CBVS: A Large-Scale Chinese Image-Text Benchmark for Real-World Short Video Search Scenarios

Xiangshuo Qiao[1,†], Xianxin Li[1,†], Xiaozhe Qu[1], Jie Zhang[1,*], Yang Liu[1], Yu Luo[1], Cihang Jin[1] and Jin Ma[2]

[1]*Tencent PCG, Beijing, China*

[2]*University of Science and Technology of China, Hefei, Anhui, China*

Abstract

Vision-Language Models pre-trained on large-scale image-text datasets have shown superior performance in downstream tasks such as image retrieval. Most of the images for pre-training are presented in the form of open domain common-sense visual elements. Differently, video covers in short video search scenarios are presented as user-originated contents that provide important visual summaries of videos. In addition, a portion of the video covers come with manually designed cover texts that provide semantic complements. In order to fill in the gaps in short video cover data, we establish the first large-scale cover-text benchmark for Chinese short video search scenarios. Specifically, we release two large-scale datasets CBVS-5M/10M to provide short video covers, and the manual fine-labeling dataset CBVS-20K to provide real user queries, which serves as an image-text benchmark test in the Chinese short video search field. To integrate the semantics of cover text in the case of modality missing, we propose UniCLIP where cover texts play a guiding role during training, however are not relied upon by inference. Extensive evaluation on CBVS-20K demonstrates the excellent performance of our proposal. UniCLIP has been deployed to Tencent online video search systems with hundreds of millions of visits and achieved significant gains. The complete dataset, code and checkpoints are available at https://github.com/QQBrowserVideoSearch/CBVS-UniCLIP.

Keywords

vision-language models, video search, contrastive learning

1. Introduction

CLIP [1] demonstrates the promise of performing contrastive learning pre-training on large-scale image-text data from the web with a data size of 400 million. In this work, visual base models represented by ViT [2] are aligned with textual base models represented by Bert [3, 4] by learning on large-scale unsupervised data. These base models can be transferred to downstream tasks such as image search [5] via natural language prompts [6].

In the field of Chinese multi-modal representation learning, previous work [7, 8, 9] supplements high-quality Chinese image-text datasets and successfully pre-trains Chinese visual language models. Most of the data are open-domain images collected from the web or multiplexed from publicly available English datasets. These images are captured by a camera and presented in the form of common-sense visual elements, including animals, buildings, activities, etc. with corresponding descriptive text.

With the rise of short videos, video search has become a popular topic [10, 11]. Previous work [12, 13, 14] create large-scale datasets for the Chinese short-video search domain and provide publicly available video frames or video features to support content-based search. Online video search systems integrate features from several information domains such as video cover, video frame, and video title. There has been academic work on video search using titles, video frames, but there is a lack of research on video search based on video covers. During the creation of short videos, creators craft video covers

Open Domain Images   Short Video Cover Images

Figure 1: Morphological differences between open domain images and short video cover images.

Table 1

Comparison of CBVS versions.

Version	#Pair	Purp.	Ann.	Text type
CBVS-20K	20,001	Test	✓	Query, OCR
CBVS-5M	4,767,435	Train	✗	Title, OCR
CBVS-10M	10,075,989	Train	✗	Title, OCR

for short videos with the aim of attracting the interest of the most relevant viewers. Therefore, in short video search scenarios, short video covers provide direct overviews and serve as crucial visual features of the videos. Besides, cover-based search has efficiency advantages over content-based search.

However, there are remarkable morphological differences between short video cover images and open domain images. As shown in Fig. 1, compared to open-domain visual elements, short video covers, as user-originated content, are mostly artificial combinations of various visual elements and may undergo post-processing such as cropping and splicing. On the other hand, many creators craft cover texts for video covers to complement or emphasize the semantic information of the video. This is a feature that open domain images do not share. Therefore, short video cover images represent a different form of data from open domain images, and the availability of large-scale cover dataset is crucial. However, available large-scale cover datasets are lacking.

