CREATE: A Benchmark for Chinese Short Video Retrieval and Title Generation

Anonymous ACL submission

Abstract

Previous works of video captioning aim to objectively describe the video's actual content, lack of subjective and attractive expression, limiting its practical application scenarios. Video titling is intended to achieve this goal, but there is a lack of a proper benchmark. In this paper, we propose CREATE, the first large-scale **C**hinese sho**R**t vid**E**o retriev**A**l and **T**itle g**E**neration benchmark, to facilitate research and application in video titling and video retrieval in Chinese. CREATE consists of a high-quality labeled 210K dataset and two large-scale 3M/10M pre-training datasets, covering 51 categories, 50K+ tags, 537K manually annotated titles and captions, and 10M+ short videos. Based on CREATE, we propose a novel model ALWIG which combines video retrieval and video titling tasks to achieve the purpose of multi-modal **AL**ignment **WI**th **G**eneration with the help of video tags and GPT pre-trained model. CREATE opens new directions for facilitating future research and applications on video titling and video retrieval in the field of Chinese short videos.

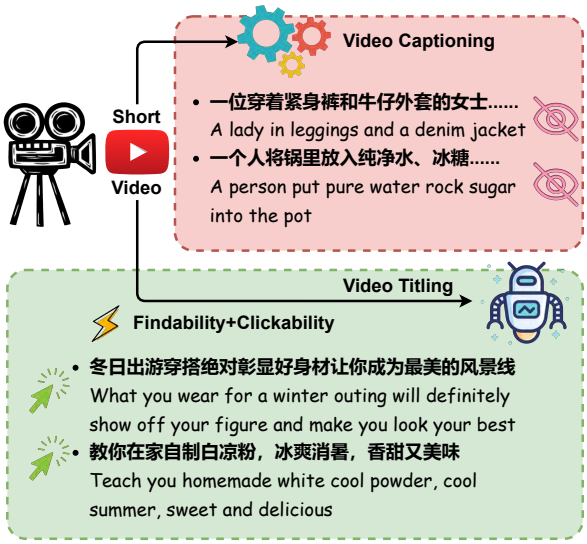


Figure 1: Imagine you are a junior video creator, you want to use the most advanced video caption generator for your video to create an attractive title, but it can only generate an objective description of the video, this is not practical. At this point, a novel video title generator can help you generate titles that have both findability and clickability simultaneously.

1 Introduction

The video captioning task is gaining increasing attention in the vision and language communities. Although many research efforts have been made in both advanced algorithms (Lei et al., 2020a; Aafaq et al., 2019; Zhang and Peng, 2019; Zhang et al., 2020c) as well as large-scale benchmarks in a variety of domains (Lei et al., 2020b; Zhou et al., 2018; Whitehead et al., 2018), there is little practical application being landed around video captioning.

The primary reason is the gap between existing benchmarks and application scenarios. Captioning intends to give an image or a video clip an appropriate title in the newspaper or other social media. However, due to the influence of existing datasets, the video captioning task has developed into an objective description of the actual content of the video without subjective factors, which is not consistent with the practical application, as shown in Figure 1.

Every successful video starts with a good video title. The two components needed for crafting the best video titles are *findability* and *clickability*. The former requires to state the primary viewpoints of the video to facilitate the text-based search, while the latter expects to add catchy expressions to hook more viewers. Therefore, an automatic video title generator with both abilities can help junior creators solve this tricky problem.

Chinese short videos play an important role in the global market, but the study of Chinese corpus is not enough. The alignment of Chinese corpus with visual content is vital for the comprehension and creation of Chinese short videos. Therefore, it is

Figure 2: **A glance at the annotations in our CREATE benchmark.** It covers 51 categories such as Lifestyle, Pet, Fashion, Gourmet, *etc.*, as well as 50K+ fine-grained tags. Each short video is annotated by an objective caption and a catchy title in an actual scenario.

necessary to pave the path for research and applications around Chinese short video titling and retrieval by establishing a new large-scale benchmark covering video titles and captions in Chinese.

To this end, we create the first **C**hinese sho**R**t vid**E**o retriev**A**l and **T**itle g**E**neration benchmark called CREATE. It contains two parts, the fine-labeled CREATE-210K and weak-labeled CREATE3M/10M. The CREATE-210K consists of 216K carefully collected short videos covering 51 categories and 15.5K tags, as illustrated in Figure 2. Each video is equipped with a high-quality title and caption to serve tasks such as video retrieval, tagging, titling and captioning. The CREATE-3M/10M are two large-scale datasets containing approximately 3M/10M videos with original titles and 53K tags. It can be used to learn vision and language alignment in the setting of weak-supervised learning through pre-training tasks.

Based on this benchmark, we propose a novel vision and language model (VLM) called ALWIG, which combines video retrieval and video titling tasks to achieve the purpose of multi-modal **AL**ignment **WI**th **G**eneration. Specifically, we utilize the tag-driven module to achieve the alignment between visual and text. We take advantage of the powerful generative capability of GPT (Radford et al., 2019) as the decoder for the textual generation. Meanwhile, we set up two popular pre-trained VLMs, *i.e.*, OSCAR (Li et al., 2020b) and UniVL (Luo et al., 2020), as baseline models. The experimental results highlight the benefits of our method.

It is worth noting that the number of videos in our CREATE-210K is 5.23 times that of VATEX (Wang et al., 2019), the largest Chinese caption dataset, and the number of annotations is 2.98 times that of T-VTD, the largest e-commerce title dataset, as shown in Table 1. Large-scale pre-training datasets CREATE-3M/10M provide more diverse training methods. In addition, the annotations are encouraged to make full reference to audio, character, speech, and other fine-grained entities, such as celebrities, locations, and popular objects in the video, to enhance the semantic representation of the model in future research.

The main contribution of this paper is three-fold:

- We establish the first large-scale benchmark CREATE for Chinese short video titling and retrieval tasks, containing over 210K fine-labeled data and 10M weak-labeled data from

Table 1: **Comparison of some relevant datasets**, the CREATE contains more open-domain videos, more annotations, and more fine-grained tag information (* indicates pre-training dataset).

