

VATeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research

vatex.org

Xin Wang^{*1} Jiawei Wu^{*1} Junkun Chen² Lei Li² Yuan-Fang Wang¹ William Yang Wang¹
¹University of California, Santa Barbara, CA, USA
²ByteDance AI Lab, Beijing, China

Abstract

We present a new large-scale multilingual video description dataset, VATeX¹, which contains over 41,250 videos and 825,000 captions in both English and Chinese. Among the captions, there are over 206,000 English-Chinese parallel translation pairs. Compared to the widely-used MSR-VTT dataset [64], VATeX is multilingual, larger, linguistically complex, and more diverse in terms of both video and natural language descriptions. We also introduce two tasks for video-and-language research based on VATeX: (1) Multilingual Video Captioning, aimed at describing a video in various languages with a compact unified captioning model, and (2) Video-guided Machine Translation, to translate a source language description into the target language using the video information as additional spatiotemporal context. Extensive experiments on the VATeX dataset show that, first, the unified multilingual model can not only produce both English and Chinese descriptions for a video more efficiently, but also offer improved performance over the monolingual models. Furthermore, we demonstrate that the spatiotemporal video context can be effectively utilized to align source and target languages and thus assist machine translation. In the end, we discuss the potentials of using VATeX for other video-and-language research.

1. Introduction

Recently, researchers in both computer vision and natural language processing communities are striving to bridge videos and natural language. For a deeper understanding of the activities, the task of video captioning/description aims at describing the video content with natural language. A few datasets have been introduced for this task and cover a variety of domains, such as cooking [15, 70], movie [45],

^{*}Equal contribution.

¹VATeX stands for Video And TEXt, where X also represents various languages.

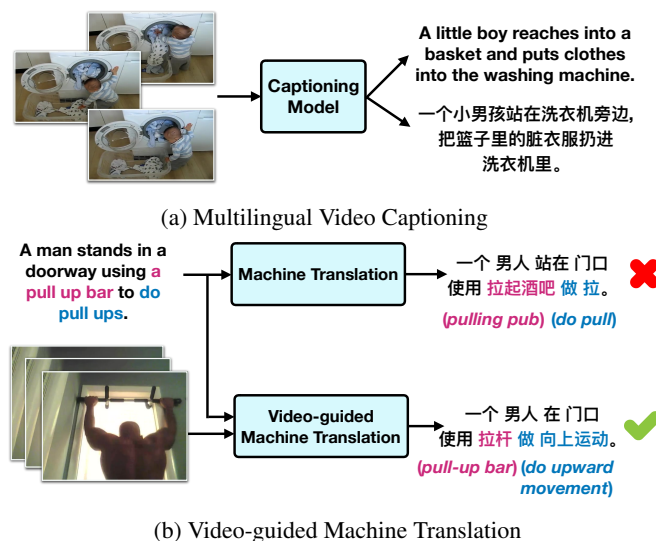


Figure 1: Demonstration of the VATeX tasks. (a) A compact unified video captioning model is required to accurately describe the video content in both English and Chinese. (b) The machine translation model mistakenly interprets “pull up bar” as “pulling pub” and “do pull ups” as “do pull” (two verbs), which are meaningless. While with the relevant video context, the English sentence is precisely translated into Chinese.

human actions [13, 64], and social media [21]. Despite the variants of this task, the fundamental challenge is to accurately depict the important activities in a video clip, which requires high-quality, diverse captions that describe a wide variety of videos at scale. Moreover, existing large-scale video captioning datasets are mostly monolingual (English only) and thus the development of video captioning models is restricted to English corpora. However, the study of multilingual video captioning is essential for a large population on the planet who cannot speak English.

To this end, we collect a new large-scale multilingual dataset for video-and-language research, VATeX, that con-

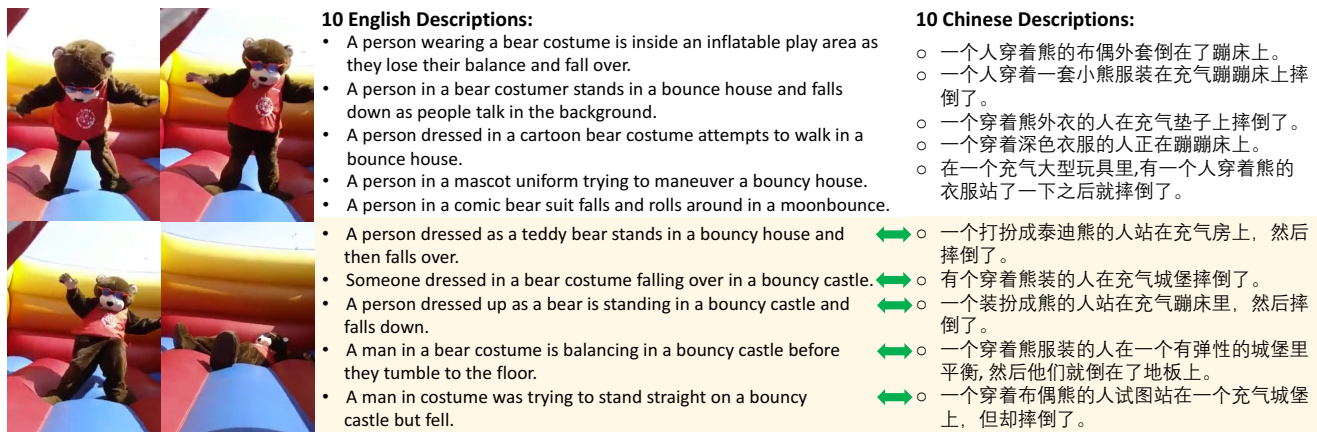


Figure 2: A sample of our VATEX dataset. The video has 10 English and 10 Chinese descriptions. All depicts the same video and thus are distantly parallel to each other, while the last five are the paired translations to each other.