In this work, we release a large-scale Chinese image-text Benchmark for short Video Search scenarios (CBVS) to fill the gap of data in real Chinese video search scenarios. CBVS is designed in three versions: the manually fine-labeled CBVS-20K and the large-scale unsupervised CBVS-5M/10M. Fig. 2 shows their data examples. Specifically, CBVS-20K contains 20K high-quality <user query-video cover> pairs, which serves as an image-text benchmark test in the field of Chinese short video search. Well-trained human experts annotate the relevance of each user query to the video cover and at least two cross-validations are performed. In addition, Optical Character Recognition (OCR) texts of cover images are provided after machine extraction and human correction. Due to the constraints of user privacy and platform rules, the large-scale CBVS-5M/10M contains about 5M/10M <video title-video cover> pairs, where the text is provided in the form of video titles and OCR texts. These data are available for visual language models to learn modal alignment in pre-training or fine-tuning tasks. The CBVS dataset includes 32 categories such as Film and animation, Character, Education, Game, Commodity, etc. to avoid data distribution bias. Tab. 1 shows a detailed comparison of the various versions of CBVS.

In short video search scenarios, cover texts complement the semantics of cover images. On one hand, CLIP lacks the ability to fuse multi-semantic signals on the visual side. On the other hand, not all cover images come with cover texts, so the modality missing problem needs to be considered. In order to effectively integrate the semantics of cover images with cover texts, we propose UniCLIP inspired by the work of OCR-free [15, 16]. Cover texts signals are unified to guide image-text contrastive learning

Query=小鹏G6和特斯拉ModelY（Xpeng G6 and Tesla Model Y）			**Query**=西红柿炒鸡蛋（Tomato and Egg Stir-fry）		
OCR text: 致敬？还是超越？特斯拉**ModelY VS 小鹏G6** Pay tribute? Or beyond? Tesla Model Y VS Xpeng G6	**OCR text**: 小鹏G6 买它！ Go for the Xpeng G6!	**OCR text**: 尼古拉·特斯拉究竟有多强？ Just how brilliant was Nikola Tesla?	**OCR text**: 无 Null	**OCR text**: 无 Null	**OCR text**: 鸡胸肉黄瓜丁 Diced Chicken Breast with Cucumber
Relevance Level: 2	**Relevance Level**: 1	**Relevance Level**: 0	**Relevance Level**: 2	**Relevance Level**: 1	**Relevance Level**: 0

Title
故宫三大殿最后一作——保和殿！一千三百个零件 历时两年 完整三大殿，不可错过！
The final masterpiece of the Forbidden City's three main halls - the Hall of Preserving Harmony! With 1,300 components and two years in the making, the complete trio of halls is a must-see!

OCR text
故宫保和殿
The Hall of Preserving Harmony in the Forbidden City

Figure 2: Top: Presentation of CBVS-20K data. Bottom: Presentation of CBVS-5M/10M data.

Figure 3: Image-text mashup data, including E-commerce poster, social media image, video cover, etc.

in an presence-guided and semantic-guided manner.

The training of UniCLIP integrates the semantics of cover text and can be generalised to the retrieval of other image-text mashup data (not limited to video covers), as shown in Fig. 3. Image-text mashup data refers to images with OCR text, such as e-commerce posters, social media images, and video covers. It is worth emphasizing that the inference process does not depend on any module related to OCR and the model is immune to the problem of missing cover text modalities. Extensive experimental evaluations demonstrate the effectiveness of our proposal.

Our contributions can be summarized as follows:

- In order to fill in the lack of cover data for short video search scenarios, we release the largest Chinese cover image-text dataset with video title texts and cover texts.
- We build a manual fine-labeling image-text benchmark test for Chinese short video search scenarios, containing real user queries from browser logs.
- We propose UniCLIP, which introduces an image classification task and an image-text matching task to guide image-text contrastive learning training. UniCLIP imposes no additional inference cost and training is immune to the modality missing problem.

2. Related Work

2.1. Chinese Video/Image-text benchmark

Compared to English multi-modal pre-training, the Chinese community is lagging behind. [7] introduces translated versions of the English multi-modal datasets [17, 18] to support Chinese multi-modal pre-training. [9] releases a large-scale Chinese dataset Wukong containing 100 million image-text pairs collected from the web to bridge the language gap. [8] further establishes large-scale Chinese cross-modal benchmarks by releasing two pre-training datasets and five fine-tuning datasets. Besides, Product1M [19] provides additions to the e-commerce domain. [12, 13, 14] supplements the Chinese video-text data, and visual modalities are provided in the form of video frames. However, large-scale video cover data is scarce.

