A. Implementation Details

Multilingual Video Captioning To preprocess the videos, we sample each video at 25 frames per second (fps) and extract the I3D features [2] from these sampled frames. The I3D model is pretrained on the original Kinetics training dataset [3] and used here without fine-tuning. Both the English and Chinese captions are truncated to a maximum of 30 words. Note that we use the segmented Chinese words rather than raw Chinese characters. The vocabularies are built with a minimum word count, resulting in around 11,000 English words and about 14,000 Chinese words.

All the hyperparameters are tuned on the validation sets and same for both English and Chinese caption training. The video encoder is a bi-LSTM of size 512 and the decoder LSTM is of size 1024. The dimensions of the word embedding layers are 512. All models are trained using MLE loss and optimized using Adam optimizer [4] with a batch size 256. We adopt Dropout for regularization. The learning rate is initially set as 0.001 and then halted when the current CIDEr score does not surpass the previous best for 4 epochs. Schedule sampling [1] is employed to train the models. At test time, we use beam search of size 5 to report the final results.

Video-guided Machine Translation The data preprocessing steps are the same as above except that we truncate the captions with a maximum length of 40 here. The baseline NMT is composed of a 2-layer bi-LSTM encoder of size 512 and a 2-layer LSTM decoder of size 1024. The dimensions of both English and Chinese word embeddings are 512. The video encoder is a bi-LSTM of size 512. MLE loss is implemented to train the model using Adam optimizer [4]. The batch size is 32 during training and early-stopping is used to choose the models. As for evaluation, we use beam search of size 5 to report the results on the BLEU-4 metric.

B. Data Collection Interfaces

We show the AMT interface for English caption collection in Figure 1. Since the Chinese captions are divided into two parts, we build two separate interfaces, one of which is to collect the captions that directly describe the video (Figure 2) and the other for collecting the Chinese translations parallel to the English captions (Figure 3).

C. More VATeX Samples

In addition to the example shown in the main paper, Figure 4 demonstrates more samples of our VATeX dataset.

D. Qualitative Results

Multilingual Video Captioning Figure 5 illustrates some qualitative examples of multilingual video captioning, where we compare both the English and Chinese results generated by the monolingual models (Base), the multilingual model that shares the video encoder for English & Chinese (Shared Enc), and the multilingual model that shares both the video encoder and the language decoder for English & Chinese (Shared Enc-Dec).

Video-guided Machine Translation (VMT) In Figure 6, we showcase the advantages of the VMT model over the base neural machine translation (NMT) model. Moreover, we further conduct the masked machine translation experiments and qualitatively demonstrate the effectiveness of VMT in recovering nouns or verbs in Figure 7.

References

[1] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with
Figure 1: The AMT interface for collecting the English captions. In each assignment, the workers are required to annotate 5 video clips. The instructions are kept visible for each clip. We provide the workers with the accepted good examples and rejected bad examples to further improve the quality of annotations. Note that the given examples are unrelated to the current video clips.

Figure 2: The interface for collecting the Chinese captions by directly describing the video content. In each assignment, the workers are required to annotate 1 video clip. The instructions are kept visible for each clip. After the first-stage annotation, each Chinese caption must be reviewed and approved by another independent worker.

Figure 3: The interface for collecting the Chinese captions by post-editing the translated reference sentences and watching the video clips. In each assignment, the workers are required to annotate 1 video clip. The instructions are kept visible for each clip. We provide the workers with three reference sentences translated by Google, Microsoft and Self-developed translation systems. Note that the order of three reference sentences is randomly shuffled for each video clip to reduce the annotation bias towards one specific translation system. After the first-stage annotation, each Chinese caption must be reviewed and approved by another independent worker.
10 English Descriptions:
- A person is parasailing above a body of water and landing on a beach.
- Someone is recording people who are parasailing and people who are watching too.
- A man is riding a parachute and a group of people are standing down and watching them.
- Someone parasailing over a lake with several men watching.
- A person is coming down from a sky riding on a balloon glide.
- Men on a beach prepare to assist an incoming parasailor.
- A person is landing with a parachute onto a beach while others are greeting him or her.
- Someone hanging from a parachute is being pulled on a line while people watch.
- Tied to the end of a long cable, someone is para sailing and comes for a landing on a sandy beach in front of others.
- A group of people help a person parasailing to the ground.