tains over 41,250 unique videos and 825,000 high-quality captions. It covers 600 human activities and a variety of video content. Each video is paired with 10 English and 10 Chinese diverse captions from 20 individual human annotators. Figure 2 illustrates a sample of our VATEX dataset. Compared to the most popular large-scale video description dataset MSR-VTT [64], VATEX is characterized by the following major unique properties. First, it contains both English and Chinese descriptions at scale, which can support many multilingual studies that are constrained by monolingual datasets. Secondly, VATEX has the largest number of clip-sentence pairs with each video clip annotated with multiple unique sentences, and every caption is unique in the whole corpus. Thirdly, VATEX contains more comprehensive yet representative video content, covering 600 human activities in total. Furthermore, both the English and Chinese corpora in VATEX are lexically richer and thus can empower more natural and diverse caption generation.

With the capabilities of the VATEX dataset, we introduce the task of multilingual video captioning (see Figure 1a), which is to train a unified model to generate video descriptions in multiple languages (e.g., English and Chinese). However, *would the multilingual knowledge further reinforce video understanding?* We examine different multilingual models where different portions of the architectures are shared for multiple languages. Experiments show that a compact unified multilingual captioning model is not only more efficient but also more effective than monolingual models.

Video captioning is designed to push forward video understanding with natural language descriptions, but *can video information help natural language tasks like machine translation in return?* To answer this question, we collect around 206K English-Chinese parallel sentences among all the captions and introduce a new task, video-guided ma-

chine translation (VMT), to translate a source language description into the target language using the video information as additional spatiotemporal context. We assume that the spatiotemporal context would reduce the ambiguity of languages (especially for verbs and nouns) and hence promote the alignment between language pairs. So we further conduct extensive experiments and verify the effectiveness of VMT. In Figure 1b, we demonstrate an example where video information can play a crucial role in translating essential information.

In summary, our contributions are mainly three-fold:

- We collect a new large-scale and high-quality multilingual video description dataset for the advance of the video-and-language research, and conduct in-depth comparisons among MSR-VTT, VATEX English corpus, and VATEX Chinese corpus.
- We introduce the task of multilingual video captioning and validate its efficiency and effectiveness of generating video descriptions in both English and Chinese with a compact, unified model.
- We are the first to propose the task of video-guided machine translation and examine the effectiveness of incorporating spatiotemporal context to improve the performance of machine translation.

2. Related Work

Video Description Datasets. Various datasets for video description/captioning have been introduced to empower different ways to describe the video content, covering a wide range of domains, such as cooking [15, 70, 43, 44], movie [54, 45, 46], social media [21], and human activities [13, 50, 64, 28]. In Table 1, we summarize existing video description datasets [1] and briefly compare their major statistics. Generally, video description tasks can mainly

Dataset	MLingual	Domain	#classes	#videos:clips	#sent	#sent/clip
TACoS[43]	-	cooking	26	127:3.5k	11.8k	-
TACoS-MLLevel[44]	-	cooking	67	185:25k	75k	3
Youcook[15]	-	cooking	6	88:-	2.7k	-
Youcook II[70]	-	cooking	89	2k:15.4k	15.4k	1
MPii MD[45]	-	movie	-	94:68k	68.3k	1
M-VAD[54]	-	movie	-	92:46k	55.9k	-
LSMDC[46]	-	movie	-	200:128k	128k	1
Charades[50]	-	indoor	157	10k:10k	27.8k	2-3
VideoStory[21]	-	social media	-	20k:123k	123k	1
ActyNet-Cap[28]	-	open	200	20k:100k	100k	1
MSVD[13]	✓	open	-	2k:2k	70k	35
TGIF[32]	-	open	-	-:100k	128k	1
VTW[67]	-	open	-	18k:18k	18k	1
MSR-VTT[64]	-	open	257	7k:10k	200k	20
VATEX (ours)	✓	open	600	41.3k:41.3k	826k	20

Table 1: Comparison of the video description datasets.

be divided into two families, single-sentence generation (*e.g.*, [13, 64]) and multi-sentence generation (*e.g.*, [28]), though they may appear as different variants due to the difference of the corpora, *e.g.*, video title generation [67] and video story generation [21]. In this work, we present a large-scale, high-quality multilingual benchmark for single-sentence generation, aiming at encouraging fundamental approaches towards a more in-depth understanding of human actions. As shown in Table 1, our VATEX dataset is the largest benchmark in terms of video coverage and the language corpora; it also provides 20 captions for each video clip to take into consideration human variance when describing the same video and hence supports more human-consistent evaluations. Moreover, our VATEX dataset contains both English and Chinese descriptions at scale, which is an order of magnitude larger than MSVD [13]. Besides, MSVD does not have any translation pairs as VATEX does. Therefore, VATEX can empower many multilingual, multimodal research that requires large-scale training.