Dataset	Domain	# Videos	# Sents	# Tags	Lang.	Annotation
VTW	Open	18K	18K	-	EN	Title
VATEX	Open	41.3K	826K	600	EN/CN	Caption
BFVD/FFVD	E-comm.	76K	76K	-	CN	Title
T-VTD	E-comm.	90K	180K	-	CN	Title
CREATE210K	Open	216K	537K	15,527	CN	Title/Caption
TGIF*	Open	100K	128K	-	EN	Title
HowTo100M*	Open	1.22M	136M	-	EN	ASR
WebVid-2M*	Open	2.5M	2.5M	-	EN	Title
Alivol-10M*	E-comm.	10.3M	11M	-	CN	Title
CREATE-10M*	Open	10M	10M	53,044	CN	Title

Table 2: **The splits of the whole CREATE dataset**, including 210K fine-tuning dataset and normal version 3M and large version 10M pre-training datasets (* indicates the title is added by the user).

CREATE	# Video	# Title	# Caption	# Tag
210K-train	210,493	210,493	210,493	15,527
210K-val	810	810×10	810×10	3,570
210K-test	5,000	5,000×10	5,000×10	1,191
3M-pretrain	3M	3M*	-	45,277
10M-pretrain	10M	10M*	-	53,044

51 categories and 50K+ tags with high-quality title and caption annotations.

- Based on CREATE, we introduce a novel VLM called ALWIG to address the above tasks. Our model bridges the gap between vision and language with video tags converting visual features to soft prompts and providing them to the GPT decoder for a generation. The experimental results highlight the advantages of our approach compared with other popular pre-trained models.
- We are the first to propose the task of video titling and video retrieval in the field of Chinese short videos. Our benchmark and baseline model can provide strong support for future multi-modal research and applications.

2 Related Work

2.1 Benchmarks for Video-and-Language

A large number of benchmarks have been introduced in recent years for video-and-language tasks, which cover in different filed, such as open scenario (Wang et al., 2019), movies (Lei et al., 2020b), news (Whitehead et al., 2018) and e-commercial (Lei et al., 2021), *etc*. The latest VALUE (Li et al., 2021) combines several datasets to test the performance of the model over multiple multi-modal tasks. These datasets always collect video captions annotated by a human. While these captions are valid for video captioning task, the practical applications have not yet been explored. Moreover, there are few special on Chinese corpus, except for VATEX-zh (Wang et al., 2019) and Poet (Zhang et al., 2020a). Therefore, we establish a benchmark, collect Chinese short videos, annotate high-quality annotated titles and captions for video titling and retrieval tasks.

2.2 Video-and-Language Pre-training

Thanks to some large-scale datasets, such as Howto100M (Miech et al., 2019) and WebVid (Bain et al., 2021), downstream VL tasks can be greatly improved by weakly supervised learning through narration-video or title-video pairs. According to the main structure of the model, the pre-trained model can be divided into one-stream and two-stream. One-stream models (Li et al., 2020a; Zhu and Yang, 2020) always design various proxy tasks, fusion multi-modals through a transformer-based model, and adapt to discriminative or generative tasks simultaneously. Two-stream models (Miech et al., 2020; Luo et al., 2021; Bain et al., 2021) leverage two separate backbones and contrastive learning to align visual and text. Our model combines the advantages of both to take alignment efficiently through the two-stream model and to generate accurately through the one-stream model.

2.3 Video Titling and Video Retrieval Tasks

The earliest work is VTW (Zeng et al., 2016), which collects video titles in the wild and combines highlight detection and title at the same time. However, the performance is not satisfactory due to the amount of data and the lack of pre-training techniques. The most relevant works are (Zhang et al., 2020b,a; Lei et al., 2021). They collect videos from Taobao and annotate titles in Chinese. Although good results have been achieved, the videos are limited to the e-commerce field. For the retrieval task, while some refined approaches (Zhang et al., 2020c, 2021) have been developed and remarkable progress has been made, there are still limitations in the efficiency for practical application.

3 CREATE Benchmark

3.1 Dataset Collection

As mentioned above, the bottleneck of the practical application of Chinese short video titling and

Table 3: **The performance of proposed simple two-steam video-text matching scorer with a variety of backbones.** We try to pre-train these models under different corpus end-to-end. The experimental results on the VATEX public-test show that using the ViT-BERT model pre-trained on the video-title of Chinese short videos can help to learn better alignment efficiently, which is used to filter out bad videos for CREATE datasets.

#	Model	Pre-trained Dataset	Finetune	Text-Video Retrieval			Video-Text Retrieval		
				R@1	R@5	R@10	R@1	R@5	R@10
1	S3DG-BERT	-	✓	1.7	15.5	8.6	5.8	15.4	21.7
2		HowTo100M-CN	×	4.2	13.2	19.3	4.4	16.1	23.7
3		HowTo100M-CN	✓	7.7	21.8	30.5	15.4	9.6	53.9
4		Random-10M	×	6.8	20.1	29.5	11.8	30.1	41.9
5		Random-10M	✓	19.6	47.7	61.7	31.2	60.4	72.1
6	TimesFormer -BERT	Random-10M	×	15.3	37.5	49.3	31.7	62.0	74.3
7		Random-10M	✓	**43.1**	**77.1**	**86.9**	60.8	88.0	94.0
8	ViT-BERT	Random-10M	×	18.8	44.0	56.4	37.2	66.6	77.5
9		Random-10M	✓	41.1	75.3	85.2	**64.2**	**89.6**	**94.3**

retrieval is the lack of appropriate dataset. In addition, it is effective to improve the performance of downstream tasks through pre-training without increasing annotations. Therefore, we construct a high-quality labeled dataset CREATE-210K and two large-scale weak-labeled datasets CREATE-3M/10M from the Tencent video platform.

High-quality Labeled CREATE-210K. We start by building a video tagging system that contains 51 categories and over 50K video tags for a wide coverage of short video content. Each video always has one category and several tags, which can be seen as the coarse-grained and fine-grained classifications of the video. Then we follow the principle of collecting at least five videos per tag to ensure enough data for training in different domains. We do not average sample videos according to the video categories since some categories, *e.g.*, military and financial are not common. Filtered by video tagging system, we try to avoid including some special tags, *e.g.*, film or television variety shows, since understanding these videos requires lots of additional information, such as stars and plots, which is difficult to annotate and model.

