IcoCap: Improving Video Captioning by Compounding Images

Yuanzhi Liang, Linchao Zhu, Xiaohan Wang, Yi Yang

***Abstract*—Video captioning is a more challenging task compared to image captioning, primarily due to differences in content density. Video data contains redundant visual content, making it difficult for captioners to generalize diverse content and avoid being misled by irrelevant elements. Moreover, redundant content is not well-trimmed to match the corresponding visual semantics in the ground truth, further increasing the difficulty of video captioning. Current research in video captioning predominantly focuses on captioner design, neglecting the impact of content density on captioner performance. Considering the differences between videos and images, there exists an another line to improve video captioning by leveraging concise and easily-learned image samples to further diversify video samples. This modification to content density compels the captioner to learn more effectively against redundancy and ambiguity. In this paper, we propose a novel approach called Image-Compounded learning for video Captioners (IcoCap) to facilitate better learning of complex video semantics. IcoCap comprises two components: the Image-Video Compounding Strategy (ICS) and Visual-Semantic Guided Captioning (VGC). ICS compounds easily-learned image semantics into video semantics, further diversifying video content and prompting the network to generalize contents in a more diverse sample. Besides, learning with the sample compounded with image contents, the captioner is compelled to better extract valuable video cues in the presence of straightforward image semantics. This helps the captioner further focus on relevant information while filtering out extraneous content. Then, VGC guides the network in flexibly learning ground truth captions based on the compounded samples, helping to mitigate the mismatch between the ground truth and ambiguous semantics in video samples. Our experimental results demonstrate the effectiveness of IcoCap in improving the learning of video captioners. Applied to the widely-used MSVD, MSR-VTT, and VATEX datasets, our approach achieves competitive or superior results compared to state-of-the-art methods, illustrating its capacity to handle the redundant and ambiguous video data.**

***Index Terms*—Video Captioning, Multi-modal understanding, Representation learning.**

I. Introduction

Video captioning is a challenging task that requires the model to learn semantics and express through natural language. The main challenge in this task is understanding the diverse visual contents in the videos. Recently, many solutions have been proposed to solve this problem, e.g., leveraging better video representations [1], [2], complex network designs [3], [4], and end-to-end learning [5], [6]. These works facilitate a better understanding of video semantics and generate coherent descriptions of the visual contents.

Despite significant improvements, understanding video semantics remains a challenging task. A major obstacle to achieving this understanding is the semantic ambiguity in videos, caused by their visual redundancy. The contents of videos are diverse and difficult to precisely trim with specific descriptions. As illustrated in Fig. 1 (a), some contents, such as irrelevant and minor events, are not described by the ground truth, and without particular descriptions, they are implicit for the network to understand. In addition, contents such as transitions are related to the events described by the ground truth but do not contain valuable semantics for network learning. These contents present a challenge for neural networks. The video captioner is always hard to generalize redundant contents or misguided by ambiguous semantics. Meanwhile, the mismatch between descriptions and visual contents further increase the difficulties in learning video semantics. All this defeats induce the captioner may be misguided by trivial and irrelevant semantics.

Comparably, image contents are more concise, and the semantics are salient, as shown in Fig. 1 (b). There are no irrelevant events or transitions apart from the valuable contents in the images. Meanwhile, the image descriptions are precise. The ground truth description can summarize most contents. This makes the image samples easier to be captioned. Empirically, results from image captioning datasets [7] are often better than video captioning in various metrics.

The primary distinction between images and videos for captioners lies in content density. The redundancy and ambiguity in video data cause the network to struggle in generalizing complex video semantics. While previous works have focused on proposing better captioner designs, improved architectures to increase network capacity, which aids in learning semantics and handling redundancy. However, an another line for improving video captioning has been overlooked: modifying content density to enhance the learning process of video captioners. In this work, we propose a novel learning method called Image-Compounded learning for video Captioners (IcoCap). IcoCap compounds concise and easily-learned image semantics into video samples, diversifying the visual contents and compelling the network to learn against redundant contents. Besides, the compounded image semantics are more easily learned compared to video semantics, which are similar to a strong competitor [8], [9] for learning video semantics. To address video captioning, the captioner must investigate valuable video cues in contrast to easily-learned image contents. This further enhances the captioner's ability

(Corresponding author: Linchao Zhu.)

Yuanzhi Liang is with the Australian Artificial Intelligence Institute, University of Technology Sydney, NSW 2007, Australia (email: yuanzhi.Liang@student.uts.edu.au). Linchao Zhu, Xiaohan Wang and Yi Yang are with the College of Computer Science and Technology, Zhejiang University, China (email: zhulinchao@zju.edu.cn; xiaohan.wang@zju.edu.cn; yangyics@zju.edu.cn).

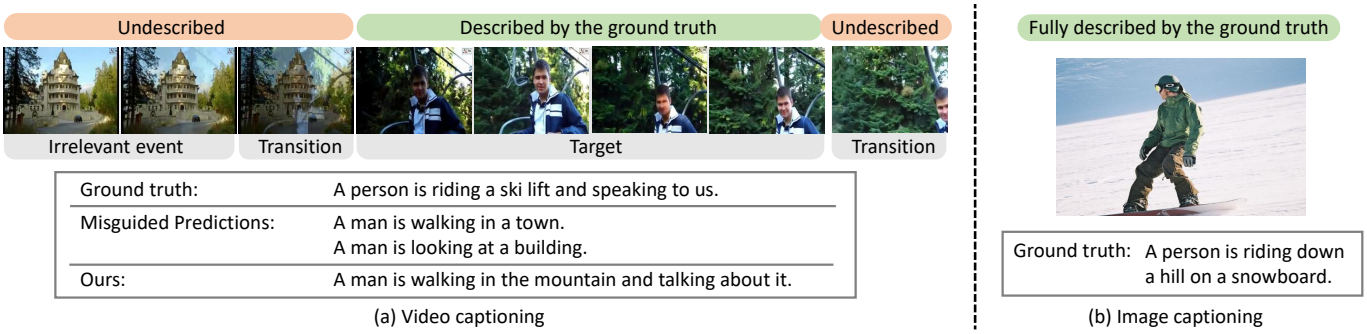

Fig. 1. Video semantics are ambiguous. Some frames contain irrelevant events or serve as transitions. They do not provide valuable contents corresponding to the ground truth in video captioning. Meanwhile, image contents are concise and explicit. The ground truth in image captioning easily summarizes all image semantics.

to learn video semantics. Additionally, IcoCap alleviates the challenges of learning from mismatched descriptions by encouraging the network to flexibly learn descriptions based on visual semantics, rather than relying on rigidly pre-assigned captions.

Specifically, IcoCap comprises two modules: the Image-Video Compounding Strategy (ICS) and Visual-Semantic Guided Captioning (VGC). In detail, ICS is designed to compound image content into video content. This approach further diversifies the video samples, guiding the video captioner to learn against redundancy. Simultaneously, the introduction of easily-learned image content compels the network to extract valuable video cues while filtering out irrelevant elements. Additionally, IcoCap addresses the issue of ambiguous video semantics by VGC, which facilitates flexible learning of semantics based on visual content. In VGC, the ground truth is selected from relevant descriptions rather than strictly corresponding to the original video ground truth. A visual-semantic consistency factor is introduced to adjust the captioning process, promoting the network to focus on the salient visual content rather than concentrating on minor and detailed contents.

The main contributions can be summarized below:

1. We propose an Image-Compounded learning for video Captioners (IcoCap). IcoCap introduces image samples and compounds the images into video contents. IcoCap impels the network to mine valuable video cues against the semantic ambiguity in videos.

2. We propose an Image-video Compounding Strategy (ICS) and Visual-semantic Guided Captioning (VGC). ICS provides a series of operations to compound images and videos, which promotes the network's ability to learn video semantics against ambiguity. VGC helps the network to flexibly learn complex video contents from ICS, rather than rigidly following the ground truth.

3. Without complicated designs or networks, our method performs favorably or outperforms the state-of-the-art methods on various datasets, including MSR-VTT, MSVD, and VATEX.

