Set Prediction Guided by Semantic Concepts for Diverse Video Captioning

Yifan Lu[1,2*], **Ziqi Zhang**[1*], **Chunfeng Yuan**[1†], **Peng Li**[3,4], **Yan Wang**[3,4], **Bing Li**[1], **Weiming Hu**[1,2,5]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Alibaba Group
[4]Zhejiang Linkheer Science and Technology Co., Ltd.
[5]School of Information Science and Technology, ShanghaiTech University
{luyifan2021, zhangziqi2017}@ia.ac.cn, {cfyuan, bli, wmhu}@nlpr.ia.ac.cn, {sanjie.lp, wy84378}@alibaba-inc.com

Abstract

Diverse video captioning aims to generate a set of sentences to describe the given video in various aspects. Mainstream methods are trained with independent pairs of a video and a caption from its ground-truth set without exploiting the intra-set relationship, resulting in low diversity of generated captions. Different from them, we formulate diverse captioning into a semantic-concept-guided set prediction (SCG-SP) problem by fitting the predicted caption set to the ground-truth set, where the set-level relationship is fully captured. Specifically, our set prediction consists of two synergistic tasks, i.e., caption generation and an auxiliary task of concept combination prediction providing extra semantic supervision. Each caption in the set is attached to a concept combination indicating the primary semantic content of the caption and facilitating element alignment in set prediction. Furthermore, we apply a diversity regularization term on concepts to encourage the model to generate semantically diverse captions with various concept combinations. These two tasks share multiple semantics-specific encodings as input, which are obtained by iterative interaction between visual features and conceptual queries. The correspondence between the generated captions and specific concept combinations further guarantees the interpretability of our model. Extensive experiments on benchmark datasets show that the proposed SCG-SP achieves state-of-the-art (SOTA) performance under both relevance and diversity metrics.

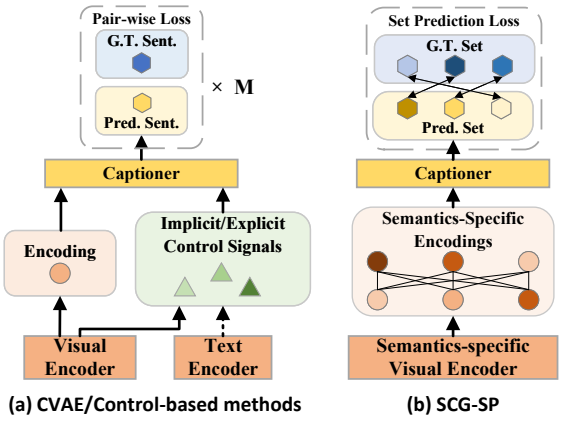

Figure 1: Difference between (a) existing CVAE/control-based diverse captioning methods and (b) our proposed SCG-SP. There is no direct interaction among generated captions in CVAE-based or control-based methods, where the loss is calculated with independent training samples. Our proposed SCG-SP generates captions based on multiple semantics-specific visual encodings with sufficient interaction and is trained by a set-level prediction loss to exploit the set-level relationship.

Introduction

Diverse video captioning (DivVC) is an emerging research branch of vision-language tasks. DivVC aims to generate a set of multiple captions that are semantically related to the given video and distinct from one another. DivVC overcomes the limitation of traditional video captioning methods, which only generate a single sentence for a video and fail to cover a wealth of visual information (Liu et al. 2022; Nie et al. 2022).

Mainstream diverse captioning methods can be classified into two categories: conditional variational encoder (CVAE) based methods (Aneja et al. 2019; Chen et al. 2019; Deb et al. 2022; Jain, Zhang, and Schwing 2017; Liu et al. 2022; Mahajan, Gurevych, and Roth 2020; Mahajan and Roth 2020; Nie et al. 2022) and control-based methods (Chen, Deng, and Wu 2022; Cornia, Baraldi, and Cucchiara 2019; Deshpande et al. 2019). As shown in Fig.1(a), both methods share a general pipeline that the diverse generation is conditioned on a visual encoding and multiple implicit or explicit control signals.

CVAE-based methods learn a latent distribution to encode diversity. Multiple latent variables, regarded as implicit control signals, are sampled from the learned distribution for diverse caption generation. However, CVAE-based methods are trained with a pair-wise loss, which ignores the intra-set relationship. This issue arises from the absence of direct interaction between generated captions for the same video and leads to inadequate capture of the diverse characteristics of captions. Moreover, CVAE-based methods have limited interpretability because of the unrevealed relationship

*These authors contributed equally.
†Corresponding author.

between the latent distribution and language patterns of the generated captions. Beyond CVAE-based methods, control-based methods apply explicit control signals to mitigate the problem of low interpretability. However, they still independently process the triplet of visual content, control signal, and caption, with the problem of ignoring intra-set relationship unsolved.

In this work, we propose a novel diverse video captioning model, **S**emantic-**C**oncept-**G**uided **S**et **P**rediction (SCG-SP). To better exploit set-level relationships, we formulate diverse captioning as a set prediction problem. The target of DivVC is achieved by fitting the predicted caption set to the ground-truth caption set. To better grasp the intra-set diversity characteristics, we consider the source of caption diversity within a set. An important observation is that the diversity of captions depends on the differences in the semantic concepts they contain (Wang and Chan 2019). These semantic concepts correspond to different objects, scenes, or actions in the visual content (Fang et al. 2015). Since different combinations of concepts lead to different interpretations of rich visual content and are suitable for serving as guidance for caption generation, we incorporate concepts into set prediction for DivVC task.