The distributions of tags and categories are illustrate in Figure 6. Sorted by the number of video categories and tags, the video contents mainly focus on "*people daily life*", "*instruction video*" and "*animals show*", *etc*. The long-tail problem is inevitable because tags are hierarchical, and some tags are subsets of larger concept tags, *e.g.*, "*monk parrot*" belongs to "*cute animals*". We limit the video time to less than 60 seconds and eventually collected over 210K short videos for next step annotations.

More than 500 workers are involved in annotation tasks to ensure data diversity. To obtain high-quality annotations, each worker has undergone rigorous training and testing, with a clear definition of the difference between video titling and captioning. Moreover, workers are provided categories and tags of each as hints and required to use information as much as possible. A word such as "*things*" should be replaced with specific objects since this could increase diversity and avoid general annotation. The word limits for video titles and captions are 15∼30 and 25∼50, respectively. Finally, it takes half a year to collect and check more than 537k annotations. More details about annotation rules and interface are shown in the appendix.

Large-scale Weak-labeled CREATE-3M/10M. To increase the generalization of the model and the extensibility of the tasks, we establish two large-scale weak-labeled datasets CREATE-3M/10M. Each video has its category, several tags and an original title. The noise within videos and titles is the most serious problem, *e.g.*, some videos or titles are of poor quality, and some video-title pairs are mismatched. In order to filter out these low-quality videos, we designed an efficient automatic filter to determine the consistency of videos and titles inspired by the work (Miech et al., 2020; Bertasius et al., 2021; Luo et al., 2021). As shown in Figure 5.a, a two-stream model leverages a visual encoder and a textual encoder to extract visual and textual features separately in an end-to-end manner. The contrastive learning is leveraged to push mismatched and pull matched features, achieving

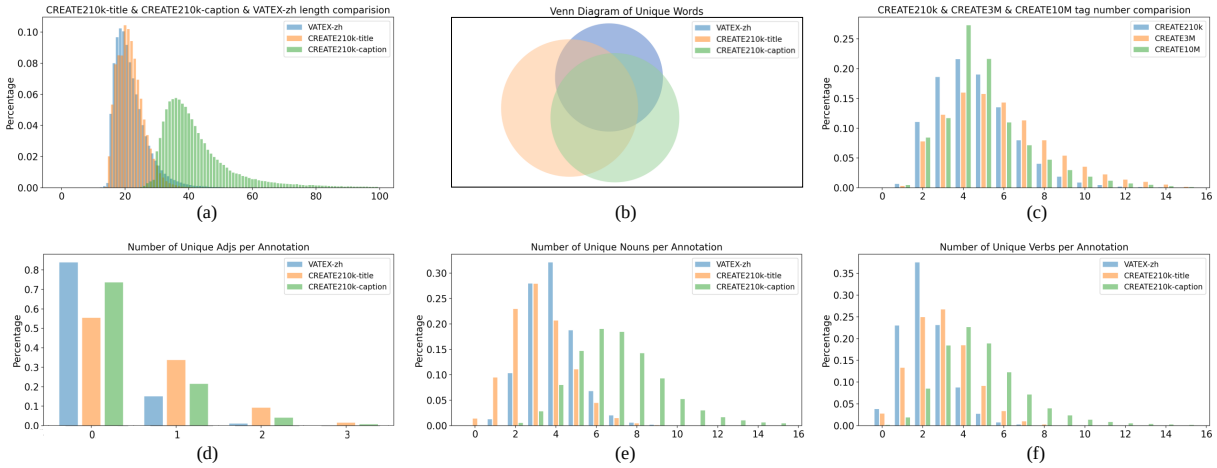


Figure 3: **Some statistics on the datasets** indicate our datasets and annotations have better diversity. (a) indicates the distribution of the annotation length in three datasets. (b) indicates the inclusion relation of unique words by Venn Diagram in three datasets. (c) represents the distribution of the number of tags. (d)-(f) shows the distribution on three part-of-speech of unique words.

alignment between both modalities.

We conduct comparison experiments on three backbones, *i.e.*, S3DG-BERT, TimeSformer-BERT and ViT-BERT, to verify which is the most suitable backbones for feature extraction, as illustrated in Table 3. We first evaluate the S3DG-BERT on the translated HowTo100M and randomly collected 10M video-title datasets following the same setting as the previous work [1]. It shows that learning with video-title pairs can achieve better alignment than video-narrations in instructional videos[2], as shown in Table 3.Line2,4. Besides, we evaluate the transformer-based models, *i.e.* TimeSformer-BERT and ViT-BERT. Compared with 3D-CNN based model, the transformer-base model can better support large-scale data during pre-training, as shown in Table 3,Line4,6,8. Moreover, compared with spatial and temporal attentions in TimeSformer, we conduct average pooling in the temporal dimension over 8 frames. Although the performance is slightly reduced, it is more efficient for calculating matching scores and extracting visual features. Eventually, we choose ViT-BERT model pre-trained on random-10M videos as the video-title matching scorer. Videos with matching scores of less than 0.3 are filtered out, and most of the remaining videos form the final pre-trained dataset, *i.e.*, the normal version CREATE-3M and large version CREATE-10M. The normal version is more convenient for algorithm iteration.

3.2 Statistics of the CREATE dataset

We analyse the CREATE dataset in terms of video information and annotations. As illustrated in Table 2, we collected a total of 210,493 videos for training, with one title and caption annotated, 810 videos for validation, and 5,000 videos for testing, each video has 10 titles and captions. For the weak-labeled dataset, we filter out 3M and 10M videos with their original titles. The average video length is around 30 seconds. The distributions of annotation length are illustrated in Figure 3.a. The average title or caption lengths of VATEX-zh, CREATE210K-caption, CREATE210K-title, CREATE3M and CREATE10M are 22.45, 43.54, 21.71, 20.03 and 23.96. The distributions of tag numbers are illustrated in Figure 3.c. The average number of tags within CREATE210K, CREATE3M and CREATE10M are 4.65, 5.73 and 5.03.