II. Related Work

Video Representation: Representation of video [5], [10]–[14] is a long-standing problem in the representation learning [15]–[18]. Numerous works have emerged, proposing diverse architectures and approaches that focus on exploiting the unique characteristics of video data to achieve effective and robust representations. In representation, the intuitive idea behind video representation is to extend the principles of image-based CNNs, which have demonstrated remarkable success in tasks such as object recognition and image classification.

One notable approach to incorporate temporal information into the original CNN framework is by introducing 3D kernels [19], [20]. These kernels extend the receptive field in the time dimension, thereby enabling the network to capture the relationships between sequential frames. This extension results in 3D Convolutional Neural Networks (3D CNNs) [20], which are specifically designed to process video data by jointly learning spatial and temporal features and have demonstrated considerable improvements in video representation tasks compared to their 2D counterparts. However, one drawback of 3D CNNs is the increased computational complexity and memory requirements, which can pose challenges in terms of scalability and efficiency. I3D [21] inflated the filters and pooling layers of 2D CNNs into 3D, enabling the network to learn richer spatio-temporal features. The I3D model achieved significant improvements in action recognition tasks and demonstrated the potential of incorporating pre-trained 2D CNN knowledge into video representation learning. More variations [22] of 3D CNNs further provide many video-based designs to boost the performances of representations in various tasks.

Moreover, recent works [10], [23]–[25] pay more attention to the large-scale pre-training. Motivated by the success of Bert [26] in NLP, many works [24], [27] propose to leverage the similar pre-training strategies to videos. Significant improvements occur in video tasks after applying the large-scale pre-training [11], [28] and various transformer-based networks [5], [25], [29]. Besides, tasks like mask-modeling [26], contrastive learning [30], [31], etc., further empower the representation ability of networks. CLIP [23], as a typical pre-training model, has also been proven that possesses remarkable ability in correlating language semantics and has already been widely used in various domains [1], [32].

These video-based designs have contributed to the evolution of video representation learning, enabling more effective and discriminative representations for various tasks. Despite the progress made thus far, video representation remains an active area of research, with ongoing efforts to develop more efficient and accurate models capable of handling the ever-increasing complexity and scale of video data. In our work, we focus on the learning method of video captioning and directly applying representation method according [1], [23], [32].

Video Captioning: Video captioning [33]–[35] is a challenging and complex task that aims to generate a natural language sentence to describe a given video sequence. Unlike image captioning, where the objective is to generate descriptions for static images, video captioning methods need to handle intricate video data that encapsulates diverse and dynamic semantics. The temporal dimension of video data adds a level of complexity that requires sophisticated approaches to capture and summarize the underlying content effectively.

In detail, the common approach in video captioning is the encoder-decoder framework, which employs a CNN to encode visual information and an RNN or LSTM to generate captions sequentially. Donahue et al. [36] proposed the Sequence-to-Sequence Video-to-Text model, which combined a 2D CNN with an LSTM to generate captions. Chen et al. [37] introduced the TDConvED network—a convolutional sequence-to-sequence learning framework, specifically tailored to enhance video captioning. Most recent works [4], [5], [38], [39] also follow this framework and present various solutions to further boost the performances. Moreover, Chen et al. [40] propose to select frames in video for video captioning. Pan et al. [41] introduce a visual semantic embedding model to specifically consider the relationship between the semantics of the entire sentence and video content unexploited.

Moreover, another line of evolution is the video representation method. Works in video captioning apply features from some pre-trained models to represent videos. Models like bottom-up [42] in image representations, 3D CNNs [21], [22] in video representation, or generic large-scale pre-training models [5], [6], [23] are applied in video captioning to represent video data. Then, various methods [2], [43]–[45] are designed to investigate the semantic cues from well-trained representations and solve video captioning. Yang et al. [46] conducted a comparative analysis between CLIP features and ImageNet pre-trained features for video captioning. Additionally, they introduced an auxiliary task designed to discern the correspondence between video content and associated concepts. Some recent works [4], [47], [48] introduce complicated structures to mine detailed information from video features and achieve significant improvements. Besides, some works [4], [5] further propose end-to-end frameworks for representing videos from scratch and exploring the detailed instances and events in the video frames.

In this paper, we propose a novel method to improve video captioning by introducing image samples to aid in video learning. The highly diverse video contents can induce ambiguity in video semantics, which can be challenging for network learning. In contrast, image samples are typically concise and explicit, making them easily learnable for the network and possessing clear semantics. Our approach compounds the image and video samples to impel the network to learn semantics from the combined samples. This leads the networks can better learn video semantics against the redundancy and ambiguity. Our experimental results demonstrate that the proposed learning approach outperforms existing methods across various datasets and metrics.

III. Image-compounded video Captioner

We propose an Image-compounded video Captioner (IcoCap) to introduce image samples to improve video captioning. IcoCap contains two parts: Image-video Compounding Strategy (ICS) and Visual-semantic Guided Captioning (VGC). ICS uses image samples to augment video samples for video captioning. It contains a series of augmentation strategies to compound image contents into video contents, as shown in Fig. 2. Moreover, VGC provides a flexible learning manner for the complex and diverse semantics from ICS.

A. Image-video Compounding Strategy

The visual contents of the video data are diverse but ambiguous. It is hard to annotate all instances and events in the frames specifically. However, semantics in image data are explicit and clear. Existing works [6], [7], [27] perform joint training to learn image and video samples. They treat the images and videos individually, which are separately provided for the network as different training samples.

However, our work aims to utilize image samples to augment video samples and compound them as one training sample. Both videos and images can occur in the same training samples. This leads to the redundancy of visual content changes according to the introduction of image samples. As shown in Fig. 2, we proposed Image-video Compounding Strategy (ICS) to produce training samples. Specifically, for a given video V with N frames, we pre-process and represent the video as frame features v following [1], [23], [32]. In IcoCap, we randomly sample M images from [49] to construct an auxiliary image set. In every training step, we select an image sample I and extract the image feature, denoted as x. $x \in \mathbb{R}^{1\times D}$. We also additionally pre-process another video V'. We denote the frame feature from V' as v'. Then, ICS takes v, v', and x as inputs and produces compounded samples h, where $h \in \mathbb{R}^{n\times D}$.

ICS consists of three steps: Intra-Video Sampling (VS), Inter-Feature Mixup (FM), and Inter-Frame Swap (FS), represented as functions f_{VS}, f_{FM}, and f_{FS}. For the current video sample, we construct a set of frame features as $\mathcal{V} = \{v_1, v_2, ..., v_N\}$, where v_i indicates a frame feature ($v_i \in \mathbb{R}^{1\times D}$). We also create an auxiliary set $\mathcal{A} = \{v'_1, v'_2, ..., v'_N, x_1, x_2, ..., x_N\}$, where x_i is a duplicate of the same image sample x, and $v'_i \in \mathbb{R}^{1\times D}$.

1. Intra-Video Sampling (VS): In VS strategy,

$$h = f_{\text{VS}}(\mathcal{V}) \tag{1}$$

The frame feature h is randomly sampled from $\mathcal{V}$ and $h \in \mathbb{R}^{n\times D}$. To be noticed, previous works [4], [5], [38] usually

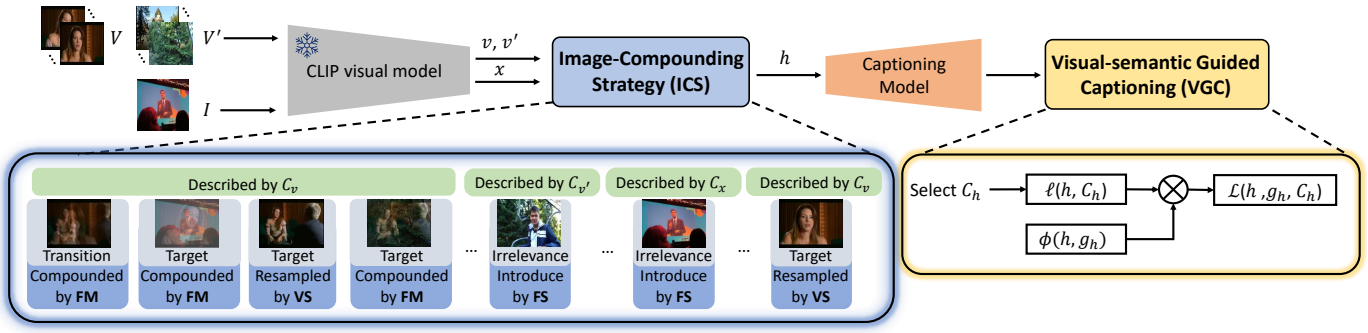

Fig. 2. Overview of Image-video Compounding Strategy (ICS). ICS introduces image samples to help the network learn ambiguous video semantics. All features are extracted by a frozen CLIP visual model. V and V' are different video samples. I is the additional image sample. v, v' and x are features for video and image samples, respectively. C_v, $C_{v'}$, C_x are the descriptions for V, V' and I, respectively.

apply uniform sampling during producing video features. Uniform sampling, with a fixed sampling interval, is insufficient in fully utilizing the information present in a video due to the redundancy of video contents. Some frames may never be sampled and used for training. Comparatively, random sampling in VS can take advantage of all frames in training and also produce more challenging and diverse samples for captioner training.

