CNVid-3.5M: Build, Filter, and Pre-train the Large-scale Public Chinese Video-text Dataset

Tian Gan[1]*, Qing Wang[2]*, Xingning Dong[1], Xiangyuan Ren[2], Liqiang Nie[3], Qingpei Guo[2]†

[1]Shandong University, [2]Ant Group, [3]Harbin Institute of Technology (Shenzhen)

gantian@sdu.edu.cn, wq176625@antgroup.com, dongxingning1998@gmail.com
xiangyuan.rxy@antgroup.com, nieliqiang@gmail.com, qingpei.gqp@antgroup.com

Abstract

Owing to well-designed large-scale video-text datasets, recent years have witnessed tremendous progress in video-text pre-training. However, existing large-scale video-text datasets are mostly English-only. Though there are certain methods studying the Chinese video-text pre-training, they pre-train their models on private datasets whose videos and text are unavailable. This lack of large-scale public datasets and benchmarks in Chinese hampers the research and downstream applications of Chinese video-text pre-training. Towards this end, we release and benchmark CNVid-3.5M, a large-scale public cross-modal dataset containing over 3.5M Chinese video-text pairs. We summarize our contributions by three verbs, i.e., "Build", "Filter", and "Pre-train": 1) To build a public Chinese video-text dataset, we collect over 4.5M videos from the Chinese websites. 2) To improve the data quality, we propose a novel method to filter out 1M weakly-paired videos, resulting in the CNVid-3.5M dataset. And 3) we benchmark CNVid-3.5M with three mainstream pixel-level pre-training architectures. At last, we propose the Hard Sample Curriculum Learning strategy to promote the pre-training performance. To the best of our knowledge, CNVid-3.5M is the largest public video-text dataset in Chinese, and we provide the first pixel-level benchmarks for Chinese video-text pre-training. The dataset, codebase, and pre-trained models are available at https://github.com/CNVid/CNVid-3.5M.

1. Introduction

Owing to well-designed large-scale datasets, video-text pre-training [15, 17, 19] has achieved superior performance in various downstream tasks, such as video-text retrieval [4, 10, 36], video question answering [27, 34, 42], and video captioning [1, 22, 30]. However, recent large-scale video-text datasets are mostly English-only (*e.g.*, Howto100M [25] and WebVid-2.5M [4]). Though some methods [14, 26, 45] turn to study the Chinese video-text pre-training, they pre-train their models on **private** datasets whose videos and text are unavailable. Therefore, the research towards Chinese video-text pre-training is still in its infancy due to the lack of large-scale **public** datasets.

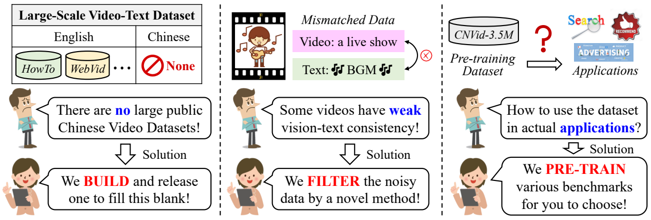

Large-Scale Video-Text Dataset
English
Chinese
HowTo
WebVid
...
None
There are no large public Chinese Video Datasets!
Solution
We BUILD and release one to fill this blank!
Mismatched Data
Video: a live show
Text: BGM
Some videos have weak vision-text consistency!
Solution
We FILTER the noisy data by a novel method!
CNVid-3.5M
Pre-training Dataset
?
Search
RECOMMEND
ADVERTISING
Applications
How to use the dataset in actual applications?
Solution
We PRE-TRAIN various benchmarks for you to choose!

Figure 1. Here presents the motivations of this paper, based on which we highly summarize our contributions with three verbs: "Build", "Filter", and "Pre-train".

Towards this problem, directly translating English text into Chinese is a simple solution. However, it may result in unacceptable performance degradation for two reasons: 1) Translation errors are inevitable. Moreover, since most of the large-scale video-text datasets employ the Automatic Speech Recognition (ASR) system to generate text, the language translator would amplify the error from the incomplete and noisy ASR text. And 2) there remains an intrinsic linguistic gap between English and Chinese. Many widely-used English idioms and slang can hardly find their Chinese counterparts, leading some translated text incomprehensible and even contrary to the original meaning.

In this paper, we aim to release and benchmark a large-scale public Chinese video-text dataset to facilitate future researchers and the community. As illustrated in Figure 1, three verbs could highly summarize our contributions, i.e., "Build", "Filter", and "Pre-train".

To **build** a large-scale Chinese video-text dataset, we collect over 4.5M videos from Chinese websites. All videos are associated with user-uploaded titles and ASR text.

*Equal contribution.
†Corresponding author.

We **filter** out the weakly-paired data by a novel method to improve the data quality. As some work [25, 26] pointed out, the pre-training performance would suffer from the noisy ASR text that fails to accurately describe the video content. Unfortunately, the problem is raised with few practical solutions. Therefore, we employ a well-trained image-text model to evaluate the video-text consistency for three reasons: 1) The text information in existing image-text datasets [11] are usually manually-written titles or captions, whose consistency is guaranteed. 2) Some video-text pre-training architectures [23, 35] are based upon image-text ones. And 3) it is cheap and efficient to "hire" a well-trained model to check millions of videos. In this way, we filter out about 1M weakly-paired videos based on the balance between the pre-training performance and efficiency, deriving the proposed CNVid-3.5M dataset.

We **pre-train** various models to benchmark our CNVid-3.5M dataset. Current video-text pre-training methods could be roughly divided into two categories: 1) feature-level pre-training methods [24, 33, 40] that employ offline video and textual feature extractors, and 2) pixel-level ones [4, 15, 36] that learn cross-modal representations end-to-end from raw videos and text. Since there remain domain gaps between pre-training datasets and frozen feature extractors, pixel-level pre-training methods usually achieve better performance and have been widely employed in recent years. However, existing Chinese video-text pre-training methods [14, 26, 45] are all feature-level ones pre-trained on **private** datasets, limiting their contributions on the development of Chinese video-text pre-training techniques. Hence, we adopt three mainstream pixel-level pre-training frameworks, which are the **first** pixel-level benchmarks for Chinese video-text pre-training.