Specifically, as shown in Fig.1(b), the intra-set reasoning is achieved by interactions between multiple semantics-specific encodings and the set-level loss. The encodings fully interact with each other, and each of them is decoded into a caption sentence accompanied by a combination of concepts. Therefore, our set prediction consists of two aspects, i.e., caption generation and the auxiliary concept combination prediction. For caption generation, we apply a GPT-2 (Radford et al. 2019) captioner to leverage the vast knowledge learned from the external corpus. The task of concept combination prediction is realized through multi-label classification, serving as an auxiliary task and providing semantic supervision. SCG-SP is optimized by a set-level prediction loss composed of both captioning and classification costs. Through set-level reasoning, SCG-SP concerns the relationship within a set of diverse captions for the same video (intra-set) as well as the relationship among sets for various videos (inter-set), achieving considerable performance on both diversity and relevance. Furthermore, each caption is guided by a particular concept combination, thus making the diverse generation interpretable.

The contributions of our work are listed as follows:

- We propose a novel diverse video captioning model named SCG-SP to formulate DivVC as a set prediction problem. By exploiting intra-set and inter-set relationships, our model achieves high captioning performance in terms of both relevance and diversity.
- We further incorporate semantic concept guidance to promote set-level reasoning through concept detection and concept combination prediction, enabling SCG-SP to generate captions with high semantic diversity in an interpretable way.
- Extensive experiments on MSVD, MSRVTT, and VATEX demonstrate that our proposed SCG-SP achieves state-of-the-art performances over existing methods.

Related Works

Diverse Captioning Diverse image captioning (DivIC) has been an essential branch of image captioning. Some of the early methods are based on beam search (Vijayakumar et al. 2018) and generative adversarial networks (Dai et al. 2017; Li et al. 2018). Recent DivIC methods (Aneja et al. 2019; Chen et al. 2019; Jain, Zhang, and Schwing 2017; Mahajan, Gurevych, and Roth 2020; Mahajan and Roth 2020) employ CVAE (Sohn, Lee, and Yan 2015) for diverse generations. Beyond CVAE-based models, Some studies (Deshpande et al. 2019; Cornia, Baraldi, and Cucchiara 2019; Chen, Deng, and Wu 2022) introduce control signals with explicit meaning, enabling diverse generations with interpretability.

In the domain of video, FLIP (Nie et al. 2022) applies CVAE with latent prior modeled by a normalizing flow. VSLAN (Deb et al. 2022) employs a CVAE to generate diverse Part-of-Speech sequences for caption generation. SMCG (Yuan et al. 2020) generates diverse captions controlled by exemplar sentences. However, these methods ignore the intra-set relationship, for there is no set-level loss or interaction among generated captions in these methods. The recent-proposed STR (Liu et al. 2022) takes the intra-set relationship into consideration for DivVC. STR clusters captions by topics and learns a latent distribution through two-stage training, respectively focusing on topics and paraphrasing. Note that our proposed SCG-SP performs set-level reasoning only requiring one-stage training, and is free of handcrafted components like clustering, making the pipeline simple yet effective.

Set Prediction Set prediction first emerges as a new paradigm for object detection. DETR (Carion et al. 2020) applies a transformer encoder-decoder framework (Vaswani et al. 2017) to directly predict the set of objects in parallel. DETR is optimized by a set prediction loss based on the Hungarian algorithm (Kuhn 1955) finding a unique element matching between the predicted and the ground-truth sets. The set prediction framework has been extended to various multi-modal tasks, such as text-conditioned object detection (Kamath et al. 2021), and spatio-temporal video grounding (Yang et al. 2022). PDVC (Wang et al. 2021) uses DETR for describing videos with multiple events. HMN (Ye et al. 2022) applies a DETR-based module to predict a set of entities in video for single sentence caption. Our proposed SCG-SP applies the set prediction framework, which is consistent with the target of DivVC. Unlike HMN or PDVC, SCG-SP predicts a set of captions for a video with a single event.

Semantic Guidance Visual concepts corresponding to objects, scenes, or actions in the visual content can encode rich and high-level semantic cues (Fang et al. 2015). Incorporating semantic concepts plays a vital role in vision-to-language tasks. Previous works (Gan et al. 2017; Fang et al. 2022; Pan et al. 2017; Perez-Martin, Bustos, and Pérez 2021) detect semantic concepts from visual content leveraging Multiple Instance Learning (MIL) (Maron and Lozano-Pérez 1997). The predicted probability distribution of concepts is used as a high-level visual feature for further caption generation. Different from previous works, SCG-SP applies