2. Inter-Feature Mixup (FM): To implement FM strategy, an auxiliary feature h' is randomly sampled from $\mathcal{A}$, and $h' \in \mathbb{R}^{n\times D}$. Then, FM strategy can be formulated as:

$$h_i = f_{\text{FM}}(h_i, h'_i, J) = \begin{cases} \alpha h_i + (1-\alpha)h'_i & \text{if} \quad i \in J \\ h_i & \text{if} \quad i \notin J \end{cases} \tag{2}$$

where i is the index number and $i \in [1, n]$. J is the set of index numbers and $J = \{j_1, ..., j_k\}$. k is a random number and $k \in (1, n)$. Moreover, α represents the mixup ratio. $\alpha \in (0, 1)$.

3. Inter-Frame Swap (FS): In FS strategy, we replace the training feature with the auxiliary features, by given the random index Q, in which Q is also a set of index number. $Q = \{q_1, ..., q_t\}$ and $t \in (1, n)$. The operation can be written as:

$$h_i = f_{\text{FS}}(h_i, h'_i, Q) = \begin{cases} h'_i & \text{if} \quad i \in Q \\ h_i & \text{if} \quad i \notin Q \end{cases} \tag{3}$$

where i is the index number and $i \in [1, n]$. In FS, each frame feature has a 50% probability of being mixed up and a 50% probability of being replaced by external features.

ICS blends the semantics of images and videos, generating an image-compounded video sample for captioner learning. Samples compounded with various video samples exhibit greater diversity than the original samples. This diversity sets a higher requirement for the network to handle redundancy, which in turn enhances the generalization ability of the captioner. Furthermore, some samples are constructed by compounding both video and image samples. By incorporating easily-learned image semantics, the video captioner is required to extract valuable video cues while disregarding easy but irrelevant cues from images. This process further improves the captioner's ability to avoid being misled by irrelevant semantics.

In addition, the network architecture used in our work is a simple transformer model comprising a four-layer transformer and a fully connected layer, following the approach described in [7]. The transformer network in our work is responsible for mining valuable contents from h and understanding complex semantics. The fully connected layer produces predictions of words at every time step.

B. Visual-semantic Guided Captioning

Features from ICS are diverse and complex, which are hard to be solved by the original captioning loss. In our work, we propose Visual-semantic Guided Captioning (VGC) to encourage the network to express semantics flexibly according to the given visual semantics.

There are two kinds of visual-semantic guidance in VGC, which are visual-semantic based description selection and the visual-semantic consistent factor.

First, the ground truth for descriptions is flexibly selected from relative descriptions guided by the visual semantics of h. The training feature h may contain contents from V, V', or I, in which all the features have corresponding ground truth and are available to be expressed. For v, v' and x, the corresponding captions are C_v, $C_{v'}$, and C_x. We utilize a pre-trained language model [23] to extract the features of the descriptions and denote them as g_v, $g_{v'}$, and g_x. All the language features are in the size of $\mathbb{R}^{1\times D}$. Then, the ground truth of the current sample can be determined by the cosine similarity between visual features (h) and language features (g_v, $g_{v'}$ and g_x). We denote the language feature with the largest similarity as g_h and select the corresponding caption C_h as the ground truth. The ground truth is the corresponding caption with the maximum similarity. Besides, to calculate similarity, g is copied n times to fit the dimension of h.

To be noticed, rather than describe all possible ground truth as in [6], the network should learn to produce C_h only, which reflects most of the visual contents in h and is the most suitable for current visual semantics. This training goal requires the network to exclude expressive but minor semantics and focus on the exploration of valuable visual semantics.

Second, we further leverage the guidance of visual semantic by designing a visual-semantic consistent factor ϕ. It regularizes the learning procedure and reduces the punishments if

the ground truth description possess lower consistency with current visual contents. Specifically, the factor encourages the predictions to be close to the salient semantics in visual contents, which can be formulated as follow:

$$\phi(h, g_h) = -log(min(S/\tau, 1)) \tag{4}$$

where τ is a temperature coefficient and S indicates the cosine similarity between h and g_h. Regardless of whether the predictions are close to C_h, if the language features of the predictions show higher similarity to most of the visual contents in h, the factor ϕ will be relatively lower.

Finally, the overall loss function can be formulated as follows:

$$\mathcal{L}(h, g_h, C_h) = \phi(h, g_h) \cdot \ell(h, C_h) \tag{5}$$

where ℓ is cross-entropy loss function which is widely used as captioning loss in [1], [33].

Instead of forcing the network to generate fixed, rigid descriptions, VGC encourages the network to learn flexibly based on the visual semantics present in the video content. This is achieved by adaptively assigning the ground truth for ambiguous video semantics originating from ICS. Simultaneously, VGC introduces a flexible factor to modulate the captioning learning process, guiding the network to generate improved descriptions that better align with the diverse visual semantics. This approach not only helps the network to better adapt to varying content densities but also ensures that the generated captions are more representative of the actual video content. All modules in IcoCap contribute to enhancing the captioner's ability to learn effectively while dealing with redundant video contents and ambiguous video semantics.

IV. Experiments

In this section, we discuss the details of our method and evaluate the captioning performances in various datasets.

A. Datasets and Implementation Details

Datasets: We evaluate our method using three established video captioning benchmarks: MSRVTT [50], MSVD [51], and VATEX [52]. In all datasets, we employ English annotations as the ground truth for experimentation.

MSR-VTT [50] is a prevalent benchmark for video captioning that consists of $10,000$ videos, each with 20 annotations. To facilitate evaluation and comparison, we adopt the standard setting used in [50], wherein the dataset is partitioned into three subsets: a training set with $6,513$ samples, a validation set with 497 samples, and a test set with $2,990$ samples.

MSVD [51] is another established benchmark in video captioning, which comprises $1,970$ YouTube videos, each with approximately 40 annotations. The dataset is partitioned into three subsets: a training set consisting of $1,200$ videos, a validation set consisting of 100 videos, and a test set consisting of 670 videos.

In addition, VATEX [52] is another widely used dataset for video captioning, sourced from the Kinetics-600 dataset [53]. VATEX contains annotations in both English and Chinese, with 10 descriptions in each language. The dataset comprises $25,991$ video clips for training, and $3,000$ and $6,000$ video clips for validation and testing, respectively.

Implementation Details: We extract features from all video frames and image samples using the CLIP visual model [23], which we utilize solely for representation and do not involve in network training. The CLIP model is a powerful representation method that has been widely applied in video captioning [1]. Unlike recent works [4], [5], [54] that utilize video-based backbones such as Vivit, C3D, and I3D, our work employs an image-based method through the CLIP model. Video-based backbones take into account the relationships between frames in chronological order, while external images lack such relationships and may confuse video-based models. In contrast, image-based methods do not consider sequential properties, providing more flexibility for image-video compounding. Therefore, we choose the CLIP model to represent visual content. Additionally, the pre-trained language model used in VGC is based on the CLIP language model.

