

# ChinaOpen: A Dataset for Open-world Multimodal Learning

Aozhu Chen  
School of Information, Renmin  
University of China  
Beijing, China

Ziyuan Wang  
School of Information, Renmin  
University of China  
Beijing, China

Chengbo Dong  
School of Information, Renmin  
University of China  
Beijing, China

Kaibin Tian  
School of Information, Renmin  
University of China  
Beijing, China

Ruixiang Zhao  
School of Information, Renmin  
University of China  
Beijing, China

Xun Liang  
School of Information, Renmin  
University of China  
Beijing, China

Zhanhui Kang  
Tencent  
Shenzhen, China

Xirong Li\*  
MoE Key Lab of DEKE, Renmin  
University of China  
Beijing, China

## ABSTRACT

This paper introduces *ChinaOpen*, a dataset sourced from *Bilibili*, a popular Chinese video-sharing website, for open-world multimodal learning. While the state-of-the-art multimodal learning networks have shown impressive performance in automated video annotation and cross-modal video retrieval, their training and evaluation are primarily conducted on YouTube videos with English text. Their effectiveness on Chinese data remains to be verified. In order to support multimodal learning in the new context, we construct *ChinaOpen-50k*, a webly annotated training set of 50k Bilibili videos associated with user-generated titles and tags. Both text-based and content-based data cleaning are performed to remove low-quality videos in advance. For a multi-faceted evaluation, we build *ChinaOpen-1k*, a manually labeled test set of 1k videos. Each test video is accompanied with a manually checked user title and a manually written caption. Besides, each video is manually tagged to describe objects / actions / scenes shown in the visual content. The original user tags are also manually checked. Moreover, with all the Chinese text translated into English, *ChinaOpen-1k* is also suited for evaluating models trained on English data. In addition to *ChinaOpen*, we propose Generative Video-to-text Transformer (GVT) for Chinese video captioning. We conduct an extensive evaluation of the state-of-the-art single-task / multi-task models on the new dataset, resulting in a number of novel findings and insights.

## CCS CONCEPTS

• Information systems → Multimedia databases; • Computing methodologies → Visual content-based indexing and retrieval.

## KEYWORDS

Chinese video dataset, multimodal learning, multi-task evaluation

## ACM Reference Format:

Aozhu Chen, Ziyuan Wang, Chengbo Dong, Kaibin Tian, Ruixiang Zhao, Xun Liang, Zhanhui Kang, and Xirong Li. 2023. ChinaOpen: A Dataset for Open-world Multimodal Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3581783.3612156>

## 1 INTRODUCTION

Online short videos, typically tens of seconds to several minutes in length, are playing an increasingly important role in information dissemination on various social media platforms and video-sharing websites. Created / re-edited by individual amateurs or professional groups, these videos cover a wide range of topics from entertainment and humor to educational and informative content in a very broad domain. For visual content-based indexing and retrieval, an open world with uncontrolled content has emerged.

While the state-of-the-art multimodal learning networks have shown impressive performance in automated video annotation [1, 35] and cross-modal video retrieval [10, 26, 27], their training and evaluation are primarily conducted on YouTube videos with English text. To what extent can these models generalize to Chinese data remains open. We aim to fill the gap with *ChinaOpen*, a new video dataset for open-world multimodal learning and evaluation.

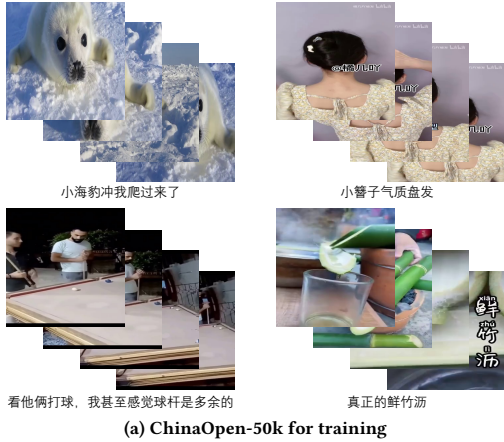
As exemplified in Fig. 1, *ChinaOpen* consists of two subsets: *ChinaOpen-50k* and *ChinaOpen-1k*. The former is a set of 50k highly selected videos with user-generated titles (and other meta data). The latter consists of 1k videos with manually checked user-generated titles / tags, manually written captions, and manual labels describing visual objects / actions / scenes present in the video content.

\*Corresponding author: Xirong Li (xirong@ruc.edu.cn)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-0108-5/23/10...\$15.00  
<https://doi.org/10.1145/3581783.3612156>



**Figure 1: Visual examples and annotations from ChinaOpen: (a) ChinaOpen-50k, a selected user-titled video set for multimodal learning and (b) ChinaOpen-1k, a manually-annotated video set for evaluating multimodal models.**

Due to the rapidly increasing need of training large video-language models, several webly-annotated video datasets have been developed, *e.g.* HowTo100M [28] and WebVid [2]. HowTo100M consists of narrative video clips collected from YouTube, with transcribed text as their annotations. WebVid was sourced from Shutterstock with professionally edited descriptions. Both datasets were from English websites with English text. By contrast, ChinaOpen was sourced from Bilibili<sup>1</sup>, a leading video-sharing website in China, with about 90 million daily active users. ChinaOpen is thus unique.

Existing manually annotated video datasets are either label based, originally developed for video classification [6, 12], or caption based [37, 38] without manual label. ChinaOpen-1k is unique, as each video has manually labeled Chinese tags that explicitly describe

objects, actions and scenes shown in the video content. Compared to existing datasets, ChinaOpen-1k has a number of novel labels related to objects, actions and scenes, see Table 1. Moreover, the video is also accompanied with a (manually checked) user-generated title and a manually written content-based caption. We argue that the title of a given video shall be sourced from its uploader as this specific user knows the context of the video and is thus in a good position to write an eye-catching title.

