VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research, Supplementary Material

vatex.org

Xin Wang*[1] Jiawei Wu*[1] Junkun Chen[2] Lei Li[2] Yuan-Fang Wang[1] William Yang Wang[1]
[1]University of California, Santa Barbara, CA, USA
[2]ByteDance AI Lab, Beijing, China

A. Implementation Details

Multilingual Video Captioning To preprocess the videos, we sample each video at $25fps$ 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[1] rather than raw Chinese characters. The vocabularies are built with a minimum word count 5, 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 prepossessing 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*).

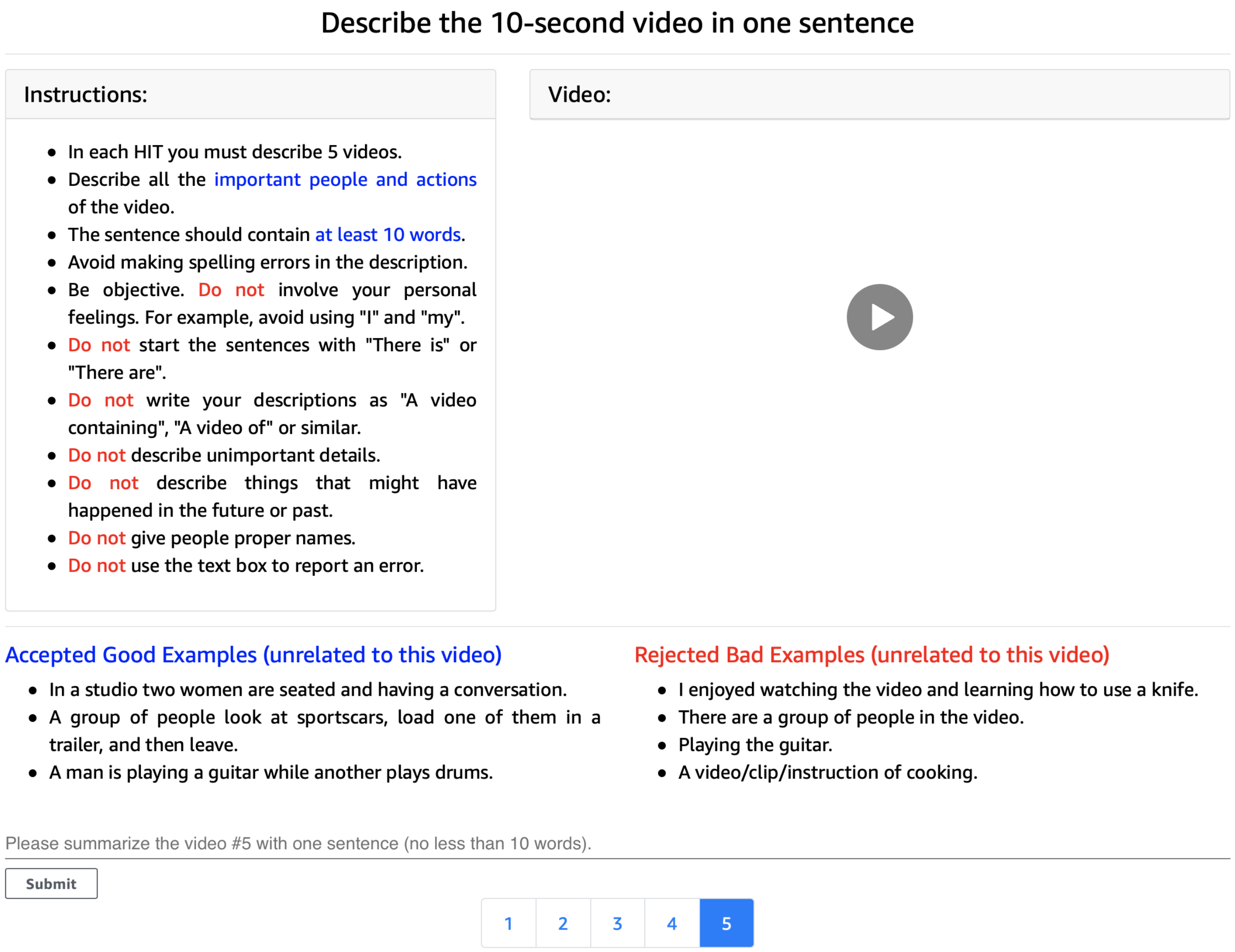
Video-guided Machine Translation (VMT) In Figure 6, we showcase the advantages of the VMT model over the base neural machine translation (NMT) model. Moreover, we further conduct the masked machine translation experiments and qualitatively demonstrate the effectiveness of VMT in recovering nouns or verbs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