Moreover, we propose the novel Hard Sample Curriculum Learning strategy to promote the pre-training performance. Since contrastive learning is a significant component in video-text pre-training, some methods [16, 18, 43] employ the hard sample mining [12, 29] strategy to promote the cross-modal alignment. However, hard sample mining would bring side effects to pre-training when the model is far from convergence. Suppose that a model is incapable of discriminating the ground-truth video-text pairs, recklessly introducing hard negatives would lead to the sub-optimal performance. Inspired by the curriculum learning [32, 37] strategy that "starts small" and gradually "learns hard", we combine these two strategies and propose the novel Hard Sample Curriculum Learning (HSCL). By gradually and smoothly emphasizing those hard samples, HSCL could effectively improve the pre-training performance.

Our contributions are summarized in four folds:

- To fill in the blank of large-scale public Chinese video-text datasets, we collect over 4.5M videos associated with titles and ASR text from the websites.
- To improve the data quality, we propose a novel method to filter out 1M weakly-paired videos, resulting in the CNVid-3.5M dataset.
- To promote the pre-training performance, we propose the novel Hard Sample Curriculum Learning strategy for better cross-modal contrastive learning.
- To the best of our knowledge, the constructed CNVid-3.5M is the **largest public** Chinese video-text dataset. Moreover, we provide the **first** Chinese pixel-level benchmarks based on CNVid-3.5M. The dataset, codebase, and benchmarks are available at https://github.com/CNVid/CNVid-3.5M.

2. Related Work

Video-Text Pre-training and Fine-Tuning. Video-text pre-training has attracted increasing interest in recent years, which could be roughly divided into two categories, feature-level pre-training methods [24, 33, 40] and pixel-level ones [15, 17, 35]. The former approaches employ offline visual and textual features extracted from frozen models, while the latter ones learn cross-modal representations from raw videos and text in an end-to-end manner. Since there remain domain gaps between pre-training datasets and frozen feature extractors, pixel-level pre-training methods usually achieve better performance than the feature-level ones.

There are three mainstream architectures in pixel-level video-text pre-training methods, *i.e.*, three-encoders-fusion (*3-E-F*) [15, 17, 30], two-encoders-fusion (*2-E-F*) [9], and twin-towers-crossed (*T-T-C*) [8, 23, 36]. *3-E-F* methods contain three separate encoders to model visual, textual, and cross-modal features; *2-E-F* methods simplify the *3-E-F* architecture by employing a shared encoder to embed textual and cross-modal features; while *T-T-C* methods are mainly based on the widely-adopted CLIP [28] architecture, which includes a couple of text and visual encoders to learn the cross-modal alignment by contrastive learning.

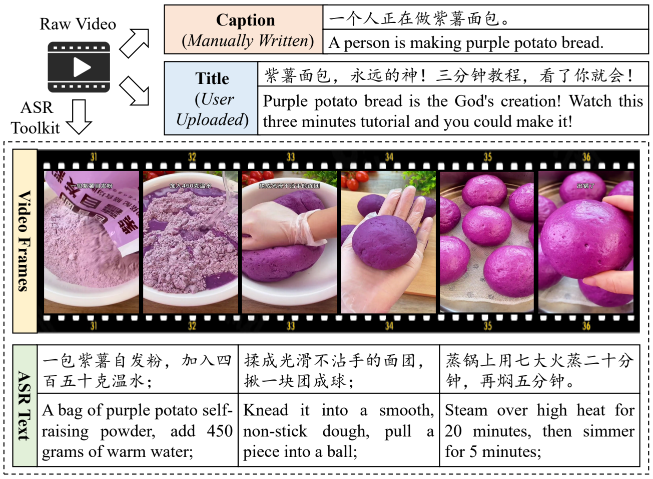
Though the pixel-level paradigm has been widely adopted recently, existing Chinese video-text models [14, 26, 45] are all feature-level ones pre-trained on private datasets. In this work, we provide the first pixel-level benchmarks pre-trained on our CNVid-3.5M dataset.

Video-Text Dataset. Well-designed large-scale datasets are a prerequisite for the successful applications of the "Pre-training & Fine-tuning" paradigm. There exist a few large-scale public English datasets (*e.g.*, Howto100M [25], WebVid-2.5M [4]) for video-text pre-training. However, existing large-scale Chinese video datasets (ALIVOL [14], Kwai-SVC [26], CREATE [45]) are all private, whose videos and text are unavailable to the public. Therefore, we aimed to release a public one to facilitate future researchers.

Hard Sample Mining in Contrastive Learning. Hard Sample Mining [12, 29] is an effective strategy to promote

Dataset Name	#(Videos)	#(Text)	Duration (hours)	Video Source	Text Type	Availability		
						Off-Feat	Raw-Vid	State
Datasets for Downstream Chinese Video-text Tasks								
VATEX [38]	41,269	825,380	115	kinetics-600	Caption	√	√	All-public
BFVD [44]	43,166	43,166	140	E-Commerce	Title	√	-	Partly-public
FFVD [44]	32,763	32,763	252	E-Commerce	Title	√	-	Partly-public
CREATE210K [45]	216,303	268,593	1,800	Open Websites	Caption	-	-	NOT-public
Kwai-SVC [26]	222,077	143,569	3,500	Open Websites	Title & ASR	√	-	Partly-public
Datasets for Chinese Video-text Pre-training								
ALIVOL-10M [14]	10,300,000	11,000,000	98,801	E-Commerce	Title	-	-	NOT-public
Kwai-SVC-11M [26]	11,075,084	3,931,879	177,200	Open Websites	Title & ASR	-	-	NOT-public
CREATE-10M [45]	10,000,000	10,000,000	83,000	Open Websites	Title	-	-	NOT-public
CNVid-3.5M (ours)	3,508,120	3,508,120	35,414	Open Websites	Title & ASR	√	√	All-public

Table 1. Data analysis of CNVid-3.5M and other Chinese video-text datasets. "*Off-Feat*" and "*Raw-Vid*" represent whether offline features and raw videos are available. As far as we know, CNVid-3.5M is the **largest public** video-text dataset in Chinese.



Caption (*Manually Written*)

一个人正在做紫薯面包。

A person is making purple potato bread.

Title (*User Uploaded*)

紫薯面包，永远的神！三分钟教程，看了你就会！

Purple potato bread is the God's creation! Watch this three minutes tutorial and you could make it!