Multilingual Visual Understanding. Numerous tasks have been proposed to combine vision and language to enhance the understanding of either or both, such as video/image captioning [17, 58, 2], visual question answering (VQA) [4], and natural language moment retrieval [24], etc. Multilingual studies are rarely explored in the vision and language domain. Gao *et al.* [20] introduce a multilingual image question answering dataset, and Shimizu *et al.* [49] propose a cross-lingual method for making use of English annotations to improve a Japanese VQA system. Pappas *et al.* [40] propose multilingual visual concept clustering to study the commonalities and differences among different languages. Meanwhile, multilingual image captioning is introduced to describe the content of an image with multiple languages [30, 55, 31]. But none of them study the interaction between videos and multilingual knowledge. Sanabria *et al.* [47] collect English→Portuguese subtitles for the automatic speech recognition (ASR) task, which however do not directly de-

Split	train	validation	public test	secret test
#videos	25,991	3,000	6,000	6,278
#captions	519,820	60,000	120,000	125,560
action label	✓	✓	-	-

Table 2: The splits of the VATEX dataset (✓ indicates the videos have publicly accessible action labels). For the secret test set, we holdout the human-annotated captions for challenge use.

scribe the video content. Therefore, we introduce the VATEX dataset and the task of multilingual video captioning to facilitate multilingual understanding of video dynamics.

Multimodal Machine Translation. The multimodal machine translation task aims at generating a better target sentence by supplementing the source sentence with extra information gleaned from other modalities. Previous studies mainly focus on using images as the visual modality to help machine translation [52, 18, 6]. The Multi30K dataset [19], which is annotated based on the image captioning dataset Flickr30K [42], is commonly used in this direction. For instance, [26, 22] consider the object features of the images, and [9, 33] import convolutional image features into machine translation. Additionally, other studies [37, 11, 38, 8] explore the cross-modal feature fusion of images and sentences. In this work, we are the first to consider videos as the spatiotemporal context for machine translation and introduce a new task—video-guided machine translation. Compared with images, videos provide richer visual information like actions and temporal transitions, which can better assist models in understanding and aligning the words/phrases between the source and target languages. Moreover, the parallel captions in VATEX go beyond of spatial relations and are more linguistically complex than Multi30K, *e.g.*, a series of actions. Last but not least, our VATEX dataset contains over 206K English-Chinese sentence pairs (5 per video), which is approximately seven times larger than Multi30K.

3. VATEX Dataset

3.1. Data Collection

For a wide coverage of human activities, we reuse a subset of the videos from the Kinetics-600 dataset [27], the largest and widely-used benchmark for action classification. Kinetics-600 contains 600 human action classes and around half a million video clips. To collect those videos, Kay *et al.* [27] first built an action list by combining previous video datasets [23, 29, 51, 3, 60], and then searched the videos from YouTube for candidates, which eventually were filtered by Amazon Mechanical Turkers. Each clip lasts around 10 seconds and is taken from a unique YouTube video. The VATEX dataset connects videos to natural language descriptions rather than coarse action labels. Notably,

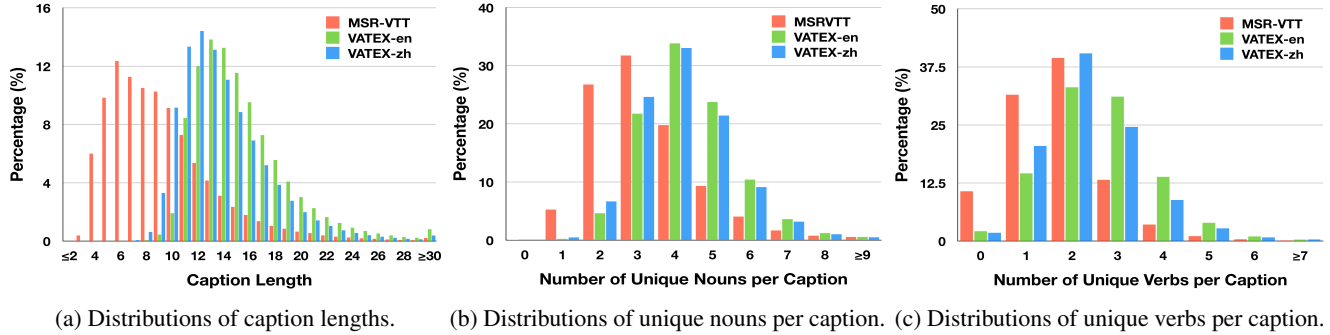


Figure 3: Statistical histogram distributions on MSR-VTT, VATEX-en, and VATEX-zh. Compared to MSR-VTT, the VATEX dataset contains longer captions, each with more unique nouns and verbs.

we collect the English and Chinese descriptions of 41,269 valid video clips from the Kinetics-600 validation and hold-out test sets, costing approximately \$51,000 in total. The data collection window is around two months. We have obtained approvals from the institutional reviewing agency to conduct human subject crowdsourcing experiments, and our payment rate is reasonably high (the estimated hourly rate is higher than the minimum wage required by law).

We split those videos into four sets as shown in Table 2. Note that the train and validation sets are split from the Kinetics-600 validation set, and the test sets are from the Kinetics-600 holdout test set. Below we detail the collection process of both English and Chinese descriptions.