References

[1] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with

*Equal contribution.

[1] We use the open-source tool Jieba for Chinese word segmentation: https://github.com/fxsjy/jieba



Describe the 10-second video in one sentence

Instructions:

- In each HIT you must describe 5 videos.
- Describe all the important people and actions of the video.
- The sentence should contain at least 10 words.
- Avoid making spelling errors in the description.
- Be objective. Do not involve your personal feelings. For example, avoid using "I" and "my".
- Do not start the sentences with "There is" or "There are".
- Do not write your descriptions as "A video containing", "A video of" or similar.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not give people proper names.
- Do not use the text box to report an error.

Video:

Accepted Good Examples (unrelated to this video)

- In a studio two women are seated and having a conversation.
- A group of people look at sportscars, load one of them in a trailer, and then leave.
- A man is playing a guitar while another plays drums.

Rejected Bad Examples (unrelated to this video)

- I enjoyed watching the video and learning how to use a knife.
- There are a group of people in the video.
- Playing the guitar.
- A video/clip/instruction of cooking.

Please summarize the video #5 with one sentence (no less than 10 words).

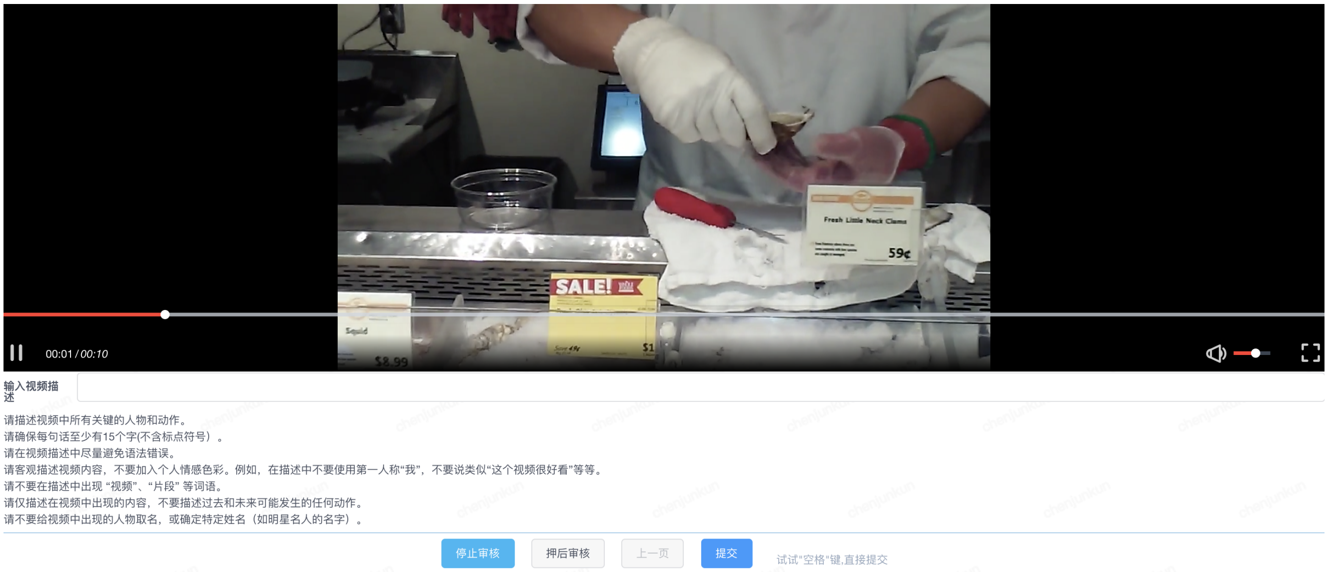
Figure 1: The AMT interface for collecting the English captions. In each assignment, the workers are required to annotate 5 video clips. The instructions are kept visible for each clip. We provide the workers with the accepted good examples and rejected bad examples to further improve the quality of annotations. Note that the given examples are unrelated to the current video clips.