一包紫薯自发粉，加入四百五十克温水；	揉成光滑不沾手的面团，揪一块团成球；	蒸锅上用七大火蒸二十分钟，再焖五分钟。
A bag of purple potato self-raising powder, add 450 grams of warm water;	Knead it into a smooth, non-stick dough, pull a piece into a ball;	Steam over high heat for 20 minutes, then simmer for 5 minutes;

Figure 2. An example of video-text pairs in our CNVid-3.5M dataset. We present three types of text information: 1) Caption (manually written), 2) Title (user uploaded), and 3) ASR text.

the performance of contrastive learning. In video-text pre-training, TACo [43] proposes a cascade sampling method to construct hard negative examples for better cross-modal alignment. However, we conjectured that hard sample mining may hinder the video-text pre-training at the beginning, where the model is far from convergence and could hardly discriminate the ground-truth video-text pairs. Therefore, we introduced curriculum learning [32, 37] and proposed the Hard Sample Curriculum Learning (HSCL) strategy. HSCL would guide the pre-training procedure from "starting small" gradually to "learning hard", mitigating the side effect of conventional hard sample mining methods.

3. The CNVid-3.5M Dataset

To fill in the blank of public Chinese video-text datasets, we release CNVid-3.5M, a large-scale cross-modal dataset containing over 3.5M Chinese video-text pairs.

3.1. Dataset Building

We collect raw videos from Douyin *, a popular Chinese social website, where millions of active users upload various types of videos every day. To enrich the data diversity, we do not limit video categories like HowTo100M [25] (mainly instructional videos) and cover various topics (*e.g.*, food, entertainment, and technology). *I.e.*, we first form up a set of keywords and then leverage them to search for videos and corresponding captions. Moreover, we span the data created time at a range of 3 years and filter the abnormal ones whose videos could not be played. In this way, we collect over 4.5M raw videos in Chinese.

Besides, the supplementary material will present more details about fairness and privacy for safety and ethics.

3.2. Captions, Titles, and ASR Text

As CREATE [45] pointed out, three types of text information are widely-employed in current Chinese video-text datasets. We show an example in Figure 2, based on which we summarize their characteristics and applications:

1) **Captions** are manually written by human annotators. They could accurately describe the video content in a formal manner. However, obtaining a large number of captions is extremely expensive. Therefore, captions are mainly employed in small-scale downstream video-text datasets like VATEX [38]. Moreover, captions are indeed global descriptions of the whole videos. Therefore, they could hardly describe the frame-level information of videos.

2) **Titles** are written and uploaded by users. Compared with captions, it is not unfeasible to obtain millions of user-uploaded titles. However, to increase the attractiveness and click rate, users may over-polish titles with some "appealing but redundant" words, which may decrease the video-text consistency. Similar to captions, titles are video-level

*Official Website: https://www.douyin.com/

descriptions rather than frame-level ones. Therefore, titles and captions are mainly employed in feature-level video-text pre-training methods rather than pixel-level ones.

3) **ASR text** is obtained by the ASR toolkit. Similar to titles, it is easy to obtain a large number of ASR text. However, some ASR text is incomplete, noisy, and even faulty, which may decrease the pre-training performance. Besides, ASR text is a frame-level description, which is widely employed in the pixel-level video-text pre-training.

Following previous methods [25, 26], we collect user-uploaded titles and obtain the ASR text of each video. Since we provide both video-level titles and frame-level ASR text, our dataset could support the video-text pre-training in both feature-level and pixel-level manner.

3.3. Weakly-paired Data Filtering

As some work [25, 26] has pointed out, some ASR text may not match with associated videos (*e.g.*, videos have background music), which would decrease the pre-training performance. However, the problem is only raised with few practical solutions. Therefore, we propose a novel weakly-paired data filtering strategy by leveraging a well-trained image-text model to automatically evaluate the video-text consistency for three reasons: 1) The text information in existing image-text datasets [11, 13] are manually-written titles or captions, whose quality and consistency are guaranteed. 2) Pixel-level pre-training methods [8, 15, 23] usually sample several frames from raw videos to serve as the visual input, whose frameworks are closely associated with image-text pre-trained models. 3) Compared with manually checking millions of videos, the proposed filtering strategy remarkably reduces the cost and improves the efficiency.

Specifically, we first pre-train an image-text model on Wukong-100M [11]. We then leverage it to calculate the similarity score between videos and ASR text, and filter out the last 1M videos by their ranks. In this way, we obtain the CNVid-3.5M dataset, based on which we provide three mainstream pixel-level pre-training benchmarks.

3.4. Comparison of Dataset Statistics

Table 1 presents the statistics of CNVid-3.5M and other Chinese video-text datasets. To the best of our knowledge, CNVid-3.5M is the **largest public** dataset for Chinese video-text pre-training, with a total of 3.5M videos that last for 35.4K hours. Though ALIVOL-10M [14], Kwai-SVC-11M [26], and CREATE-10M [45] contain more Chinese videos than ours, yet they do not release raw videos and text information to the public. Moreover, these three methods only provide conventional feature-level pre-trained models, whose performance highly depends on frozen feature extractors. Towards this end, we provide the first pixel-level benchmarks based on our CNVid-3.5M dataset, which would pave the way for future research on Chinese video-text pre-training.

Moreover, we present additional statistics of our CNVid-3.5M dataset in the supplementary material, including the distributions of topics, keywords, video durations, and Part-of-Speech (POS) tags of the ASR text.

4. Pixel-level Pre-training Paradigm

Since mainstream pixel-level video-text pre-training methods [17, 23, 35] extract cross-modal representations from raw data directly, they usually outperform the conventional feature-level ones [2, 24, 33]. Therefore, we benchmark our CNVid-3.5M dataset with three mainstream pixel-level architectures. As illustrated in Figure 3, we split the pixel-level pre-training process into three steps, namely: 1) data pre-processing to process raw videos and text into patches and tokens, 2) model architectures to generate cross-modal representations, and 3) proxy tasks to determine overall pre-training objectives.

