mathematics

MDPI

Article

Towards Human-Interactive Controllable Video Captioning with Efficient Modeling

Yoonseok Heo [1,†], Taehoon Kim [2], Seunghwan Kim [2], Jungyun Seo [2] and Juae Kim [3,*]

[1] Department of Computer Science and Engineering, Sogang University, Seoul 04107, Republic of Korea; ysheo419@sogang.ac.kr
[2] LG AI Research, Seoul 07796, Republic of Korea; taehoon.kim@lgresearch.ai (T.K.); sh.kim@lgresearch.ai (S.K.); jungyun.seo@lgresearch.ai (J.S.)
[3] Department of English Linguistics and Language Technology, Division of Language & AI, Hankuk University of Foreign Studies, Seoul 02450, Republic of Korea
* Correspondence: juaekim@hufs.ac.kr
† Work done while interning at LG AI Research.

Abstract: Video captioning is a task of describing the visual scene of a given video in natural language. There have been several lines of research focused on developing large-scale models in a transfer learning paradigm, with major challenge being the tradeoff between scalability and performance in limited environments. To address this problem, we propose a simple yet effective encoder–decoder-based video captioning model integrating transformers and CLIP, both of which are widely adopted in the vision and language domains, together with appropriate temporal feature embedding modules. Taking this proposal a step further, we also address the challenge of human-interactive video captioning, where the captions are tailored to specific information desired by humans. To design a human-interactive environment, we assume that a human offers an object or action in the video as a short prompt; in turn, the system then provides a detailed explanation regarding the prompt. We embed human prompts within an LSTM-based prompt encoder and leverage soft prompting to tune the model effectively. We extensively evaluated our model on benchmark datasets, demonstrating comparable results, particularly on the MSR-VTT dataset, where we achieve state-of-the-art performance with 4% improvement. In addition, we also show potential for human-interactive video captioning through quantitative and qualitative analysis.

Keywords: video captioning; controllable video captioning; human-interactive; multimodal representation learning

MSC: 68T07; 68T45; 68T50

Citation: Heo, Y.; Kim, T.; Kim, S.; Seo, J.; Kim, J. Towards Human-Interactive Controllable Video Captioning with Efficient Modeling. *Mathematics* **2024**, *12*, 2037. https://doi.org/10.3390/math12132037

Academic Editor: Jose Luis Vicente Villardon

Received: 27 April 2024
Revised: 3 June 2024
Accepted: 26 June 2024
Published: 30 June 2024

1. Introduction

With the success of large-scale image caption models [1–7] using large amounts of image–text pairs in a transfer learning paradigm, interest has recently turned to the video domain. Video captioning [8–15] is the task of describing the visual scene of a given video in natural language. It is much more challenging than image captioning, which only addresses a static piece of information, in that it requires a comprehensive understanding of multiple frames over the entire video. In fact, from a visual perspective, the key is that the model should have the ability to understand both the static information and the dynamics presented over multiple frames. Moreover, from a linguistic point of view, it should be able to generate coherent descriptions.

It is not surprising that only high-capacity models can handle such multi-modality. As can be estimated from zero-shot image caption works, it inevitably requires millions of video–text pairs, resulting in two problems. First of all, creating a large-scale video caption dataset from scratch is a formidable task. While image captioning studies have

shown promising results in zero-shot settings, video captioning requires tens to hundreds of thousands of video–text pairs that can only be created through the meticulous efforts of human annotators. However, compared to the image captioning task, labeling high-quality video data is more challenging in terms of quality rather than quantity. Videos contain multiple scene transitions and captions can have a wide range of variations depending on the annotator's perspective, making it difficult to achieve a certain level of semantic unity for a single video. This semantic variation poses a significant challenge in generating high-quality video captions that accurately reflect the content of the video. MV-GPT [16] proposed an algorithm that substitutes captions with speech transcriptions to train their model. These characteristics imply that an ideal video captioning model should have the ability to generate captions from various angles through human interaction. In this paper, we refer to this as human-interactive video captioning. We describe this problem in Section 3.3 and introduce a caption tuning method using the prompt tuning approach as a stepping stone for future research.

More importantly, as a threshold problem there is a strong dependency on resources to obtain advanced video understanding capabilities. In fact, ViViT [17], which has shown considerable performance on a number of video understanding tasks, includes Vision Transformer [18] and BERT [19] as its backbone. However, each model is a large-capacity transformer-based model specialized in the visual and language domains, and the individual models have a significant load as well. Even the MV-GPT [16] model, which has the highest capacity in the current video caption model, includes the ViViT [17] model. Therefore, as it contains a great deal of prior knowledge about the behavior appearing in the video, it shows significant results in caption generation, including video understanding; however, the learning process, which is accompanied by considerable resources, is indispensable. Therefore, optimal design of the video caption model is required, as its performance represents a trade-off with resources due to the scalability of the model.

To address this issue, we propose a simple yet effective encoder–decoder-based video captioning model. Unlike previous works [16,20–22] that used pretrained 3D features that express motion or pretrained with a large amount of video data, the proposed model does not require any prior knowledge of video. Instead, we use CLIP, which is widely used in the image domain, for the encoder and Visual GPT, a GPT2 series specializing in text generation, for the decoder. Our model uses a simple transformer block to encode feature information over time. The encoder is divided into two steps: a keyframe selection process using CLIP, and dynamic information recognition. First, instead of relying on human heuristics, we selectively sample frames that contain semantically meaningful scene transitions using CLIP and use them as keyframes for video encoding. In addition, in order to ensure that the video is able to capture semantic information appearing throughout the frame, our model encodes dynamic information using only the structure of the transformer block, including self-attention. In the decoder, we utilize VisualGPT, which facilitates the convergence of vision and text knowledge, as the backbone instead of the conventional GPT2. Similar to GPT2, VisualGPT utilizes cross-attention to incorporate visual information as input. However, it includes a dynamic controller that selectively processes the reflection of caption context information and visual information, enabling more effective multimodal integration. The proposed model uses CLIP only during the keyframe sampling process, and freezes it during the training process. The training of our model is performed only with GPT2 and a few transformer blocks, resulting in a significant reduction in resource dependence.