3.1.1 English Description Collection

Towards large-scale and diverse human-annotated video descriptions, we build upon Amazon Mechanical Turk (AMT)² and collect 10 English captions for every video clip in VATEX, where each caption from an individual worker. Specifically, the workers are required to watch the video clips and describe the corresponding captions in English. In each assignment, the workers are required to describe 5 videos. We show the instructions that the workers should describe all the important people and actions in the video clips with the word count in each caption no less than 10. The AMT interface can be found in the supplementary material, which contains more details.

To ensure the quality of the collected captions, we employ only workers from the English-speaking countries, including Australia, Canada, Ireland, New Zealand, UK, and USA. The workers are also required to complete a minimum of 1K previous tasks on AMT with at least a 95% approval rate. Furthermore, we daily spot-check the captions written by each worker to see if they are relevant to the corresponding videos. Meanwhile, we run scripts to check the captions according to the following rules: (1) whether the captions

are shorter than 8 words; (2) whether there are repeated captions; (3) whether the captions contain sensitive words; and (4) whether the captions are not written in English. We reject all the captions that do not achieve the requirements and block the workers consistently providing low-quality annotations. The rejected captions are re-collected until all captions strictly follow the requirements. In preliminary experiments, we find that the workers may struggle to write good captions with only the instructions. Hence, we further provide some accepted good examples and rejected bad examples (both are unrelated to the current video clips) for workers’ reference. We observe that this additional information brings in evident quality improvement on the collected captions. Overall, 2,159 qualified workers annotate 412,690 valid English captions.

3.1.2 Chinese Description Collection

Similar to the English corpus, we collect 10 Chinese descriptions for each video. But to support the video-guided machine translation task, we split these 10 descriptions into two parts, five directly describing the video content and the other five are the paired translations of 5 English descriptions for the same video. All annotations are conducted on the Bytedance Crowdsourcing platform³. All workers are native Chinese speakers and have a good education background to guarantee that the video content can be correctly understood and the corresponding descriptions can be accurately written.

For the first part that directly describes the video content, we follow the same annotation rules as in the collection process of the English captions, except that each Chinese caption must contain at least 15 Chinese characters.

As for the second part, we aim to collect 5 English-Chinese parallel pairs for each video to enable the VMT task. However, direct translation by professional transla-

²<https://www.mturk.com>

³A public Chinese crowdsourcing platform: <https://zc.bytedance.com>

Dataset	sent length	duplicated sent rate		#unique n -grams				#unique POS tags			
		intra-video	inter-video	1-gram	2-gram	3-gram	4-gram	verb	noun	adjective	adverb
MSR-VTT	9.28	66.0%	16.5%	29,004	274,000	614,449	811,903	8,862	19,703	7,329	1,195
VaTeX-en	15.23	0	0	35,589	538,517	1,660,015	2,773,211	12,796	23,288	10,639	1,924
VaTeX-zh	13.95	0	0	47,065	626,031	1,752,085	2,687,166	20,299	30,797	4,703	3,086

Table 3: We demonstrate the average sentence length, the duplicated sentence rate within a video (intra-video) and within the whole corpus (inter-video), the numbers of unique n -grams and POS tags. Our VaTeX dataset is lexically richer than MSR-VTT in general. Note that the Chinese POS tagging rules follow the Penn Chinese Treebank standard [63], which is different from English due to different morphemes. For instance, VaTeX-zh has more nouns and verbs but fewer adjectives than VaTeX-en, because the semantics of many Chinese adjectives are included in nouns or verbs [69]⁴.

tors is costly and time-consuming. Thus, following previous methods [7, 66] on collecting parallel pairs, we choose the post-editing annotation strategy. Particularly, for each video, we randomly sample 5 captions from the annotated 10 English captions and use multiple translation systems to translate them into Chinese reference sentences. Then the annotation task is, given the video and the references, the workers are required to post-edit the references and write the parallel Chinese sentence following two rules: (1) the original sentence structure and semantics need be maintained to guarantee the alignment to the corresponding English sentence, and (2) lost or wrong entities and actions could be corrected based on the video content to eliminate the errors from the translation systems. To further reduce the annotation bias towards one specific translation system, here we use three advanced English→Chinese translation systems (Google, Microsoft, and self-developed translation systems) to provide the workers with machine-translated sentences as references for each English caption.

In order to ensure the quality of the Chinese captions, we conduct a strict two-stage verification: every collected description must be reviewed and approved by another independent worker. Workers with less than 90% approval rate are blocked. The interfaces for Chinese caption collection can be found in the supplementary material. Eventually, 450 Chinese workers participate in these two tasks and write 412,690 valid Chinese captions. Half of the captions are English-Chinese parallel sentences, so we have 206,345 translation pairs in total.

3.2. Dataset Analysis

In Table 1, we briefly compare the overall statistics of the existing video description datasets. In this section, we conduct comprehensive analysis between our VaTeX dataset and the MSR-VTT dataset [64], which is the widely-used benchmark for video captioning and the closest to VaTeX in terms of domain and scale. Since MSR-VTT only has

⁴For example, the segmented Chinese word 长发 (“long hair”) is labeled as one noun in Chinese, but an adjective (“long”) and a noun (“hair”) in English.