4.1. Data Pre-processing

Given a mini-batch (denoted as $\mathcal{B}$) of videos $\{V_i\}_{i=1}^{|\mathcal{B}|}$ and their corresponding ASR text $\{T_i\}_{i=1}^{|\mathcal{B}|}$, pixel-level pre-training methods would first sparsely (and randomly) sample N_v frames from each video. We then slice each frame into patches, obtaining visual patch embeddings $\mathbf{P} \in \mathbb{R}^{N_v * N_p * d}$, where N_p is the number of sliced patches, and d is the dimension of the embedding. Simultaneously, a BERT embedder is employed to process the text T into fixed-length word embeddings $\mathbf{W} = [\mathbf{w}^{cls}, \mathbf{w}^1, \mathbf{w}^2, \cdots, \mathbf{w}^{N_t-1}]$, where $\mathbf{W} \in \mathbb{R}^{N_t * d}$, N_t is the length of text tokens.

4.2. Model Architectures

There are three mainstream pixel-level video-text pre-training architectures, namely three-encoders-fusion [15, 17, 30], two-encoders-fusion [9], and twin-towers-crossed [8, 23, 36]. We benchmark CNVid-3.5M with these three pixel-level pre-training architectures.

Three-encoders-fusion (*3-E-F*) pre-training methods contain a visual encoder E_{vis}, a text encoder E_{txt}, and a cross-modal encoder E_{mul}. After obtaining visual features $\mathbf{V} = E_{vis}(\mathbf{P})$ and textual features $\mathbf{T} = E_{txt}(\mathbf{W})$, we concatenate these two-modal features into $[\mathbf{T}, \mathbf{V}] = [\mathbf{t}^{cls}, \mathbf{t}^1, \mathbf{t}^2, \cdots, \mathbf{t}^{N_t-1}, \mathbf{v}^{cls}, \mathbf{v}^1, \mathbf{v}^2, \cdots, \mathbf{v}^{N_v}]$. Following CLIP4Clip [23], we set position embeddings to zero, and obtain cross-modal features $\mathbf{M} = E_{mul}([\mathbf{T}, \mathbf{V}])$. This forwarding process could be formulated as follows:

$$\mathbf{M} = E_{mul}([E_{txt}(\mathbf{W}), E_{vis}(\mathbf{P})]), \tag{1}$$

where $\mathbf{M} = [\mathbf{m}^{cls}, \mathbf{m}^1, \mathbf{m}^2, \cdots, \mathbf{m}^{N_t+N_v}]$, and $[\cdot, \cdot]$ denotes the concatenation operation.

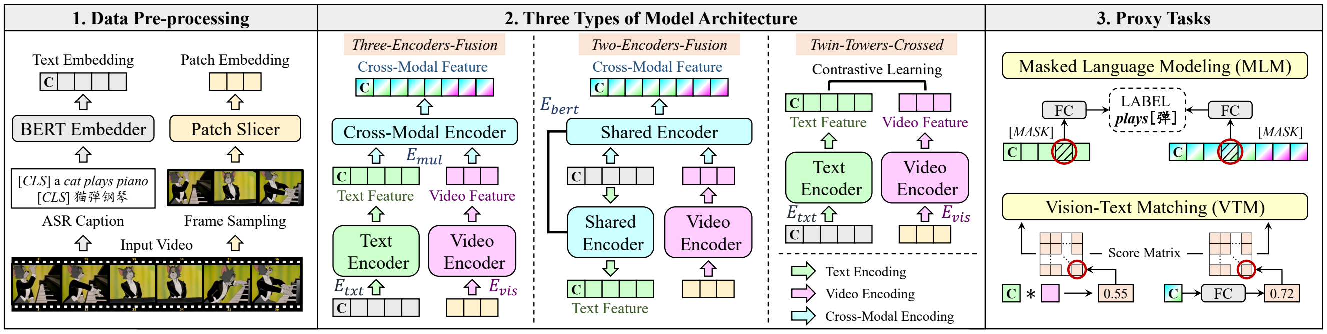


Figure 3. The pipeline of mainstream pixel-level video-text pre-training methods, which contains three key steps: 1) Data pre-processing, 2) model architectures, and 3) proxy tasks. Note that we employ three mainstream architectures to enrich pixel-level pre-training benchmarks.

Two-encoders-fusion (*2-E-F*) pre-training methods simplify the *3-E-F* architecture by employing a shared BERT-type encoder E_{bert} to embed textual and cross-modal features simultaneously. The pipeline of obtaining textual features $\mathbf{T}$ and cross-modal features $\mathbf{M}$ could be formulated as follows:

$$\begin{cases}\mathbf{T} = E_{bert}(\mathbf{W}), \\ \mathbf{M} = E_{bert}([\mathbf{W}, E_{vis}(\mathbf{P})]).\end{cases} \tag{2}$$

Twin-towers-crossed (*T-T-C*) pre-training methods mainly follow the conventional CLIP [28] architecture. After obtaining visual and textual features, *T-T-C* methods employ contrastive learning to optimize the twin encoders E_{vis} and E_{txt} for better cross-modal alignment. Note that *T-T-C* methods do not generate cross-modal video representations like *3-E-F* and *2-E-F*.

4.3. Proxy Tasks

Proxy tasks are crucial for video-text pre-training as they directly determine the final optimization objectives. Following the conventional protocol [4, 15, 43], we adopt two widely-employed proxy tasks, *i.e.*, Masked Language Modeling (MLM) and Video-Text Matching (VTM).

MLM first masks out a certain percentage of words in a given sentence, and then forces the model to restore these clozes according to visual and textual cues. MLM is calculated twice for textual features $\mathbf{T}$ ($\mathcal{L}_1$) and cross-modal features $\mathbf{M}$ ($\mathcal{L}_2$) as follows:

$$\mathcal{L}_1 = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathcal{L}_{CE}(y^q, \Theta_1(\mathbf{t}^q)), \tag{3}$$

$$\mathcal{L}_2 = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathcal{L}_{CE}(y^q, \Theta_2(\mathbf{m}^q)), \tag{4}$$

where $\mathcal{Q}$ records the location of masked tokens, $|\cdot|$ denotes the length of a given set, y^q denotes the ground-truth token label, Θ is a Multi-Layer Perception (MLP), and $\mathcal{L}_{CE}$ is the regular Cross-Entropy cost function.