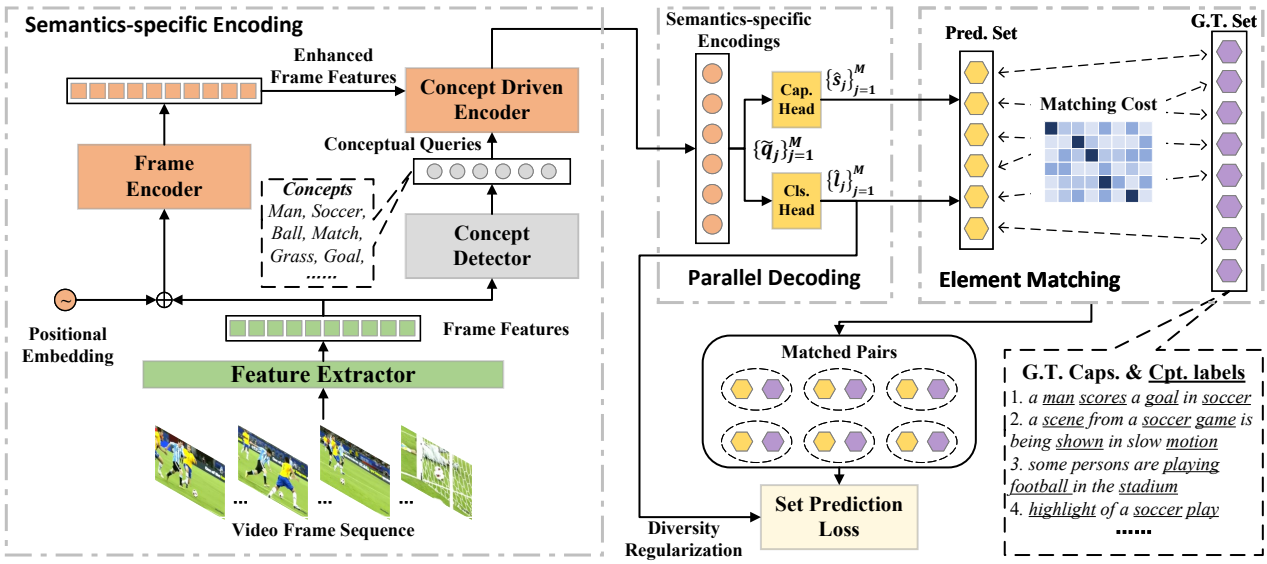

Figure 2: Overview of the proposed SCG-SP. Based on pre-extracted video frame features, we first employ a temporal encoder, a concept detector, and a concept driven encoder to obtain multiple semantics-specific encodings for the input video. In the parallel decoding stage, we apply a caption head and a classification head to respectively decode each encoding into a caption sentence and a concept combination label, which together form the prediction set. By performing element matching between the predicted set and the ground-truth set, the set prediction loss is calculated over matched element pairs. Note that the ground-truth concept combination labels are assigned by taking nouns and verbs of high word frequency from the captions.

semantic concept to guide the Set Prediction process via two novel ways of conceptual queries and supervision from the classification task.

Methodology

Problem Formulation

In the proposed SCG-SP, we formulate diverse video captioning as a set prediction problem. Given a video frame sequence $\mathcal{V}$, SCG-SP directly predicts a caption set with capacity M, which is denoted as $\hat{\mathcal{C}} = \{\hat{c}_j | \hat{c}_j = (\hat{\mathbf{s}}_j, \hat{\mathbf{l}}_j)\}_{j=1}^{M}$. For concept guidance, each caption sentence $\hat{\mathbf{s}}_j$ is attached with a concept combination label $\hat{\mathbf{l}}_j$. Similarly, the ground-truth caption set with capacity M' is denoted as $\mathcal{C} = \{c_j | c_j = (\mathbf{s}_j, \mathbf{l}_j)\}_{j=1}^{M'}$. The learning target is to fit the predicted set $\hat{\mathcal{C}}$ to the ground-truth set $\mathcal{C}$. Fig.2 illustrates the architecture of SCG-SP, which consists of three main blocks: semantics-specific encoding, parallel decoding, and loss calculation. Details are discussed in the following subsections.

Semantics-specific Encoding

In the semantics-specific encoding stage, we represent an input video with multiple encodings $\tilde{\mathcal{Q}}$, each of which is subsequently decoded into a distinct caption. The encodings are derived from iterative attention on conceptual queries $\mathcal{Q}$ and temporal enhanced frame features $\tilde{\mathcal{F}}$, which infuse specific semantics into the encodings. The underlying interactions between encodings also facilitates the interactions between their corresponding generated captions.

Temporal Encoding Given a sequence of video frames $\mathcal{V}$, we first employ pretrained 2D CNNs and 3D CNNs to extract the appearance and motion features of the video, respectively. By sampling keyframes and temporal concatenation, we get the frame features denoted as $\mathcal{F} = \{\mathbf{f_i}\}_{i=1}^{N}$ for further encoding. Then, we apply a Temporal Encoder composed of multiple self-attention layers. By performing self-attention on the frame feature sequence $\mathcal{F}$, the Temporal Encoder exploits inter-frame relationships and extract temporal-enhanced frame features denoted as $\tilde{\mathcal{F}} = \{\tilde{\mathbf{f}}_i\}_{i=1}^{N}$.

Concept Detection We design a Concept Detector to detect visual concepts in a video and generate conceptual queries $\mathcal{Q} = \{\mathbf{q}_j\}_{j=1}^{M}$. We firstly build the concept vocabulary by selecting N_c nouns or verbs with the highest word frequency. Subsequently, we tag each caption $\mathbf{s}_j$ with a 0/1 concept combination label $\mathbf{l}_j \in \mathbb{R}^{N_c}$, where 1 is assigned to existing concepts and 0 for non-existing ones. Finally, the pseudo ground-truth concept label of a video $\mathbf{l}_V$ is obtained by taking the bit-wise *OR* operation: $\mathbf{l}_V = \mathbf{l}_1 \mid \mathbf{l}_2 \mid ... \mid \mathbf{l}_{M'}$, which means the k-th concept exists in the video if it appears in any of the M' ground-truth captions.

The Concept Detector performs multi-label classification task with $\mathbf{l}_V$ as the ground truth. It comprises a multi-layer perceptron (MLP) and predicts $\hat{\mathbf{l}}^V$, the probabilities of the N_c concepts appearing in a video based on mean-pooled frame features. The M conceptual queries $\mathcal{Q} = \{\mathbf{q}_j\}_{j=1}^{M}$ are obtained by taking the linear-projected GloVe embeddings (Pennington, Socher, and Manning 2014) of the M concepts with the highest probabilities. Since the operation of taking the maximum is not differentiable, the detector is

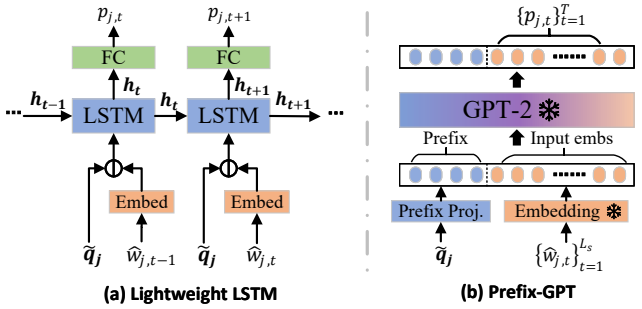

Figure 3: Illustration of (a) the lightweight LSTM captioner and (b) the prefix-GPT captioner.

trained offline from the entire SCG-SP pipeline.