We evaluated the performance of our proposed model by conducting comparative experiments with various baseline models using the video caption benchmark. Our model achieves similar results to the latest studies or even sets a new state-of-the-art performance. Additionally, we conducted a quantitative analysis of the generated video captions, with the results suggesting the possibility of human-interactive video captioning. Section 2 provides previous studies related to video captions and human interaction, Section 3 describes our proposed model in detail, and Section 4 analyzes its results. Finally, Section 5 concludes the paper.

2. Related Work

In this section, we describe the research on video captioning in terms of model capacity, dynamic feature embedding, and human-interactive video captioning.

Video Captioning. At the beginning of the related research, rule-based approaches [23–27] were proposed to produce sentences using a fixed set of predefined templates, specifically, a triple consisting of subject, verbs, and objects. Despite high grammatical correctness, there are strong limitations in terms of the low complexity of the sentence construction rules and low generalization ability. With the rise of deep learning, encoder–decoder-based architectures have been exploited in various ways. SCN [28] is a semantic-concept detecting method that obtains the probabilities of concepts appearing in the video from a CNN. It incorporates concept-dependent information in an LSTM to compose the semantic representations. SGLSTM [29] introduces the way of jointly exploring visual and semantic features using two semantic guiding layers by adopting different levels of semantics as guidance to control the language model in order to generate sentences.

Unlike image captioning, which only handles static moments, video captioning should address ways to understand the dynamics appearing across multiple frames in the video. SemSynAN [20] introduces a way of strengthening the understanding of temporal composition by mapping visual concepts to their corresponding part-of-speech tags in the descriptions. In addition, Ref. [21] introduces a recurrent region-based attention mechanism and motion-guided information-controlling method to adaptively capture the temporal relations. With the success of transformer-based architectures in most vision-language tasks, SwinBERT [22] proposes a way of incorporating transformer-based video feature extractor and transformer-based encoder, showing considerable performance on video captioning tasks. Moreover, MVGPT [16] introduces a large-scale video captioning model in a pretraining–finetuning strategy. It contains a large-scale video understanding model [17] and transformer-based decoder [30] as a part of the backbone. As the model capacity is the largest of all the works reviewed here, it has been able to show considerable results, though at the same time it has high dependency on resources.

Human-interactive video captioning. As a video contains richer visual content and semantic details than an image, different people may have different points of view when it comes to understanding the context. Therefore, it is challenging to meet the consensus by generating one caption with one specific point of view for a given video. To address this problem, ViCap [31] proposes a new task of video interactive captioning for the first time. Given a round of dialog interaction between a human and a captioning agent, the agent first shows an initial caption for a given video. When it does not meet expectations, the human provides the agent with a short prompt, i.e., the first two seed words of the caption to be generated. The agent then produces a more engaging caption. While this study is meaningful enough in that it proposes a way to colorfully transform captions that can meet user needs for a single video, there is much room for further exploration in that the methodology depends on human heuristics. In this paper, we introduce a way to adapt captions according to human prompts using the soft prompting method [32–34], as described in the following section.

3. Proposed Model

As shown in Figure 1, we propose a simple yet effective video captioning model consisting of a video encoder and a caption decoder. First, the video encoder consists of a CLIP-based keyframe sampler and a temporal embedding layer. This selectively embeds only the key information necessary for video understanding, efficiently conveying the necessary information for caption generation. Next, the caption decoder, which is from the GPT2 family, fuses the output of the video encoder with the context information to produce the optimal caption. Taking this a step further, we address the challenge of human-interactive video captioning, where the captions generated are tailored to the specific information desired by humans as opposed to providing a general description of the entire video. Motivated by [31], we assume that a human offers an object or action in the video

as a short prompt; in turn, the system then provides a detailed explanation. We embed human prompts with an LSTM-based prompt encoder and leverage soft prompting to tune the model effectively. We describe each module in more detail in the following subsections.

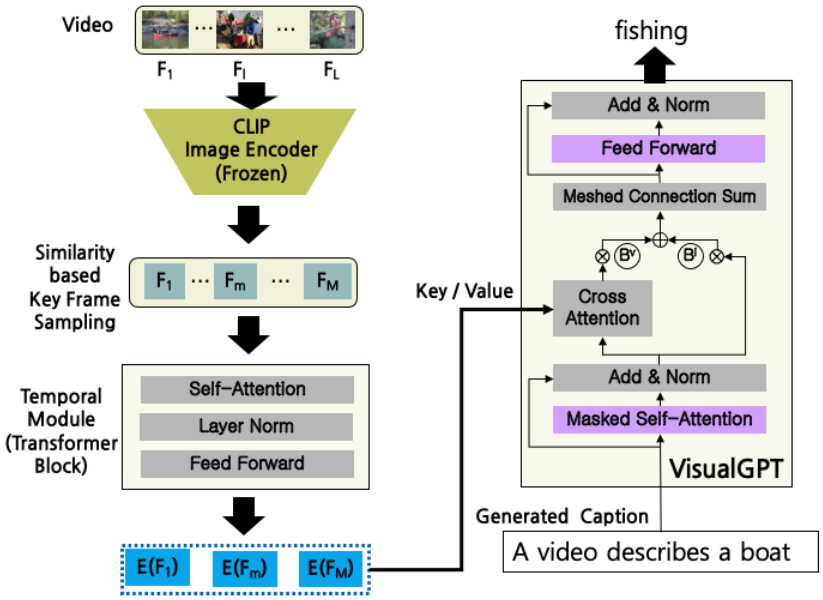

Figure 1. An illustration of our proposed model.