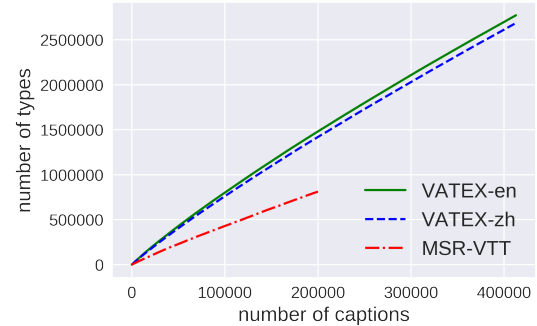


Figure 4: Type-caption curves. Type: unique 4-gram. VaTeX has more lexical styles and caption diversity than MSR-VTT.

English corpus, we split VaTeX into the English corpus (VaTeX-en) and the Chinese corpus (VaTeX-zh) for comparison. VaTeX contains 413k English and 413k Chinese captions depicting 41.3k unique videos from 600 activities, while MSR-VTT has 200k captions describing 7k videos from 257 activities. In addition to the larger scale, the captions in both VaTeX-en and VaTeX-zh are longer and more detailed than those in MSR-VTT (see Figure 3). The average caption lengths of VaTeX-en, VaTeX-zh, and MSR-VTT are 15.23, 13.95, and 9.28.

To assess the linguistic complexity, we compare the unique n -grams and part-of-speech (POS) tags (*e.g.*, verb, noun, adverb *etc.*) among MSR-VTT, VaTeX-en and VaTeX-zh (see Table 3), which illustrates the improvement of VaTeX over MSR-VTT and the difference between the English and Chinese corpora. Evidently, our VaTeX datasets represent a wider variety of caption styles and cover a broader range of actions, objects, and visual scenes.

We also perform in-depth comparisons of caption diversity. First, as seen in Table 3, MSR-VTT faces a severe duplication issue in that 66.0% of the videos contains some exactly same captions, while our VaTeX datasets are free of this problem and guarantee that the captions within the same video are unique. Not only within videos, but the captions in our VaTeX datasets are also much more diverse even

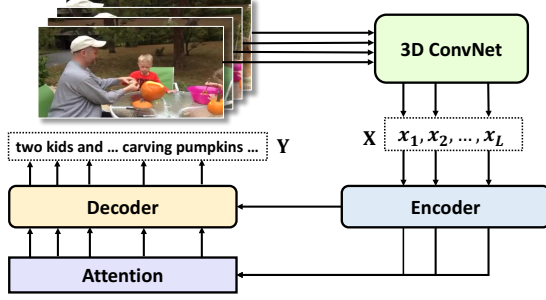


Figure 5: Monolingual video captioning model.

within the whole corpus, which indicates that our VATeX can also be a high-quality benchmark for video retrieval.

For a more intuitive measure of the lexical richness and caption diversity, we then propose the *Type-Caption Curve*, which is adapted from the type-token vocabulary curve [65] but specially designed for the caption corpora here. The total number of captions and the number of distinct vocabulary words (types) are computed for each corpus. So we plot the number of types against the number of captions for MSR-VTT, VATeX-en, and VATeX-zh (see Figure 4 where we choose 4-grams as the types). From these type-caption curves, inferences are drawn about lexical style or caption diversity (vocabulary use), as well as lexical competence (vocabulary size), so our VATeX datasets are shown to be more linguistically complex and diverse.

4. VATeX Tasks

4.1. Multilingual Video Captioning

Multilingual video captioning is the task of describing the content of a video using more than one language such as English and Chinese. Below we first introduce a baseline model for monolingual video captioning and then present three different models for multilingual video captioning.

4.1.1 Models

We begin with the well-known attention-based encoder-decoder model for video captioning. As illustrated in Figure 5, there are three main modules to this architecture:

- A 3D convolutional neural network (3D ConvNet) that learns the spatiotemporal features of the video and outputs a sequence of segment-level features $X = \{x_1, x_2, \dots, x_L\}$.
- A video encoder module f_{enc} that encodes X into video-level features $V = \{v_1, v_2, \dots, v_L\}$ by modeling long-range temporal contexts.
- An attention-based language decoder module f_{dec} that produces a word y_t at every time step t by consider-

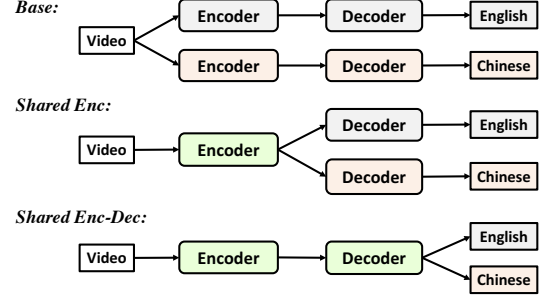


Figure 6: Multilingual video captioning models.

ing the word at previous step y_{t-1} , the visual context vector c_t learned from the attention mechanism.

We instantiate the captioning model by adapting the model architectures from the state-of-the-art video captioning methods [41, 59]. We employ the pretrained I3D model [12] for action recognition as the 3D ConvNet to obtain the visual features X , Bidirectional LSTM [48] (bi-LSTM) as the video encoder f_{enc} , and LSTM [25] as the language decoder f_{dec} . We also adopt the dot-product attention, so at the decoding step t , we have

$$y_t, h_t = f_{dec}([y_{t-1}, c_t], h_{t-1}), \quad (1)$$

where h_t is the hidden state of the decoder at step t and

$$c_t = \text{softmax}(h_{t-1} W V^T) V, \quad (2)$$

where W is a learnable projection matrix.