We capture all the video frames and extract the features by CLIP visual model [23]. The image samples are also processed similarly. We only use CLIP model to represent video and image data, which does not participate in the network training. Besides, CLIP model is a powerful representation method and has already been widely applied in video captioning [1]. Unlike recent works [4], [5], [54] that utilize video-based backbones [21], [22], [25], our work employs an image-based method through the CLIP model. Video-based backbones take into account the relationships between frames in chronological order, while external images lack such relationships and may confuse video-based models. In contrast, image-based methods do not consider sequential properties, providing more flexibility for image-video compounding. Therefore, we choose the CLIP model to represent visual content. Additionally, the pre-trained language model used in VGC is also based on the CLIP language model.

The dimension of input features D is 512. As our captioning model, we employ a simple transformer network [55], and all training settings follow [7]. For the external image subset, we randomly select samples from MSCOCO [49], with a default length of $10,000$. In addition, we set the hyper-parameters $n = 32$ and $\tau = 0.5$. Image samples, original video samples, and additional video samples have a 24.71%, 51.79%, and 23.5% probability of being the ground truth, respectively. All input samples are potentially salient content and can be learned for captioning. During evaluation, we uniformly sample frames from videos in accordance with [1], [7]. Further experiments and ablations will be presented in the next section. To ensure a fair comparison, we evaluate our method and report results for other methods based on the official test split of the corresponding datasets.

B. Performance Comparison

As shown in Table I, we evaluate our method on the MSR-VTT dataset using various captioning metrics. Our method outperforms the current state-of-the-art SWINBERT [5] without bells and whistles. Using only the CLIP ViT-B/32 model,

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE TEST SPLIT OF MSRVTT. † INDICATES THE RESULTS FROM THE OFFICIAL IMPLEMENTATION OF [5] TAKING 32 FRAMES AS INPUTS. VIT-B/32 AND VIT-B/16 STAND FOR CLIP VIT-B/32 AND CLIP VIT-B/16 MODELS, RESPECTIVELY. CLIP BASELINE ONLY USES THE VIDEO FEATURES EXTRACTED BY CLIP MODEL AND DOES NOT APPLY OUR METHOD. JOINT BASELINE INDICATES BOTH VIDEO AND IMAGE SAMPLES ARE JOINTLY TRAINED WITH CLIP BASELINE.

Method	Feature	MSR-VTT			
		BLEU-4	METEOR	ROUGE-L	CIDEr
GRU-EVE [44]	InceptionResNetV2 + C3D	38.3	28.4	60.7	48.1
STG-KD [2]	ResNet101 + I3D	40.5	28.3	60.9	47.1
POS-CG [56]	InceptionResNetV2	42.0	28.2	61.6	48.7
POS-VCT [57]	InceptionResNetV2 + C3D	42.3	29.7	62.8	49.1
ORG-TRL [4]	InceptionResNetV2 + C3D	43.6	28.8	62.1	50.9
SAAT [58]	InceptionResNetV2 + C3D	39.9	27.7	61.2	51.0
OpenBook [3]	InceptionResNetV2 + C3D	33.9	23.7	50.2	52.9
Revnet [59]	Inception-V4	42.4	28.1	62.3	53.2
HMN [60]	InceptionResNetV2 + C3D	43.5	29.0	62.7	51.5
SWINBERT† [5]	VidSwin	41.9	29.8	62.1	53.7
CLIP4Clip [1]	ViT-B/32	46.1	30.7	63.7	57.7
CLIP Baseline	ViT-B/32	43.1	29.3	61.9	54.8
Joint Baseline	ViT-B/32	43.5	29.4	62.4	55.2
Ours	ViT-B/32	46.1	30.3	64.3	59.1
Ours	ViT-B/16	**47.0**	**31.1**	**64.9**	**60.2**

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE TEST SPLIT OF MSVD. † INDICATES THE RESULTS FROM THE OFFICIAL IMPLEMENTATION OF [5] TAKING 32 FRAMES AS INPUTS.

Method	Feature	MSVD			
		BLEU-4	METEOR	ROUGE-L	CIDEr
GRU-EVE [38]	InceptionResNetV2 + C3D	47.9	35.0	71.5	78.1
POS-CG [48]	InceptionResNetV2	52.5	34.1	71.3	88.7
POS-VCT [49]	InceptionResNetV2 + C3D	52.8	36.1	71.8	87.8
SAAT [50]	InceptionResNetV2 + C3D	46.5	33.5	69.4	81.0
STG-KD [2]	ResNet101 + I3D	52.2	36.9	73.9	93.0
ORG-TRL [4]	InceptionResNetV2 + C3D	54.3	36.4	73.9	95.2
HMN [52]	InceptionResNetV2 + C3D	59.2	37.7	75.1	104.0
SWINBERT† [5]	VidSwin	55.7	**39.6**	75.7	109.4
CLIP Baseline	ViT-B/32	55.5	38.0	74.4	95.5
Joint Baseline	ViT-B/32	57.2	37.5	74.6	96.7
Ours	ViT-B/32	56.3	38.9	75.0	103.8
Ours	ViT-B/16	**59.1**	39.5	**76.5**	**110.3**

our method improves by 5.1, 1.3, 2.8, and 5.4 in BLEU-4 [61], METEOR [62], ROUGE-L [63], and CIDEr [64], respectively, which is significant in MSR-VTT. Notably, our method does not employ multi-modal features [56], [58] or features from detectors [2], [4]. We also do not apply complex network design [54], [60], costly end-to-end training [5], or assemble operations [1].

Furthermore, we present the results of CLIP baseline (which only uses CLIP features without ICS and VGC) and joint training baselines (where CLIP baseline is jointly trained with image features without ICS and VGC) implemented with CLIP ViT-B/32 model. CLIP is a powerful representation method, and even the CLIP baseline, which utilizes only CLIP features and our network, achieves relatively higher performance than recent methods [4], [5], [54]. Additionally, introducing image samples and joint training with MSR-VTT data slightly improves the performance of the CLIP baseline. However, due to the domain gap between video and images, joint training video samples with image samples yields only marginal improvement. In comparison, our method with image samples yields a significant improvement in video captioning, with the value of the CIDEr metric increasing by 4.3.

In addition, due to the effectiveness of ICS and VGC, our method outperforms the CLIP4Clip [1] method, which is also based on the CLIP ViT-B/32 model and utilizes CLIP features. With the same features as inputs, our method achieves significantly better performance than CLIP4Clip, with gaps of 1.1, 1.2, and 3.5 in BLEU, METEOR, and CIDEr, respectively.

As presented in Table II, we conducted experiments on the MSVD dataset, which further demonstrates the effectiveness of our method. In comparison with the improvements observed between our method and the joint training baseline on MSR-VTT, our method's superiority is further highlighted. Specifically, our method yields a significant increase of 7.1 in the CIDEr metric, which is a dramatic improvement compared to the joint training baseline.

Moreover, we present further results on the VATEX dataset [52], which includes more complex and diverse descriptions compared to MSR-VTT and MSVD. In addition to exploring video content, VATEX sets a higher requirement for language diversity in predictions. To achieve better performance in evaluation, the predictions should be more vivid and diverse, which is relatively challenging for the simple captioning network utilized in our method. As demonstrated in Table III, we achieve comparable performance to state-of-the-art methods in VATEX. Despite not aiming to generate vivid descriptions with high linguistic complexity, IcoCap still outperforms many recent methods with more complex designs.

Additionally, the input features in our work can be summarized as image-based representations, which extract features from each frame individually. In this section, we also evaluate video-based representations, which represent multiple sequential frames as a single feature.