In addition to basic information, we pay more attention to the richness of content covered by the annotations. We use the Venn Diagram to depict the approximate inclusion relation of unique words in VATEX-zh, CREATE210k-title and CREATE210k-caption, as shown in Figure 3.b. Our CREATE210k-caption covers 72.85% of the vocabulary of VATEX-zh, and 62.68% of the unique words do not appear in VATEX-zh. Moreover, we analyze parts of speech(POS) of annotations, *i.e.*, of the above three datasets. As shown in Figure 3.d-f, as can be seen from the distributions of the three POS, our datasets contain more

[1]https://github.com/antoine77340/MIL-NCE_HowTo100M

[2]Note: This does not exclude the reason for the increased noise introduced by translation.

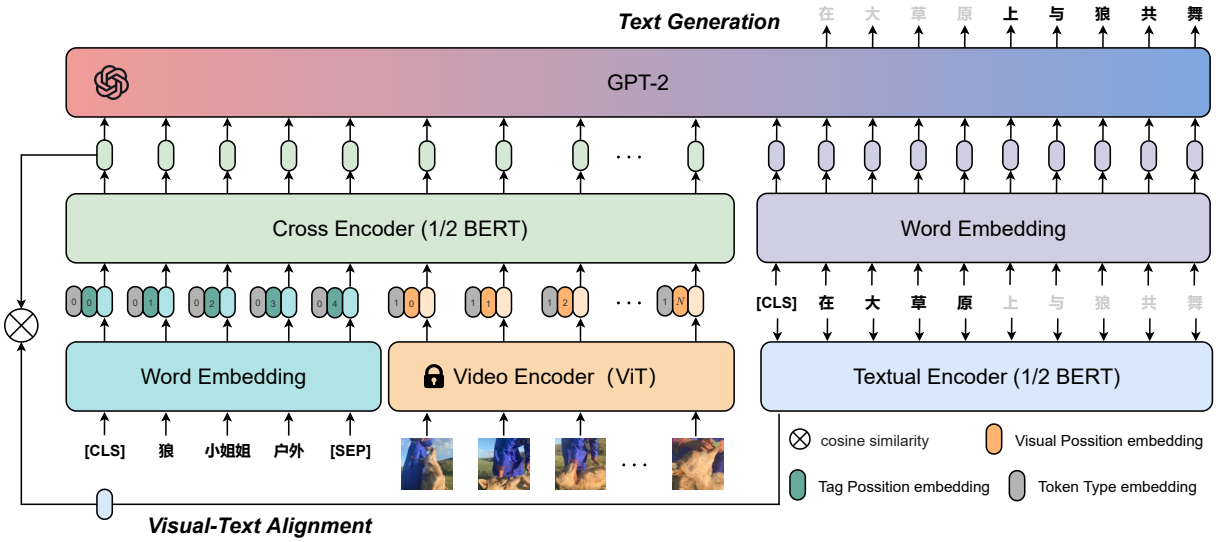


Figure 4: **Overall framework of our proposed ALWIG model.** ALWIG consists of a tag-driven video-text alignment module and a GPT-based generation module for video titling and retrieval tasks.

information in each annotation.

4 ALWIG Method

ALWIG consists of a tag-driven video-text alignment module and a GPT-based generation module for video titling and retrieval tasks, as shown in Figure 4. We use a 12-layer transformers ViT-B/16 as the video feature extractor, and initialized it with the weights from CLIP model [3]. The video clip is encoded into a sequence of N video features $V = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$. We get M tag embeddings $O = \{\boldsymbol{o}_{\text{cls}}, \boldsymbol{o}_1, \cdots, \boldsymbol{o}_M, \boldsymbol{o}_{\text{sep}}\}$ using the word embedding in BERT.

Tag-driven video-text alignment module. We concatenate tag embeddings and video features into $\{\boldsymbol{o}_{\text{cls}}, \boldsymbol{o}_1, \cdots, \boldsymbol{o}_M, \boldsymbol{o}_{\text{sep}}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$, where $\boldsymbol{o}_{\text{cls}}$ and $\boldsymbol{o}_{\text{sep}}$ are embeddings of [CLS] and [SEP] tokens. Furthermore, we use two independent 6-layer transformers as the cross-encoder and the textual-encoder. Both encoders are initialized with the first six layers transformer of Bert model [4]. The cross-encoder is utilized to integrate video features and tag embeddings into fusion embeddings $F = \{\boldsymbol{f}_{\text{cls}}, \cdots, \boldsymbol{f}_{\text{M+N}}\}$, where $\boldsymbol{f}_{\text{cls}}$ can be regarded as the fused video-tag representation driven by the tags. Textual encoder embeds a text input T into a sequence of L token embeddings $W = \{\boldsymbol{w}_{\text{cls}}, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_L\}$, where $\boldsymbol{w}_{\text{cls}}$ represents the whole representation of textual embedding. The video-text alignment is to learn a similar function $s(F, W) = \phi(\boldsymbol{f}_{\text{cls}})^{\text{T}}\psi(\boldsymbol{w}_{\text{cls}})$ between visual and textual embeddings via contrastive learning, where $\phi(\cdot)$ and $\psi(\cdot)$ are the linear functions with normalization to map each embedding into the common semantic space. We follow the infoNCE loss function as shown in Equation 1, where τ is a learnable temperature coefficient, and W_+ and F_+ are positive samples within batch.

$$\mathcal{L}_{align} = -\log \frac{\exp\left(s(F, W_+)/\tau\right)}{\sum_{i=1}^{K} \exp\left(s(F, W_i)/\tau\right)} - \log \frac{\exp\left(s(W, F_+)/\tau\right)}{\sum_{i=1}^{K} \exp\left(s(W, F_i)/\tau\right)}. \quad (1)$$

GPT-based generation module. One of the downstream tasks we are most interested in is video titling. It requires to express not only the general meaning of the video but also subjective expressions to attract the audiences' interests. Therefore, we leverage the power of GPT as the decoder to help introduce external linguistic knowledge to reduce the difficulty of textual generation and improve the generalization of the model (Zhang et al., 2020c). The input of the decoder is the fusion embeddings F mentioned above, and the output is the ground-truth text. We utilize the typical auto-regressive training method to train the model following cross-entropy loss as shown in Equation 2, where F can be seen as a soft-prompts for the generation.