VTM aims to promote the cross-modal alignment between videos and text, which is calculated in a parameter-free ($\mathcal{L}_3$) and parameter-employed ($\mathcal{L}_4$) way:

$$\mathcal{L}_3 = -\sum_{i=1}^{|\mathcal{B}|}\log\frac{\sum_{k=1}^{N_v}\exp^{\langle \mathbf{v}_i^k, \mathbf{t}_i^{cls}\rangle}}{\sum_{k=1}^{N_v}(\exp^{\langle \mathbf{v}_i^k, \mathbf{t}_i^{cls}\rangle} + \sum_{j\neq i}\exp^{\langle \mathbf{v}_j^k, \mathbf{t}_i^{cls}\rangle})}, \tag{5}$$

$$\mathcal{L}_4 = -\sum_{i=1}^{|\mathcal{B}|}\log\frac{\exp^{\Theta_4(\mathbf{f}_{i,i})}}{\exp^{\Theta_4(\mathbf{f}_{i,i})} + \sum_{j\neq i}\exp^{\Theta_4(\mathbf{f}_{j,i})}}, \tag{6}$$

where $|\mathcal{B}|$ is the length of a mini-batch, $\langle\cdot,\cdot\rangle$ denotes the matrix multiplication operation, and $\mathbf{f}_{j,i}$ is global-pooling cross-modal features of the video-text pair (V_j, S_i).

For *3-E-F* and *2-E-F* models, the objective functions $\mathcal{L}_{3EF}/\mathcal{L}_{2EF}$ are calculated as follows:

$$\mathcal{L}_{3EF}(\mathcal{L}_{2EF}) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4. \tag{7}$$

While for *T-T-C* models, since they do not contain a cross-modal encoder to generate video representations, their objective function $\mathcal{L}_{TTC}$ is calculated as follows:

$$\mathcal{L}_{TTC} = \mathcal{L}_1 + \mathcal{L}_3. \tag{8}$$

4.4. Fine-tuning

In order to thoroughly evaluate the pre-training performance, we fine-tune our models on the Text-to-Video Retrieval (TVR) task. TVR is a widely-employed downstream video-text task, aiming to retrieve the most relevant videos according to text queries. Since there has only one Chinese TVR dataset (VATEX [38]), we translate two widely-adopted English datasets (MSRVTT [41] and DiDemo [3]) into Chinese to verify the superiority and robustness of the constructed dataset and proposed methods.

Specifically, we fine-tune our pre-trained models by reusing Video-Text Matching objectives ($\mathcal{L}_3$ in Eq. 5 and $\mathcal{L}_4$ in Eq. 6, we only employ $\mathcal{L}_3$ for *T-T-C* models).

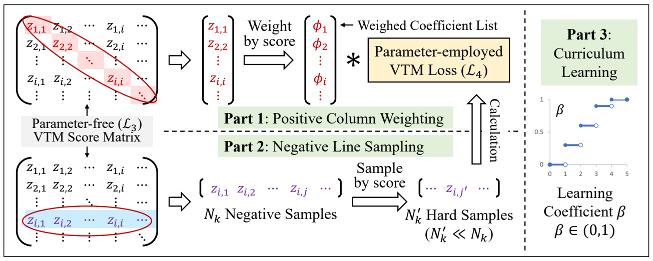

$z_{1,1}$ $z_{1,2}$ ⋯ $z_{1,i}$ ⋯
$z_{2,1}$ $z_{2,2}$ ⋯ $z_{2,i}$ ⋯
$z_{i,1}$ $z_{i,2}$ ⋯ $z_{i,i}$ ⋯
Parameter-free ($\mathcal{L}_3$) VTM Score Matrix
Weight by score
ϕ_1 ϕ_2 ϕ_i
Weighed Coefficient List
Parameter-employed VTM Loss ($\mathcal{L}_4$)
Part 1: Positive Column Weighting
Part 2: Negative Line Sampling
Calculation
$z_{i,1}$ $z_{i,2}$ ⋯ $z_{i,j}$ ⋯
N_k Negative Samples
Sample by score
⋯ $z_{i,j'}$ ⋯
N'_k Hard Samples ($N'_k \ll N_k$)
Part 3: Curriculum Learning
β
1
0.5
0
0 1 2 3 4 5
Learning Coefficient β
$\beta \in (0,1)$

Figure 4. The pipeline of Hard Sample Curriculum Learning (HSCL), which contains three components to promote the performance of video-text pre-training.

5. Hard Sample Curriculum Learning

We propose Hard Sample Curriculum Learning (HSCL), a model-agnostic strategy to promote the video-text pre-training performance. HSCL is based on two motivations. 1) The negative samples involved in the calculation of parameter-employed VTM loss ($\mathcal{L}_4$ in Eq. 6) are not representative. Suppose that the batch size is set to be N_k, a model needs to process $N_k * N_k$ video-text pairs for calculating $\mathcal{L}_4$, whose calculation cost is hardly affordable. Therefore, previous methods [9, 15, 22] tend to cut the number of negative samples to $N_k^{'}$ by random sampling. However, since $N_k^{'}$ is usually much smaller than N_k, it may result in the sub-optimal performance. 2) If a model is far from convergence, directly introducing hard negative samples would amplify the burden of learning, which would decrease the pre-training performance.

Therefore, inspired by hard sample mining and curriculum learning, we propose the HSCL strategy. By leveraging the similarity score calculated by the parameter-free VTM loss ($\mathcal{L}_3$ in Eq. 5), HSCL could gradually and smoothly introduce hard samples for better cross-modal alignment. As illustrated in Figure 4, HSCL contains the following three key components:

Positive Column Weighting (PCW) aims to decrease the learning coefficient ϕ of well-learned positive samples, forcing the model to emphasize those hard positive ones. Specifically, we re-weigh half of samples within the batch according to their similarity scores z calculated by $\mathcal{L}_3$, which could be formulated as follows:

$$z_{j,i} = \frac{1}{N_v} \sum_{k=1}^{N_v} \exp^{\langle \mathbf{v}_j^k, \mathbf{t}_i^{cls} \rangle}, \tag{9}$$

$$\phi_i = \mathrm{Min}(\frac{\mathrm{Med}(Z_{diag}) - \mathrm{Min}(Z_{diag})}{z_{i,i} - \mathrm{Min}(Z_{diag}) + e_0}, 1.0), \tag{10}$$

where $Z_{diag} = [z_{1,1}, z_{2,2}, \cdots, z_{i,i}, \cdots]$, $\mathrm{Med}(\cdot)$ and $\mathrm{Min}(\cdot)$ represent the median and minimum result of the list, e_0 is a pre-defined coefficient to avoid a zero denominator.