3.1. Video Encoder

CLIP-based Keyframe Sampling. Given a video, we leverage the powerful CLIP [35] image encoder to determine keyframes using the semantic gap between two consecutive frames. First, across the entire video, we select every three frames per second: the first frame, the middle frame, and the last frame. Then, the frames are represented as embeddings using a pretrained CLIP encoder [35]. Assuming that the first frame of each video is always a keyframe, we calculate the cosine similarity between the embeddings of two consecutive frames starting from the first frame. If the difference between the second frame and the first keyframe is greater than the threshold, we determine the second frame as a keyframe; otherwise, the second frame is regarded to be semantically similar, and we discard it. This process is repeated iteratively to select the keyframes of the video. During the model training process, the CLIP encoder is frozen. The threshold value is experimentally determined; the specific value is introduced in Section 4.

Temporal Feature Embedding. While CLIP-based embedding can capture static information in a single frame, it is not suitable for representing dynamic information that appears across multiple frames, such as running up a set of stairs or canoeing along a riverbank. In light of [36], we utilize a set of transformer blocks to learn temporal features that capture this dynamic information when given a sequence of M keyframes. Specifically, we use three transformer blocks that consist of a multi-head self-attention layer, layer normalization, and feedforward layer, which is similar to the existing transformer encoder structure [19,37]. We use CLIP embeddings of keyframes directly as token embeddings within the blocks, and we use the frame order as a learnable positional embedding. Additionally, we enable batch training by using zero-padding for the keyframes.

3.2. Caption Decoder

As shown in Figure 1, given temporal features from the video encoder, a caption decoder autoregressively generates the output sentence conditioned on this context. We adopt VisualGPT [38] from the GPT2 family [30], as it is more optimal for fusing visual and textual information to generate the subsequent output [38]. In fact, the scalar dot product-based attention operation in the cross-attention layer, in which visual features

act as the key and value and context information acts as the query, is the only option that can fuse multimodality into GPT2. However, it has been said that it is suboptimal to fuse this knowledge due to the fact that GPT2 is trained solely on a massive amount of textual knowledge and no visual information [38]. In contrast, VisualGPT has the ability to fuse visual and textual information using additional components in a block. There are two distinct components: a self-resurrecting unit and a meshed connection sum. The self-resurrecting unit refers to gates that are controllable by manipulating the visual features and textual information. Next, the meshed connection sum layer has full connections between all decoder blocks and all encoder blocks. For each block, this layer enables fusion between contextual information and the different features of visual information implied in each layer of the video encoder block. As shown in Figure 1, only the masked self-attention layer and the feedforward layer are initialized with the weights of the pretrained GPT2, with all others being trained from scratch. The whole architecture is fine tuned in an end-to-end manner by applying the cross-entropy loss function in a teacher-forcing decoding strategy:

$$\mathcal{L} = -\sum_{t=0}^{K} \log p(Y_t | Y_{0:t-1}, V_X) \tag{1}$$

where $p(*)$ is the output probability of the predicted token (Y_t) given the previously generated captions ($Y_{0:t-1}$) and visual features (V_X). The optimal parameters are determined on the validation dataset using the CIDEr metric. During inference, our model takes a sequence of keyframes as input and outputs a natural language sentence. We generate the output sentence in an autoregressive manner using a beam-search decoding strategy. We perform generation until our model outputs a predefined ending token (EOS) or reaches the maximum output length.

3.3. Prompt Encoder for Human-Interaction

Next, we attempt to address the ambiguity inherent in labeling a single caption for a given video. Figure 2 shows three sampled frames in which a semantic scene transition occurs in a video, with short descriptions provided below each frame. The descriptions of the first frame (A) and third frame (C) describe the motion of an object (e.g., canoe, man) over several frames. In contrast, the description of the second frame (B) captures a transient moment, akin to an image caption. These three descriptions offer different perspectives, depicting separate parts of the video; therefore, it is plausible that any of these descriptions could serve as the gold standard caption for the video. In other words, it is inappropriate to determine a semantically consistent caption for video content that changes over time. Instead, various captions should be created based on the viewer's perspective. Ultimately, video captions should be able to reflect multiple viewpoints, allowing the system to share a perspective with the viewer.

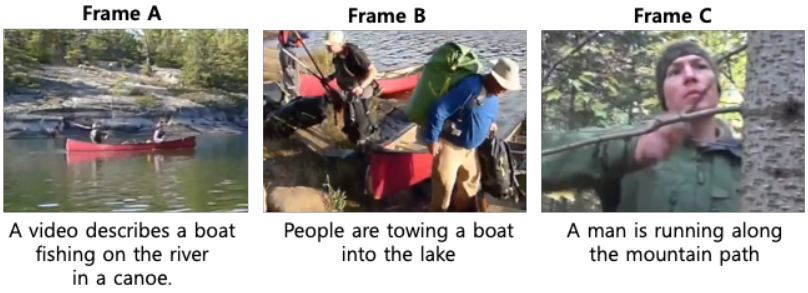

Figure 2. Three sampled frames in a video with their corresponding descriptions.

We adopt a similar assumption as [31], where in the interaction mode a human provides a short prompt containing two or three words. The key difference is that we do not use a human prompt with the initial seed words of the caption; instead, a human

prompt serves as implicit guidance to refine the caption. Inspired by [32], we employ a soft prompting method that allows the model to adjust the captions using the prompt. As shown in Figure 3, K pseudo tokens are added to the front and back of the human prompt (e.g., "Shania Twain"). In this work, we set K to 5, which can be determined experimentally. A bidirectional LSTM is used as the prompt encoder. The prompt is passed to the prompt encoder, which ultimately makes the embedding of the pseudotoken reflect the context information according to the human prompt. The embedding ($E(P_k)$) for the pseudotoken (P_k) is generated by concatenating the hidden layers of each LSTM layer. These are directly used as the token embeddings of the caption decoder. We exploit a teacher-forcing training strategy using the cross-entropy loss between the generated captions and the correct captions. During inference, we employ a beam-search decoding method to obtain adaptive captions. The search begins with the start symbol token (SOS) after feeding the prompts.