Concept Driven Encoding The Concept Driven Encoder takes the conceptual queries $\mathcal{Q}$ and the temporal-enhanced frame sequence $\tilde{\mathcal{F}}$ as input to generate semantics-specific encodings $\tilde{\mathcal{Q}} = \{\tilde{\mathbf{q}}_j\}_{j=1}^{M}$. Built on the transformer-decoder architecture (Vaswani et al. 2017), it first groups concepts into combinations using self-attention on conceptual queries, resulting in representations encoding specific semantics. In the subsequent cross-attention layer, these concept-level representations act as queries, with frame-level video features serving as keys and values, to generate context-aware encodings that capture the relationship between specific concepts and video frames. The encodings are further refined through successive self-attention layers.

Parallel Decoding

In the parallel decoding stage, the semantics-specific encodings are decoded into elements of the prediction set. A captioning head generates captions $\{\hat{\mathbf{s}}_j\}_{j=1}^{M}$ for each encoding. Besides, a classification head predicts concept combination labels $\{\hat{\mathbf{l}}_j\}_{j=1}^{M}$, performing an auxiliary task that provides extra semantic supervision. The two heads share the semantics-specific encodings as input.

Captioning Head We apply two kinds of captioning heads to show the compatibility of our framework with both lightweight LSTM and large-scale pretrained GPT-2, as shown in Fig.3. Specifically, the j-th predicted caption $\hat{\mathbf{s}}_j$ consists of predicted words, i.e. $\hat{\mathbf{s}}_j = (\hat{w}_{j,1}, \hat{w}_{j,2}, \dots \hat{w}_{j,T})$. At time step t, $\hat{w}_{j,t}$ is conditioned on previous words $\hat{w}_{j,1:t-1}$ and the j-th encoding $\tilde{\mathbf{q}}_j$, i.e. $p(\hat{w}_{j,t}|\tilde{\mathbf{q}}_j, \hat{w}_{j,1:t-1})$ (denoted as $p_{j,t}$ for simplicity). $p_{j,t}$ is modelled by the captioning head as shown in eq.1 and eq.2 respectively for LSTM and GPT-2:

$$p_{j,t} = \text{Softmax}\Big(\text{LSTM}\big(\big[\tilde{\mathbf{q}}_j, \text{Emb}(\hat{w}_{j,t-1})\big]; \mathbf{h}_{j,t-1}\big)\Big), \tag{1}$$

$$p_{j,t} = \text{Softmax}\Big(\text{GPT2}\big(\big[\text{MLP}(\tilde{\mathbf{q}}_j), \text{Emb}(\hat{w}_{j,1:t-1})\big]\big)\Big), \tag{2}$$

where $[\cdot,\cdot]$ denotes concatenation; Emb is the word embedding layer; $\mathbf{h}_{j,t-1}$ denotes the previous hidden state of LSTM. The MLP projects $\tilde{\mathbf{q}}_j$ into word embedding space to serve as a prefix prompt. It is efficient to adapt the GPT-2 captioner to the captioning task by learning a prefix while freezing the GPT-2.

Classification Head The classification head is composed of an MLP. For the j-th element, it predicts the probability distribution of concepts $\hat{\mathbf{l}}_j \in \mathbb{R}^{N_c}$, i.e., the concept combinations, based on the encoding $\tilde{\mathbf{q}}_j$.

Semantic Concept Guided Set Prediction Loss

The proposed SCG-SP is optimized by a set prediction loss where the predicted and ground-truth set are treated as a whole to achieve set-level reasoning. The set prediction loss starts with a semantic concept guided element matching. The deterministic and optimal matching allows the loss to more accurately measure the differences between sets, aiding in better optimization. After that, the set prediction loss is calculated by pair. Besides caption cost, we consider classification loss of concept combination labels for extra semantic supervision. We further design a diversity regularization term on concepts to encourage the model to generate semantically diverse captions with various concept combinations.

Specifically, the element matching finds the optimal element assignment $\hat{\sigma}$ between $\hat{\mathcal{C}}$ and $\mathcal{C}$ with the lowest matching cost based on concept combination labels. Assuming $\hat{\mathcal{C}}$ has a capacity no larger than $\mathcal{C}$, i.e., $M \leq M'$, each predicted element is assigned to a unique ground-truth element, i.e., $c_{\hat{\sigma}(j)}$ and $\hat{c}_j$ are matched into a pair. The optimal element assignment $\hat{\sigma}$ is obtained by using Hungarian algorithm (Kuhn 1955) to solve:

$$\hat{\sigma} = \arg\min_{\sigma \in \Omega_{M'}} \sum_{j=1}^{M} L_{fl}\big(\mathbf{l}_{\sigma(j)}, \hat{\mathbf{l}}_j\big), \tag{3}$$

where $\Omega_{M'}$ is the searching space for M' elements; $L_{fl}(\cdot,\cdot)$ refers to the focal loss (Lin et al. 2017).