Negative Line Sampling (NLS) aims to sample $N_k^{'}$ hard negative examples for better contrastive learning. Specifically, for each text T_j, we choose the Nearliest $N_k^{'}$ aligned negative videos according to their similarity scores in the list $Z_j = [z_{j,1}, z_{j,2}, \cdots, z_{j,i}, \cdots]$. *I.e.*, Pick the minimum $N_k^{'}$ samples in the list $Z_j^{'} = \{|z_{j,i} - z_{j,j}|\}_{j \neq i}$.

Curriculum Learning (CL) aims to mitigate the side effect of hard sample mining when the model is under convergence. We set a gradually increasing coefficient $\beta \in (0, 1)$. β is initialized with zero at the beginning of pre-training. It will gradually increase by a fixed number β_0 for each N_d iteration until reaching 1.0. For PCW, the actual coefficient ϕ^* is calculated as follows:

$$\phi_i^* = (1.0 - \beta) + \beta * \phi_i. \tag{11}$$

For NLS, we first generate a random number $rand \in (0, 1)$, and only perform the NLS step if $rand < \beta$. In this way, the model would conduct conventional pre-training at the beginning, and smoothly change to learn hard negative samples to pursue better pre-training performance.

6. Experiments and Benchmarks

6.1. Pre-training Datasets

In Section 3.3, the Chinese image-text model that helps to filter out the weakly-paired data is pre-trained on *Wukong-100M* [11] (100M image-text pairs). The Chinese video-text models are pre-trained on our *CNVid-3.5M* dataset, which contains 3.5M video-text pairs.

To verify the generality of the proposed Hard Sample Curriculum Learning (HSCL) strategy, we also pre-trained our models on four large-scale English datasets, they are 1) *COCO* [20] (0.6M image-text pairs), 2) *VG* [13] (5.4M image-text pairs), 3) *CC* [31] (3.1M image-text pairs), and 4) *WebVid-2.5M* (2.5M video-text pairs).

6.2. Fine-tuning Datasets

We fine-tuned our pre-trained models on the Text-to-Video Retrieval (TVR) task, which includes three widely-adopted datasets as follows: 1) *VATEX* [38] contains 28K video clips, and each video is associated with 10 Chinese and 10 English diverse captions. 2) *MSRVTT* [41] contains 10K video clips associated with 200K English captions. Following [15, 22, 43], we used 7K videos for training and randomly selected 1K videos from the remaining ones for testing (7K-1K split). 3) *DiDemo* [3] contains 10K Flickr videos associated with 40K English captions.

Moreover, for a comprehensive performance evaluation of Chinese video-text pre-trained models, we employ Google [†] to translate all captions in MSRVTT and DiDemo from English to Chinese.

Metrics. We employed Recall@K (R@K, K=1/5/10) and Median Rank (MdR) to measure the TVR performance.

[†] Official Website: https://translate.google.com/

No.	Model Type	Data Volumn	Masking Strategy	Matching Strategy	VATEX (Chinese)	MSRVTT (Translated)	DiDemo (Translated)
					R@1/5/10 ↑ (MdR ↓)	R@1/5/10 ↑ (MdR ↓)	R@1/5/10 ↑ (MdR ↓)
For "Building": Ablation Study of the constructed CNVid-3.5M dataset							
A1	2-E-F	No Pre-training			36.4 / 75.4 / 85.2 (2)	15.9 / 42.1 / 54.1 (8)	8.0 / 24.8 / 39.2 (18)
A2	2-E-F	3.5M	MLM	VTM	**39.9 / 77.2 / 87.0** (2)	**20.7 / 47.9 / 61.2** (**6**)	**13.3 / 34.6 / 45.9** (**13**)
For "Filtering": Ablation Study of Weakly-Paired Data Filtering							
B1	2-E-F	4.5M	MLM	VTM	38.8 / 76.4 / 86.4 (2)	19.4 / 45.6 / 58.6 (6)	10.6 / 32.7 / 45.0 (14)
B2	2-E-F	4.0M	MLM	VTM	39.5 / 77.0 / 86.5 (2)	19.8 / 47.2 / 60.3 (6)	12.4 / 32.8 / 45.6 (14)
B3	2-E-F	3.5M	MLM	VTM	39.9 / **77.2** / **87.0** (2)	**20.7** / **47.9** / 61.2 (6)	**13.3** / **34.6** / 45.9 (**13**)
B4	2-E-F	3.0M	MLM	VTM	**40.5** / 76.8 / 86.5 (2)	20.6 / 47.3 / **61.3** (6)	12.2 / 34.3 / 45.6 (14)
B5	2-E-F	2.5M	MLM	VTM	39.5 / 76.6 / 86.9 (2)	20.1 / 47.4 / 60.8 (6)	12.1 / 33.6 / **46.0** (14)
For "Pre-training": Ablation Study of Hard Sample Curriculum Learning							
C1	2-E-F	3.5M	MLM	VTM	39.9 / 77.2 / 87.0 (2)	20.7 / 47.9 / 61.2 (6)	13.3 / **34.6** / 45.9 (13)
C2	2-E-F	3.5M	MLM	VTM+HSM	40.7 / 76.8 / 86.9 (2)	21.1 / 47.3 / 60.1 (6)	13.4 / 34.3 / 46.4 (12)
C3	2-E-F	3.5M	MLM	VTM+HSCL	**41.5** / **78.2** / **87.2** (2)	**23.3** / **48.0** / **61.2** (6)	**13.6** / 34.4 / **47.3** (**12**)

Table 2. For **Chinese** video-text pre-training: Ablation study of Weakly-Paired Data Filtering and Hard Sample Curriculum Learning (HSCL) on the Text-to-Video Retrieval task of three datasets (VATEX-Chinese, MSRVTT-Translated, and DiDemo-Translated). "HSM" denotes the conventional Hard Sample Mining strategy without curriculum learning. Note that *A2*, *B3*, and *C1* are the same model.