10 Chinese Descriptions:
- 一群人看另一个人从降落伞上准备落下。
- 一个人拿着降落伞从空中落了下来。
- 一个人在一个滑翔伞上滑翔，几个人把他拽了下来。
- 一个人乘着降落伞将降落到沙滩上，沙滩上的人们在对他挥手。
- 在一个晴朗的天气，有一个人飘在空中，旁边有一些人在看着。
- 在海滩上的人都在准备协助降落伞的掉落。
- 一个人拿着降落伞降落在海滩上，而其他人正在围着他。
- 挂在降落伞上的人被用绳子拉着，而人们则在旁边观看。
- 一个人在一条长线的末端并和别人面前降落在沙滩上。
- 在室外，有一群人正在帮助一个伞跳到地面。

10 English Descriptions:
- A person is walking around in an outdoor field with a can that is on fire.
- A man holds a beer bottle that is on fire and tries two times to blow on it to make the flame bigger.
- A man is holding a burning bottle and then he spits flames from it in the air.
- A man holding a flaming beer being coaxed by others to spit into the flame.
- Someone holds a bottle with a flame and blows on it to make the flame even larger.
- A man is cheered on by others as demonstrated fire spitting.
- A man is holding a torch with a fire and spitting a liquid on it.
- A man is holding something on fire as he blows in to it to make a large flame.
- A crowd cheers on “go go go” as a boy holds a bottle on fire and blows to make flames.
- A man holding a flame in his hands tries to unsuccessfully blow it out.

10 Chinese Descriptions:
- 一个男人正在一片绿色的草地上玩喷火。
- 一个男人在草地上拿着点燃的瓶子给周围人表演喷火。
- 一个男人拿着火把进行杂技表演。
- 一个穿着短袖的人在户外草坪上玩火。
- 一个男人手中拿着燃烧的燃料瓶，并用嘴吹了第一下喷火，吹了第二下时没喷火。
- 一个男人在别人的手底下对着火把吐火。
- 一名男子手持火炬，然后在上面喷了一口液体，表演喷火。
- 当一个男人吹火的时候，手里拿着东西着火了，形成一个火的火焰。
- 当一个男孩拿着一个瓶子着火并吹起火焰时，一群人欢呼。
- 一个男人手里拿着一个带火焰的物体，他用嘴使劲吹，但是火焰变得更大了。

Figure 4: More samples of our VATEX dataset. Each 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.
Figure 5: Qualitative comparison among different methods of multilingual video captioning on the VaTeX dataset. Both the English and Chinese results are shown. For each video sample, we list a human-annotated caption and the generated results by three models, Base, Shared Enc, and Shared Enc-Dec. The multilingual models (Shared Enc and Shared Enc-Dec) can generate more coherent and informative captions than the monolingual model (Base).
English:
a young girl does a cartwheel in her
homes living room.

Ground Truth:
一个年轻女孩在她的起居室里做侧手翻。

NMT:
一个年轻女孩在她房间里做车轮。

VMT:
一个年轻女孩在她的房间里翻筋斗。

Figure 6: Qualitative comparison between neural machine translation (NMT) and
document understanding (VMT) on the VATEX dataset. For each video sample, we list the
original English description and the translated sentences by the base NMT model and
our VMT model. The NMT model mistakenly interprets some words and phrases, while
the VMT model can generate more precise translation with the corresponding video context.

English:
a woman with blonde hair is giving a
[M] a[M].

[M]: dog, haircut.

Ground Truth:
一位金发女士正在用电剪刀给理发。

NMT:
一个金发女人正在给指示一个人。

VMT:
一个金发女人正在给一只狗理发。

Figure 7: Qualitative comparison between masked neural machine translation (NMT) and
masked video-guided machine translation (VMT) on the VATEX dataset. The nouns/verbs in
English captions are randomly replaced by a special token [M]. For each video sample, we list the
original English description and the translated sentences by the base NMT model and
our VMT model. The NMT model struggles to figure out the correct nouns/verbs because of the
scarce parallel pairs, while the VMT model can rely on the video context to recover the masked nouns/verbs.