In detail, we first apply the ICS strategies directly to the original video frames and then extract the features using the VideoSwin Transformer following [5], [25]. Since the VGC requires calculating similarity between visual and language features, we only adapt ICS with the VideoSwin Transformer and use our captioning model for comparison. As in our work, the parameters of the VideoSwin Transformer are also fixed during training. However, after applying the video-based representations, the results on MSR-VTT are only 38.5, 27.3, 59.0, and 45.3 for BLEU-4, METEOR, ROUGE-L, and CIDEr, respectively. The performance of the ICS strategies with the VideoSwin Transformer decreases by 8.5 in the CIDEr metric compared with the reported value from SWINBERT [5]. More significantly, the gap between applying ICS with the VideoSwin Transformer and our method is 14.9. The significant drop in the captioning performance indicates that introducing augmentation strategies into video-based representation is not feasible. Methods such as the VideoSwin Transformer need to model temporal information specifically. Complex augmentations can break the connections and relationships between the original frames, causing representation methods to become confused and fail to produce valuable features for video captioning.

We also evaluate feature-level augmentations based on the VideoSwin Transformer [5], [25]. First, we extract frame features using a frozen VideoSwin Transformer and then

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE TEST SPLIT OF VATEX. † INDICATES THE RESULTS FROM THE OFFICIAL IMPLEMENTATION OF [5] TAKING 32 FRAMES AS INPUTS.

Method	Feature	VATEX			
		BLEU-4	METEOR	ROUGE-L	CIDEr
VATEX [52]	bi-LSTM + I3D	28.4	21.7	47.0	45.1
ORG-TRL [4]	InceptionResNetV2 + C3D	32.1	22.2	48.9	49.7
Support-set [65]	ResNet152	32.8	24.4	49.1	51.2
OpenBook [3]	InceptionResNetV2 + C3D	33.9	23.7	50.2	57.5
SWINBERT† [5]	VidSwin	37.8	26.1	53.0	71.6
CLIP Baseline	ViT-B/32	35.9	24.0	52.1	57.3
Joint Baseline	ViT-B/32	35.5	24.4	51.6	60.1
Ours	ViT-B/32	36.9	24.6	52.5	63.4
Ours	ViT-B/16	37.4	25.7	53.1	67.8

TABLE IV
COMPARISON FOR COMBINATIONS OF DATA SAMPLES AND STRATEGIES IN ICS. VS, FM, AND FS ARE SHORT FOR INTRA-VIDEO SAMPLING, INTER-FEATURE MIXUP AND INTER-FRAME SWAP, RESPECTIVELY. ALL STRATEGIES AND ADDITIONAL DATA ARE USEFUL AND THE COMBINATIONS LEAD TO HIGHER PERFORMANCES.

ICS Strategy	Data Sample	BLEU-4	METEOR	ROUGE-L	CIDEr
VS + FM + FS	v, v'	56.5	37.8	73.9	97.5
VS + FM + FS	v, x	57.5	37.8	74.1	100.4
VS	v, v', x	53.3	37.5	74.1	97.6
FM	v, v', x	53.5	37.8	74.5	97.9
FS	v, v', x	59.1	38.7	75.5	101.2
VS + FM	v, v', x	59.0	38.6	74.7	99.1
VS + FS	v, v', x	59.4	38.2	75.3	102.3
FM + FS	v, v', x	60.5	38.9	75.0	103.2

augment these features using the same strategies as in ICS. However, the resulting CIDEr metric is only 20.5. Augmenting feature-level representations in video-based methods leads to a decline in performance. This is because features from video-based methods naturally contain sequential relationships and contexts. Although we can obtain features for specific frames and apply augmentations, it may be difficult for the captioning model to understand the augmented features with broken and confusing inter-frame relationships after compounding. By comparison, our image-based representations do not face this issue. Every frame is represented individually, making them flexible to various augmentations and achieving better performance.

C. Ablation Studies

In this section, we conduct a comprehensive ablation study for several details in our method. All experiments are operated based on CLIP ViT-B/32 and MVSD test split.

Comparison of Data Combinations: Our work utilizes video data v' from the current training dataset and external image samples x to augment training features. In Table IV, we compare the different combinations of data samples. Both combinations, v with v' or x, are useful for improving video captioning. Additional samples expand the training data and lead to better results. However, improvements from introducing image samples are more significant, with an increase of 2.9 in the CIDEr metric. Image samples possess concise contents and precise descriptions, which are easily learnable for the network. Introducing image samples formulates the more challenging samples. This forces the network to mine video semantics against the easy image semantics

TABLE V
ABLATION FOR DIFFERENT PARTS IN OUR METHOD. OURS WITH ALL DESCRIPTIONS INDICATES TAKING ALL RELATIVE DESCRIPTIONS C_v, $C_{v'}$, AND C_x AS THE GROUND TRUTH FOR h AT THE SAME TIME. IN COMPARISON, ALL MODULES IN OUR WORK ARE HELPFUL IN IMPROVING THE PERFORMANCE OF VIDEO CAPTIONING.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr
Ours w/o VGC	60.4	38.7	75.1	100.8
Ours w/ all descriptions	60.1	38.4	75.3	100.0
VGC w/o ϕ	59.5	38.2	75.3	102.1
Ours w/o ICS	59.2	37.8	74.5	95.9
CLIP Baseline	55.5	38.0	74.4	95.5
Joint Baseline	57.2	37.5	74.6	96.7
Ours	56.3	38.9	75.0	103.8

Comparison of different strategies in ICS: We conducted an ablation study on the strategies in ICS, and the results are presented in Table IV. All strategies effectively diversify the training samples and improve performance. Inter-frame swap was found to be the most helpful strategy as it directly replaces the visual contents with additional data, offering maximum influence on the semantics of the training samples compared to other strategies. Additionally, image samples are easier to learn than video samples, and inter-frame swap introduces expressive image samples into video samples, requiring the network to ignore irrelevant semantics and refocus on the video contents. In our work, we utilized all three ICS strategies, and the combination of these strategies with additional data samples resulted in significant improvement in video captioning performance. With CLIP ViT-B/32 features, we achieved a CIDEr score of 103.8, which is an 8.3 improvement over the baseline.

Efficacy of ICS and VGC: To evaluate the efficacy of VGC, we separately evaluate our captioning loss $\ell(h, C_h)$ and the factor ϕ. Firstly, we use all descriptions as the ground truth and experiment with the loss function $\mathcal{L}(h, g_h, C_v, C_{v'}, C_x) = \phi(h, g_h) \cdot (\ell(h, C_v) + \ell(h, C_{v'}) + \ell(h, C_x))$. As shown in Table V, the result with all the descriptions (corresponding to VGC w/ all descriptions) is relatively lower. Giving all possible descriptions to the network at the same time may confuse the network and result in worse performance of 100.8 in CIDEr. Meanwhile, the factor ϕ is also useful in improving performance. Without the modulation of factor ϕ, the performance decreases by 1.7 in CIDEr.

Additionally, ICS serves to diversify and expand the training samples. When ICS is applied to the CLIP baseline, the network can also be significantly improved, as evidenced by a CIDEr score of 102.1. However, this improvement is not as significant as the results obtained using our method, which combines both ICS and VGC (103.8 in CIDEr). These comparisons suggest that both ICS and VGC are effective in helping the network learn useful visual content.

Moreover, Results in Table V also reveal the efficacy of VS in ICS. When employing the ICS strategy exclusively with VS, our method surpasses the baseline by 1.1% in the CIDEr metric. Conversely, when utilizing the complete ICS strategy excluding VS, there is a performance decline of 0.6%

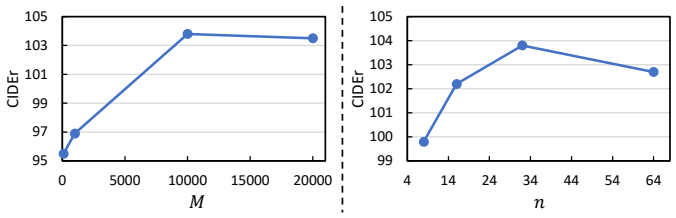

Fig. 3. Comparison of the values of CIDEr metrics for different sizes of the external image set (M) and different numbers of frames in the video samples (n).

TABLE VI
ABLATION OF τ. τ CHANGES THE INFLUENCES OF VGC AND WE SET $\tau = 0.5$ AS DEFAULT.