No.	Pre-training Dataset	Masking Strategy	Matching Strategy	VATEX (English)	MSRVTT	DiDemo
				R@1/5/10 ↑ (MdR ↓)	R@1/5/10 ↑ (MdR ↓)	R@1/5/10 ↑ (MdR ↓)
D0	No Pre-training			39.5 / 78.3 / 88.6 (2)	19.2 / 46.2 / 59.5 (6)	14.2 / 34.1 / 46.1 (12)
D1	COCO+VG+CC	MLM	VTM	45.5 / 81.5 / 91.0 (2)	28.6 / 55.5 / 66.1 (4)	25.3 / 50.3 / 62.4 (5)
D2	COCO+VG+CC	MLM	VTM+HSCL	**48.7 / 83.1 / 91.1** (2)	**31.6 / 56.5 / 66.7** (4)	**27.1 / 53.8 / 63.8** (**4**)
D3	CC+WebVid	MLM	VTM	50.5 / 85.6 / 92.5 (1)	29.4 / 56.9 / 68.3 (4)	29.0 / 57.3 / 67.8 (4)
D4	CC+WebVid	MLM	VTM+HSCL	**53.4 / 86.2 / 92.9** (1)	**32.6 / 58.8 / 69.7** (**3**)	**29.8 / 58.0 / 68.9** (**3**)

Table 3. For **English** video-text pre-training: Ablation study of Hard Sample Curriculum Learning (HSCL) on the Text-to-Video Retrieval task of three datasets (VATEX-English, MSRVTT, and DiDemo). All models are based on the two-encoders-fusion (*2-E-F*) architecture.

6.3. Experimental Settings

Settings for Weakly-paired Data Filtering. We pre-trained the image-text model on the Wukong-100M [11] dataset, whose pre-training details are presented in the supplementary material. We then employed the pre-trained model to obtain the parameter-free video-text consistency score ($z_{i,i}$ in Eq. 9) of 4.5M video-text pairs. To improve the reliability, we randomly sampled four frames to calculate the consistency score, and repeated this step for three times. The final score of each video is the average of three results calculated under different random seeds. Ultimately, we sorted the consistency score in descending order, and filtered out the last 1M videos to refine the dataset.

Settings for Chinese Video-Text Pre-training. For model architectures, we employed the Video Swin Transformer [21] to serve as the visual encoder, which is initialized with parameters pre-trained on ImageNet [6]. The text, cross-modal, and shared encoders belong to the BERT-Base [7] model, which are initialized with parameters privdied by Hugging Face [‡]. We sparsely sampled 4 (N_v) frames from raw videos during pre-training. For the MLM task, the masking rate is 15%. For the proposed Hard Sample Curriculum Learning, the coefficient β would increase by 0.15 (β_0) for each 5000 (N_d) iterations until reaching 1.0.

For hyper-parameters, we set the length of text tokens $N_t = 30$, and the dimension of the hidden state $d = 768$. All models are pre-trained by the Adam optimizer with a momentum of 0.9. The total pre-training stage lasts for 10 epochs with a batch size of 128. The initial learning rate is 5e-5 and is decayed by the factor of 10 after 5 epochs. The whole pre-training on CNVid-3.5M takes about 4 days to complete on 8 NVIDIA V100 GPUs.

Settings for Fine-tuning on Downstream Chinese Video-text Tasks. The optimizer and hyper-parameters in fine-tuning remain the same as the pre-training configuration. The total fine-tuning stage lasts for 25,000 steps. The batch size is set to 128. The initial learning rate is set to 1e-5.

6.4. Ablation Study

As aforementioned, we summarized our work with three verbs, *i.e.*, "Build", "Filter", and "Pre-train". To evaluate the effectiveness of these three contributions, we took the *2-E-F* architecture as an example model in the ablation study.

Ablation Study of "Building". We compared the performance obtained by models with or without pre-training. As illustrated in Table 2 (*A1 vs. A2*), models pre-trained on

[‡] Official Website: https://huggingface.co/bert-base-chinese

No.	Model Type	Dataset	Pre-training Strategy	VATEX (Chinese)	MSRVTT (Translated)	DiDemo (Translated)
				R@1/5/10 ↑ (MdR ↓)	R@1/5/10 ↑ (MdR ↓)	R@1/5/10 ↑ (MdR ↓)
E1	3-E-F	CNVid-3.5M	MLM+VTM	39.9 / **78.2** / **87.6** (2)	20.6 / 47.1 / 60.0 (6)	**13.8** / 34.5 / **46.6** (13)
E2	2-E-F	CNVid-3.5M	MLM+VTM	39.9 / 77.2 / 87.0 (2)	20.7 / **47.9** / **61.2** (6)	13.3 / **34.6** / 45.9 (13)
E3	T-T-C	CNVid-3.5M	MLM+VTM	**40.5** / 77.3 / 87.0 (2)	**21.1** / 46.5 / 58.9 (6)	13.0 / 32.6 / 44.6 (15)

Table 4. We benchmark CNVid-3.5M with three mainstream pixel-level pre-training architectures, which is the first pixel-level benchmarks for Chinese video-text pre-training.

our CNVid-3.5M dataset (*A2*) would achieve a remarkable performance gain compared with those without pre-training (*A1*), which demonstrates the effectiveness of building a large-scale Chinese video-text dataset.

Ablation Study of "Filtering". We compared the performance obtained by models pre-trained under different numbers of video-text pairs. For example, "4.5M", "4.0M", and "3.5M" denotes that we employ the 4.5M full set, the Top-4.0M split set, and the Top-3.5M split set (CNVid-3.5M) according to the sorted video-text consistency scores calculated by weakly-paired data filtering. As illustrated in Table 2 (*B1-B5*), we have two conclusions as follows:

1) The proposed weakly-paired data filtering strategy could effectively improve the quality of datasets. The fine-tuning performance on three downstream datasets first increases rapidly when filtering out videos with low video-text consistency scores (*B1* → *B2* → *B3*). It would remain stable (*B3* → *B4*) or drop sightly (*B4* → *B5*) when continuously removing videos after the Top-3.5M split.

2) In general, employing the Top-3.5M split (*B3*) achieves the best performance. The results among all ablation models (*B1-B5*) may prove that, for the video-text pre-training, it is not "more data is better", but "more good data is better". Therefore, we filtered out the last 1M videos, resulting in the CNVid-3.5M dataset.

