	59.3	57.9	48.9
EFF.	38.4	26.8	59.4	58.3	49.4
INT.	**38.6**	**26.8**	**59.4**	**58.7**	**49.6**

TABLE IV
COMPARISON OF PERFORMANCES USING DIFFERENT TYPES OF COMMONSENSE KNOWLEDGE IN CAVAN ON V2C TESTING SPLIT.

effectiveness of $\mathcal{L}_{cms}$. Specifically, when equipped with $\mathcal{L}_{cms}$, CIDEr of the baseline is obviously increased from 47.3 to 48.4. Similar trend can be observed on all other metrics, verifying that the use of commonsense anchor brings comprehensive benefits to captioning tasks. We notice that both object-level feature and commonsense knowledge make improvements on SE score. This is because object-level features provide more semantic information and commonsense knowledge put more semantic constrains for the generation. The performance of CAVAN is shown at the last row of Table III where both DFM and commonsense entailment loss are utilized.

E. Effects of Types of Knowledge

We investigate the benefit of using different types of knowledge annotated in V2C. The results are presented in Table IV. We can observe that each type of knowledge all can produce positive impact on the caption generation. Among them, using intention-type knowledge gives the best performance for CAVAN. This observation also aligns with the conclusion in [4], where the intention-type descriptions lead best generation scores. We analyze that this relates to the annotating bias in ATOMIC dataset, where intentions of human activities are more likely to be annotated correctly.

F. Human Evaluation

Human evaluation is critical to verify the quality of queried commonsense knowledge and the performance of CAVAN. We conduct human evaluations by crowdsourcing ratings from workers on Amazon Mechanical Turk (AMT). To evaluate the quality of retrieved commonsense-knowledge, the workers are provided with the ground-truth caption and retrieved knowledge and asked to rate whether the retrieved knowledge entails the caption from a scale of 1-5 (the higher the better, 1 denotes irrelevant and 3 means valid.). We get an average score 3.6, 3.3, 3.1 for retrieved intention, attribute and effect respectively on MSR-VTT, this verifies the extracted commonsense-knowledge from ATOMIC is highly relevant to the video content. To validate the performance of CAVAN, given the videos and generations from CAVAN, the AMTurkers are required to watch and rate how well the generated caption describes the video content from 1-5. The skilled workers report that CAVAN achieves 3.65 on average versus 3.50 from SOTA methods.

V. DISCUSSIONS OF CMS IN GENERATIONS

Figure 4 shows examples of generations from CAVAN comparing with the baseline model without CMS constrain. As illustrated in the left example, the model without CMS constrain generates amusing descriptions, whose keywords marked as red are totally misaligned with the ones in ground-truth caption. CAVAN however has the capacity to rectify wrong semantics and hit the correct keywords marked in blue. Moreover, even when both models generate semantically correct descriptions, the probabilities of keywords marked in orange are improved after applying the CMS constrain (right example in Figure 4).

VI. CONCLUSION

We present CAVAN, a novel training schema for captioning leveraging commonsense knowledge as anchors during model learning. CAVAN is among the first which measures sentence-level semantics using inferential-knowledge, and incorporate it over an end-to-end training as a supervision signal. We conduct extensive experiments to verify the effectiveness of CAVAN on MSR-VTT and V2C dataset, where CAVAN achieves new state-of-the-art results respectively. The observed success of CAVAN confirms the exciting research avenue by adopting commonsense knowledge for high level cognitive vision tasks, including but not limited to image/video captioning, Visual Question Answering, visual navigation, etc.

Acknowledgement. This work was supported by the National Science Foundation under Grant IIS-2132724, IIS-1750082 and CNS-2038666.

REFERENCES

[1] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3027–3035.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[3] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language." IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/

[4] Z. Fang, T. Gokhale, P. Banerjee, C. Baral, and Y. Yang, "Video2commonsense: Generating commonsense descriptions to enrich video captioning," *Empirical Methods in Natural Language Processing*, 2020.

[5] X. Wang, J. Wu, J. Chen, L. Li, Y. Wang, and W. Y. Wang, "VATEX: A large-scale, high-quality multilingual dataset for video-and-language research," *CoRR*, vol. abs/1904.03493, 2019. [Online]. Available: http://arxiv.org/abs/1904.03493

[6] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *CoRR*, vol. abs/1411.5726, 2014. [Online]. Available: http://arxiv.org/abs/1411.5726

[7] 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 on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[8] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," *CoRR*, vol. abs/1607.08822, 2016. [Online]. Available: http://arxiv.org/abs/1607.08822

[9] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Yang Wang, "Video captioning via hierarchical reinforcement learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4213–4222.

[10] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 873–881.

[11] L. Zhang, F. Sung, F. Liu, T. Xiang, S. Gong, Y. Yang, and T. M. Hospedales, "Actor-critic sequence training for image captioning," *arXiv preprint arXiv:1706.09601*, 2017.

[12] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 290–298.

[13] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence – video to text," 2015.

[14] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," 2015.

[15] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2016.

[17] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," 2015.

[18] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[19] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Optimization of image description metrics using policy gradient methods," *CoRR*, vol. abs/1612.00370, 2016. [Online]. Available: http://arxiv.org/abs/1612.00370

[20] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. C. Courville, and Y. Bengio, "An actor-critic algorithm for sequence prediction," *ArXiv*, vol. abs/1607.07086, 2017.

[21] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao, "Self-critical n-step training for image captioning," *CoRR*, vol. abs/1904.06861, 2019. [Online]. Available: http://arxiv.org/abs/1904.06861

[22] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.

[23] 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.

[24] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," *CoRR*, vol. abs/1705.07750, 2017. [Online]. Available: http://arxiv.org/abs/1705.07750

[25] 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.

[26] 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 (CVPR)*, June 2020.

[27] N. Xu, A.-A. Liu, Y. Wong, Y. Zhang, W. Nie, Y. Su, and M. Kankanhalli, "Dual-stream recurrent neural network for video captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2482–2493, 2018.

[28] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5267–5275.

[29] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," *CoRR*, vol. abs/1506.07285, 2015. [Online]. Available: http://arxiv.org/abs/1506.07285

[30] G. Li, H. Su, and W. Zhu, "Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks," *CoRR*, vol. abs/1712.00733, 2017. [Online]. Available: http://arxiv.org/abs/1712.00733

[31] R. Pasunuru and M. Bansal, "Reinforced video captioning with entailment rewards," *arXiv preprint arXiv:1708.02300*, 2017.

[32] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL 2004*, 2004.

[33] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," 01 2005.

[34] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," *CoRR*, vol. abs/1803.11438, 2018. [Online]. Available: http://arxiv.org/abs/1803.11438

[35] 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.

[36] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y. Tai, "Memory-attended recurrent network for video captioning," *CoRR*, vol. abs/1905.03966, 2019. [Online]. Available: http://arxiv.org/abs/1905.03966

[37] 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.

[38] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," *CoRR*, vol. abs/1902.10322, 2019. [Online]. Available: http://arxiv.org/abs/1902.10322

[39] S. Chen and Y.-G. Jiang, "Motion guided spatial attention for video captioning," in *AAAI*, 2019.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[41] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[42] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[44] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497
[45] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *CoRR*, vol. abs/1602.07332, 2016. [Online]. Available: http://arxiv.org/abs/1602.07332