recurrent neural networks. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*, 2015. 1

[2] 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. 1

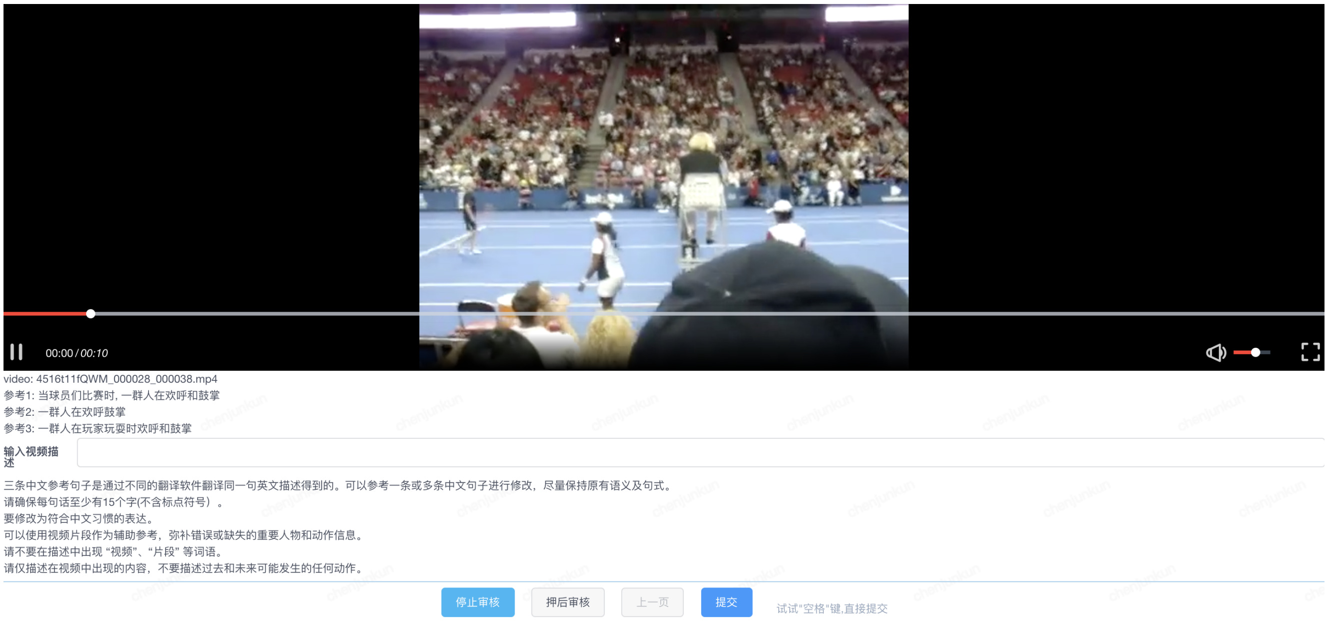
[3] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*, 2015. 1



请描述视频中所有关键的人物和动作。
请确保每句话至少有15个字(不含标点符号）。
请在视频描述中尽量避免语法错误。
请客观描述视频内容，不要加入个人情感色彩。例如，在描述中不要使用第一人称“我”，不要说类似“这个视频很好看”等等。
请不要在描述中出现 “视频”、“片段” 等词语。
请仅描述在视频中出现的内容，不要描述过去和未来可能发生的任何动作。
请不要给视频中出现的人物取名，或确定特定姓名（如明星名人的名字）。