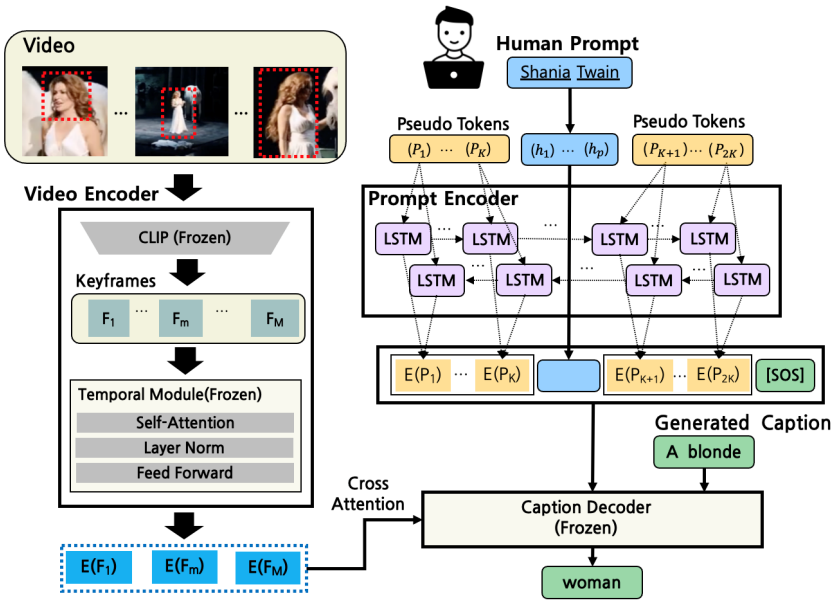

Figure 3. An illustration of the proposed human-prompted video captioning model. The rectangle highlighted by the red dashed line in the video frames refers to the object representing the human prompt (Shania Twain); SOS refers to the start token symbol, (P_*) stands for the pseudotoken, and E(P_*) refers to the embedding of the pseudotoken. Best viewed in color.

To establish the human-interactive environment, we define human prompts by heuristically extracting noun phrases from the gold standard captions. The potential of the tuning capabilities is qualitatively described in the following section.

4. Experiment

In this section, we describe the effectiveness of our proposed model in diverse ways. First, in Section 4.1, we introduce the experimental setup, including the benchmark datasets, evaluation metrics, and implementation details. In Section 4.2, we describe the baseline models. Then, we introduce the quantitative and qualitative analyses in Section 4.3.

4.1. Experimental Settings

Datasets. We used the MSR-VTT [39] and MSVD [40] benchmark datasets to evaluate our proposed method.

- MSR-VTT [39] is a large-scale dataset with 10,000 open-domain video clips taken from sources such as video games and TV shows, including 20 ground truth captions. We used the standard split, which has 6513/497/2990 clips for training/validation/testing, respectively.

- MSVDchen-dolan-2011-collecting contains 1970 video clips of 9.6 s on average. Each clip has 40 gold standard captions. We used the standard split, which has 1200/100/670 clips for training/validation/testing, respectively.

Evaluation Metrics. For quantitative evaluation, we adopted four automatic evaluation metrics widely used on generation tasks, i.e., CIDEr [41], BLEU [42], METEOR [43], and ROUGE [44]. Specifically, CIDEr measures the TF-IDF metric to aggregate the matching score for N-grams between the generated and gold standard captions; BLEU-N measures the ratios of the co-occurrence of N-grams between the generated and gold standard captions; ROUGE measures the N-gram-based recall between the generated and gold standard captions; and METEOR measure F-scores for the co-occurrence of N-grams between the generated and gold standard captions. All these evaluation metrics can be officially processed via the open-source toolkit at (https://github.com/tylin/coco-caption, accessed on 4 January 2023). In this paper, we adopt N = 4 for checking the similarity of sentences.

Training Details. For all three datasets used in the experiments, we preprocessed the videos using the same pretrained CLIP model based on a Vision Transformer architecture equivalent to ViT-B/32 (https://huggingface.co/openai/clip-vit-base-patch32, accessed on 4 January 2023). In CLIP-based frame sampling, we set the similarity threshold value as 0.85. The video encoder has three transformer blocks with a hidden size of 512D and a dropout layer with 0.1. We added one MLP layer with layer normalization into the cross-attention layer to match the decoder's hidden size to the size of the encoder output. We trained the model with scales corresponding to those of GPT2, such as GPT2-small and GPT2-medium. We utilized the two different pretrained weights from the Huggingface hub. For the small-size model, we trained it using a batch size of 20 and learning rate of 1×10^{-4} with the Adam Optimizer and early-stopping for five iterations using the CIDEr score on the validation dataset. For the medium-size model, we trained it using a batch size of 2 and learning rate of 5×10^{-5} with the Adam Optimizer, a gradient accumulation step of 2, and early stopping for ten iterations, using the CIDEr score on the validation dataset. All experiments were performed on two V100 GPUs (2×16 GB) with distributed training using the Huggingface library. When fine-tuned on MSR-VTT and MSVD, respectively, we concatenated five captions into one gold standard caption with multiple sentences instead of using one single caption for one video, as this approach showed more robust performance during training (Table 1).

Table 1. Hyperparameters for training models.