The captioning loss L_{cap} is based on the Cross-Entropy loss of words:

$$L_{cap} = -\sum_{j=1}^{M}\sum_{t=1}^{T} \delta(w_{\hat{\sigma}(j),t})^{\top} \log p_{j,t}, \tag{4}$$

where $\delta(w_{\hat{\sigma}(j),t})$ denotes the one-hot vector for the j-th word in $\hat{\sigma}(j)$-th ground-truth caption; L_s is the length of the sentence. The classification loss is based on focal loss:

$$L_{cls} = \sum_{j=1}^{M} L_{fl}\big(\mathbf{l}_{\hat{\sigma}(j)}, \hat{\mathbf{l}}_j\big), \tag{5}$$

The diversity regularization term L_{div} is based on standard deviation (StdDev):

$$L_{div} = -\sum_{k=1}^{N_c} \text{StdDev}(\hat{l}_{1,k}, \hat{l}_{2,k}, \dots, \hat{l}_{M,k}), \tag{6}$$

where $\hat{l}_{j,k}$ is the probability of the k-th concept for the j-th element. Minimizing L_{div} encourages SCG-SP to exploit more concept combinations. At last, the set prediction loss is obtained by taking the weighted sum of the captioning loss, the classification loss, and the diversity regularization term:

$$L_{sp} = L_{cap} + \lambda L_{cls} + \lambda_d L_{div}. \tag{7}$$

Experiments

In this section, we evaluate our proposed model on three public video datasets: MSVD (Chen and Dolan 2011), MSRVTT (Xu et al. 2016), and VATEX (Wang et al. 2019), under both relevance and diversity metrics. We present the results compared with state-of-the-art models and report main ablation studies to demonstrate the effectiveness of our methods. More results are in **Supplementary Material**.

Experimental Setup

Datasets **MSVD** contains 1970 video from YouTube. Each video is annotated with 41 captions on average. We follow the split of 1200/100/670 for training, validation, and test. **MSRVTT** contains 10000 open domain videos. Each video is annotated with 20 captions. We follow the split of 6513/497/2990 for training, validation, and test. **VATEX** contains 34991 videos, each with 10 English captions. We follow the split of 25991/3000/6000 for training, validation, and test.

Preprocessing For each video, 8 frames/clips are sampled for feature extraction. We employ pre-extracted ResNet-101 (He et al. 2016) and C3D (Hara, Kataoka, and Satoh 2018) features for MSVD and MSRVTT videos, and I3D (Carreira and Zisserman 2017) for VATEX. For captions sentences, we use NLTK toolkit (Bird, Klein, and Loper 2009) for part-of-speech tagging. We apply the 300-d GloVe embedding of detected concepts as conceptual queries. The size of the ground-truth caption set, M', is set to 20/20/10 for MSVD/MSRVTT/VATEX.

Implementation Details The size of the concept vocabulary N_c is set to 1000. The capacity of the predicted set and the conceptual query number M is set to 20. We implement two versions of our proposed model, i.e., SCG-SP-LSTM and SCG-SP-Prefix, using Lightweight LSTM and GPT-2 with prefix as captioners, respectively. For the Prefix-GPT captioner, we employ the smallest official version of GPT-2. The prefix length is set to 10. The weights of loss terms are set as $\lambda = 1$ and $\lambda_d = 0.5$. We apply AdamW as the optimizer. The learning rate and batch size are set to $8e^{-5}$ and 32 for SCG-SP-LSTM, $1e^{-5}$ and 8 for SCG-SP-Prefix. We use beam search with size 3 for generation at the inference stage. The model is implemented with PyTorch, and all the experiments are conducted on 1 RTX 3090 GPU.

Evaluation Metrics The relevance of the captions is evaluated with metrics including BLEU@4 (B@4) (Papineni et al. 2002), METEOR (M) (Banerjee and Lavie 2005), ROUGE-L (R-L) (Lin 2004), and CIDEr (C) (Vedantam, Lawrence Zitnick, and Parikh 2015). We report **oracle** scores, where only the top-1 caption from each set is selected, demonstrating the upper-bound performance.

The diversity of the captions is evaluated with metrics including Div-n (D-n), m-BLEU (m-B) (Aneja et al. 2019), and self-CIDEr (s-C) (Wang and Chan 2019). Div-n is the average ratio of distinct n-grams within each predicted caption set. We report Div-1 and Div-2 in our experiments. M-BLEU computes BLEU@4 for each diverse caption with the remaining captions in the set. Self-CIDEr calculates the ratio of the largest eigenvalue of the kernel matrix composed of CIDEr values between all pairs of captions in the set. **Higher** Div-n, self-CIDEr, and **lower** m-BLEU indicate more sentence diversity.

Models	MSVD				MSRVTT			
	B@4	M	R-L	C	B@4	M	R-L	C
*top-1 over **20** sentences*								
Div-BS†	39.1	45.5	75.2	85.5	47.3	41.9	68.3	60.5
SeqCVAE‡	50.7	57.8	81.0	113.4	44.9	43.2	69.7	64.5
COSCVAE‡	45.9	52.8	78.9	105.3	41.8	41.8	68.5	63.6
DML†	54.9	49.8	79.7	105.7	48.5	37.5	68.2	56.8
STR	54.5	57.2	81.5	115.2	47.2	44.2	71.0	67.2
SCG-SP-LSTM	**58.4**	**60.5**	**82.9**	**120.2**	**54.5**	47.2	**72.4**	67.0
SCG-SP-Prefix	56.5	58.9	81.0	114.7	54.4	**47.5**	72.1	**67.7**
*top-1 over **10** sentences*								
VSLAN	**57.4**	36.9	75.6	98.1	46.5	32.8	62.4	55.8
SCG-SP-LSTM	52.4	**57.7**	**81.1**	**110.8**	**48.1**	**43.8**	**69.8**	**61.7**
SCG-SP-Prefix	51.6	56.1	79.7	109.9	46.4	43.4	69.2	61.5

Table 1: The oracle relevance scores on MSVD and MSRVTT. † indicates DivIC models re-implemented for DivVC by us. ‡ indicates DivIC models re-implemented by authors of STR.