To enable multilingual video captioning, we examine three methods (see Figure 6): (1) Two **Base** models, which are two monolingual encoder-decoder models (as described in Figure 5) trained separately for either English or Chinese; (2) A **Shared Enc** model, which has a shared video encoder but two language decoders to generate English and Chinese; (3) A **Shared Enc-Dec** model, where there are just one encoder and one decoder, both shared by English and Chinese, and the only difference is that the word embedding weight matrices are different for different languages.

4.1.2 Experimental Setup

Implementation Details. We train the models on the VATeX dataset following the splits in Table 2. To preprocess the videos, we sample each video at 25fps and extract the I3D features [12] from these sampled frames. The I3D model is pretrained on the original Kinetics training dataset [27] and used here without fine-tuning. More details about data preprocessing and implementation can be found in the supplementary material.

Evaluation Metrics. We adopt four diverse automatic evaluation metrics: BLEU [39], Meteor [16], Rouge-L [34], and CIDEr [56]. We use the standard evaluation code from MSCOCO server [14] to obtain the results.

Model	#Params	English				Chinese			
		BLEU-4	Meteor	Rouge-L	CIDEr	BLEU-4	Meteor	Rouge-L	CIDEr
Base w/o WT	52.5M	28.1 \pm 0.38	21.7 \pm 0.15	46.8 \pm 0.18	44.3 \pm 0.98	24.4 \pm 0.86	29.6 \pm 0.30	51.3 \pm 0.43	34.0 \pm 0.11
Base	39.7M	28.1 \pm 0.32	21.6 \pm 0.19	46.9 \pm 0.16	44.3 \pm 0.10	24.9 \pm 0.20	29.7 \pm 0.21	51.5 \pm 0.28	34.7 \pm 0.47
Shared Enc	34.9M	28.4 \pm 0.21	21.7 \pm 0.65	47.0 \pm 0.09	45.1 \pm 0.25	24.9 \pm 0.26	29.7 \pm 0.11	51.6 \pm 0.20	34.9 \pm 0.40
Shared Enc-Dec	26.3M	27.9 \pm 0.50	21.6 \pm 0.55	46.8 \pm 0.19	44.2 \pm 0.23	24.9 \pm 0.25	29.8 \pm 0.23	51.7 \pm 0.09	35.0 \pm 0.18

Table 4: Multilingual video captioning. We report the results of the baseline models in terms of BLEU-4, Meteor, and Rouge-L, and CIDEr scores. Each model is trained for five times with different random seeds and the results are reported with a confidence level of 95%. WT: weight tying, which means the input word embedding layer and the softmax layer share the same weight matrix.

4.1.3 Results and Analysis

Table 4 shows the results of the three baseline models on both English and Chinese test sets. The performances of the multilingual models (*Shared Enc* and *Shared Enc-Dec*) are consistently (though not significantly) improved over the monolingual model (*Base*). It indicates that multilingual learning indeed helps video understanding by sharing the video encoder. More importantly, the parameters of the *Shared Enc* and *Shared Enc-Dec* are significantly reduced by 4.7M and 13.4M over the *Base* models. These observations validate that a compact unified model is able to produce captions in multiple languages and benefits from multilingual knowledge learning. We believe that more specialized multilingual models would improve the understanding of the videos and lead to better results. Furthermore, incorporating multimodal features like audio [61] would further improve the performance, which we leave for future study.

4.2. Video-guided Machine Translation

In this section, we discuss the enabled new task, Video-guided Machine Translation (VMT), to translate a source language sentence into the target language using the video information as additional spatiotemporal context. This task has various potential real-world applications, e.g., translating posts with the video content in social media.

4.2.1 Method

In VMT, the translation system takes a source sentence and the corresponding video as the input, and generates the translated target sentence. To effectively utilize the two modalities, text and video, we design a multimodal sequence-to-sequence model [53, 57] with the attention mechanism [5, 36] for VMT. The overview of our model is shown in Figure 7, which mainly consists of the following three modules.

Source Encoder. For each source sentence represented as a sequence of N word embeddings $S = \{s_1, s_2, \dots, s_N\}$, the source encoder f_{enc}^{src} transforms it into the sentence features $U = \{u_1, u_2, \dots, u_N\}$.

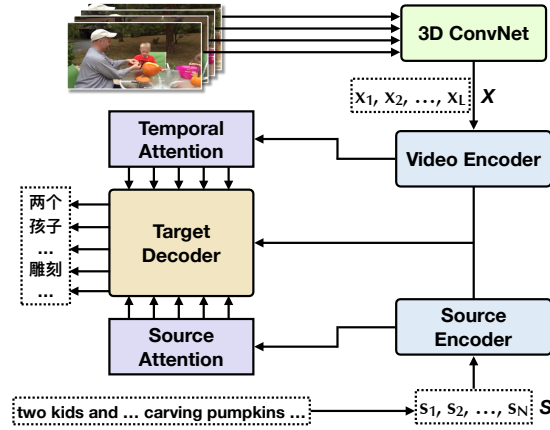


Figure 7: Video-guided machine translation model.

Video Encoder. Similar in Section 4.1, we use a 3D ConvNet to convert each video into a sequence of segment-level features X . Then we employ a video encoder f_{enc}^{vi} to transform X into the video features $V = \{v_1, v_2, \dots, v_L\}$.