2.2. Image-text Matching

The image-text matching task aims to measure the semantic similarity of different modalities in the same embedding space[20]. Existing implementations fall into two categories: the first is embedding-based, i.e., encoding global representations of vision and text separately, and then performing similarity computation [21, 22, 23]. The second is score-based, i.e., performing cross-modal interactions locally and calculating cumulative scores [24, 25, 26]. Due to the advantages of performance and efficiency, embedding-based methods have attracted the attention of researchers [27]. In particular, CLIP [1] provides new ideas for multi-modal representation learning. A series of studies following CLIP have been applied to downstream tasks including image text matching. For example, GCLIP[28], CLIP-Adapter[29] and SLIP[30] raise the upper performance limit of CLIP, AltCLIP[31], CN-CLIP[7], and TaiyiCLIP[32] expand the CLIP language domain.

3. CBVS Dataset

3.1. Comparison

A comparison with other Chinese image-text/video-text datasets is shown in Tab. 2. On the one hand, CBVS provides cover images, user queries, and is larger in size compared to publicly available video datasets such as VATEX [33], BFVD/FFVD [34], and CNVid-3.5M [35]. CREATE [12], Kwai-SVC [13], ALIVOL-10M [36] with similar scale are access restricted. On the other hand, compared to image-text datasets such as Wukong [9], Product1M [19], M6-Corpus [37], ZERO-Corpus/R2D2 [8], etc., the biggest advantage of CBVS lies in the uniqueness of the video cover image and the cover text specific to the cover image. To the best of our knowledge, CBVS is the largest publicly available Chinese image-text dataset providing cover images.

3.2. Data Collection

In order to provide real video search data, we capture the user query logs of mainstream mobile browsers and divide user queries into two parts. We retrieve more than 8M videos from the Chinese video website BiliBili[1] through the first part of user queries. To avoid data distribution bias due to a single platform, we retrieve more than 5M videos from Tencent Video[2] as a supplement in the same way, and finally obtain more than 13M <cover-title> pairs as a data source for CBVS-5M/10M. In this process, there are videos with inconsistent cover and content in the original data. The video cover-video frame consistency is filtered by the video long click signal and the video completion rate signal. Slight data noise is allowed. Besides, we manually select more than 2K high-quality user queries from the second part, and collect 20K high-quality <user query-cover image> pairs in the same way, as the data source of CBVS-20K. The number of cover images under each user query is controlled in [5, 30).

---

[1]https://www.bilibili.com
[2]https://v.qq.com

Table 2
Statistics of Chinese image-text/video-text datasets.

Dataset	#Vision	#Text	Source	Vision type	Text type	Availability
			Chinese Video-Text Datasets			
VATEX	41,269	825,380	Kinetics-600	Video	Caption	✓
BFVD	43,166	43,166	E-Commerce	Feature	Title	✓
FFVD	32,763	32,763	E-Commerce	Feature	Title	✓
CREATE-210K	216,303	268,593	Open Websites	Video	Title, Caption	✗
CREATE-3M	3,000,000	3,000,000	Open Websites	Video	Title	✗
CREATE-10M	10,000,000	10,000,000	Open Websites	Video	Title	✗
Kwai-SVC	222,077	143,569	Video Search	Feature	Title, OCR, ASR	✗
Kwai-SVC-11M	11,075,084	3,931,879	Video Search	Video	Title, ASR	✗
CNVid-3.5M	3,508,120	3,508,120	Open Websites	Video	ASR, Title	✓
ALIVOL-10M	10,300,000	11,000,000	E-Commerce	Video, Image	Title, Abstract	✗
Youku-mPLUG	10,000,000	10,000,000	Open Websites	Video	Title	✓
			Chinese Image-Text Datasets			
Wukong	101,483,885	101,483,885	Open Websites	Image	Caption	✓
Wukong-Test	33,365	33,365	Open Websites	Image	Caption	✓
Product1M	1,182,083	1,182,083	E-Commerce	Image	Caption	✓
M6-Corpus	60,500,000	60,500,000	Open Websites	Image	Caption	✗
ZERO-Corpus	250,000,000	750,000,000	Image Search	Image	Title, Content, Query	✓
R2D2-ICR	200,000	200,000	Image Search	Image	Caption	✓
R2D2-IQR	200,000	200,000	Image Search	Image	Query	✓
CBVS-20K	20,001	20,001	Video Search	Cover Image	OCR, Query	✓
CBVS-5M	4,767,435	4,767,435	Video Search	Cover Image	OCR, Title	✓
CBVS-10M	10,075,989	10,075,989	Video Search	Cover Image	OCR, Title	✓