[3] https://github.com/openai/CLIP
[4] https://huggingface.co/bert-base-chinese

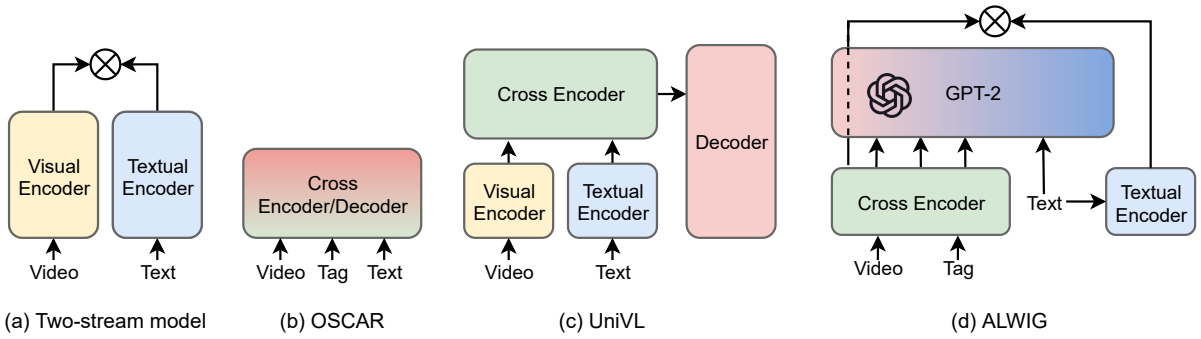


Figure 5: **The mean structure of the models.** (a) indicates the two-stream model, which we used as a video-text matching scorer for filtering the 10M pre-training dataset. The model is trained in an end-to-end manner. (b) shows the OSCAR model, which is a simple cross encoder/decoder structure. We replace the object features with frames features as input. (c) shows the UniVL model, a typical encoder and decoder structure, and it is a popular video-text pre-trained model. (d) indicates our ALWIG model, which leverages the tag-driven fusion via contrastive learning to achieve alignment, and the power of GPT-2 to achieve generation.

$$\mathcal{L}_{gen} = -\sum_{l=1}^{L} \log p_\theta(T_l | T_{<l}, F). \quad (2)$$

In summary, we utilize a tag-driven cross-encoder with the help of contrastive learning to align the modalities and take advantage of a GPT-based decoder to generate text. The full pre-training objective of ALWIG is:

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{gen}. \quad (3)$$

5 Experiments

5.1 Baseline Models

We benchmark several representative vision and language pre-trained models, *i.e.*, OSCAR and UniVL, and our ALWIG on the proposed CREATE dataset. Both models are pre-trained on the large-scale pre-training dataset through multiple proxy tasks, such as mask tokens prediction, contrastive learning on visual-textual pairs, *etc*, then finetuned on many downstream tasks, such as cross-modal retrieval, captioning, VQA, *etc*. In the model structure, both models adopt the general transformer-based structure.

As shown in Figure 5.b, OSCAR leverages an integrated encoder-decoder to fuse visual and textual features from beginning to end. It controls generation or discriminative tasks by setting different types of masks. Moreover, the most instructive thing in OSCAR is using object tags as anchor points to align the image and language modalities in a shared semantic space.

Instead, UniVL is a flexible model for most of the multimodal downstream tasks considering both efficiency and effectiveness, as illustrated in Figure 5.c. It utilizes two independent encoders to enhance the representation of each modal at the beginning and leverages cross-encoder to fuse each other. Besides, a separate encoder-decoder can explicitly handle generation tasks, which is more flexible.

5.2 Experimental Setting

Our model consists of two half a $\text{BERT}_{\text{base}}$ with 123.7M parameters and a GPT_{base} with 154.5M parameters. We pre-train the model for 30 epochs using a batch size of 32 on 48 NVIDIA A100 GPUs. We use the AdamW optimizer with a weight decay of 0.02. The learning rate is warmed-up to $1e^{-5}$ in the first 10 epochs and decayed to $1e^{-6}$ following a cosine schedule. Before training, we extract video features 1fps via ViT pre-trained by CLIP. The dimension of the hidden state in BERT and GPT is 768, and the output features of the two-stream structure are mapped to 512. We utilize beam-search (beam size=3) for the generation.

We leverage the standard captioning metrics, *i.e*, BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015) and Rouge-L (Lin, 2004), to measure the performance of video titling and captioning. We split the words using Jieba Chinese text segmentation [5], and we count 4 continuous words as 4-grams. It can be noted that we are not using Meteor as the metric since Mereor (Denkowski and Lavie, 2014) defaults to calculate the relationship of English synonyms but does not contain the Chi-

[5] https://github.com/fxsjy/jieba

Table 4: The experimental results of UniVL, OSCAR and proposed ALWIG. All the models are pre-trained on two large-scale weak-labeled dataset (3M=CREATE3M, 10M=CREATE10M) and fine-labeled on CREATE210K for three tasks: video retrieval, titling and captioning. The bottom three lines represent the ablation studies.