Target Decoder. The sentence embedding from the source language encoder f_{enc}^{src} and the video embedding from the video encoder f_{enc}^{vi} are concatenated and fed into the target language decoder f_{dec}^{tgt} . To dynamically highlight the important words of the source sentence and the crucial spatiotemporal segments in the video, we equip the target decoder f_{dec}^{tgt} with two attention mechanisms. Thus, at each decoding step t , we have

$$y_t, h_t = f_{dec}^{tgt}([y_{t-1}, c_t^{src}, c_t^{vi}], h_{t-1}), \quad (3)$$

where h_t is the hidden state of the decoder at step t . c_t^{vi} is the video context vector that is computed with the temporal attention of the video segments (see Equation 2), and c_t^{src} is the source language context vector:

$$c_t^{src} = \text{softmax}(h_{t-1} W^{src} U^T) U, \quad (4)$$

where W^{src} is a learnable projection matrix.

Model	English→Chinese	Chinese→English
Base NMT w/o VI	26.85	24.31
+ Average VI	26.97 (+0.12)	24.39 (+0.08)
+ LSTM VI w/o Attn	27.43 (+0.58)	24.76 (+0.45)
+ LSTM VI w/ Attn (VMT)	29.12 (+2.27)	26.42 (+2.11)

Table 5: Video-guided Machine Translation. Results are reported on the BLEU-4 scores. VI: video features from the pretrained I3D model. Attn: temporal attention mechanism.

4.2.2 Experimental Setup

Baselines. We consider the following three baselines to compare: (1) *Base NMT Model*: We only consider the text information for machine translation and adopt the encoder-decoder model with the source attention mechanism. (2) *Average Video Features*: We average the segment-level features X of each video as \bar{x} . The average video feature \bar{x} is then concatenated with each word embedding s_t in S . The model structure is the same as the base NMT model. (3) *LSTM Video Features*: This is our VMT model without the temporal attention for videos in the decoder.

4.2.3 Results and Analysis

VMT. We first show in Table 5 the results of four different models on *Chinese→English* and *English→Chinese* translations. The marginal improvements by the *Average video Features* and the *LSTM Video Features* reveal that, passively receiving and incorporating the video features is ineffective in helping align source and target languages. However, we can observe that the translation system achieves much better performance when using the *LSTM Video Features* with temporal attention (our full VMT model) to dynamically interact with the video features. It is because that with the attention mechanism, the language dynamics are used as a query to highlight the relevant spatiotemporal features in the video, and then the learned video context would assist the word mapping between source and target language spaces. This also validates that extra video information can be effectively utilized to boost machine translation systems.

Masked VMT. Videos contain rich information on subject/object nouns and action verbs. Therefore, we conduct noun/verb masking experiments [10] to investigate to what extent the video information can help machine translation. We randomly replace 0%/25%/50%/75%/100% nouns or verbs of the English captions with a special token [M], and then train the NMT and VMT models on the *Chinese→English* translation task with different masking rates. This experimental design is to evaluate the capability of VMT in recovering the missing information of the source sentence with the help of the video context.

In addition to the BLEU-4 metric, we propose to use the noun/verb recovery accuracy, which is the percentages of

Rate	BLEU-4					Accuracy (%)				
	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%
Noun Masking										
NMT	26.9	20.2	13.0	8.5	4.1	70.2	53.7	35.4	15.6	10.1
VMT	29.1	24.7	19.3	16.9	14.3	76.4	65.6	50.8	43.2	39.7
Verb Masking										
NMT	26.9	23.3	15.4	11.6	7.2	65.1	57.4	40.9	33.6	19.8
VMT	29.1	26.8	22.0	19.3	16.5	70.4	63.6	54.2	48.7	40.5

Table 6: Video-guided machine translation on *English→Chinese* with different noun/verb masking rates. We evaluate the results using the BLEU-4 score and noun/verb recovery accuracy.

the correctly translated nouns/verbs in the target sentences, to precisely evaluate the impact of additional video information on recovering nouns/verbs. The results with different masking rates are shown in Table 6. First, the VMT model consistently outperforms the NMT model with different masking rates on both metrics. Moreover, as the masking rate increases, the NMT model struggles to figure out the correct nouns/verbs because of the scarce parallel caption pairs; while the VMT model can rely on the video context to obtain more useful information for translation, and thus the performance gap on the recovery accuracy increases dramatically. It shows that in our VMT model, video information can play a crucial role in understanding subjects, objects, and actions, as well as their relations.

5. Discussion and Future Work

In this paper, we introduce a new large-scale multilingual dataset for video-and-language research. In addition to (multilingual) video captioning and video-guided machine translation, there are also some other potentials of this dataset. For example, since the natural language descriptions in VATEX are unique, one promising direction is to use multilingual descriptions of our dataset as queries to retrieve the video clip from all videos [35] or even localize it within an untrimmed long video [68]. Meanwhile, VATEX has 600 fine-grained action labels, so we can hold-out certain action classes to evaluate the generalizability of different video captioning models to support zero-/few-shot learning [62]. Furthermore, our dataset can contribute to other research fields like Neuroscience. For instance, when describing the same videos, the focus points of people using different languages can be reflected by their written captions. By analyzing multilingual captions, one can likely infer the commonality and discrepancy on the brain attention of people with different cultural and linguistic backgrounds. In general, we hope the release of our VATEX dataset would facilitate the advance of video-and-language research.