After that, we design the data cleaning program from two aspects: data quality and image-text relevance. First, we filter the video covers with low resolution and scale disproportion, and eliminate the dead links. After that, we score the relevance of video covers and titles in 13M data with the open-source Chinese image-text model QA-CLIP[3], and filter the trailing 3M data to obtain CBVS-10M. We randomly sample 1K data from 13M and 10M data, respectively, and human experts evaluate whether the video cover is relevant to the title. The evaluation conclusions show that the relevance is improved from 75.6% to 93.0% after data cleaning. Finally, CBVS-5M is obtained by sampling in CBVS-10M.

3.3. Data Annotation

The data annotation of CBVS-20K is performed by trained experts in the field of video search in a two-stage, cross-validated approach. First, they annotate whether the user query reveals a clear intent for video, and filter out query terms related to pornography and violence, and finally take the queries with a need for video as candidates for annotation in the second stage. After that, the annotators mark the degree of relevance for each <user query-cover image> pair, which is categorized into three grades: strongly relevant, weakly relevant, and irrelevant. The semantics of cover images and cover texts are required to be considered together. Meanwhile, the annotators correct the OCR text extracted by the machine.

The following criteria are used to determine whether the user query reveals a clear intent for video:

- With video intent: queries that explicitly need to be satisfied by video resources, or better (with
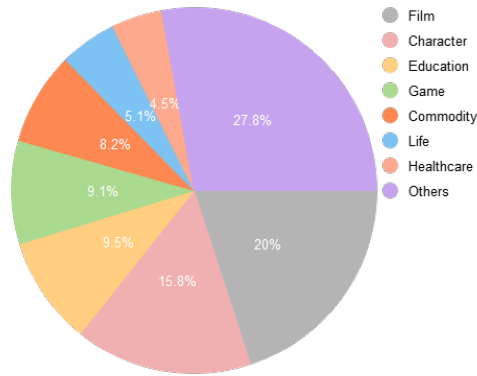
---

[3]https://github.com/TencentARC-QQ/QA-CLIP

Figure 4: Distribution of categories of user queries in CBVS-20K.

visual informativeness gain) by video resources. For example, "Football Match", "Tomato and Egg Stir-fry" (Fig.2 in the paper), etc.

- No-video intent: queries that explicitly cannot be met by video resources, e.g., "Microsoft Recruitment Official Website", "Computer Wallpaper Download", etc.

We exclude the controversial data, and end up with 20,001 image-text pairs consisting of 2,486 unique queries and 19,648 unique images. The percentage of strongly relevant, weakly relevant and irrelevant data is 29.74%, 30.80% and 39.46% respectively. The average length of user queries is 7.0 and the average length of OCR texts is 14.5. 33.41% of cover images come with OCR texts. Fig. 4 shows the distribution of categories of these user queries, where the categories are labelled manually.

4. Methodology

4.1. Image-Text Contrastive Learning

We follow CLIP [1] to co-train the image encoder with the text encoder, taking InfoNCE Loss [38] as the Image-Text Contrastive (ITC) loss L_{ITC}, as shown in Fig. 5. Specifically, we maximize the similarity scores of the matched image and text embeddings in terms of batch. We adopt ViT [2] and RoBERTa [39] as the visual and textual skeletons, respectively, and introduce the weight initialization of QA-CLIP.