Parameter τ	BLEU-4	METEOR	ROUGE-L	CIDEr
$\tau = 0.1$	55.0	37.8	74.5	100.0
$\tau = 0.3$	57.2	38.3	75.3	102.8
$\tau = 0.5$	56.3	38.9	75.0	103.8
$\tau = 0.7$	55.8	38.0	75.0	103.0
$\tau = 0.9$	56.6	38.3	75.3	102.4

in the CIDEr metric. These comparative analyses underscore the efficacy of VS within the ICS strategy.

Ablation on frame number: The number of frames, denoted as n, plays a crucial role in determining the diversity of the training samples. As illustrated in Fig. 3, larger n values generally lead to better performances. However, this also increases the amount of noise introduced into the model. Interestingly, we observed that when n was set to 64, the results were slightly lower than those obtained with $n = 32$. We hypothesize that this is because larger n values result in more diverse and complex samples from the ICS strategy, making the training samples more difficult for the model to learn. Thus, the size of the training samples should not be too large. In our work, we set $n = 32$ to strike a balance between diversity and complexity.

In addition, we conducted an ablation study on the number of auxiliary image sets. The features x extracted from the image set provide additional visual content and semantics, which effectively enhance the video samples. As shown in Fig. 3, increasing the amount of image data results in better performance for our method. However, the improvements become marginal after introducing more than $10,000$ image samples. Therefore, the default value of M in our method is set to $10,000$.

D. Ablation on τ

Factor τ influences the modulation degree of VGC in IcoCap and should be properly set. As shown in Table VI, the results for $\tau = 0.5$ achieve the highest performances. Both larger and smaller values of τ lead to a decrease in performance. Moreover, a smaller value of τ causes S/τ in VGC to be closer to 1, resulting in lower punishments. This reduces the modulation from the consistent factor ϕ and the efficacy of our VGC. In experiments, a smaller value of τ also performs worse. These results further prove the efficacy of our method.

TABLE VII
ABLATION OF THE MIXUP RATIO α. THE RATIO INFLUENCES SAMPLES AFTER AUGMENTATIONS, WHICH SHOULD BE SET APPROPRIATELY.

Parameter α	BLEU-4	METEOR	ROUGE-L	CIDEr
$\alpha = 0.01$	55.4	38.2	75.0	102.2
$\alpha = 0.05$	56.3	38.9	75.0	103.8
$\alpha = 0.5$	58.2	37.7	74.1	98.3
Random α	58.5	37.4	74.0	97.8

TABLE VIII
ABLATION FOR SWAPPED FRAME RATIO s IN FS ON MSR-VTT.

s	75%	50%	25%	10%	0%
CIDEr	58.1	59.1	58.5	57.7	57.4

E. Ablation on Mixup Ratio α

We conducted extensive experiments to analyze the impact of different mixup ratios α. As demonstrated in Table VII, the value of this ratio needs to be carefully determined. When setting $\alpha = 0.05$, the performance surpasses other values, yielding the best results. Furthermore, we observed that using a random value for α leads to the creation of more challenging compounded samples. It is important to note that more difficult samples do not always guarantee improved performance and may potentially confuse the networks during the learning process. The results obtained with a random ratio are lower than those achieved with $\alpha = 0.05$, exhibiting a decrease of 5.0 in CIDEr.

F. Ablation on Swap Ratio in FS

In assessing the impact of content sampling in IcoCap, we introduce a swap ratio, denoted as s, to represent the proportion of content replaced by FS. The ablation study concerning the swap ratio s is shown in Tab. VIII, showcasing the CIDEr results on MSR-VTT.

When $s = 0$, it implies that no frames have been swapped within the visual content. On the other hand, $s = 50\%$ means that each frame has a 50% chance of being replaced by randomly selected visual content. The results illustrate that an optimal value for s diversifies inputs and enhances video learning. A lower value of s provides more straightforward input samples, reducing the need for generalization. In contrast, a very high value of s introduces more unrelated visual content, making it challenging for the network to process. In our research, we opted for $s = 50\%$, as it demonstrated the best results in our ablations.

Ablation on other baseline: We have extended our method to another baseline, VALOR [66], a large-scale pre-training model tailored for visual language tasks. Specifically, in line with the other baselines, we employed both the visual and text branches from VALOR and proceeded to fine-tune video captioning on the MSR-VTT dataset, following the settings for VALOR-B model. Through our implementation, we achieve a CIDEr metric of 61.05. Then, by integrating IcoCap with VALOR on the MSR-VTT dataset, we observed a further enhancement in performance, which are 61.53 in the CIDEr metric. With the large-scale pre-training, VALOR already

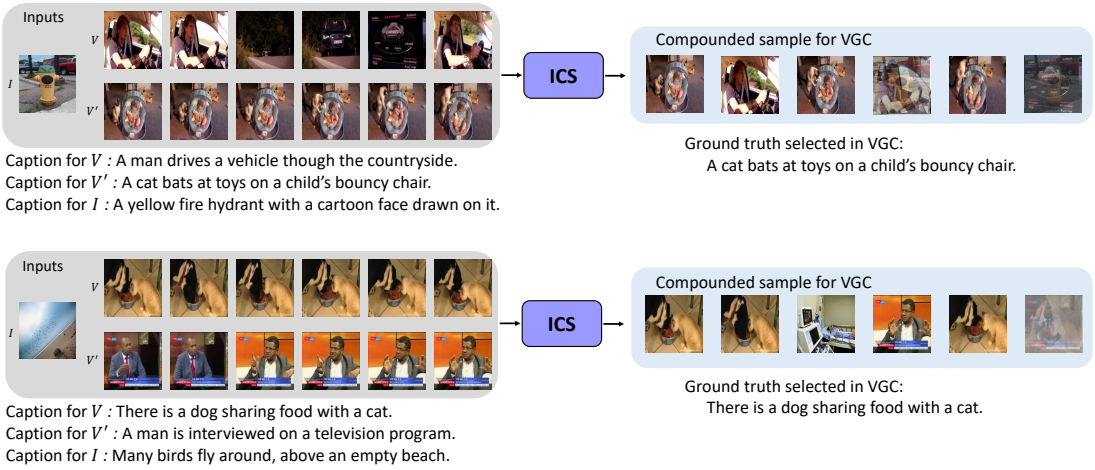

Fig. 4. Examples for input videos and images, compounded video samples, corresponding captions, and ground truth selected by VGC in IcoCap.

TABLE IX
PERFORMANCE OF IMAGE CAPTIONING. THE EXPERIMENTS ARE BASED ON THE TEST SET OF [49]. ONLY IMAGE SET INDICATES ONLY TRAINING WITH IMAGE SET AND WITHOUT VIDEO SET. THE FRAME NUMBER IS SET AS $n = 1$.

Data	BLEU-4	METEOR	ROUGE-L	CIDEr
Only Image Set	31.9	25.3	54.1	99.3
MSR-VTT + Image Set	28.0	23.5	50.5	89.1
MSVD + Image Set	32.4	25.8	54.4	101.1
VATEX + Image Set	23.3	21.4	46.4	73.9

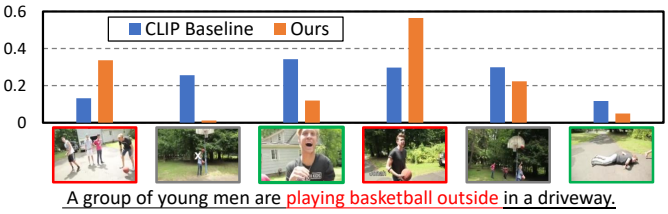

Fig. 5. Comparison of attention weights in the captioner. The video captioner is a standard transformer model. We provide a comparison of the attention weights of the last attention layer for the video frames. keyframes, transitions, and irrelevant frames are marked with red, gray, and green borders, according to the caption below. IcoCap produces larger attention weights for the keyframes and lower weights for transitions and irrelevant frames.

exhibits remarkable improvements. Meanwhile, our methodology improves the captioning capability even further. The improvements underscore the generalization of our IcoCap.

G. Performance in Image Captioning

Since IcoCap introduces additional image data, we also report its performance in image captioning. To ensure a fair comparison, we train IcoCap with all the available training data and evaluate it on the test set of the MSCOCO dataset [49].