Temporal Module		**Caption Decoder (GPT2-Small / GPT2-Medium)**		**P-Tuning**	
Maximum number of frames	25	Batch Size	20/2	number of pseudo tokens	3
Number of transformer blocks	3	Accumulation Step	1/2	hidden size (LSTM)	512
Attention heads	8	Early Stopping Criteria	CIDEr/CIDEr	batch size	48
Hidden size	512	Learning Rate	$1 \times 10^{-4}/5 \times 10^{-5}$	learning rate	4×10^{-5}
dropout ratio	0.1	Decay Rate	0.98/0.98	Decoding Strategy	Beam-search
		Decoding Strategy	Beam-search	Beam size	3
		Beam size	3/3	max sequence length	35
		max sequence length	40/40		

4.2. Baselines

For our comparison of caption generation capabilities, we selected the following four models:

- ViCap [31], which is based on CNN-GRU architecture without additional 3D features or any pretraining phase on video datasets.
- VNS-GRU [45], which exploits 3D-CNN visual features with a GRU-based decoder. It adopts novel regularization and training strategies and does not have any pretraining phase on video datasets.
- SemSynAN [20], which exploits 3D-CNN visual features with a LSTM-based decoder. It does not have any pretraining phase on video datasets.
- MGRMP [21], which exploits 3D-CNN visual features to intensify the motion guidance. It does not have any pretraining phase on video datasets.
- SwinBERT [22], which consists of Video-Swin Transformer [46] as the feature extractor and BERT for frame-level fusion. In particular, the Video-Swin Transformer includes strong prior knowledge of the video domain.
- MV-GPT [16], which includes ViViT [17], a powerful video encoder, and GPT2 [30] as the decoder. In addition, it has been pretrained on a massive amount of video data.
- ZS-Vicap [47], which is a zero-shot video captioning model which does not require any training step for video captioning tasks. Instead, it requires only a few caption refinement steps in the inference phase.

4.3. Performance Comparison for Video Captioning

Overall, the experimental results showed different outcomes depending on the characteristics of the proposed model and the datasets, as shown in Table 2, representing the performance on both the MSR-VTT and MSVD datasets. For MSR-VTT, our proposed model demonstrates state-of-the-art performance on most evaluation metrics, in particular achieving 4% improvement on the CIDEr metric, except for the MV-GPT model. However, for MSVD the VNS-GRU based on GRU shows the highest performance. First, when it comes to SwinBERT, which exploits a large-scale model pretrained on the video domain as a feature extractor, it shows the second-best performance on MSR-VTT using the CIDEr metric. Despite the proposed model having no prior knowledge of dynamic motion and a smaller model capacity, it demonstrates comparable results using only CLIP embeddings for keyframes and temporal embeddings. Moreover, it shows consistently better results on the MSR-VTT dataset. As a more advanced model, MV-GPT has a much higher capacity than the proposed model, as it uses pretrained models and video caption data with a much more complex structure. Therefore, it does not make sense to evaluate both models on the same criteria. Rather, the proposed model shows comparable results despite its very simple structure. In particular, for the small-sized version of the proposed model, it is possible to obtain figures very close to the reported results even with only one 16 GB GPU. Therefore, the proposed model could be a better option in circumstances with limited resources.

Next, when it comes to VNS-GRU, SemSynANm and MGRMP which exploit additional 3D features from pretrained models such as ResNeXt, 3D-CNN, and C3D, our proposed model demonstrates better results on MSR-VTT in terms of almost all evaluation metrics. This suggests the effectiveness of using CLIP embeddings with the temporal embedding layer as an alternative to pretrained 3D feature extractors. Next, we look at ViCap, which similar to the proposed model does not use any additional 3D features or pretrained video feature extractors. Notably, the ViCap model contains some of its engineering heuristics in the caption generation process, resulting in a significant performance gap compared to the latest models.

Finally, we compare the results with a zero-shot video captioning model denoted as ZS-ViCap. Although it has lower figures on the automatic evaluation metrics, the captioning quality shows meaningful results. However, when it comes to inference time our proposed model takes 1.81 s per video on average, compared with ZS-ViCap requiring approximately 90 s for final caption refinement. This implies that our model has better availability in real-world scenarios.

Table 2. Performance comparison on the MSR-VTT and MSVD datasets; '3D features' indicates motion features such as 3D CNN, C3D, etc., while 'pretrained on videos' indicates that the model has been pretrained on video datasets. The details are described in the reference papers. C, B, M, R respectively indicate CIDEr, BLEU-4, METEOR, and ROUGE. Numbers in bold indicate the best results, while underlined numbers represent the second-best result. In addition, the figures marked with * represent results with a considerably larger capacity than the proposed model.

Models	Year	3D Features	Pretrained on Videos	MSR-VTT				MSVD			
				C	B	M	R	C	B	M	R
ViCap	2019	X	X	0.460	0.044	0.092	0.245	-	-	-	-
VNS-GRU	2020	O	X	0.530	0.453	0.299	0.634	**1.215**	**0.665**	**0.421**	**0.797**
SemSynAN	2021	O	X	0.510	**0.464**	0.304	0.647	1.115	0.644	0.419	0.795
MGRMP	2021	O	X	0.514	0.417	0.289	0.621	0.985	0.558	0.369	0.745
SwinBERT	2022	O	O	0.538	0.419	0.299	0.621	1.206	0.582	0.413	0.775
ZS-ViCap	2022	X	X	0.113	0.030	0.146	0.277	0.174	0.030	0.178	0.314
Ours(small)	2023	X	X	**0.561**	0.435	**0.305**	**0.657**	0.893	0.477	0.367	0.723
Ours(medium)	2023	X	X	0.524	0.407	0.294	0.610	0.962	0.508	0.374	0.740
MV-GPT	2022	O	O	0.600 *	0.489 *	0.386 *	0.640 *	-	-	-	-

As a threshold problem, the proposed model did not show significant performance gain on the MSVD dataset compared to other baselines. We conjecture that this is due to the correlation between dataset size and model capacity. In fact, the proposed model was trained from scratch, excluding only some layers of VisualGPT. In addition, the amount of data corresponded to only 10% that of MSR-VTT. Therefore, the volume of data can be considered insufficient for training the temporal embedding layer of the video encoder and VisualGPT. Furthermore, the VNS-GRU model, despite being a relatively old-fashioned model consisting of a CNN and GRU, still achieves state-of-the-art performance on MSVD data by leveraging its prior knowledge of kinetics. This implies that encoding methods for dynamic motion or movement appearing over multiple frames still need to be explored. We leave this as future work.