Models	MSVD				MSRVTT			
	D-1	D-2	m-B↓	s-C	D-1	D-2	m-B↓	s-C
***20** sentences*								
Div-BS	20.0	41.9	**29.9**	**82.5**	22.1	42.0	**49.4**	**80.9**
DML	18.9	34.2	70.8	60.3	20.1	36.1	69.9	63.0
SCG-SP-LSTM	21.0	33.7	63.9	54.1	21.4	37.0	65.1	62.6
SCG-SP-Prefix	**27.4**	**45.9**	50.2	64.8	**24.3**	**43.1**	58.4	67.9
***10** sentences*								
VSLAN	32.0	36.0	62.0	-	30.0	33.0	58.0	-
STR	28.2	48.5	60.1	-	33.4	**58.4**	**46.3**	-
SCG-SP-LSTM	25.7	39.0	62.6	54.5	30.5	47.7	56.9	64.6
SCG-SP-Prefix	**33.3**	**52.4**	**58.4**	**67.9**	**34.0**	54.0	48.2	**70.3**

Table 2: The diversity scores on MSVD and MSRVTT.

Performance Comparison with SOTA

To evaluate the effectiveness of SCG-SP, we compare our model with SOTA methods for DivVC, i.e., Div-BS (Vijayakumar et al. 2018), SeqCVAE (Aneja et al. 2019), COSCVAE (Mahajan and Roth 2020), DML (Chen, Deng, and Wu 2022), STR (Liu et al. 2022), and VSLAN (Deb et al. 2022). Note that Div-BS, SeqCVAE, COSCVAE, and DML are re-implemented based on corresponding DivIC methods. For a fair comparison with VSLAN and STR, we implement a version of SCG-SP with 10 generated captions per video on MSVD and MSRVTT.

Tab.1 shows the oracle relevance scores of DivVC models on MSVD and MSRVTT. On MSVD, SCG-SP-LSTM achieves the best performance on METEOR, ROUGE-L, and CIDEr. Though with lower BLEU@4, both SCG-SP-LSTM and SCG-SP-Prefix outperform VSLAN on the rest

Models	Relevance Scores				Diversity Scores			
	B@4	M	R-L	C	D-1	D-2	m-B ↓	s-C
Div-BS	30.6	24.7	51.0	53.5	19.1	**36.8**	69.7	**64.8**
DML	**36.1**	24.8	54.1	56.2	15.9	29.2	80.2	48.7
SCG-SP-LSTM	32.0	25.8	52.8	56.4	10.4	18.5	91.5	52.4
SCG-SP-Prefix	34.5	**26.1**	**54.1**	**61.2**	**24.4**	34.6	**63.8**	60.1

Table 3: Relevance and diversity scores on VATEX.

Models	MSRVTT			
	B@4	M	R-L	C
ORG-TRL	43.6	28.8	62.1	50.9
SGN	40.8	28.3	60.8	49.5
Open-Book	42.8	29.3	61.7	52.9
SemSynAN	46.4	30.4	64.7	51.9
HMN	43.5	29.9	62.7	51.5
TextKG	43.7	29.6	62.4	52.4
MV-GPT§	48.9	38.7	64.0	60.0
SwinBERT§	45.4	30.6	64.1	55.9
Vid2Seq§	-	30.8	-	64.6
SCG-SP-LSTM	**54.5**	47.2	**72.4**	67.0
SCG-SP-Prefix	54.4	**47.5**	72.1	**67.7**

Table 4: Comparison with the SOTA methods for single sentence video captioning on MSRVTT. § indicates models pretrained on large-scale datasets

of the metrics by a large margin. On MSRVTT, our method achieves the best performance on all the metrics.

Tab.2 shows the diversity scores on MSVD and MSRVTT. Except for Div-BS, SCG-SP-Prefix scores the highest on MSVD and keeps the diversity on par with STR while outperforming the rest on MSRVTT. STR, as mentioned previously, also achieves set-level perception. The high diversity scores of STR and SCG-SP from the side prove the significance of the set-level relationship for diverse captioning. Note that Div-BS performs the best on two sentence-level diversity metrics because of its characteristic of generating all unique sentences. However, Div-BS performs poorly under relevance metrics.

Tab.3 shows the relevance and diversity scores on VATEX. Our proposed SCG-SP-Prefix has the best overall performance. To summarize, our model achieves considerable performance on the three benchmarks under both metrics, verifying the effectiveness of our proposed methods.

We also compare our model with SOTA methods for traditional single sentence video captioning, including ORG-TRL (Zhang et al. 2020), SGN (Ryu et al. 2021), OpenBook (Zhang et al. 2021), SemSynAN (Perez-Martin, Bustos, and Pérez 2021), HMN (Ye et al. 2022), TextKG (Gu et al. 2023), MV-GPT (Seo et al. 2022), SwinBERT (Lin et al. 2022), and Vid2Seq (Yang et al. 2023), as shown in Tab.4. For our model, we use oracle relevance scores derived from the best of 20 generated captions per video. For traditional models, scores are calculated using the single predicted caption. The significant improvement in scores of our model demonstrates its superior ability to generate high-quality captions,

Configurations	Relevance Scores				Diversity Scores			
	B@4	M	R-L	C	D-1	D-2	m-B ↓	s-C
LSTM-Base	51.1	45.4	71.2	64.7	17.3	29.3	75.9	53.9
- w/o cpt. queries	49.7	45.3	70.7	63.5	13.8	22.8	86.5	46.2
- w/o cls. head	38.5	40.2	66.8	56.7	8.4	11.1	96.2	19.8
- w/ L_{div}	**54.5**	**47.2**	**72.4**	**67.0**	**21.4**	**37.0**	**65.1**	**62.6**
Prefix-Base	51.6	46.3	71.2	65.9	22.5	39.8	63.1	64.4
- w/o cpt. queries	49.8	46.5	70.9	66.2	16.9	28.1	82.2	49.8
- w/o cls. head	48.8	45.2	70.0	64.3	18.6	31.1	75.3	54.3
- w/ L_{div}	**54.4**	**47.5**	**72.1**	**67.7**	**34.0**	**54.0**	**48.2**	**70.3**

Table 5: Ablative results on Semantic Concept Guidance.