In particular, to bridge the gap between the data morphology of the title and the user query, we employ a Chinese word-splitting component, Lexical Analysis of Chinese (LAC)[4]. In order to simulate the morphological distribution of the user query, the result of the lexical segmentation is composed into a string using spaces as the splice character. For the case of failed word splitting, the original title is employed. This setting takes effect for all fine-tuning tasks unless otherwise specified.

4.2. Presence-guided Encoder

Video cover images differ from open domain images in that they partially carry cover texts. One option is to outsource the cover text understanding task to an external OCR engine and fuse the cover image with the cover text on the image side in an ALBEF [40] manner. However, the cost of image-text similarity inference becomes expensive. In addition, the image-text contrastive learning becomes highly dependent on the accuracy of the OCR engine, which creates an obstacle for model generalization.

Differently, we are inspired by [15, 16] to design UniCLIP in an OCR-free form. One idea is to guide the ViT to perceive the cover texts during the training process through agent tasks, so that it relies on no module related to the OCR function in the inference process. Since the presence of cover text is
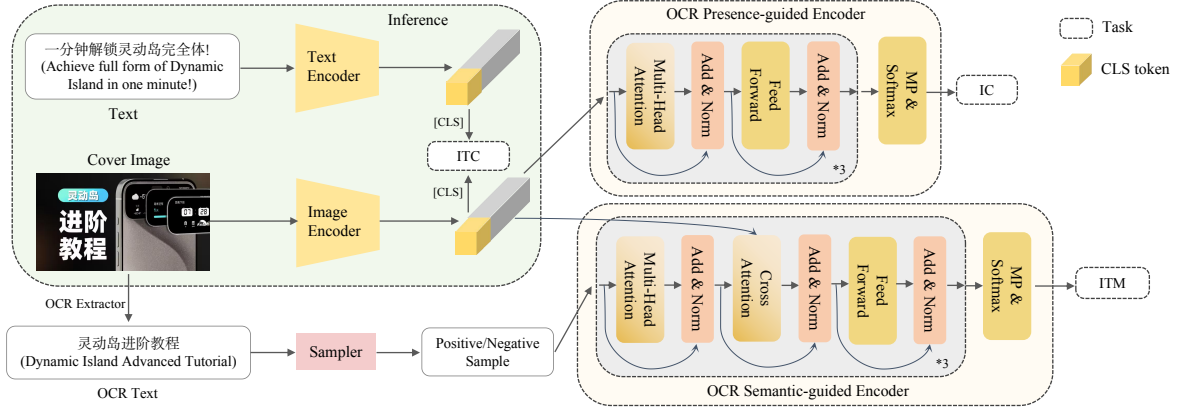
---

[4]https://github.com/baidu/lac

Figure 5: Model structure of UniCLIP. When the model performs inference, only the green area works. ITC stands for "Image-Text Contrastive" (Sec. 4.1), IC stands for "Image Classification" (Sec. 4.2), and ITM stands for "Image-Text Matching" (Sec. 4.3). L_{ITC} and L_{ITM} are computed in the same way as ALBEF[40]. L_{IC} is realised by the binary cross entropy function. L_{ITC} is the core task of the image-text contrastive learning, L_{IC} and L_{ITM} are used for guidance, and Tab. 4 shows the ablation results of the latter two.

uncertain, we propose the presence-guided encoder, where the first agent task of UniCLIP is set as an Image Classification (IC) task: "To determine whether an image carries cover texts".

Specifically, as shown in Fig. 5, the presence-guided encoder takes the output tokens from the last layer of ViT as input. These tokens go through a 3-layer, 8-header Transformer [3] structure, after which they are fed into a MLP layer for predicting the presence or absence of cover texts. The loss in this part is L_{IC}.

4.3. Semantic-guided Encoder

The presence-guided encoder directs the cover image encoder to focus on the cover texts, but does not involve semantic information. We further propose the semantic-guided encoder, which sets the second agent task as an Image-Text Matching (ITM) task: "To determine whether the specified text is consistent with the text on the cover image", encouraging the ViT to incorporate gains from the semantics of the cover texts. This design is motivated by the notion that visual tokens from the ViT contain the semantic information of cover texts, if they are successfully employed to discriminate the consistency of the cover text with a given text.