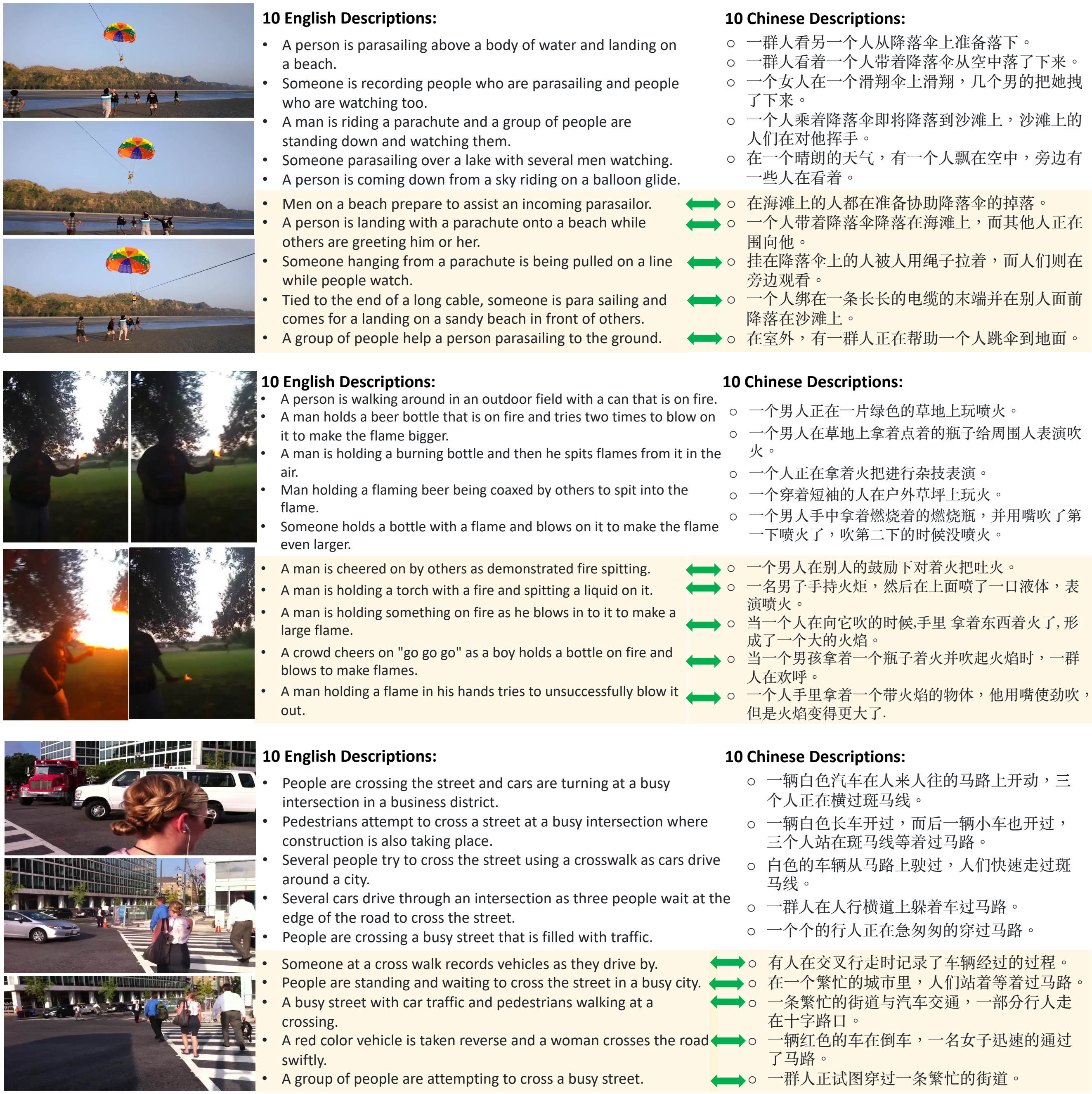
Figure 2: The interface for collecting the Chinese captions by directly describing the video content. In each assignment, the workers are required to annotate 1 video clip. The instructions are kept visible for each clip. After the first-stage annotation, each Chinese caption must be reviewed and approved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