#	Model	Pre-trained Dataset	Task1: Video Retrieval		Task2: Video Titleing			Task3: Video Captioning		
			T2V Recall@1/5/10	V2T Recall@1/5/10	CIDEr	BLEU-4	Rouge-L	CIDEr	BLEU-4	Rouge-L
1	UniVL	3M	59.3 / 83.9 / 90.4	73.6 / 90.7 / 94.8	13.3	6.4	26.1	18.9	14.0	33.2
2		10M	61.7 /85.1 / 91.3	76.8 / 92.1 / 95.9	13.8	6.9	26.5	22.9	14.5	33.4
3	OSCAR	3M	61.3 / 84.6 / 90.7	74.9 / 91.5 / 95.1	35.8	8.4	29.8	34.7	14.2	33.4
4		10M	62.1 / 85.5 / 91.3	75.2 / 91.5 / 95.5	**36.3**	9.0	30.7	35.2	15.2	33.8
5	ALWIG	3M	60.7 / 85.0 / 91.0	75.3 / 91.9 / 96.0	35.5	9.3	31.0	32.2	14.9	34.6
6		10M	**65.6 / 87.7 / 92.7**	**79.3 / 93.9 / 96.8**	36.1	**9.7**	**31.6**	**35.9**	**16.3**	**35.5**
	Ablation Study									
7	Baseline	3M	**60.7 / 85.0 / 91.0**	**75.3 / 91.9 / 96.0**	**35.5**	**9.3**	**31.0**	**32.2**	**14.9**	**34.6**
8	w/o Tag	3M	51.7 / 79.0 / 86.8	67.1 / 87.5 / 92.7	15.6	6.5	27.2	13.9	12.0	31.9
9	w/o GPT	3M	58.7 / 83.5 / 90.1	72.2 / 90.2 / 94.8	26.7	6.1	28.1	29.1	12.8	34.3
10	w/o pretrain	-	43.0 / 72.3 / 81.8	56.6 / 80.9 / 88.2	21.4	6.1	28.1	23.8	13.9	33.7

nese thesaurus. Furthermore, we utilize metrics Recall at K (Recall@K) to measure video-text retrieval performance. R@K measures the proportion of correct targets retrieved from K samples.

5.3 Results and Analysis

We adapt three pre-trained models *i.e.*, UniVL, OSCAR and ALWIG to three tasks, as shown in 4. Three models are all pre-trained on CREATE3M/10M with their best performances. Compared with UniVL, our model has significantly improved in all three tasks, especially the performance of ALWIG improve by 62.5% on CIDEr in the video titling task, indicating that many novel words have been generated, which is attributed to the use of tags through tag-driven fusion module. Compared to the automatically detected tags in OSCAR, the manually collected tags in our work are more rich and high-quality. Instead of focusing on how to get these tags, we try to use them directly as a bridge for the alignment between multiple modalities in pre-training or external knowledge for downstream tasks. Compared with OSCAR on retrieval task, our model is superior, which also illustrates the validity of the GPT model

In addition, to illustrate the effectiveness of our proposed two modules in more detail, we conducted multiple ablation experiments. When the tag-driven fusion module is removed, the performance of CIDEr is significantly reduced by 56% and 56.8% respectively on the video titling and captioning tasks, and 14.8% reduces Recall@1 on the text-video retrieval task. It is worth noting that the decline of other metrics in the video description task is not particularly severe, which indicates that it has a small impact on the overall sentence pattern and also reflects that the description task pays more attention to the main content of the video rather than some novel expression. When the GPT module is removed, the retrieval performance slightly decreases, while the BLEU-4 of the two generation tasks decreases by 34.4% and 14.1% respectively, indicating that GPT is of great help to the learning of basic sentence patterns and can reduce the difficulty of generation. Meanwhile, we also experiment with the model without pre-training, and all indicators are far lower than the pre-training model's, indicating that the pre-training and fine-tuning paradigm can effectively improve the model's performance.

6 Conclusion

In this paper, we establish the first Chinese short video retrieval and title generation benchmark, *i.e.*, CREATE to facilitate research and application for video retrieval and titling tasks. The CREATE contains a high-quality fine-labeled dataset and two large-scale pre-training datasets. A large number of statistics indicate that our datasets have richer visual content and annotations. Based on CREATE, we propose a novel model ALWIG to better accomplish the video retrieval and generation tasks with the power of tag-driven fusion and the GPT model. Extensive experiments verify the validity of our model and provide some good baselines for future research. The rules-based metrics seem unable to reflect the quality of the video titling model for findability and clickability, but the 10-annotations per video alleviates this problem to some extent, and the learnable metrics are left for the future.

References

Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. 2019. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12487–12496. Computer Vision Foundation / IEEE.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. *CoRR*, abs/2104.00650.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 376–380. The Association for Computer Linguistics.

Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. 2021. Understanding chinese video and language via contrastive multimodal pre-training. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 2567–2576. ACM.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L. Berg, and Mohit Bansal. 2020a. MART: memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2603–2614. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020b. TVR: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, volume 12366 of *Lecture Notes in Computer Science*, pages 447–463. Springer.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020a. HERO: hierarchical encoder for video+language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2046–2065. Association for Computational Linguistics.

Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, Tamara Lee Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. 2021. VALUE: A multi-task benchmark for video-and-language understanding evaluation. *CoRR*, abs/2106.04632.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.

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

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *CoRR*, abs/2002.06353.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of CLIP for end to end video clip retrieval. *CoRR*, abs/2104.08860.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9876–9886. Computer Vision Foundation / IEEE.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International*

Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 4580–4590. IEEE.

Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare R. Voss. 2018. Incorporating background knowledge into video description generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3992–4001. Association for Computational Linguistics.

Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. 2016. Title generation for user generated videos. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 609–625. Springer.

Junchao Zhang and Yuxin Peng. 2019. Object-aware aggregation with bidirectional temporal graph for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8327–8336. Computer Vision Foundation / IEEE.

Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020a. Poet: Product-oriented video captioner for e-commerce. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1292–1301. ACM.

Shengyu Zhang, Ziqi Tan, Zhou Zhao, Jin Yu, Kun Kuang, Tan Jiang, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020b. Comprehensive information integration modeling framework for video titling. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 2744–2754. ACM.

Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. 2021. Open-book video captioning with retrieve-copy-generate network. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9837–9846. Computer Vision Foundation / IEEE.

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020c. Object relational graph with teacher-recommended learning for video captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13275–13285. Computer Vision Foundation / IEEE.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8743–8752. Computer Vision Foundation / IEEE.

A Annotation Details

In this section, we introduce our annotation details for CREATE-210K to deepen the user's understanding of the dataset. This introduction is in the actual labeling process requires the workers to read in advance, and after a multi-questions test to meet the standards to enter the formal labeling link.

A.1 Video Captioning Details

1. **About the degree of labeling.** Highlight the key elements of the video, such as the key event, people and objects. The general description should be avoided.