References

- [1] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: a survey of methods, datasets and evaluation metrics. *arXiv preprint arXiv:1806.00186*, 2018.
- [2] Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. *arXiv preprint arXiv:1812.08658*, 2018.
- [3] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.
- [5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*, 2015.
- [6] Loïc Barrault, Fethi Bougares, Lucia Specia, Chirag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT)*, pages 304–323, 2018.
- [7] Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. A human judgement corpus and a metric for arabic mt evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, 2014.
- [8] Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. Lium-cvc submissions for wmt17 multimodal translation task. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, pages 432–439, 2017.
- [9] Ozan Caglayan, Loïc Barrault, and Fethi Bougares. Multimodal attention for neural machine translation. *arXiv preprint arXiv:1609.03976*, 2016.
- [10] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.
- [11] Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 992–1003, 2017.
- [12] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.
- [13] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011.
- [14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [15] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the 26th IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2634–2641, 2013.
- [16] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the 9th workshop on statistical machine translation*, pages 376–380, 2014.
- [17] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:677–691, 2015.
- [18] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, pages 215–233, 2017.
- [19] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, 2016.
- [20] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*, 2015.
- [21] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–974, 2018.
- [22] Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, et al. The memad submission to the wmt18 multimodal translation task. In *Proceedings of the 3rd Conference on Machine Translation (WMT)*, 2018.
- [23] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nibbles. Activitynet: A large-scale video benchmark for human activity understanding. *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.
- [24] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. *Proceedings of the*

- 2017 *IEEE International Conference on Computer Vision (ICCV)*, pages 5804–5813, 2017.
- [25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.
- [26] Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the 1st Conference on Machine Translation (WMT)*, pages 639–645, 2016.
- [27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [28] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017.
- [29] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. *Proceedings of the 2011 International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011.
- [30] Weiyu Lan, Xirong Li, and Jianfeng Dong. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM international conference on Multimedia (ACM-MM)*, 2017.
- [31] Xirong Li, Xiaoxu Wang, Chaoxi Xu, Weiyu Lan, Qijie Wei, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning and retrieval. *IEEE Transactions on Multimedia*, 2019.
- [32] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650, 2016.
- [33] Jindřich Libovický and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 196–202, 2017.
- [34] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.
- [35] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. Visual semantic search: Retrieving videos via complex textual queries. *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2657–2664, 2014.
- [36] Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015.
- [37] Mingbo Ma, Dapeng Li, Kai Zhao, and Liang Huang. Osu multimodal machine translation system report. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, pages 465–469, 2017.
- [38] Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. Sheffield multimt: Using object posterior predictions for multimodal machine translation. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, pages 470–476, 2017.
- [39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2001.
- [40] Nikolaos Pappas, Miriam Redi, Mercan Topkara, Brendan Jou, Hongyi Liu, Tao Chen, and Shih-Fu Chang. Multilingual visual sentiment concept matching. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR)*, 2016.
- [41] Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.
- [42] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the 2015 IEEE international conference on computer vision (ICCV)*, pages 2641–2649, 2015.
- [43] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36, 2013.
- [44] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Proceedings of the German Conference on Pattern Recognition*, pages 184–195. Springer, 2014.
- [45] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the 28th IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3202–3212, 2015.
- [46] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision (IJCV)*, 123(1):94–120, May 2017.
- [47] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*, 2018.
- [48] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, Nov. 1997.
- [49] Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 2018.
- [50] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the 2016 European Conference on Computer Vision (ECCV)*, pages 510–526, 2016.

- [51] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [52] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the 1st Conference on Machine Translation (WMT)*, pages 543–553, 2016.
- [53] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceedings of the 28th Conference on Neural Information Processing Systems (NeurIPS)*, pages 3104–3112, 2014.
- [54] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.
- [55] Satoshi Tsutsui and David Crandall. Using artificial tokens to control languages for multilingual image caption generation. *arXiv preprint arXiv:1706.06275*, 2017.
- [56] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.
- [57] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4534–4542, 2015.
- [58] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.
- [59] Xin Wang, Wenhui Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [60] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions transformations. *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2658–2667, 2016.
- [61] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018.
- [62] Xin Wang, Jiawei Wu, Da Zhang, and William Yang Wang. Learning to compose topic-aware mixture of experts for zero-shot video captioning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019.
- [63] Fei Xia. The part-of-speech tagging guidelines for the penn chinese treebank (3.0). *IRCS Technical Reports Series*, page 38, 2000.
- [64] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [65] Gilbert Youmans. Measuring lexical style and competence: The type-token vocabulary curve. *Style*, 24(4):584–599, 1990.
- [66] Wajdi Zaghrouani, Nizar Habash, Ossama Obeid, Behrang Mohit, Houda Bouamor, and Kemal Oflazer. Building an arabic machine translation post-edited corpus: Guidelines and annotation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1869–1876, 2016.
- [67] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Generation for user generated videos. In *Proceedings of the 2016 European conference on computer vision (ECCV)*, pages 609–625, 2016.
- [68] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019.
- [69] Ruihua Zhang. *Sadness expressions in English and Chinese: Corpus linguistic contrastive semantic analysis*. Bloomsbury Publishing, 2016.
- [70] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018.