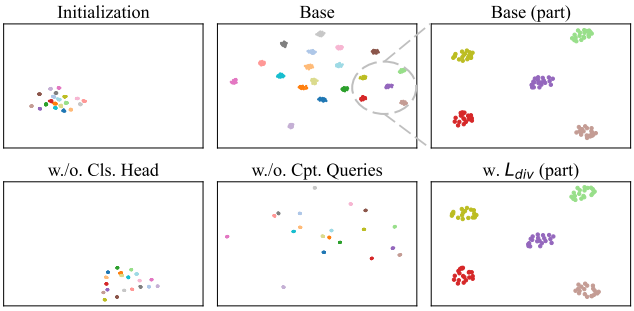

Figure 4: Distribution of semantics-specific encodings. Different colors stand for encodings of different videos. We take the same part from the distribution maps of *Base* and *w/* L_{div} for clear comparison. Best viewed in color.

even when compared to models pretrained on large-scale data. This advantage is attributed to the fact that learning to describe a video with multiple sentences prevents mode collapse, as discussed in (Chen, Deng, and Wu 2022).

Ablation Studies

Evaluation on Semantic Concept Guidance As shown in Tab.5, we design ablative experiments on MSRVTT to evaluate the three components of Semantic Concept Guidance (SCG). The *Base* configuration is obtained by removing the diversity regularization term L_{div} from the full model. The other three settings are obtained by replacing conceptual queries with random vectors, removing the classification head, and adding diversity regularization, respectively. Quantitative results show that for two versions, all three components of SCG bring performance gain on both relevance and diversity performance of diversity caption.

We further present the explanation through visualization of semantics-specific encodings distribution shown in Fig.4. We randomly sample 20 videos and generate 20 sets of encodings with models (LSTM version) trained under the four configurations. We plot t-SNE (Van der Maaten and Hinton 2008) reduced encodings in different colors representing different sets. The initialized encodings are also plotted. The inter-set distance illustrates the distinction between videos with different visual content and represents the relevance performance of the model, while the intra-set distance indi-

G.T. Caps & Cpts:
{a band is giving a live concert, a person preforms a song on stage, a band is performing in front of a live audience, people are cheering while a man sings on stage, people play music on a stage in front of a large crowd, ··· }

Pred. Cpt. Combinations:
1. **music**, **concert**, **video**, performance
2. people, **cheer**, crowd, stage
3. **dance**, **man**, people, concert
4. **perform**, **stage**, **band**, audience
5. **play**, **band**, guitar, **music**

Pred. Caps:
1. a **music video** of a **concert**
2. a **crowd** is **cheering** for a band
3. a **man** sings and **dances** in front of a crowd
4. a **band performing** on **stage**
5. a **band** is **playing music** on stage

Figure 5: An example of generations by SCG-SP-Prefix on MSRVTT. Concepts both showed in predicted combinations and captions are highlighted in red. Best viewed in color.

Training style	C	m-B ↓	s-C	# Params	SPE
From scratch	64.3	72.9	57.1	169.1M	424
Fine-tuning	**66.3**	68.3	60.3	169.1M	424
Prefix-tuning	65.9	**63.1**	**64.4**	44.6M	**325**

Table 6: Ablative results on GPT-2 training style.

cates the diversity performance. Therefore, the distribution of the encodings is highly related to the quantitative performance of relevance and diversity.

As shown in Fig.4, initialized encodings and encodings by the model without classification head are with lower intra-set and inter-set distances compared with the encodings by the base model. Adding a diversity regularization term L_{div} can increase the intra-set distance. The model without conceptual queries can produce encodings with appropriate inter-set distance but low intra-set distance.

Combining both quantitative and qualitative results, we conclude that SCG improves diverse captioning performance by providing distinctive encoding sets and unique encodings within each set. Specifically, it is achieved by three means: 1) semantic information in conceptual queries; 2) deterministic element alignment and extra semantic supervision from concept combination prediction; 3) diversity preference brought by concept-based regularization.

Analysis on Captioners As shown in Tab.1-3, both SCG-SP-LSTM and SCG-SP-Prefix outperform SOTA methods. These results show the compatibility of our proposed SCG-SP framework for different captioners.

We also find that SCG-SP-Prefix outperforms the SCG-SP-LSTM under diversity metrics. For relevance metrics, though the Prefix version scores lower than the LSTM one on MSVD, the gap narrows on MSRVTT, and the Prefix version even scores better on VATEX. These results demonstrate that as the complexity of the evaluation dataset increases, the advantages of using external knowledge become more pronounced.

By comparing the two halves in Tab.5, the ablative results of the two versions show consistency regarding effectiveness on SCG components. However, the Prefix version has a lower requirement for conceptual guidance. For example, the LSTM version drops 8.0 and 34.1 on CIDEr and self-CIDEr, respectively, after removing the classification head, but for the Prefix version, the dropping is 1.6 and 10.1. We can also attribute this to external knowledge.

Training of GPT We employ three training styles for SCG-SP-Prefix, including training the GPT-2 model from scratch (i.e., without loading pretrained parameters), fine-tuning based on pretrained parameters, and prefix-tuning. For models trained by three styles within 100 epochs, captioning metrics on MSRVTT are listed with the number of trainable parameters and seconds per training epoch (SPE) in Tab.6. It is pretty difficult to fully train a GPT-2 from scratch with limited data and time. Thus the model trained from scratch performs the worst. Though the model with fine-tuned GPT-2 performs the best on CIDEr, the model with prefix-tuning, which is proven effective, achieves the highest diversity with 73.6% less trainable parameters 23.3% less training time and only 0.4 drop on CIDEr. The results prove the efficiency of prefix-tuning.