2. **About the objectivity.** Describe the content of the video objectively without mixing any personal emotions and comments, such as "*that's awesome*", "*it's really*", "*it/that looks*".

3. **About the source of information.** Try to describe what you see. Narration, subtitles, and conversations are only used as supporting information. Do not describe background music.

4. **About the usage of tags.** Use the tags provided as much as possible. Refer to the tags for some objects if you don't recognize them. Avoid general meaningless expressions such as "*something*", "*object*", "*liquid*". Fine-grained tags of objects should be used, *e.g.*, tags provide "*agates*" that cannot simply be described as "*stone*".

5. **About stars and varieties.** Should be specifically stated the name of the person and the name of the variety, if the label has relevant tips.

6. **About wearing description.** If wearing is not the focus of the video please do not describe wearing throughout, while avoiding the use of templates, such as a large number of the expressions with "*A person wearing*", will be considered invalid labels.

7. **Describe directly.** The descriptions like "*I/We can see*", "*the video is the description of*", "*a person is facing the camera/phone*" are not available.

8. **About hands.** In some videos, there is no person but a pair of hands instead, such as cooking, please infer the gender or occupation of the person, do not say what one hand is doing.

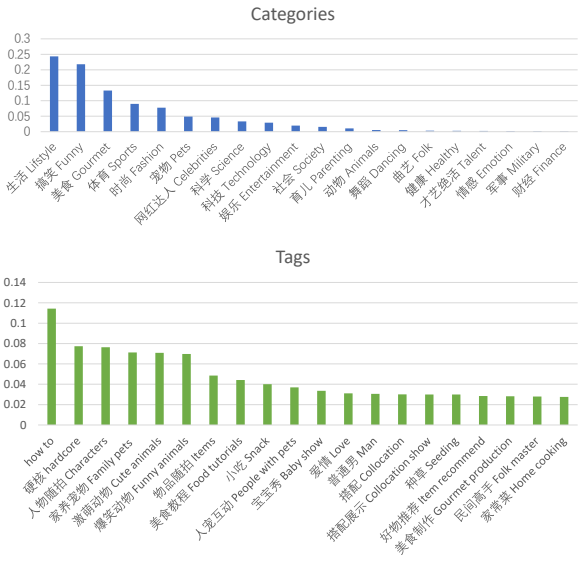


Figure 6: The distributions of categories and tags in the CREATE dataset.

A.2 Video Titling Details

1. **Reject the clickbaits.** Avoid exaggerate content, such as "*Shock!*", no actual content "*LOL*", and adult content.

2. **About attraction.** Attraction needs express interesting content through the title, not touts, such as "*Take a look at it!*"

3. **About plagiarism.** The title can not be copied, must be original. The use of video audio or subtitles in some words is acceptable, but not more than 80% of the original video content.

A.3 The interface of annotation

In this section, we demonstrate the interface during the labelling process. Workers are provided with the video content, original video title and video tags to help to label. At the same time as labelling, it is necessary to check which tags are used to force workers to use them as much as possible, thus improving the quality of labelling.

B Result Details

In this section, we demonstrate some cases sampled from the result on video retrieval, titling and captioning tasks, as shown in Figure 8, 9 and 10.