**Table 1: ChinaOpen-1k vs public datasets in terms of their label sets. Novel labels are those unique in ChinaOpen-1k.**

Dataset	Modality	Labels in common	Novel labels
<b>Objects</b>			
VidOR [32]	vid	43	537
MSCOCO [24]	img	45	535
Objects365 [33]	img	131	449
OpenImages [16]	img	188	392
HVU [6]	vid	226	354
LVIS [8]	img	229	351
VisualGenome [14]	img	231	349
COCO-CN [22]	img	362	218
<b>Actions</b>			
UCF-101 [34]	vid	9	457
Sports-1M [11]	vid	16	450
Kinetics-600 [3]	vid	130	336
Kinetics-400 [12]	vid	139	327
Kinetics-700 [4]	vid	139	327
HVU	vid	146	320
<b>Scenes</b>			
HVU	vid	44	86
Places365 [41]	img	78	52

In sum, this paper makes the following contributions:

- **Data.** We build *ChinaOpen*, with ChinaOpen-50k for multimodal learning and ChinaOpen-1k for multimodal model evaluation. To the best of our knowledge, ChinaOpen is the first of its kind<sup>2</sup>.
- **Model.** We propose Generative Video-to-text Transformer (GVT) for Chinese video captioning. GVT improves over GIT [35] with a simple visual-token reduction layer that effectively scales up the number of input video frames, resulting in better performance.
- **Evaluation.** We evaluate up to 15 SOTA models (11 English and 4 Chinese), see Table 2. Our evaluation covers up-to-date developments, *e.g.* ERNIE-ViL2 [31], CN-CLIP [39] and Taiyi [40] for open-set video tagging, X-CLIP [27] for text-to-video retrieval, Flamingo [1], GIT, mPLUG [17] and BLIP-2 [18] for video captioning.

The rest of the paper is organized as follows. We discuss related work in Sec. 2. ChinaOpen is detailed in Sec. 3, followed by GVT in Sec. 4 and evaluation in Sec. 5. Conclusions are given in Sec. 6.

## 2 RELATED WORK

Webly-annotated video data is crucial for developing large multimodal learning models. Meanwhile, manually-annotated video data is a must for properly evaluating the developed models. We briefly review progress in these two subjects, explaining accordingly how ChinaOpen uniquely contributes to the field.

**Progress on webly-annotated video datasets.** Due to the growing need of training large multimodal models for video-language related tasks, there have been good efforts on harvesting weakly-annotated videos from the web [2, 28]. HowTo100M [28], consisting

<sup>1</sup><https://www.bilibili.com/>

<sup>2</sup>ChinaOpen is available at <https://ruc-aimec-lab.github.io/ChinaOpen/>

of hundred million narrated video clips, has been used for learning text-video embedding. Despite the extremely large scale, the transcribed texts from narrated video clips are over noisy that specifically designed noise-tolerant learning algorithms have to be used. A more recent dataset, WebVid, contains 10M short videos with their textual descriptions sourced from stock footage sites [2]. In particular, the authors of [2] have used a 2.5M subset of WebVid to train a deep video-text matching model. We observe that these descriptions tend to be carefully worded to describe the video content in a concise manner, making them differ substantially from common user-generated titles. Also note that both HowTo100M and WebVid were sourced from English websites. In this context, our ChinaOpen-50k, which consists of user-titled videos from a popular Chinese video-sharing website, is unique. Moreover, we develop automated data cleaning to exclude videos that are either with low-quality annotations or lacking meaningful visual elements. As such, ChinaOpen-50k, while being relatively small-scale, already shows a good potential in our experiments.

**Progress on manually-annotated video datasets.** Existing manually-annotated video datasets mostly focus on a specific task, including human action recognition (HMDB [15] and UCF-101 [34]), sports-related activities (Sports-1M [11]) or a broader range of actions (Kinetics-400 [12]). The Holistic Video Understanding (HVV) dataset expands the labels from actions to scenes, objects, attributes and concepts [6]. However, the annotations of the above datasets are in the form of labels, making them unsuited for video-language tasks such as text-to-video retrieval which matches videos and natural-language text and video captioning that generates textual descriptions of the video content. Meanwhile, current video-text datasets such as MSVD [5], MSR-VTT [38], and VaTeX [37] has no manual label. Our ChinaOpen-1k dataset is unique as each video has manually labeled Chinese tags that explicitly describe objects, actions and scenes shown in the video content. Moreover, the video is also accompanied with a (manually checked) user-generated title and a manually-written content-based caption.

Table 2: SOTA models evaluated on ChinaOpen-1k.

Model	#params	Vision	Lang.	Tagging	Retrieval	Captioning
ResNet-P365 [41]	24M	img	EN	✓	✗	✗
SwinB-K400 [25]	88M	vid	EN	✓	✗	✗
CLIP4Clip [26]	164M	vid	EN	✗	✓	✗
X-CLIP [27]	164M	vid	EN	✗	✓	✗
CLIP-32/B [30]	151M	img	EN	✓	✓	✗
CN-CLIP [39]	188M	img	CN	✓	✓	✗
ERNIE-ViL2 [31]	204M	img	CN	✓	✓	✗
Taiyi [40]	254M	img	CN	✓	✓	✗
CLIP-L/14@336px [30]	427M	img	EN	✓	✓	✗
OFA-Chinese [36]	160M	img	CN	✗	✗	✓
GIT [35]	161M	img/vid	EN	✗	✗	✓
BLIP [19]	247M	img	EN	✗	✗	✓
mPLUG [17]	574M	img	EN	✗	✗	✓
Flamingo [1]	1,138M	img/vid	EN	✗	✗	✓
BLIP-2 [18]	3,745M	img	EN	✗	✗	✓
GVT (this paper)	146M	vid	CN	✗	✗	✓

### 3 THE ChinaOpen DATASET

As Fig. 2 shows, the ChinaOpen dataset is constructed in three stages. That is, raw data gathering from Bilibili, automated data cleaning to obtain ChinaOpen-50k (for multimodal learning), and lastly manual video annotation to produce ChinaOpen-1k (for multi-task evaluation). We depict each stage in the following.

#### 3.1 Raw Data Gathering

In order to obtain a representative and diverse subset of Bilibili short videos, we randomly generated many video ids as candidates. For this study, we gathered nearly 100k videos uploaded between May 2010 and Sep. 2021. Besides the MP4 video files, we also downloaded varied meta data, including titles, tags, descriptions, comments and danmaku (a.k.a. flying comments), if available. The duration of the downloaded videos ranges from 2 seconds to 608 seconds, with a mean value of 28.4 seconds and median of 25. The number of Chinese characters per video title ranges from 1 to 80, with a mean value of 16.4 and median of 14. The Bilibili platform organizes user-uploaded videos in channels, which are Bilibili-defined keywords that describe the videos at a very high level. Videos in ChinaOpen were collected from nearly 100 channels, where the top 10 channels are *daily*, *funny*, *celebrity*, *society*, *film* and *TV editing*, *general*, *beauty and skincare*, *body building*, *cat*, and *outfit*.