Qualitative Analysis

We present an example of generations by our proposed SCG-SP-Prefix in Fig.5. The generated captions are with considerable diversity and are highly related to the predicted concept combinations. The results demonstrate the effectiveness of guiding diverse generations with different combinations of concepts and show that the diverse generation of the proposed SCG-SP is highly interpretable. More examples are given in **Supplementary Material**.

Conclusion

In this paper, we have proposed a novel model for diverse video captioning named SCG-SP, which stands for semantic-concept-guided set prediction. Through set-level reasoning, SCG-SP has captured the linguistic characteristics of the caption corpus. We have further incorporated semantic concept guidance of concept detection and concept combination prediction to improve the semantic diversity of captions and achieve interpretable generation. Our proposed model has achieved the state-of-the-art performance on MSVD, MSRVTT, and VATEX benchmarks. Extensive quantitative and qualitative experiments have demonstrated the effectiveness of our methods.

Acknowledgments

This work is supported by National Key R&D Program of China (No. 2022ZD0118501), Beijing Natural Science Foundation (Grant No. JQ21017, L223003, 4224093), the Natural Science Foundation of China (Grant No. 61972397, 62036011, 62192782, U2033210, 62202470).

References

Aneja, J.; Agrawal, H.; Batra, D.; and Schwing, A. 2019. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4261–4270.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

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

Chen, D.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 190–200.

Chen, F.; Ji, R.; Ji, J.; Sun, X.; Zhang, B.; Ge, X.; Wu, Y.; Huang, F.; and Wang, Y. 2019. Variational structured semantic inference for diverse image captioning. *Advances in Neural Information Processing Systems*, 32.

Chen, Q.; Deng, C.; and Wu, Q. 2022. Learning Distinct and Representative Modes for Image Captioning. In *Advances in Neural Information Processing Systems*.

Cornia, M.; Baraldi, L.; and Cucchiara, R. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8307–8316.

Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE international conference on computer vision*, 2970–2979.

Deb, T.; Sadmanee, A.; Bhaumik, K. K.; Ali, A. A.; Amin, M. A.; and Rahman, A. 2022. Variational stacked local attention networks for diverse video captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4070–4079.

Deshpande, A.; Aneja, J.; Wang, L.; Schwing, A. G.; and Forsyth, D. 2019. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10695–10704.

Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1473–1482.

Fang, Z.; Wang, J.; Hu, X.; Liang, L.; Gan, Z.; Wang, L.; Yang, Y.; and Liu, Z. 2022. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18009–18019.

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

Gu, X.; Chen, G.; Wang, Y.; Zhang, L.; Luo, T.; and Wen, L. 2023. Text with Knowledge Graph Augmented Transformer for Video Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18941–18951.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6546–6555.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jain, U.; Zhang, Z.; and Schwing, A. G. 2017. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6485–6494.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Li, D.; Huang, Q.; He, X.; Zhang, L.; and Sun, M.-T. 2018. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, K.; Li, L.; Lin, C.-C.; Ahmed, F.; Gan, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17949–17958.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings*

of the IEEE international conference on computer vision, 2980–2988.

Liu, Z.; Wang, T.; Zhang, J.; Zheng, F.; Jiang, W.; and Lu, K. 2022. Show, Tell and Rephrase: Diverse Video Captioning via Two-Stage Progressive Training. *IEEE Transactions on Multimedia*.

Mahajan, S.; Gurevych, I.; and Roth, S. 2020. Latent Normalizing Flows for Many-to-Many Cross-Domain Mappings. In *International Conference on Learning Representations*.

Mahajan, S.; and Roth, S. 2020. Diverse image captioning with context-object split latent spaces. *Advances in Neural Information Processing Systems*, 33: 3613–3624.

Maron, O.; and Lozano-Pérez, T. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10.

Nie, L.; Qu, L.; Meng, D.; Zhang, M.; Tian, Q.; and Bimbo, A. D. 2022. Search-oriented Micro-video Captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3234–3243.

Pan, Y.; Yao, T.; Li, H.; and Mei, T. 2017. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6504–6512.

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

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Perez-Martin, J.; Bustos, B.; and Pérez, J. 2021. Improving video captioning with temporal composition of a visual-syntactic embedding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3039–3049.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ryu, H.; Kang, S.; Kang, H.; and Yoo, C. D. 2021. Semantic grouping network for video captioning. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2514–2522.

Seo, P. H.; Nagrani, A.; Arnab, A.; and Schmid, C. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17959–17968.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

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

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

Vijayakumar, A.; Cogswell, M.; Selvaraju, R.; Sun, Q.; Lee, S.; Crandall, D.; and Batra, D. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1.

Wang, Q.; and Chan, A. B. 2019. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4195–4203.

Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; and Luo, P. 2021. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6847–6857.

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

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

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16442–16453.

Yang, A.; Nagrani, A.; Seo, P. H.; Miech, A.; Pont-Tuset, J.; Laptev, I.; Sivic, J.; and Schmid, C. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10714–10726.

Ye, H.; Li, G.; Qi, Y.; Wang, S.; Huang, Q.; and Yang, M.-H. 2022. Hierarchical modular network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17939–17948.

Yuan, Y.; Ma, L.; Wang, J.; and Zhu, W. 2020. Controllable video captioning with an exemplar sentence. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1085–1093.

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

Zhang, Z.; Shi, Y.; Yuan, C.; Li, B.; Wang, P.; Hu, W.; and Zha, Z.-J. 2020. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13278–13288.