As shown in Table IX, training with samples compounded with image samples also empowers the model's ability in image captioning. All models trained with IcoCap can solve image captioning. However, due to the differences in video sets, the performances on the image set vary. Among the different video sets, training with MSVD [51] leads to the highest results across various metrics. On the other hand, due to the domain gap between image and video data, the performances of models trained with MSR-VTT [50] and VATEX [52] are lower. The larger scale and complexity of MSR-VTT and VATEX may make it challenging for the network to learn complex video cues, thereby limiting the model's ability to improve semantic understanding.

H. Qualitative Analysis

Visualization for compounded samples: We provide detailed examples illustrating the application of ICS, exemplified in Fig. 4. This figure comprehensively presents illustrative instances, encompassing both input videos and images, the compounded samples by ICS, corresponding captions aligned with each input, and the definitive ground truth selected by VGC. These exemplars effectively spotlight two attributes of our approach: 1. The compounded samples showcase amplified diversity and reduced redundancy in comparison to the original inputs. These characteristics impose more demanding prerequisites on the captioner's learning process, thereby propelling the network to delve deeper into the realm of intricate visual content. 2. In IcoCap, the ground truth captions can be flexibly adapted based on the visual context. This phenomenon underscores the efficacy of our VGC in flexibly learning intricate visual contents.

Visualization for attention weights: In Fig. 5, we present a comprehensive visualization of the video frames, along with their corresponding attention weights after normalization. This illustration provides valuable insights into the attention mechanisms employed by our proposed method. Moreover, we offer a comparison of the attention weights for the video frames, specifically focusing on the last attention layer. Based on the caption provided below the figure, we have marked the keyframes, transitions, and irrelevant frames with red, gray, and green borders, respectively. Upon closer examination, it can be observed that IcoCap effectively assigns larger attention weights to keyframes, which are crucial for understanding the

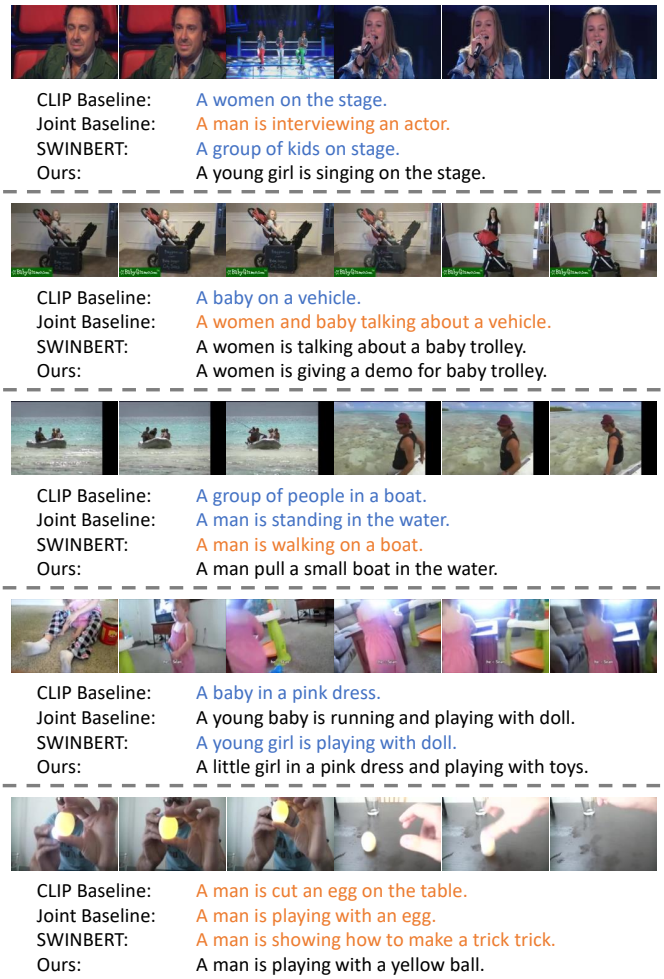

Fig. 6. Comparison of generated captions on MSR-VTT dataset. To better illustrate the difference, we mark some results in blue, which only describe the detailed and minor semantics of the overall video. Some incorrect descriptions for the visual contents are marked in orange. Our method shows better performances against the diverse contents and ambiguous semantics in videos.

content, while assigning lower weights to transition frames and irrelevant frames. This demonstrates the ability of our method to effectively capture and emphasize the most relevant aspects of the video content, ultimately leading to better captioning performance.

Visualization of captioning results: Results for the baselines, SWINBERT, and our method are shown in Fig. 6. Due to the complexity of visual content in videos, models may be biased and produce sentences that do not holistically describe the overall content. Some results, marked in blue, only express a part of the content in the videos, which may relate to detailed and minor events in the video data but fail to describe the major and valuable events. Additionally, some inaccurate descriptions are generated due to ambiguous semantics, marked in orange, that try to describe and summarize the content but are misled by the complex content and do not correctly reflect the semantics in video frames. The diverse semantics in video samples may confuse the network, making it difficult for the network to understand video content and exclude irrelevant content. In comparison, our method

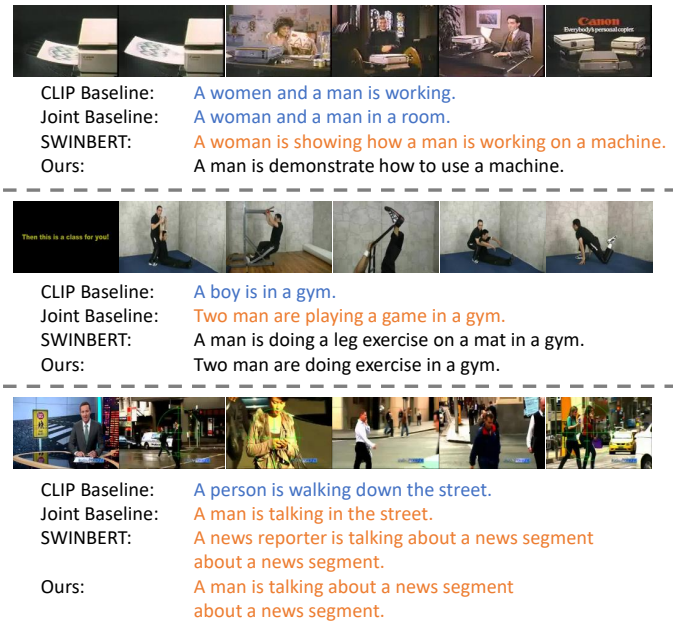

Fig. 7. Comparison of generated captions on VATEX dataset.

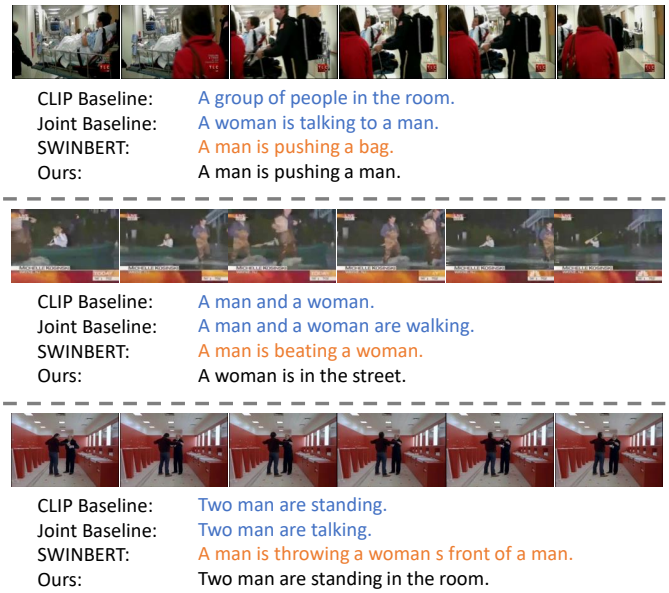

Fig. 8. Comparison of generated captions on MSVD dataset.

effectively improves the performance of handling complex visual content. The captioning results from our method can more precisely describe the video semantics.