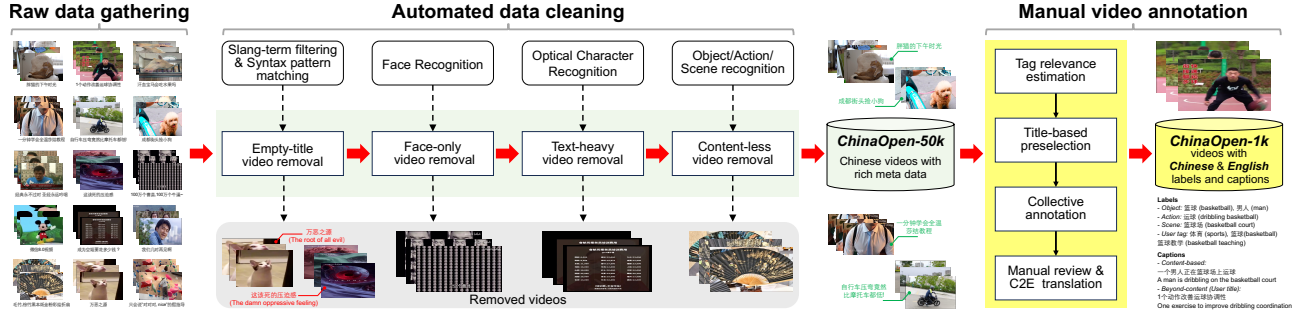
#### 3.2 Automated Data Cleaning

As the raw data is quite diverse with varied annotation quality, automated data cleaning is necessary to remove videos which are either with low-quality annotations or lacking meaningful visual content for a broad audience. With manual screening on the raw data, we empirically find that unwanted videos can be largely attributed to the following four categories, *i.e.* *empty-title*, *face-only*, *text-heavy*, and *content-less*. To that end, we develop as follows a multimodal method to identify videos of the four categories step-by-step, and remove them accordingly.

**3.2.1 Empty-title video removal.** We consider a video title empty if it has no verb-noun phrase (VNP) and thus tells little about the video content. In order to determine the presence of VNPs, we parse the given title with HanLP [9], an open-source Chinese NLP toolbox<sup>3</sup>, to obtain a syntactical representation of the title. Accordingly, syntax pattern matching is performed to find verb-object structures, nominal phrases with modifier-head constructions, and subject-verb-object constructions within the representation.

Note that due to the Chinese video-sharing culture, many titles are mainly comprised of slang terms such as “名场面” (iconic scene), “打卡挑战” (daily attendance), and “跟着UP主创作吧” (follow the uploader to create it). These terms are so frequently used that they tell little about the video content, and thus form the basis of our stopword list. We further expand the list to cover mental verbs such as “觉得” (think), “知道” (know) and “建议” (suggest) and non-Chinese characters such as punctuation, emoticons and Japanese / Korean characters. The title of a given video is filtered with the stopword list, followed by syntax pattern matching. If no VNP is found, the given video will be removed.

<sup>3</sup><https://github.com/hankcs/HanLP>



**Figure 2: Conceptual diagram of the construction of the proposed *ChinaOpen* dataset.** Given a set of 100k *Chinese* videos randomly gathered from Bilibili (Sec. 3.1), we perform automated data cleaning to remove videos either with low-quality annotations or lacking meaningful visual content (Sec. 3.2). This leads to *ChinaOpen-50k*, a weby-labeled set of 50k videos for multimodal learning. For a multi-faceted evaluation, we build a ground-truthed test set *ChinaOpen-1k* (Sec. 3.3). Each test video is accompanied with a manually-checked user title, a manually-written content-based caption, manually-checked user tags, and a number of labels describing visual objects / actions / scenes present in the video content. With all the Chinese text translated to English, *ChinaOpen-1k* is also suited for directly evaluating multimodal models trained on English data.

**3.2.2 Face-only video removal.** Face-only videos show mostly faces, with little background or other visual elements. More specifically, we observe two major patterns, *i.e.* *talking head* and *face mosaic*. A talking-head video shows only a person’s head (and shoulders), while a face-mosaic video contains frames showing a collage of many (small) faces. We therefore resort to frame-wise face detection. We adopt InsightFace<sup>4</sup>, an open-source deep face analysis library, with its default detection model (SCRFD-10GF). For a given video, we sample its frames uniformly. Face detection is performed per frame. A frame is classified as talking-head if a detected face region is over 50% of the image area. Accordingly, the given video talking-head is treated as talking-head if more than 75% of its frames are talking-head. To determine if the video is face-mosaic, we count the maximum number of faces detected per frame. If the number exceeds a given threshold (which is empirically set to 8), the video is labeled as face-mosaic. As such, we remove face-only videos.

**3.2.3 Text-heavy video removal.** Video with many texts on their frames typically lack visual elements of common interest. In order to filter out such text-heavy videos, we conduct Optical Character Recognition (OCR) on frames. In particular, we adopt PaddleOCR<sup>5</sup>, a public and leading OCT tool that recognizes multilingual texts from images. A frame is considered as text-heavy, if the number of OCR-detected characters exceeds 50. Similarly, we consider a video text-heavy if more than 75% of its frames are text-heavy.

**3.2.4 Content-less video removal.** We consider a given video content-less if it lacks recognizable object, action or scene. To that end, we employ existing visual recognition models to estimate if there is any object / action / scene present in the given video. For scene recognition, we employ a ResNet-152 network trained on the Places-365 image dataset [41] (ResNet-P365). As the input of ResNet-P365 shall be an image, we simply take the middle frame of the given video. For action recognition, we utilize a Video Swin Transformer (Swin-B as its backbone) [25] trained on the Kinetics-400 video dataset

[12], which we term SwinB-K400. Note that the 365 scene / 400 action classes defined in Places365 / Kinetics-400 are insufficient to cover the rich content of the Chinese videos. ResNet-P365 tends to incorrectly categorize videos of dogs or cats as “veterinarians of-fice”, while SwinB-K400 tends to mistakenly label videos of cooking as “cooking chicken”. Despite such biases, their predictions remain instructive to filter out content-less videos.