Lastly, we carried out a qualitative analysis of the MSR-VTT test. First, we analyzed the video captioning capabilities of our model. Figure 4 presents two cases, each of which contains two ground truth (GT) captions and the caption generated by our model (Ours). In the first case, our model was able to generate a sentence with a complete description of the video. More surprisingly, for the second case our model generated a long and grammatically correct caption. Compared with the ground truth captions, our model shows a more fine-grained description (e.g., unfinished, hardwood).

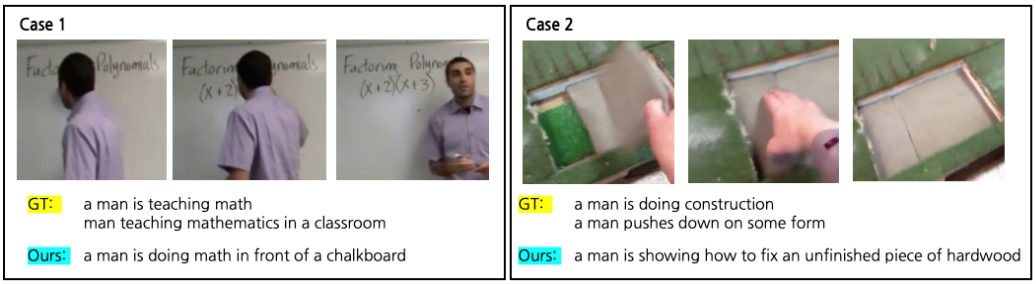

Figure 4. Qualitative analysis for video captioning on MSR-VTT test.

Effects of Temporal Embedding Layer. We investigated the effectiveness of the temporal embedding layer on MSR-VTT using the small-sized model. Instead of using transformer blocks, we first used naive CLIP-based frame embeddings as the output of

the video encoder (i.e., the number of layers in Table 3 is zero). The results show the effectiveness of the temporal embedding layer. Without the temporal embedding layer, the performance drops by 10% on Cider. We set the optimal number of layers based on the experimental results.

Table 3. Ablation study on the effectiveness of the temporal embedding layer.

# of Layers	C	B	M	R
0	0.503	0.384	0.241	0.540
3 (Ours)	0.561	0.435	0.305	0.657

4.4. Probing for the Potential of Human-Interactive Video Captioning

Following [31] as a baseline, we adopted the MSR-VTT dataset [39]. For a fair comparison, we divided the dataset into two halves, one for training caption generation and the other for human-interactive video captioning. We first fine-tuned the video captioning model mentioned in Section 3.3 on the former, then exploited the latter for the soft prompting method. We used the first 3000 video clips for training, the next 500 for validation, and the last 1000 for evaluation. We removed duplicate annotated captions for each video clip. Without the publicly available test data from the original paper, for each one of the annotations of each video clip on the test split we regarded the first two words as a human prompt. An exception was that we considered the first three words as prompts only if the first word was a preposition. Unlike the original paper, in order to assume an unconstrained environment for caption generation we did not use such words as a starting prefix.

Although the baseline uses human prompts directly as a prefix in the decoder, our proposed model still outperforms the baseline by a large margin on all metrics, as shown in Table 4. This is due to the gap between the models' original caption generation capabilities. In addition, we investigated the effectiveness of the prompt encoder. Instead of using the prompt encoder for soft prompting, we only attached the prompts to the seed words with a special token (SEP) without additional training. We found only a marginal performance drop due to the disparity between the patterns and the original ones.

Table 4. Performance comparison for human-interactive video captioning; C, B, M, R respectively indicate CIDEr, BLEU-4, METEOR, and ROUGE. The numbers in bold indicate the best results.

Models	MSR-VTT			
	C	B	M	R
ViCap [31]	0.060	0.56	0.048	0.149
Ours (small)	1.199	10.30	0.158	0.389
(−) prompt encoder	0.897	7.71	0.109	0.166

Lastly, as shown in Figure 5, we analyzed two video captioning results including human-interactive cases. The first case is part of a stage play. As can be seen, our model generates more fine-grained captions compared with the ground truth (GT). In particular, our model can handle lengthy captions in a robust way, which remains a challenging issue in video captioning tasks. More surprisingly, when we fed the model 'Shania Twain' (i.e., a woman in a white dress) as a human prompt, our model generated a more descriptive caption regarding the woman. This implies that if prompt embeddings were handled in a more complex way, the model would be able to generate a more engaging caption that better meets the user's intention. In addition, for the second case our model generated a more lengthy and descriptive caption than the ground truth. Unlike the first case, for the human-interactive caption it produced phrases (like a keyword), not a complete sentence; however, we leave further steps for future work. Additional examples from the human-

interactive video captioning test are provided in Figure 6 to help readers better understand the diverse aspects of our results.

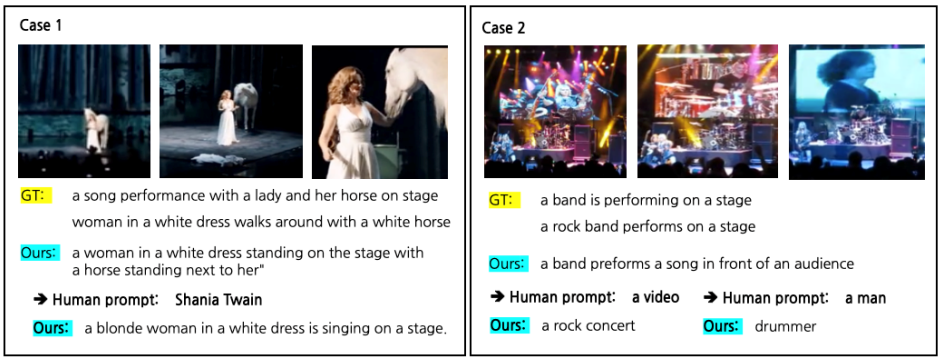

Figure 5. Qualitative analysis including human-interactive video captioning.