We design negative samples by nearest neighbor lookup. Take the training on CBVS-5M dataset as an example. First, we adopt RoBERTa-wwm-Base as the encoder and load the checkpoints released by QA-CLIP to encode all the 2.0M valid OCR texts. The Hierarchical Navigable Small World Algorithm (HNSW) [41] is applied to retrieve the Top-K OCR texts that are most semantically similar but not identical to the anchor, and one of them is randomly selected as the negative sample. For covers without OCR texts, only negative samples are considered. For covers with OCR texts, the percentage of positive samples is set to 70%.

The semantic-guided encoder accepts two inputs, i.e., tokens from the last layer of the ViT, and the embeddings of positive or negative samples. As shown in Fig. 5, the module is a 3-layer structure, where each layer consists of a self-attention, a cross-attention with an MLP, and includes residual connections. The embeddings of the samples are updated layer by layer. The loss of this process is L_{ITM}.

4.4. Training and Inference

Pre-training for UniCLIP starts with the checkpoints released by QA-CLIP. Fine-tuning is then performed on the CBVS-5M/10M dataset with a total loss of:

$$L_{total} = \lambda_1 L_{ITC} + \lambda_2 L_{IC} + \lambda_3 L_{ITM}, \tag{1}$$

where λ_1, λ_2 and λ_3 are hyperparameters.

It is worth noting that the fine-tuning of UniCLIP relies on positive samples, negative samples and ground truths related to OCR texts, but the inference process does not rely on any OCR-related components. As shown in Fig. 5, the presence-guided encoder and semantic-guided encoder guide the training of the image-text alignment task in UniCLIP, but do not participate in the inference process. Therefore, UniCLIP infers in a manner consistent with CLIP.

5. Experiments

5.1. Evaluation Metrics

5.1.1. Recall Metrics

Recall is a widely adopted retrieval metric. We report on the Recall (R) and Mean Recall (MR) of the models.

5.1.2. Rank Metrics

Positive-to-Negative Ratio (PNR) measures the consistency of predicted results with ground truth. Formally, PNR is defined as:

$$PNR = \frac{\sum_k \sum_{i,j \in S_k} \mathbb{I}\{y_{ki} > y_{kj}\} \cdot \mathbb{I}\{\hat{y}_{ki} > \hat{y}_{kj}\}}{\sum_k \sum_{i,j \in S_k} \mathbb{I}\{y_{ki} > y_{kj}\} \cdot \mathbb{I}\{\hat{y}_{ki} < \hat{y}_{kj}\}}, \tag{2}$$

where $\mathbb{I}$ is the indicator function. The result of the indicator function is 1 if the internal expression is true, and 0 otherwise. S_k represents the set of all documents under query k. y_{ki}, $\hat{y}_{ki}$ are the true and predicted labels of <image i, text k>, respectively. In particular, we compute PNR only for documents under the same user query.

Normalized Discounted Cumulative Gain (NDCG) is a widely adopted metric in the field of search ranking that encourages higher rankings for better matching documents. Formally, DCG is defined as:

$$DCG@k = \sum_{pos=1}^{k} \frac{2^{y_{pos}} - 1}{log(1 + pos)}, \tag{3}$$

where y_{pos} is the ground truth label of the document at position pos. Further, IDCG is the DCG value of the ideal sort. NDCG is obtained by dividing by IDCG [42]:

$$NDCG@k = \frac{DCG@k}{IDCG@k}. \tag{4}$$

In addition to PNR and NDCG, we also report on the Mean Average Precision (MAP) metric to fully evaluate our model.

5.2. Implementation Details

For the comparison model, we adopt the hyper-parameters suggested in their open source project. For UniCLIP, we follow the model architecture setup in OpenAI CLIP with ViT-B/16 as the visual backbone network. We have tried a number of hyper-parameter combinations and report the best one. The text encoder is the 12-layer architecture of RoBERTa-wwm-Base. Both are implemented by

12-layer 12-head Transformers with 768 encoding dimensions and eventually mapped linearly to 512 dimensions. The vocabulary size of text tokenizer is consistent with CN-CLIP. The initialized weights of the visual encoder and text encoder are from QA-CLIP as described in Sec. 4.4. In addition, the weights of presence-guided encoder and semantic-guided encoder are randomly initialized with normal distribution, and K for the semantic-guided encoder is set to 10. The weights of L_{ITC}, L_{IC}, and L_{ITM} are set to 0.8, 0.1, and 0.1, respectively. The data enhancement module LAC is configured by default. Other hyper-parameter settings that are not detailed are consistent with those of the open-sourced Chinese CLIP.