For object recognition, we curate a set of 3,841 object labels by merging object classes from MSCOCO [24], VisualGenome [14], Objects365 [33], LVIS [8] and OpenImages [16]. In order to predict the relevance of these objects *w.r.t.* the video, we use a pre-trained CLIP model (ViT-B/32) [30] for zero-shot tagging on the middle frame. With the English labels manually translated to Chinese, we further employ CN-CLIP [39], a Chinese version of CLIP, to tag the video with the Chinese object labels.

Object-wise, we consider a video not content-less if the video is predicted with at least one highly confident label (cutoff at the 75th percentile) or two moderately confident labels (cutoff at the 50th percentile). Action and scenes labels are postprocessed in a similar manner. A video is treated as content-less if no label is emitted.

With the video cleaning process described above, we obtain a cleaned set of 50k videos, termed *ChinaOpen-50k*. Video duration is between 3 seconds to 608 seconds, with a mean value of 29.8 and median of 27. File size per video is between 81.9KB and 20.9MB, with a mean value of 1.4MB and a median value of 1.2MB. *ChinaOpen-50k* has 431.2 hours and 69.1 GB of videos in total. With 19.2 characters on average, video titles are longer than those in the raw data (16.4 characters on average). More importantly, compared to the raw data, *ChinaOpen-50k* provides a better starting point for both multimodal learning and fine-grained manual annotation.

In addition to the user-generated titles, we enrich the annotations of *ChinaOpen-50k* by auto captioning. We adopt an existing model [7], trained on a joint set of MSR-VTT [38], VaTeX [37], TGIF [23] and Action-GIF [29]. As the generated captions are in English, we use machine translation<sup>6</sup> to convert them to Chinese.

<sup>4</sup><https://insightface.ai/>

<sup>5</sup><https://github.com/PaddlePaddle/PaddleOCR>

<sup>6</sup><https://fanyi-api.baidu.com/>

### 3.3 Manual Video Annotation

As aforementioned, we aim to build a ground-truthed Chinese video dataset to support multi-task evaluation. The tasks include general-purpose video content recognition (objects, actions, and scenes), assisted user tagging / captioning, and video retrieval by natural-language text. Since manual annotation is known to be expensive and thus much limited, video preselection is necessary to make the manual annotation process well pay off.

**3.3.1 Video preselection.** User tags are known to be subjective and personalized [20]. In order to find from ChinaOpen-50k videos that are likely to be accompanied with content-relevant tags, we adopt the classical neighbor voting algorithm [20]. Per video, we retrieve its 200 neighbors from the dataset in terms of cosine similarity between the video-level CLIP features. A user tag associated with the query video is deemed to be visually relevant if the tag also appears in the user-tag list of the neighbor videos. Next, from the videos with at least one content-related tag, we randomly sample 10k videos for title-based preselection as follows.

In contrast to a machine-generated caption trying to objectively describe what is visible, a user-generated title tends to be more eye-catching, providing readers with a beyond-content interpretation of the video. To strike a proper balance between relevance and attractiveness, we prefer to choosing videos with relevant titles such that a common user can easily relate the titles to the video content. Following this criterion, a review board of three experienced annotators read the titles of the 10k sampled videos, accordingly selecting 3k videos for collective annotation. To reduce the annotation workload, videos exceeding 60 seconds are excluded beforehand.

**3.3.2 Collective annotation.** Our annotation team consists of 16 members who are staffs and students in our lab. Each annotator has been instructed to annotate a given video in a coarse-to-fine manner. Firstly, the annotator is asked to check again if the user title is indeed content-relevant. If the answer is negative, the video will be skipped. Second, the annotator writes a caption that shall faithfully describe the gist of the video content. Next, the annotator describes with Chinese labels what objects / actions / scenes are shown in the video. Lastly, the annotator checks if the user-provided tags are content-relevant. To ensure the annotation richness, a video with zero label in a specific aspect (objects, actions, scenes or user tags) will be discarded.

**3.3.3 Manual review.** After the collective annotation stage, the review board performs a double check on the annotations for two purposes. That is, to fix labeling issues occasionally made by individual annotators and to translate the Chinese captions and labels to English assisted by machine translation. In total, we have 1,092 videos manually annotated with 1,092 user-generated titles, 1,092 content-based captions, 7,910 Chinese tags and 7,856 English tags<sup>7</sup> in total. The number of distinct Chinese / English tags is 2,100 / 2,030. Compared to existing (Chinese) video captioning datasets, e.g. VaTeX-CN, which have content-based captions only, the availability of user titles allows us to evaluate models in a novel beyond-content track. While targeted at Chinese models, the availability of

<sup>7</sup>Chinese tags such as “喵星人”, “猫”, “猫咪” are translated to “cat”, so the number of English tags is relatively smaller.

English annotations also allows us to evaluate English models. We term the new testset ChinaOpen-1k.

The testset has 9.9 hours and 1.6 GB of videos in total. Playback duration per video is between 5 seconds to 60 seconds, with a mean value of 32.5 seconds and a median value of 30 seconds. File size is between 0.2MB to 3.4MB, with mean value of 1.5MB and median value of 1.4MB. The size of the Chinese object / action / scene / user-tag vocabulary is 580 / 466 / 130 / 924. Table 3 shows annotation statistics in details. Compared with existing label-based datasets, ChinaOpen-1k has hundreds of unique labels, see Table 1. There is no overlap between ChinaOpen-50k and ChinaOpen-1k videos.

**Table 3: Annotation statistics of ChinaOpen-1K.**

Annotations	Min	Max	Mean	Median
<i>Number of Chinese labels per video:</i>				
Objects	1	9	2.45	2
Actions	1	6	1.37	1
Scenes	1	3	1.10	1
Verified user-tags	1	9	2.33	2
<i>Number of characters per Chinese caption:</i>				
User-generated title	6	79	18.88	17
Content-based caption	5	38	14.23	14

## 4 MULTIMODAL LEARNING ON CHINAOPEN

As a showcase of multimodal learning on ChinaOpen-50k, we describe in this section how to train a Transformer-based Chinese video captioning model on the webly-annotated dataset.