video: 4516t11fQWM_000028_000038.mp4
参考1: 当球员们比赛时, 一群人在欢呼和鼓掌
参考2: 一群人在欢呼鼓掌
参考3: 一群人在玩家玩耍时欢呼和鼓掌

三条中文参考句子是通过不同的翻译软件翻译同一句英文描述得到的。可以参考一条或多条中文句子进行修改，尽量保持原有语义及句式。
请确保每句话至少有15个字(不含标点符号）。
要修改为符合中文习惯的表达。
可以使用视频片段作为辅助参考，弥补错误或缺失的重要人物和动作信息。
请不要在描述中出现 “视频”、“片段” 等词语。
请仅描述在视频中出现的内容，不要描述过去和未来可能发生的任何动作。

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。
- 一群人看着一个人带着降落伞从空中落了下来。
- 一个女人在一个滑翔伞上滑翔，几个男的把她拽了下来。
- 一个人乘着降落伞即将降落到沙滩上，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。
- 一个人带着降落伞降落在海滩上，而其他人正在围向他。
- 挂在降落伞上的人被人用绳子拉着，而人们则在旁边观看。
- 一个人绑在一条长长的电缆的末端并在别人面前降落在沙滩上。
- 在室外，有一群人正在帮助一个人跳伞到地面。

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.
- 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着的瓶子给周围人表演吹火。
- 一个人正在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着的燃烧瓶，并用嘴吹了第一下喷火了，吹第二下的时候没喷火。
- 一个男人在别人的鼓励下对着火把吐火。
- 一名男子手持火炬，然后在上面喷了一口液体，表演喷火。
- 当一个人在向它吹的时候,手里 拿着东西着火了,形成了一个大的火焰。
- 当一个男孩拿着一个瓶子着火并吹起火焰时，一群人在欢呼。
- 一个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.

10 English Descriptions:

- People are crossing the street and cars are turning at a busy intersection in a business district.
- Pedestrians attempt to cross a street at a busy intersection where construction is also taking place.
- Several people try to cross the street using a crosswalk as cars drive around a city.
- Several cars drive through an intersection as three people wait at the edge of the road to cross the street.
- People are crossing a busy street that is filled with traffic.
- Someone at a cross walk records vehicles as they drive by.
- People are standing and waiting to cross the street in a busy city.
- A busy street with car traffic and pedestrians walking at a crossing.
- A red color vehicle is taken reverse and a woman crosses the road swiftly.
- A group of people are attempting to cross a busy street.

10 Chinese Descriptions:

- 一辆白色汽车在人来人往的马路上开动，三个人正在横过斑马线。
- 一辆白色长车开过，而后一辆小车也开过，三个人站在斑马线等着过马路。
- 白色的车辆从马路上驶过，人们快速走过斑马线。
- 一群人在人行横道上躲着车过马路。
- 一个个的行人正在急匆匆的穿过马路。
- 有人在交叉行走时记录了车辆经过的过程。
- 在一个繁忙的城市里，人们站着等着过马路。
- 一条繁忙的街道与汽车交通，一部分行人走在十字路口。
- 一辆红色的车在倒车，一名女子迅速的通过了马路。
- 一群人正试图穿过一条繁忙的街道。

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.

The Boy learning to cast a bait caster.
The Boy learning to cast a bait caster.
The Boy learning to cast a bait caster.

English Captions

Human:
a young man is getting set and then throws a frisbee across an open field .

Base:
a man is throwing a frisbee in a field .

Shared Enc:
a man is throwing a frisbee across a grassy field .

Shared Enc-Dec:
a man is standing in a field and throws a frisbee into the air .