In the fine-tuning stage, we perform random-size cropping and AutoAugment [43] on the input image. All parameters of the image encoder and text encoder are allowed to be updated. Training is executed on 8 NVIDIA A100 GPUs for 20 epochs, with a learning rate of 2e-5. The maximum length of the text encoder is 12 and the weight decay is 1e-3. We leverage all-gather communications across GPU workers to compute L_{ITC} on the global batch. The batch size is set to 1,520 due to GPU memory limitations. Mixed-precision training is activated. We save checkpoints for each epoch and report results with the highest Positive-to-Negative Ratio (PNR) on CBVS-20K. Due to computational resource constraints, we report the results of training on the CBVS-5M dataset. We simultaneously release the CBVS-10M dataset for subsequent studies.

It needs to be noted that UniCLIP is specifically designed for the short video search domain. Since CBVS is the first and only large-scale dataset in this field that provides video covers and real queries (which is one the contributions of this work), there is a lack of other data benchmarks to compare it with. In addition, although UniCLIP is not limited to Chinese, there is a lack of large-scale video cover-text training datasets for other languages. Therefore, in this work, we only evaluate the performance of each model on the CBVS dataset. We look forward to subsequent work to establish more video cover-text benchmarks for.

5.3. Comparisons

To demonstrate the performance of our proposal and the value of the CBVS dataset, we extensively evaluate advanced Chinese image-text models on CBVS-20K. The competing models include CN-CLIP [7], R2D2 [8], Wukong [9], TaiyiCLIP [32], Ernie-ViL2.0 [44] and AltCLIP [31]. The results are shown in Tab. 3.

The results show that, for example, WuKong, although it achieves competitive performance on other public datasets, it lacks the ability of image-text matching on CBVS-20K. The same conclusion appears in most of the other competitors, e.g., the MR of Taiyi-CLIP$_{ViT-B}$ is only 0.407, which much lower than UniCLIP's 0.692. The generalization performance of models trained on large-scale open domain images is generally low on the CBVS-20K, which demonstrates the domain uniqueness of video cover images. Vision-Language Models that perform well in the open domain do not migrate to the video cover domain as expected.

It is worth noting that although R2D2-250M$_{ViT-L}$'s recall metrics are significantly behind UniCLIP, its ranking metrics are close to those of UniCLIP. In particular, NDCG@1 slightly outperforms UniCLIP with a maximum of 0.789. We infer that the experimental result is due to the fact that R2D2-250M$_{ViT-L}$ employs a more powerful visual architecture and enjoys a training corpus size of 250M. We encourage the incorporation of CBVS-10M into the training corpus on the one hand, and the adoption of ViT-L as a visual skeleton on the other hand, to facilitate further improvement of UniCLIP performance in subsequent studies.

Compared to the pre-trained QA-CLIP, the fine-tuning on the CBVS-5M dataset comprehensively improves the metrics, especially the PNR by 3.67%, and R@1 by 18.25%. Performing fine-tuning on the publicly released CN-CLIP$_{ViT-B}$, consistent findings are observed with a 7.21% improvement in PNR and a 22.66% improvement in R@1. Performing fine-tuning on the R2D2-250M$_{ViT-L}$, significantly higher recall metrics are observed, as well as largely comparable rank metrics. These results demonstrate that fine-tuning on a large-scale cover dataset can improve the performance of the model in the video search domain. In addition, UniCLIP achieves state-of-the-art performance with the highest metrics compared

Table 3
Evaluation on the CBVS-20K dataset. Our proposal achieves SOTA performance.