Ablation Study of "Pre-training" (HSCL). To prove the effectiveness of Hard Sample Curriculum Learning (HSCL), we pre-trained several ablation models on Chinese video-text datasets (*C1-C3* in Table 2) and English ones (*D1-D4* in Table 3). We have two conclusions as follows:

1) HSCL is a model-agnostic strategy, which could evidently improve the pre-training performance. In Table 2, compared with the baseline (*C1*), HSCL (*C3*) achieves an improvement of 1.6%, 2.6%, and 0.3% at R@1 of VATEX, MSRVTT, and DiDemo dataset, respectively. While in Table 3, HSCL outperforms the baseline by 3.2 / 3.0 / 1.8% (*D1 vs. D2* on COCO+VG+CC) and 2.9 / 3.2 / 0.8% (*D3 vs. D4* on CC+WebVid) at R@1 of three TVR datasets.

2) The proposed Curriculum Learning (CL) strategy in HSCL could effectively mitigate the side effect of conventional Hard Sample Mining (HSM) when the model is far from convergence. As illustrated in Table 2 (*C1 vs. C2*), models equipped with HSCL perform better than HSM.

Moreover, we present the detailed quantitative analyses of PCW and NLS in the supplementary material.

6.5. Chinese Video-text Pre-training Benchmarks

Based on the constructed CNVid-3.5M dataset, We provided the **first** pixel-level benchmarks for Chinese video-text pre-training. Following existing video-text pre-training methods, we adopted three mainstream pixel-level architectures, *i.e.*, three-encoders-fusion (*3-E-F*) [15, 17, 30], two-encoders-fusion (*2-E-F*) [9], and twin-towers-crossed (*T-T-C*) [8, 23, 36]. Detailed performance comparisons of three pixel-level benchmarks are presented in Table 4.

7. Conclusion

In this work, we release the largest public Chinese video-text dataset, *i.e.*, CNVid-3.5M. Besides, we provide the first pixel-level benchmarks for Chinese video-text pre-training. Moreover, we propose a novel weakly-paired data filtering method to improve the quality of datasets. We also design a novel Hard Sample Curriculum Learning strategy to promote the pre-training performance. In conclusion, we "build", "filter", and "pre-train" the large-scale Chinese video-text dataset. We believe this work would pave the way for future research on Chinese video-text pre-training.

8. Limitations and Future Work

Address the Difference between English and Chinese. One potential limitation of our work is that we adopt pixel-level architectures proved effective in English video-text pre-training, ignoring the linguistic gap between Chinese and English. *E.g.*, it may be better to employ the Whole Word Masking (WWM) [5, 46] strategy rather than the conventional MLM for Chinese video-text pre-training. We plan to explore more Chinese-specific methods in the future.

Need More Downstream Chinese Video-Text Datasets. Though we release a large-scale Chinese video-text dataset, there still lack some downstream datasets (*e.g.*, MSRVTT [41], MSVD-QA [39] in English) to comprehensively evaluate the performance of Chinese video-text pre-trained models. Currently, there only exists VATEX [38] for Chinese cross-modal retrieval and some captioning datasets [26, 45]. Therefore, building more downstream Chinese video datasets is a promising future direction.

Acknowledgments. This work is supported by the National Natural Science Foundation of China, No.: 62176137; Ant Group.

References

[1] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, 52(6):1–37, 2019. 1

[2] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021. 4

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 5, 6

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1728–1738, 2021. 1, 2, 5

[5] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021. 8

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7

[8] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2, 4, 8

[9] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2, 4, 6, 8

[10] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 1

[11] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*, 2022. 2, 4, 6, 7

[12] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020. 2

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 4, 6

[14] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the ACM International Conference on Multimedia*, pages 2567–2576, 2021. 1, 2, 3, 4

[15] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 2, 4, 5, 6, 8

[16] Chunyuan Li, Xiujun Li, Lei Zhang, Baolin Peng, Mingyuan Zhou, and Jianfeng Gao. Self-supervised pre-training with hard examples improves visual representations. *arXiv preprint arXiv:2012.13493*, 2020. 2

[17] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 1, 2, 4, 8

[18] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[19] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 1

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6

[21] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 7

[22] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 1, 6

[23] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2, 4, 8

[24] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Hongyang Chao, and Tao Mei. Coco-bert: Improving video-language pre-training with contrastive cross-modal matching and denoising. In *Proceedings of the ACM International Conference on Multimedia*, pages 5600–5608, 2021. 2, 4

[25] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic.

Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2630–2640, 2019. 1, 2, 3, 4

[26] Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. Search-oriented micro-video captioning. In *Proceedings of the ACM International Conference on Multimedia*, pages 3234–3243, 2022. 1, 2, 3, 4, 8

[27] Devshree Patel, Ratnam Parikh, and Yesha Shastri. Recent advances in video question answering: A review of datasets and methods. In *International Conference on Pattern Recognition*, pages 339–356. Springer, 2021. 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5

[29] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020. 2

[30] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 1, 2, 4, 8

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018. 6

[32] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, pages 1–40, 2022. 2, 3

[33] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019. 2, 4

[34] Guanglu Sun, Lili Liang, Tianlin Li, Bo Yu, Meng Wu, and Bolun Zhang. Video question answering: a survey of models and datasets. *Mobile Networks and Applications*, 26(5):1904–1937, 2021. 1

[35] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 2, 4

[36] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2022. 1, 2, 4, 8

[37] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3

[38] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4581–4591, 2019. 3, 5, 6, 8

[39] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the ACM International Conference on Multimedia*, pages 1645–1653, 2017. 8

[40] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 2

[41] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016. 5, 6, 8

[42] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1686–1697, 2021. 1

[43] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 2, 3, 5, 6

[44] Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. Poet: Product-oriented video captioner for e-commerce. In *Proceedings of the ACM International Conference on Multimedia*, pages 1292–1301, 2020. 3

[45] Ziqi Zhang, Yuxin Chen, Zongyang Ma, Zhongang Qi, Chunfeng Yuan, Bing Li, Ying Shan, and Weiming Hu. Create: A benchmark for chinese short video retrieval and title generation. *arXiv preprint arXiv:2203.16763*, 2022. 1, 2, 3, 4, 8

[46] Wei Zhu. Mvp-bert: Multi-vocab pre-training for chinese bert. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 260–269, 2021. 8