Chinese Captions

Human:
一个 男人 站 在 一个 长满 草 的 小丘 上 把 飞盘 扔出去 了 。

Base:
一个 男人 在 草地 上 扔 飞盘 ， 然后 把 它 扔 了 出去 。

Shared Enc:
一个 穿着 黑色 衣服 的 男人 在 草地 上 扔 飞盘 。

Shared Enc-Dec:
一个 男人 正在 室外 的 空地 上 拿 着 飞盘 扔 了 出去 。

English Captions

Human:
a boy is casting a fishing line into the river .

Base:
a man is standing in the water with a fishing pole .

Shared Enc:
a young boy is standing in the water and he is casting a fishing pole into the water .

Shared Enc-Dec:
a young boy is standing in the water and casting his fishing line into the water .

Chinese Captions

Human:
一个 穿着 蓝色 马甲 的 男孩 在 河边 ， 把 鱼竿 甩出去 。

Base:
一个 男人 站 在 河边 ， 手里 拿 着 鱼竿 在 钓鱼 。

Shared Enc:
一个 男人 在 河边 拿 着 鱼竿 在 钓 鱼 。

Shared Enc-Dec:
一个 穿着 蓝色 衣服 的 小男孩 在 河边 拿 着 鱼竿 钓鱼 。

English Captions

Human:
two teams of women play netball on an outdoor court in the evening .

Base:
a group of girls are playing a game of basketball on an outdoor court .

Shared Enc:
a group of women are playing a game of basketball on a court .

Shared Enc-Dec:
a group of people are playing a game of basketball on a basketball court .

Chinese Captions

Human:
两队 女子 在 室外 篮球场 进行 篮球 比赛 。

Base:
一群 人 在 室外 的 篮球场 上 进行 着 激烈 的 篮球 比赛 。

Shared Enc:
一群 人 在 一个 大 的 篮球场 上 打篮球 。

Shared Enc-Dec:
一群 人 在 篮球场 上 进行 着 激烈 的 篮球 比赛 。

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 她家 的 起居室 里 做 侧手 翻 。
NMT:
一个 年轻 女孩 在 她 的 房间 里 做 车轮 。
VMT:
一个 年轻 女孩 在 她 的 房间 里 翻筋斗 。

English:
a boy hits his head on a wall and knocks himself out .
Ground Truth:
一个 男孩 的 头 撞 在 墙上，把 自己 撞 倒 了 。
NMT:
一个 男孩 撞 他 的 头 在 墙上， 然后 敲 自己 出去 。
VMT:
一个 男孩 他 的 头 撞 在 墙上， 然后 自己 撞倒 了 。

English:
a girl shows how to apply eyeliner, describing how to use strokes .
Ground Truth:
一个 女孩 展示 了 如何 使用 眼线笔， 讲述 如何 画眉 。
NMT:
一个 女孩 展示 了 如何 使用 眼线笔， 描述 了 如何 使用 笔画 。
VMT:
一个 女孩 展示 了 如何 使用 眼线笔， 描述 了 如何 画眼线 。

Figure 6: Qualitative comparison between neural machine translation (NMT) and video-guided machine translation (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 狗 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 。

English:
a [M] is putting on a [M] performance while using large string instruments .
[M]: band, rock.
Ground Truth:
一支 乐队 正在 使用 大型 弦乐器 表演 摇滚乐 。
NMT:
一个 男人 正在 使用 大型 弦乐器 表 演 。
VMT:
一支 乐队 正在 使用 大型 弦乐器 表演 摇滚 。

English:
a child, together with mom is [M] around as they [M] in the room .
[M]: playing, laugh.
Ground Truth:
一个 孩子 和 妈妈 在 房间 里 玩 的 时候 在 笑 。
NMT:
一个 孩子 和 妈妈 在 房间 里 跳舞 同时 笑 。
VMT:
一个 孩子 和 妈妈 在 房间 里 正在 玩 同 时 笑 。

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.