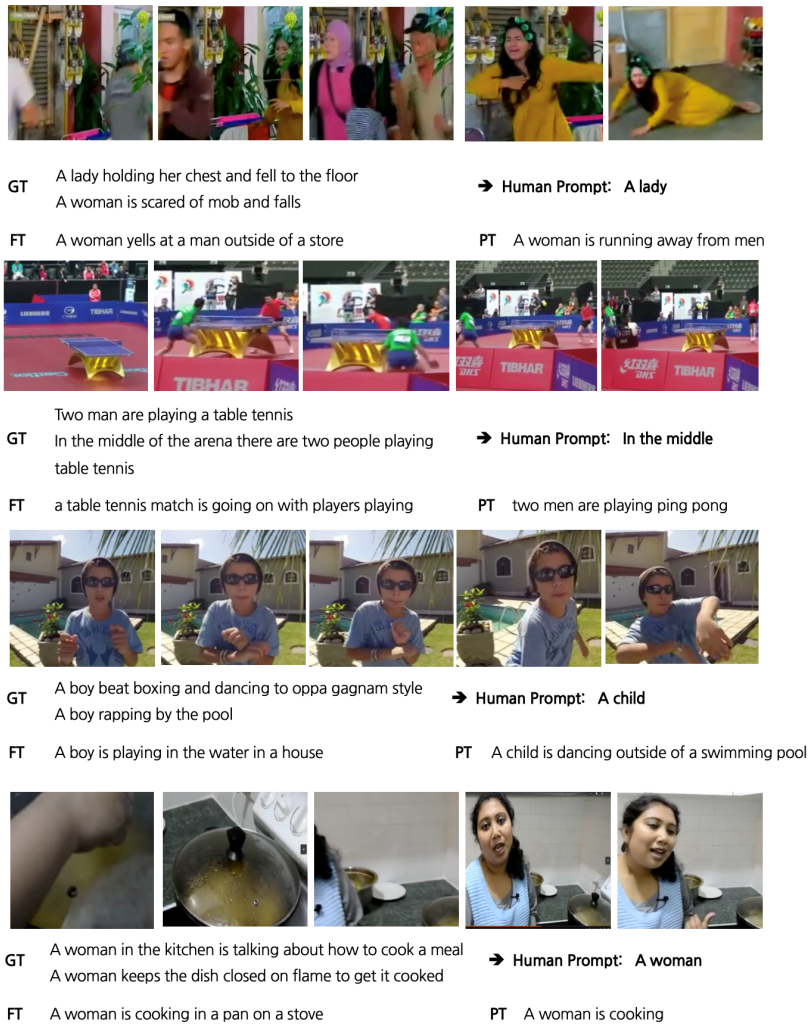

Figure 6. Examples of human-interactive video captioning. We only added two ground-truth (GT) captions for visualization. FT refers to the caption generated by the fine-tuned model, while PT refers to the caption generated by the human-interactive video captioning model.

5. Conclusions

In this paper, we have proposed a simple yet effective encoder–decoder-based video captioning model which does not require extensive model capacity or additional dynamic

motion features from pretrained models. Instead, our proposed model only requires a pretrained CLIP model as the frame feature extractor, a set of transformer blocks for temporal feature embedding, and a VisualGPT model from the GPT2 family specializing in multimodal fusion. Our proposed model demonstrates comparable or state-of-the-art performance on the MSR-VTT dataset. In addition, we introduce a strategy for human-interactive video captioning incorporating prompt tuning. Despite limitations in generating lengthy captions, we have demonstrated its potential for adaptively meeting users' needs. In future work, we plan to include effective methods for capturing dynamic features across multiple frames. In addition, we will investigate ways of adaptively capturing and utilizing visual features with respect to user interaction.

Author Contributions: Conceptualization, Y.H., T.K., S.K., J.S., and J.K.; funding acquisition, J.K.; investigation, Y.H.; methodology, Y.H. and T.K.; project administration, S.K. and J.S.; supervision, S.K. and J.S.; validation, Y.H.; writing—original draft, Y.H.; writing—review and editing, Y.H., T.K., and J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Hankuk University of Foreign Studies Research Fund (of 2024).