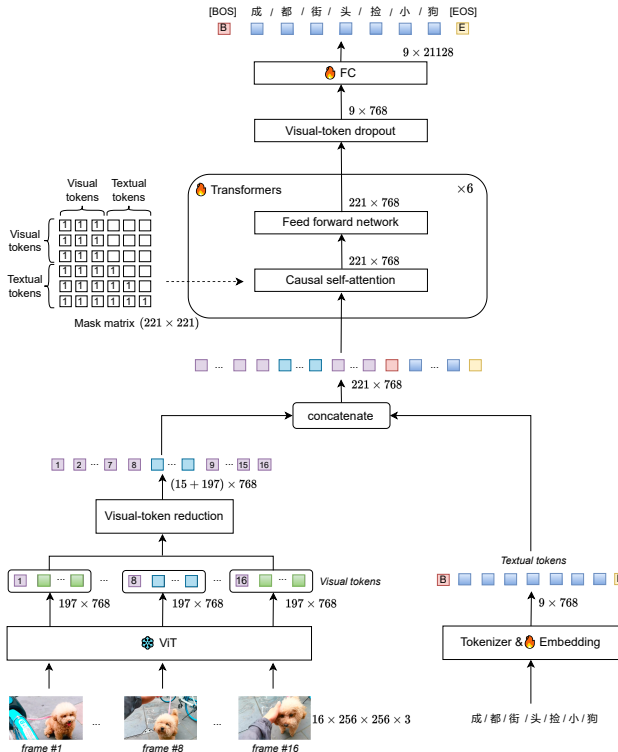
We depart from the Generative Image-to-text Transformer (GIT) [35], a state-of-the-art model on multiple vision-to-language generation tasks including image captioning, video captioning, and VQA. Consider image captioning for instance. At the training stage, given an input image of size  $256 \times 256$  and a reference caption of  $m$  words, GIT uses a Vision Transformer (ViT) to encode the input image, generating an array of  $14 \times 14 = 196$  visual tokens plus a special [CLS] token, each with a 768-d embedding vector. Meanwhile, the caption is also tokenized and represented by an array of  $(m + 2)$  embeddings, where the two extra tokens indicate the beginning and the end of the sentence, a.k.a. [BOS] and [EOS]. The visual and textual tokens are then concatenated and fed into a language Transformer for text generation. Note that in order to prevent information leakage during decoding, causal self-attention is used in the language Transformer such that each textual token is only permitted to “see” its preceding tokens, i.e. all visual tokens and the textual tokens before the current token. For video captioning, GIT simply concatenates the visual tokens of the input video frames. Given  $k$  frames as input, such a strategy will yield a large number of  $k \times 197$  visual tokens. Consequently, GIT has to set  $k$  to be a small number (which is 6) to make the computation feasible. However, the small  $k$  means sparse sampling of the video frames, inevitably causing much loss in the visual information.

Note that for a specific frame, its [CLS] token has been updated by the other tokens of the frame within ViT. Hence, the [CLS] token represents the frame to a large extent. Also note that for a given video, its middle frame is typically more representative than the other frames. Hence, selectively combining all the tokens from the middle frame and the [CLS] tokens of the other frames seems



reasonable. We implement this idea with a simple visual-token reduction (VTR) layer, see Fig. 3. With VTR, the number of visual tokens to be fed into the language Transformer is substantially reduced from  $k \times 197$  to  $k + 196$ . Such a minor tweak<sup>8</sup> allows us to effectively scale up the number of input video frames from 6 to 16. With the VTR layer, GIT is tailored to video captioning. We term the improved model Generative Video-to-text Transformer (GVT).

The vision and language Transformers of GVT are initialized by a pre-trained GIT\_BASE<sup>9</sup>. In order to cope with Chinese, the language-specific layers, *e.g.* text tokenizer, token embedding and the last FC layer, are re-trained from scratch if applicable.



**Figure 3: Proposed Generative Video-to-text Transformer (GVT) for video captioning. GVT improves over GIT with a simple visual-token reduction layer that effectively scales up the number of input video frames, from 6 in GIT to 16.**

## 5 EVALUATION

### 5.1 Common Setup

Subject to the availability of a model’s PyTorch training / inference code and our computation capacity (8×NVIDIA RTX 3090 GPUs), we collect the following SOTA models:

- **Video tagging (2):** *ResNet-P365* (ResNet-152 trained on Places365) [41] and *SwinB-K400* [25] (Video Swin Transformer with Swin-B as its backbone trained on the Kinetics-400 video action dataset).

<sup>8</sup>We also try adding all tokens from the first and the last frames. Accordingly, the number of visual tokens increases from  $k+196$  to  $k+588$ . Adding more frames marginally improves the performance, yet with noticeably increased computational overhead.

<sup>9</sup>[https://publicgit.blob.core.windows.net/data/output/GIT\\_BASE/snapshot/model.pt](https://publicgit.blob.core.windows.net/data/output/GIT_BASE/snapshot/model.pt)

- **Text-to-video retrieval (2):** *CLIP4Clip* [26] (Transferring CLIP to the video domain with a sequential Transformer for temporal modeling) and *X-CLIP* [27] (improving over CLIP4Clip with fine-grained cross-modal matching).

- **Tagging & retrieval (5):** *CLIP-B/32* [30] (ViT-B/32 as visual encoder and GPT-2 as text encoder), *CLIP-L/14@336px* [30] (ViT-L/14-336px as visual encoder and GPT-2 as text encoder), *Taiyi* [40] (ViT-B/32 as visual encoder and Chinese-Roberta-wwm-base as text encoder), *CN-CLIP* [39] (ViT-B/16 as its visual encoder and Chinese-Roberta-wwm-base as text encoder), and *ERNIE-ViL2* [31] (ViT-B/16 as visual encoder and ERNIE-3.0-base as text encoder).

- **Video captioning (6):** *OFA-Chinese* [36] (ResNet-101 as visual encoder and Transformer as text encoder), *GIT* [35] (ViT-B/16 as visual encoder), *BLIP* [19] (ViT-B/16 as visual encoder and BERT-base as text encoder), *BLIP-2* [18] (ViT-L/14 as visual encoder and Query Transformer as text encoder), *mPLUG* [17] (ViT-B/16 as visual encoder and 6-layer Transformer as text encoder) and *Flamingo* [1] (ViT-L/14 as visual encoder and OPT-350m as text encoder).