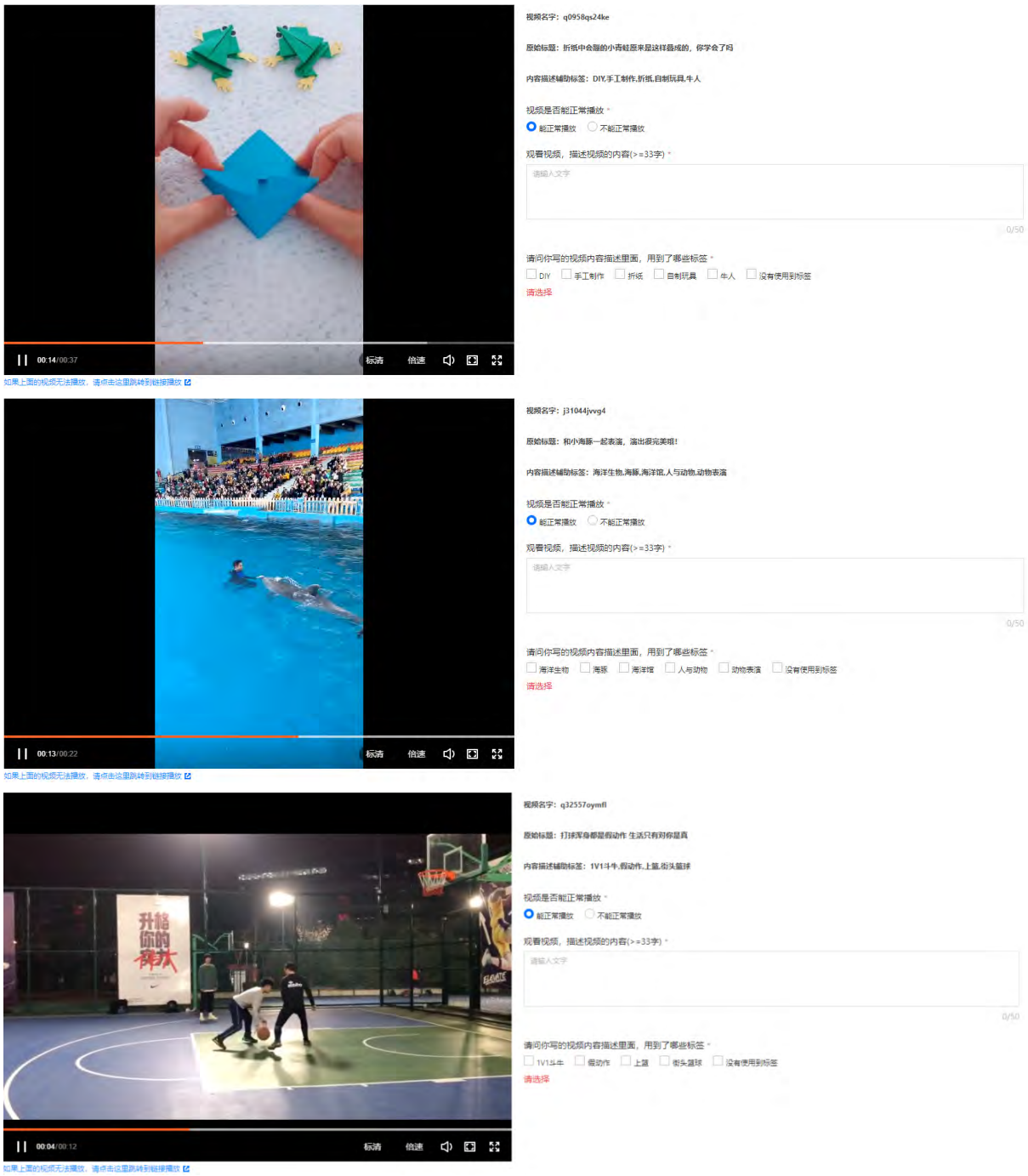


Figure 7: The demonstration of the interface during the actual labeling process.

房间的地上摆满了软垫，一个穿着印花毛衣运动鞋的男子动作娴熟的连续做了两个空翻。

The floor of the room is covered with cushions, and a man wearing a printed sweater and sneakers skilfully performed two somersaults in a row.

西装革履的绅士大哥跳搞笑舞蹈摩托摇，双臂上下摆动，一脚踢踏一脚旋转，时而抖肩，时而握摩托车把的架势。

The elder gentleman in a suit and leather shoes dances a funny dance pretending on a motorcycle, swings his arms up and down, kicks and spins, sometimes shakes his shoulders, and sometimes holds the motorcycle handlebar.

夜市市场里，并排有很多的小屋，小屋里都在卖吃的，有很多人正在这里逛着。

In the night market, there are many huts side by side. The huts are selling food, and many people are shopping here.

Figure 8: The visualization of text-to-video retrieval.

Caption

工地上，一个身穿白色衣服的男子，手里拿着一把铲子，在砌着墙。

On the construction site, a man in white clothes is laying a wall with a shovel in his hand.

Title

工地上的砌墙高手，一天能砌多少墙呢?

How many walls can be built in a day by the master wall builder on the construction site?

Caption

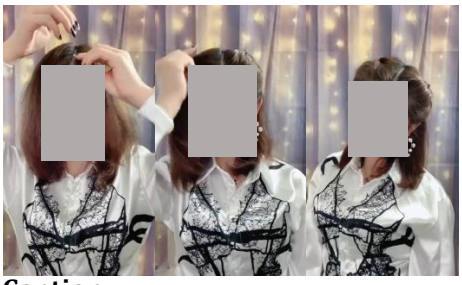
鲫鱼汤的做法，先将鲫鱼清洗干净，放入锅中煎至两面金黄，然后加入葱姜料酒和水，盖上锅盖焖煮，最后撒上葱花即可。

For crucian carp soup, first clean the crucian carp, put it in a pan and fry it until golden on both sides, then add green onion ginger cooking wine and water, cover the pot and simmer, and finally sprinkle with chopped green onion.

Title

奶白色的鲫鱼汤，营养又美味，你学会了吗?

The milky white crucian carp soup is nutritious and delicious. Have you learned it?



Caption

一名身穿白色上衣的女子，用手将自己的头发扎成了一个古风发型。

A woman wearing a white shirt tied her hair into an antique hairstyle with her hands.

Title

小姐姐教你一款简单易学的古装发型，你学会了吗?

Miss sister teaches you a simple and easy-to-learn period hairstyle. Have you learned it?

Figure 9: The generated video captions and titles.

在鱼塘里，一位年轻的钓客扯高钓鱼杆，一条大鱼在水里翻滚激起水花，旁边黑衣男子拿起网兜在捕捞。

Beside the fishpond, a young angler pulls a fishing rod, a big fish rolls in the water and splash, and a man in black picks up a net to catch it.

鱼塘边有好多人在钓鱼，男人钓着一条鱼，拉不上来，另一个男人拿着渔网捞鱼。

There are many people fishing by the fishpond. The man catch a fish and couldn't pull it up. Another man take a fishing net to catch the fish.

鱼塘边，几个钓客在钓鱼，其中一个男人钓到了一条大鱼，另一个男人用抄网想帮他把鱼捞起来。

By the fishpond, several anglers are fishing. One of the men catch a big fish, and the other man use a dip net to help him catch the fish.

一位农村老人打开汉堡的纸袋，把里面的汉堡递给了坐在一旁的老伴，老伴接过并一起吃了起来。

An old rural man opens the paper bag of hamburgers and hands the hamburger to his wife who is sitting aside. His wife takes it and eats it together.

老大爷从腿上的纸包里拿出一个汉堡递给坐在旁边的大娘 然后两个老人一起吃起了汉堡。

The old man takes out a hamburger from the paper bag on his lap and hands it to the aunt sitting next to him. Then the two old men eat the hamburger together.

一对老人坐在庭院的木头堆上，老爷爷拆开汉堡后第一时间就递给老奶奶，后来他们一起吃汉堡。

A pair of old people sit on a pile of wood in the courtyard. The grandfather hand the burger to the grandmother as soon as he opens the burger, and then they eat the burger together.

一个印度大叔蹲在地上干活，旁边有一个火炉，他从桶里舀了一些水然后洒在模具里。

An Indian man squats on the ground to work. There is a stove next to him. He scoops some water from the bucket and sprinkles it in the mold.

一个印度大叔蹲在泥巴前，他手上拿着一个锅用工具敲了敲，然后从一旁的桶里面舀出泥水用手洒在锅上。

An Indian man squats in front of the mud. He holds a pot in his hand and taps it with a tool. Then he scoops out the mud from the bucket on the side and sprinkles it on the pot by hand.

一位印度小伙蹲在地上，右手拿着夹子夹住不锈钢锅，左手拿着金属条在锅上敲打了几下，又在桶内拿出罐子。

An Indian guy squats on the ground, holding a clamp in his right hand to clamp a stainless steel pot, a metal strip in his left hand and banging on the pot a few times, then takes out the pot from the barrel

Figure 10: The visualization of video-to-text retrieval.