Data Availability Statement: The MSR-VTT dataset can be found at https://www.kaggle.com/datasets/vishnutheepb/msrvtt (accessed on 4 January 2023). The MSVD dataset can be found at https://paperswithcode.com/sota/video-captioning-on-msvd-1 (accessed on 4 January 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mokady, R.; Hertz, A.; Bermano, A.H. ClipCap: CLIP Prefix for Image Captioning. *arXiv* **2021**, arXiv:2111.09734.
2. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021 ; pp. 5579–5588.
3. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 121–137.
4. Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; Wang, L. Git: A generative image-to-text transformer for vision and language. *arXiv* **2022**, arXiv:2205.14100.
5. Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; Wang, L. Scaling up vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17980–17989.
6. Yang, X.; He, S.; Wu, J.; Yang, Y.; Hou, Z.; Ma, S. Exploring Spatial-Based Position Encoding for Image Captioning. *Mathematics* **2023**, *11*, 4550. [CrossRef]
7. Omri, M.; Abdel-Khalek, S.; Khalil, E.M.; Bouslimi, J.; Joshi, G.P. Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Captioning. *Mathematics* **2022**, *10*, 288. [CrossRef]
8. Aafaq, N.; Akhtar, N.; Liu, W.; Gilani, S.Z.; Mian, A. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12487–12496.
9. Li, L.; Lei, J.; Gan, Z.; Yu, L.; Chen, Y.C.; Pillai, R.; Cheng, Y.; Zhou, L.; Wang, X.E.; Wang, W.Y.; et al. Value: A multi-task benchmark for video-and-language understanding evaluation. *arXiv* **2021**, arXiv:2106.04632.
10. Liu, S.; Ren, Z.; Yuan, J. Sibnet: Sibling convolutional encoder for video captioning. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1425–1434.
11. Pan, B.; Cai, H.; Huang, D.A.; Lee, K.H.; Gaidon, A.; Adeli, E.; Niebles, J.C. Spatio-temporal graph for video captioning with knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10870–10879.
12. Pei, W.; Zhang, J.; Wang, X.; Ke, L.; Shen, X.; Tai, Y.W. Memory-attended recurrent network for video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8347–8356.
13. Shi, B.; Ji, L.; Niu, Z.; Duan, N.; Zhou, M.; Chen, X. Learning semantic concepts and temporal alignment for narrated video procedural captioning. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 4355–4363.
14. Zhang, Z.; Qi, Z.; Yuan, C.; Shan, Y.; Li, B.; Deng, Y.; Hu, W. Open-book video captioning with retrieve-copy-generate network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 9837–9846.

15. Chen, S.; Yao, T.; Jiang, Y.G. Deep Learning for Video Captioning: A Review. In Proceedings of the IJCAI, Macao, China, 10–16 August 2019; Volume 1, p. 2.
16. Seo, P.H.; Nagrani, A.; Arnab, A.; Schmid, C. End-to-End Generative Pretraining for Multimodal Video Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17959–17968.
17. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A Video Vision Transformer. In Proceedings of the International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
20. Perez-Martin, J.; Bustos, B.; Perez, J. Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Virtual, 5–9 January 2021; pp. 3039–3049.
21. Chen, S.; Jiang, Y.G. Motion Guided Region Message Passing for Video Captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 1543–1552.
22. Lin, K.; Li, L.; Lin, C.C.; Ahmed, F.; Gan, Z.; Liu, Z.; Lu, Y.; Wang, L. SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. In Proceedings of the CVPR, New Orleans, LA, USA, 18–24 June 2022.
23. Das, P.; Xu, C.; Doell, R.F.; Corso, J.J. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2634–2641.
24. Kojima, A.; Tamura, T.; Fukunaga, K. Natural language description of human activities from video images based on concept hierarchy of actions. *Int. J. Comput. Vis.* **2002**, *50*, 171–184. [CrossRef]
25. Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R.; Darrell, T.; Saenko, K. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2712–2719.
26. Krishnamoorthy, N.; Malkarnenkar, G.; Mooney, R.; Saenko, K.; Guadarrama, S. Generating natural-language video descriptions using text-mined knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, Bellevue, WA, USA, 14–18 July 2013; Volume 27, pp. 541–547.
27. Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; Schiele, B. Translating video content to natural language descriptions. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 433–440.
28. Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; Deng, L. Semantic Compositional Networks for Visual Captioning. In Proceedings of the CVPR, Honolulu, HI, USA, 21–36 July 2017.
29. Yuan, J.; Tian, C.; Zhang, X.; Ding, Y.; Wei, W. Video Captioning with Semantic Guiding. In Proceedings of the 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), Xi'an, China, 13–16 September 2018; pp. 1–5. [CrossRef]
30. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *Openai Blog* **2019**, *1*, 9.
31. Wu, A.; Han, Y.; Yang, Y. Video Interactive Captioning with Human Prompts. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, Macao, China, 10–16 August 2019; pp. 961–967. [CrossRef]
32. Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; Tang, J. GPT Understands, Too. *arXiv* **2021**, arXiv:2103.10385.
33. Yang, K.; Liu, D.; Lei, W.; Yang, B.; Xue, M.; Chen, B.; Xie, J. Tailor: A Prompt-Based Approach to Attribute-Based Controlled Text Generation. *arXiv* **2022**, arXiv:2204.13362. https://doi.org/10.48550/ARXIV.2204.13362.
34. Senadeera, D.C.; Ive, J. Controlled Text Generation using T5 based Encoder-Decoder Soft Prompt Tuning and Analysis of the Utility of Generated Text in AI. *arXiv* **2022**, arXiv:2212.02924. https://doi.org/10.48550/ARXIV.2212.02924.
35. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020. https://doi.org/10.48550/arXiv.2103.00020.
36. Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; Xie, W. Prompting Visual-Language Models For Efficient Video Understanding. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XXXV; Springer: Berlin/Heidelberg, Germany, 2022; pp. 105–124. [CrossRef]
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762.
38. Chen, J.; Guo, H.; Yi, K.; Li, B.; Elhoseiny, M. VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18030–18040.

39. Xu, J.; Mei, T.; Yao, T.; Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5288–5296.
40. Chen, D.; Dolan, W. Collecting Highly Parallel Data for Paraphrase Evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 190–200.
41. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
42. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
43. Banerjee, S.; Lavie, A. 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, Ann Arbor, MI, USA, 25–30 June 2005 ; pp. 65–72.
44. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, 21–26 July 2004 ; pp. 74–81.
45. Chen, H.; Li, J.; Hu, X. Delving Deeper into the Decoder for Video Captioning. *arXiv* **2020**, arXiv:2001.05614. https://doi.org/10.48550/arXiv.2001.05614.
46. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin Transformer. *arXiv* **2021**, arXiv:2106.13230.
47. Tewel, Y.; Shalev, Y.; Nadler, R.; Schwartz, I.; Wolf, L. Zero-Shot Video Captioning with Evolving Pseudo-Tokens. *arXiv* **2022**, arXiv:2207.11100.