By default, each model is evaluated using the ground truth of its own language, unless stated otherwise.

### 5.2 Task I: Open-Set Video Tagging

**5.2.1 Task setup.** In open-set video tagging, a model is asked to tag a given video with an ad-hoc vocabulary that the model is not specifically tuned for. Recall that ChinaOpen-1k has tags along four dimensions, *i.e.* objects, actions, scenes and user tags. Evaluating video tagging per dimension reveals how good the model is at recognizing objects / actions / scenes and assisting user tagging.

By prompt-based label embedding, large multimodal models such as CLIP, CN-CLIP, Taiyi and ERNIE-ViL2 are naturally applicable for the open-set setting. However, the setting is challenging for ResNet-P365 and SwinB-K400 trained with a fixed vocabulary. Note for instance ChinaOpen-1k has 52 novel scene labels compared to Places365, see Table 1. To resolve the label mismatch, we convert the prediction of ResNet-P365 as follows. Per test video, we use ResNet-P365 to predict top-5 labels for the middle frame. For each (English) tag in the ChineseOpen-1k scene vocabulary, its relevance score to the given video is calculated by summing up the scores of the predicted labels weighed by their BERT similarity scores to the tag. In a similar vein we handle SwinB-K400.

Among the three pre-trained Chinese models (CN-CLIP, ERNIE-ViL2, and Taiyi), CN-CLIP is the only model that has PyTorch training code released. We thus choose to fine-tune this model with ChinaOpen-50k and a learning rate of 1e-5. Since CN-CLIP is an image model, we simply take the middle frame per video. We also try the same fine-tuning strategy with VaTeX-CN (VaTeX with Chinese captions). For the ease of reference, we use the notation  $\{model\} [\{dataset\}]$  to indicate  $\{model\}$  trained on  $\{dataset\}$ .

**5.2.2 Performance metric.** We compute Average Precision (AP) per test image, as commonly used to evaluate multi-label classification [21]. The overall performance is measured by mean AP.

**5.2.3 Results.** The performance of different models is reported in Table 4. Among the four English models, the CLIP series clearly surpass ResNet-P365 and SwinB-K400, showing the superiority of the large multimodal models in the open-set tagging scenario.

Meanwhile, CLIP-L/14@336px is noticeably better than CLIP-B/32. Note the main difference between the two models that the former is equipped with a much larger ViT. Similar results are also observed among the Chinese models, where CN-CLIP which uses ViT-B/16 outperforms Taiyi which uses the smaller ViT-B/32. CN-CLIP fine-tuned on ChinaOpen-50k is better than the original.

**Table 4: Performance of open-set video tagging. Metric: AP.**

Model	Objects	Actions	Scenes	User-tags	Mean
<i>English models:</i>					
SwinB-K400	-	7.7	-	-	-
ResNet-P365	-	-	8.7	-	-
CLIP-B/32	34.6	32.8	32.4	27.9	31.9
CLIP-L/14@336px	<b>44.0</b>	<b>42.7</b>	<b>36.7</b>	<b>35.7</b>	<b>39.8</b>
<i>Chinese models:</i>					
Taiyi	38.8	38.2	45.7	35.7	39.6
ERNIE-ViL2	40.8	40.0	47.4	35.3	40.9
CN-CLIP [VaTeX-CN]	40.7	41.7	<b>48.3</b>	34.4	41.3
CN-CLIP	39.2	<b>43.0</b>	47.1	36.4	41.4
CN-CLIP [ChinaOpen-50k]	<b>42.6</b>	42.2	45.8	<b>37.3</b>	<b>42.0</b>

Along the four dimensions, the result suggests that the top English model (CLIP-L/14@336px) recognizes objects the best, while the three Chinese models are relatively consistent, all recognizing scenes the best. For both English and Chinese models, their performance in the user-tag dimension is relatively the worst, suggesting that assisted user tagging is more challenging.

### 5.3 Task II: Text-to-Video Retrieval

**5.3.1 Task setup.** Text-to-video retrieval is to rank videos in terms of their cross-modal similarity to a given textual query. Recall that each test video is associated with a user-generated title and a manually-written caption. This allows us to setup two evaluation tracks: a content-based track which uses the manual captions as test queries and a beyond-content track which uses the user titles as test queries.

For the large image-text models (CLIPs, CN-CLIP, Taiyi and ERNIE-ViL2), their video-level feature is obtained by mean pooling over the corresponding frame-level features. As for X-CLIP and CLIP4Clip originally developed for text-to-video retrieval, we use ViT-B/32 as their visual encoder and have them trained on four popular (English) video-text datasets, *i.e.* MSVD [5], MSR-VTT [38], VaTeX [37] and ActivityNet-Caption [13], respectively.

**5.3.2 Evaluation criteria.** We report the commonly used Recall at Rank N ( $R@N$ ,  $N=1, 5, 10$ ) and their summation, denoted as SumR.

**5.3.3 Results.** Text-to-video retrieval performance of the individual models are summarized in Table 5. Among the English models, CLIP-L/14@336px is again the best, showing the importance of using a larger ViT. Nevertheless, the performance gap between CLIP-L/14@336px and CLIP-B/32 (233.2 versus 210.3 in SumR in the content-based track) can be effectively reduced by training a task-specific network on many video-text pairs, see X-CLIP [VaTeX] with SumR of 229.7. The superior performance of the English models as compared to their Chinese counterparts is largely due to the use of much larger ViT, see CLIP-L/14@336px. Given model size (#parameters) at the same level, *c.f.* Table 2, CN-CLIP (with 188M parameters) and CLIP B/32 (with 151M parameters) are largely

comparable (SumR 213.2 vs 210.3). The performance of CN-CLIP is improved by fine-tuning on ChinaOpen-50k.

Comparing the two tracks, we observe that the performance gain of X-CLIP [VaTeX] over CLIP-B/32 in the beyond-content track is much less than its counterpart in the content-based track (7.4 *versus* 19.4 in SumR). The result indicates a clear discrepancy between manually-written captions and user-generated titles. For all models, their performance in the beyond-content track is consistently lower than in the content-based track. We conclude from the result that querying by user-titles is more difficult.

Comparing the three Chinese models, while CN-CLIP is the best for the video tagging task, ERNIE-ViL2 now outperforms CN-CLIP and Taiyi. Given that ERNIE-ViL2 and CN-CLIP use the same visual encoder (ViT-B/16) but different text encoders (ERNIE-3.0 *versus* Chinese-Roberta-wwm), the result suggests that ERNIE-3.0 provides a better textual-query representation.