We present a comparison of the generated results in MSVD and VATEX datasets, as shown in Fig. 7. VATEX dataset is more linguistically complex, with more diverse and complex descriptions than MSVD. Although our IcoCap does not specifically address this issue, it still achieves comparable results to state-of-the-art methods.

Moreover, benefiting from the compounded samples in our work, the network performs better with some complex video contents. For videos with multiple scenarios and characters (e.g., first row in Fig. 7 and first row in Fig. 8), our method is

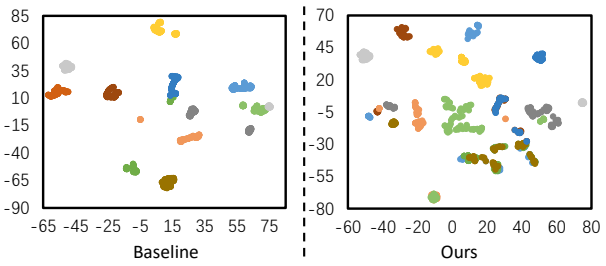

Fig. 9. Visualization for features in baseline and IcoCap.

not misled by the complicated semantics and provides accurate descriptions.

Visualization of features: We employed t-SNE to visualize the features generated by ICS, as depicted in Fig. 9. On the left, we present the baseline features, which are from the original CLIP features. Conversely, the right showcases the features within IcoCap, formulated by the ICS. Given the inherent redundancy in the original video frames, the baseline features tend to be more compact, which are easier for network learning. Such compactness might lead a captioner towards overfitting and pose challenges in learning with intricate semantics. In contrast, the features presented in IcoCap are more diversified and intricate. Their distribution also poses a higher level of complexity compared to the baseline. This demands a more rigorous learning paradigm from the captioner, urging it to achieve enhanced generalization for intricate visual semantics.

V. Conclusion

In this paper, we propose the Image-compounded video Captioner (IcoCap), a method that introduces image samples into the training procedure of video captioning to address the issue of ambiguous semantics in video data. Due to the complexity and diversity of video contents, it is difficult for the network to learn valuable video semantics. In contrast, image samples possess concise visual contents and clear semantics, making them easier to learn. The video samples compounded with image samples possess more difficult semantics. The network should learn to mine valuable video cues to solve the complex semantics. Specifically, In IcoCap, we propose Image-Compounding Strategy (ICS), which compounds video samples with images. ICS leads the network to handle complicated visual contents better and mine the valuable contents for captioning further. Besides, IcoCap also includes Visual-semantic Guided Captioning (VGC), which leads the network to learn the diverse video semantics flexibly. Experiments in various datasets prove the efficacy of our method. With a simple transformer network, we achieve comparable and even better performances in video captioning than the state-of-the-art methods.

Acknowledgment

This work was supported in part by the Australian Research Council (ARC) under Grant DP200100938.

References

[1] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[2] 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 *CVPR*, 2020.

[3] 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 *CVPR*, 2021.

[4] 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.

[5] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 949–17 958.

[6] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 959–17 968.

[7] Y. Li, Y. Pan, J. Chen, T. Yao, and T. Mei, "X-modaler: A versatile and high-performance codebase for cross-modal analytics," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3799–3802.

[8] X. Wu and H. Yu, "Mars-fl: Enabling competitors to collaborate in federated learning," *IEEE Transactions on Big Data*, pp. 1–11, 2022.

[9] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cognitive science*, vol. 9, no. 1, pp. 75–112, 1985.

[10] L. Zhu, H. Fan, Y. Luo, M. Xu, and Y. Yang, "Temporal cross-layer correlation mining for action recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 668–676, 2021.

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

[12] X. Wang, L. Zhu, Z. Zheng, M. Xu, and Y. Yang, "Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision," *IEEE Transactions on Multimedia*, pp. 1–11, 2022.

[13] Y. Li, R. Quan, L. Zhu, and Y. Yang, "Efficient multimodal fusion via interactive prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2604–2613.

[14] Y. Han, B. Wang, R. Hong, and F. Wu, "Movie question answering via textual memory and plot graph," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 875–887, 2019.

[15] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 661–669, 2012.

[16] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, and A. G. Hauptmann, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 572–581, 2012.

[17] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE transactions on multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[18] A. Wu, Y. Han, L. Zhu, and Y. Yang, "Instance-invariant domain adaptive object detection via progressive disentanglement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4178–4193, 2021.

[19] S. Mittal *et al.*, "A survey of accelerator architectures for 3d convolution neural networks," *Journal of Systems Architecture*, vol. 115, p. 102041, 2021.

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

[21] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.

[22] P. W. Dempsey, M. E. Allison, S. Akkaraju, C. C. Goodnow, and D. T. Fearon, "C3d of complement as a molecular adjuvant: bridging innate and acquired immunity," *Science*, vol. 271, no. 5247, pp. 348–350, 1996.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable

visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[24] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8746–8755.

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

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," pp. 4171–4186, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1423

[27] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7464–7473.

[28] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 437–10 446.

[29] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9758–9770, 2020.

[30] X. Chen, H. Fan, R. B. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *CoRR*, vol. abs/2003.04297, 2020. [Online]. Available: https://arxiv.org/abs/2003.04297

[31] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3024–3033.

[32] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li, and X. Li, "Clip4caption: Clip for video caption," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4858–4862.

[33] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–37, 2019.

[34] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE transactions on multimedia*, vol. 22, no. 1, pp. 229–241, 2019.

[35] S. Chen, Q. Jin, J. Chen, and A. G. Hauptmann, "Generating video descriptions with latent topic guidance," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2407–2418, 2019.

[36] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[37] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, and T. Mei, "Temporal deformable convolutional encoder-decoder networks for video captioning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8167–8174.

[38] S. Chen and Y.-G. Jiang, "Motion guided region message passing for video captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1543–1552.

[39] X. Li, B. Zhao, X. Lu *et al.*, "Mam-rnn: multi-level attention model based rnn for video captioning." in *IJCAI*, vol. 2017, 2017, pp. 2208–2214.

[40] 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.

[41] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4594–4602.

[42] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[43] S. Liu, Z. Ren, and J. Yuan, "Sibnet: Sibling convolutional encoder for video captioning," *IEEE TPAMI*, 2020.

[44] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *CVPR*, 2019.

[45] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical boundary-aware neural encoder for video captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1657–1666.

[46] B. Yang, T. Zhang, and Y. Zou, "Clip meets video captioning: Concept-aware representation learning does matter," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2022, pp. 368–381.

[47] W. Zhao, X. Wu, and J. Luo, "Multi-modal dependency tree for video captioning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6634–6645, 2021.

[48] H. Wang, Y. Xu, and Y. Han, "Spotting and aggregating salient regions for video captioning," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1519–1526.

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[50] 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.

[51] D. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 190–200. [Online]. Available: https://aclanthology.org/P11-1020

[52] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[53] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017. [Online]. Available: http://arxiv.org/abs/1705.06950

[54] S. Chen, W. Jiang, W. Liu, and Y.-G. Jiang, "Learning modality interaction for temporal sentence localization and event captioning in videos," in *ECCV*, 2020.

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

[56] 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 *ICCV*, 2019.

[57] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, "Joint syntax representation learning and visual cue translation for video captioning," in *ICCV*, 2019.

[58] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in *CVPR*, 2020.

[59] H. Li, D. Song, L. Liao, and C. Peng, "Revnet: Bring reviewing into video captioning for a better description," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1312–1317.

[60] H. Ye, G. Li, Y. Qi, S. Wang, Q. Huang, and M.-H. Yang, "Hierarchical modular network for video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 939–17 948.

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

[62] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.

[63] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

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

[65] M. Patrick, P.-Y. B. Huang, Y. M. Asano, F. Metze, A. Hauptmann, J. F. Henriques, and A. Vedaldi, "Support-set bottlenecks for video-text representation learning," vol. abs/2010.02824, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222142276

[66] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu, "Valor: Vision-audio-language omni-perception pretraining model and dataset," *ArXiv*, vol. abs/2304.08345, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258179576