**Table 5: Performance of text-to-video retrieval. Models per language are sorted by their overall performance.**

Model	Content-based track				Beyond-content track			
	R@1	R@5	R@10	SumR	R@1	R@5	R@10	SumR
<i>English models:</i>								
CLIP-B/32	49.5	75.9	84.9	210.3	32.3	59.0	68.3	159.6
X-CLIP [MSR-VTT]	50.7	79.9	86.9	217.5	32.5	58.8	69.9	161.2
CLIP4CLIP [MSVD]	51.3	77.3	86.6	215.2	34.6	60.2	70.6	165.4
CLIP4CLIP [ActivityNet]	53.5	79.9	87.7	221.1	34.1	58.3	68.8	161.2
CLIP4CLIP [MSR-VTT]	52.3	79.9	88.5	220.7	33.9	59.9	68.9	162.7
X-CLIP [MSVD]	53.7	79.3	86.6	219.6	34.1	61.8	71.7	167.6
CLIP4CLIP [VaTeX]	56.0	82.5	88.9	227.4	34.0	60.3	70.6	164.9
X-CLIP [ActivityNet]	55.0	81.3	88.6	224.9	35.6	62.5	71.2	169.3
X-CLIP [VaTeX]	56.9	<b>83.3</b>	89.5	229.7	35.0	61.0	71.0	167.0
CLIP-L/14@336px	<b>59.5</b>	83.2	<b>90.5</b>	<b>233.2</b>	<b>45.2</b>	<b>69.4</b>	<b>78.8</b>	<b>193.4</b>
<i>Chinese models:</i>								
Taiyi	48.4	77.5	85.8	211.7	41.1	68.2	79.8	189.2
CN-CLIP	48.2	77.6	87.5	213.2	43.8	72.2	80.5	196.4
ERNIE-ViL2	53.6	81.8	89.4	224.7	46.1	72.5	80.8	199.4
CN-CLIP [VaTeX-CN]	59.3	<b>87.2</b>	92.2	238.7	42.0	70.1	78.3	190.4
CN-CLIP [ChinaOpen-50k]	<b>62.5</b>	86.4	<b>92.7</b>	<b>241.5</b>	<b>52.2</b>	<b>79.2</b>	<b>88.0</b>	<b>219.4</b>

### 5.4 Task III: Video Captioning

**5.4.1 Task setup.** Video captioning is to generate a natural-language sentence for video content description. Similar to the text-to-video retrieval task, we also setup two tracks. The content-based track uses the manually-written captions as ground truth, while the beyond-content track uses the user-generated title.

For the four image captioning models, *i.e.* mPLUG, BLIP, BLIP-2, and OFA-Chinese, we use the middle frame as their visual input. As for Flamingo, we follow the original paper [1], uniformly sampling 8 frames per video as its visual input.

As GVT is derived from GIT, the latter is a direct baseline to the former. We thus fine-tune both models on ChinaOpen-50k and VaTeX-CN, respectively. To validate the necessity of our proposed data cleaning pipeline, we randomly selected an equivalent-sized dataset from the raw data, referred to as Bilibili-50k. Moreover, we try to implement a multi-task version of GVT by adding a token-based binary classification head to predict if a given video-text pair is relevant. The head takes as input the [EOS] token, which has seen all preceding tokens. With prompt-based label embedding, open-set video tagging can be performed. Given the caption generation

loss ( $\ell_g$ ) and the video-text-matching loss ( $\ell_m$ ), a combined loss is formed as  $w \times \ell_g + (1 - w) \times \ell_m$ , with weight  $w \in [0, 1]$ .

**5.4.2 Evaluation criteria.** We adopt three common metrics, *i.e.* BLEU4, METEOR, and CIDEr, with their mean for overall comparison. The calculation related to the Chinese captions is conducted at a word-level, using Jieba<sup>10</sup> for Chinese word segmentation.

**5.4.3 Results.** As Table 6 shows, with the ability to utilize a pre-trained LLM in a frozen-weights manner, BLIP-2 clearly outperforms the other English models. In the content-based track, the better performance of GIT [ChinaOpen-50k] against GIT [VaTeX-CN] (42.2 *versus* 23.1) and that of GVT [ChinaOpen-50k] against GVT [VaTeX-CN] (44.4 *versus* 31.3) shows that ChinaOpen-50k leads to better Chinese video captioning models. The same conclusion can also be drawn from the beyond-content track. Given that ChinaOpen-50k is auto-constructed, these results are encouraging.

**Table 6: Performance of video captioning.**

Model	Content-based track				Beyond-content track			
	BLEU4	METEOR	CIDEr	Mean	BLEU4	METEOR	CIDEr	Mean
<b>English models:</b>								
Flamingo	8.7	7.9	27.3	14.6	3.2	3.5	8.8	5.2
mPLUG	10.9	12.8	33.7	19.1	<b>3.9</b>	<b>4.6</b>	10.0	6.2
GIT	9.7	9.2	43.0	20.6	2.2	3.1	9.2	4.8
BLIP	17.9	13.3	62.0	31.1	2.8	3.5	10.0	5.4
BLIP-2	<b>19.5</b>	<b>15.2</b>	<b>79.3</b>	<b>38.0</b>	3.6	4.2	<b>14.3</b>	<b>7.4</b>
<b>Chinese models:</b>								
OFA-Chinese	3.8	6.3	13.6	7.9	1.1	3.0	4.3	2.8
GIT[VaTeX-CN]	10.7	18.4	40.1	23.1	<b>1.7</b>	4.3	4.9	3.6
GVT[VaTeX-CN]	<b>18.5</b>	18.4	56.9	31.3	1.6	4.5	4.8	3.6
GIT[Bilibili-50k]	14.9	18.4	67.8	33.7	0.9	3.9	6.5	3.8
GIT[ChinaOpen-50k]	17.0	<b>19.1</b>	90.1	42.1	1.2	4.5	9.2	5.0
GVT[ChinaOpen-50k]	17.7	<b>19.1</b>	96.3	44.4	1.5	4.6	9.1	5.1

GIT [ChinaOpen-50k] is better than its counterpart trained on Bilibili-50k, 42.1 *versus* 33.7 in the content-based track and 5.0 *versus* 3.8 in the beyond-content track. The result justifies the necessity of data cleaning. The performance of multi-task GVT is shown in Table 7. For tagging / retrieval, we see a clear performance gap between the multi-task GVT and the SOTA. Developing a unified model is nontrivial, necessitating further investigation.

For both English and Chinese models, their performance scores in the beyond-content track are much lower than their content-based counterparts. Clearly, there is a big gap between what the SOTA video captioning models can generate and what a real user wants his or her videos to be titled. See Fig. 4 for qualitative results.

## 6 SUMMARY AND CONCLUDING REMARKS

We develop ChinaOpen, a new video dataset for open-world multi-modal learning. The dataset consists of ChinaOpen-50k, a webly

<sup>10</sup><https://github.com/fxsjy/jieba>



**Figure 4: Qualitative results by leading models, *i.e.* CLIP-L/14@336px and CN-CLIP for tagging and BLIP-2, GIT and GVT for captioning. Middle frames are with red borders.**

annotated video set for training, and ChinaOpen-1k, a manually annotated bilingual video set for testing. Fifteen SOTA models and the proposed GVT have been evaluated, leading to conclusions as follows. For open-set video tagging, the best English / Chinese model is CLIP-L/14@336px / CN-CLIP. Predicting user tags is more difficult than recognizing objects, actions and scenes. For text-to-video retrieval, CLIP-L/14@336px is again the best English model, while EARNIE-ViL2 is the winning Chinese model. For video captioning, BLIP-2 generates the best English captions, while GVT trained on ChinaOpen-50k generates the best Chinese captions. For both retrieval and captioning tasks, the beyond-content track appears to be more challenging than the content-based track. ChinaOpen has demonstrated a new opportunity for future research.

**Acknowledgments.** This work was supported by NSFC (62172420), Tencent Marketing Solution Rhino-Bird Focused Research Program, and Public Computing Cloud, Renmin University of China. The first author thanks H. Doughty and C. Snoek from UvA for helpful discussion on the topic.

**Table 7: Performance of multi-task GVT on video tagging, retrieval and captioning.**

w	Open-set video tagging					Text-to-video retrieval								Video captioning							
	Object	Action	Scene	User-tags	Mean	Content-based track				Beyond-content track				Content-based track				Beyond-content track			
						R@1	R@5	R@10	SumR	R@1	R@5	R@10	SumR	BLEU4	METEOR	CIDEr	Mean	BLEU4	METEOR	CIDEr	Mean
0.8	16.0	17.1	17.1	11.7	15.5	35.3	65.6	79.9	180.8	15.8	41.4	54.9	112.1	<b>16.6</b>	<b>18.8</b>	<b>87.5</b>	<b>41.0</b>	<b>1.5</b>	<b>4.5</b>	<b>9.1</b>	<b>5.0</b>
0.5	16.8	18.5	<b>20.8</b>	12.8	17.2	39.0	70.3	81.9	191.2	19.4	47.3	61.3	128.0	16.2	18.5	83.9	39.5	1.4	4.4	8.8	4.9
0.2	<b>20.0</b>	<b>21.6</b>	19.3	<b>16.1</b>	<b>19.3</b>	<b>40.2</b>	<b>71.3</b>	<b>83.3</b>	<b>194.8</b>	<b>22.3</b>	<b>51.6</b>	<b>64.0</b>	<b>137.9</b>	15.6	18.4	77.6	37.2	1.2	4.2	8.4	4.6



## REFERENCES

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A visual language model for few-shot learning. In *NeurIPS*.
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *ICCV*.
- [3] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. arXiv:1808.01340
- [4] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2022. A Short Note on the Kinetics-700 Human Action Dataset. arXiv:1907.06987
- [5] David Chen and William Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *CVPR*.
- [6] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhofen, and Luc Van Gool. 2020. Large scale holistic video understanding. In *ECCV*.
- [7] Chengbo Dong, Xinru Chen, Aozhu Chen, Fan Hu, Zihan Wang, and Xirong Li. 2021. Multi-Level Visual Representation with Semantic-Reinforced Learning for Video Captioning. In *ACMMM*.
- [8] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *CVPR*.
- [9] Han He and Jinho D. Choi. 2021. The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders. In *EMNLP*.
- [10] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. 2022. Lightweight Attentional Feature Fusion: A New Baseline for Text-to-Video Retrieval. In *ECCV*.
- [11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.
- [12] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The Kinetics human action video dataset. arXiv:1705.06950
- [13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Nieves. 2017. Dense-Captioning Events in Videos. In *ICCV*.
- [14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV* 123, 1 (2017), 32–73.
- [15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: A large video database for human motion recognition. In *ICCV*.
- [16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV* (2020).
- [17] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In *EMNLP*.
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597
- [19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.
- [20] Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2009. Learning Social Tag Relevance by Neighbor Voting. *TMM* 11, 7 (2009), 1310–1322.
- [21] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. 2016. Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval. *CSUR* 49, 1 (2016), 14:1–14:39.
- [22] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for Cross-Lingual Image Tagging, Captioning and Retrieval. *TMM* 21, 9 (2019), 2347–2360.
- [23] Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2015. TGIF: A New Dataset and Benchmark on Animated GIF Description. In *CVPR*.
- [24] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.
- [25] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video Swin Transformer. In *CVPR*.
- [26] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.
- [27] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In *ACMMM*.
- [28] 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 *ICCV*.
- [29] Yingwei Pan, Yehao Li, Jianjie Luo, Jun Xu, Ting Yao, and Tao Mei. 2022. Auto-captions on GIF: A Large-scale Video-sentence Dataset for Vision-language Pre-training. In *ACMMM*.
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- [31] Bin Shan, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. ERNIE-ViL 2.0: Multi-view Contrastive Learning for Image-Text Pre-training. arXiv:2209.15270
- [32] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *ICMR*.
- [33] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *ICCV*.
- [34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402
- [35] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A generative image-to-text transformer for vision and language. arXiv:2205.14100
- [36] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.
- [37] 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 *ICCV*.
- [38] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video And Language. In *CVPR*.
- [39] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. arXiv:2211.01335
- [40] Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. arXiv:2209.02970
- [41] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *TPAMI* 40, 6 (2017), 1452–1464.