Mode	Method	Recall Metrics				Rank Metrics				
		R@1	R@5	R@10	MR	PNR	NDCG@1	NDCG@5	NDCG@10	MAP
Zero-shot	CN-CLIP$_{ViT-B/16}$	0.384	0.628	0.704	0.572	2.718	0.768	0.835	0.885	0.764
	CN-CLIP$_{ViT-L/14}$	0.434	0.685	0.756	0.625	2.812	0.773	0.840	0.889	0.775
	WuKong$_{ViT-B/32}$	0.197	0.356	0.439	0.331	2.000	0.702	0.791	0.858	0.712
	WuKong$_{ViT-L/14}$	0.311	0.503	0.583	0.466	2.229	0.739	0.811	0.872	0.738
	Taiyi-CLIP$_{ViT-B}$	0.251	0.445	0.525	0.407	2.142	0.718	0.800	0.861	0.727
	Taiyi-CLIP$_{ViT-L}$	0.269	0.492	0.577	0.446	2.278	0.726	0.805	0.866	0.733
	Ernie-ViL2.0$_{ViT-B}$	0.413	0.660	0.742	0.605	2.759	0.764	0.835	0.886	0.768
	R2D2-23M$_{ViT-L/14}$	0.258	0.407	0.436	0.367	2.312	0.733	0.810	0.868	0.738
	R2D2-250M$_{ViT-L/14}$	0.356	0.512	0.535	0.468	2.829	**0.789**	0.842	0.891	0.775
	AltCLIP$_{ViT-L}$	0.162	0.284	0.336	0.261	1.869	0.669	0.771	0.842	0.701
	QA-CLIP$_{ViT-B/16}$	0.400	0.652	0.724	0.592	2.804	0.774	0.838	0.888	0.770
Fine-tuning	CN-CLIP$_{ViT-B/16}$	0.471	0.721	0.796	0.663	2.914	0.767	0.838	0.888	0.767
	R2D2-250M$_{ViT-L/14}$	0.418	0.605	0.650	0.558	2.934	0.780	0.844	0.891	0.774
	QA-CLIP$_{ViT-B/16}$	0.473	0.711	0.783	0.656	2.907	0.778	0.841	0.890	0.771
	ALBEF-CLIP$_{ViT-B/16}$	0.468	0.731	0.794	0.664	2.906	0.771	0.839	0.889	0.769
	UniCLIP$_{ViT-B/16}$	**0.503**	**0.754**	**0.820**	**0.692**	**3.069**	0.784	**0.846**	**0.893**	**0.779**

Table 4
Results of ablation study of UniCLIP.

L_{IC}	L_{ITM}	Recall Metrics				Rank Metrics				
		R@1	R@5	R@10	MR	PNR	NDCG@1	NDCG@5	NDCG@10	MAP
		0.473	0.711	0.783	0.656	2.907	0.778	0.841	0.890	0.771
✓		0.491	0.747	0.818	0.685	2.991	0.776	0.843	0.890	0.772
	✓	0.499	**0.754**	0.812	0.688	3.006	0.783	0.845	**0.893**	**0.779**
✓	✓	**0.503**	**0.754**	**0.820**	**0.692**	**3.069**	**0.784**	**0.846**	**0.893**	**0.779**

to its competitors and does not introduce additional inference cost compared to the simple and efficient CN-CLIP.

5.4. Ablation Study

We implement versions of the model with and without L_{IC} and L_{ITM}, respectively. The weights of L_{ITC} and L_{IC} (or L_{ITM}) are set to 0.8 and 0.2. Tab. 4 shows the results of the ablation study.

If both the presence-guided encoder and the semantic-guided encoder are removed and the cover text is discarded, the model degenerates into a fine-tuned version of QA-CLIP. The PNR is reduced from 3.069 to 2.907 and the MR from 0.692 to 0.656. In addition, removing any of the two encoders results in varying degrees of degradation in model performance. Removing the presence-guided encoder reduces the PNR of UniCLIP by 2.05%. Besides, removing the semantic-guided encoder reduces the PNR by 2.54%. Interestingly, when only the semantic-guided encoder is employed, the NDCG@1/5/10, MAP, and R@1/5 of the model are basically the same as the final scheme, which indicates that the gain of cover text is mainly in the semantic information. If two encoders are employed at the same time, i.e., the proposed two agent tasks are considered at the same time, the highest metrics are achieved in all aspects. This result is ample proof of the validity of our proposal. We encourage follow-up studies to
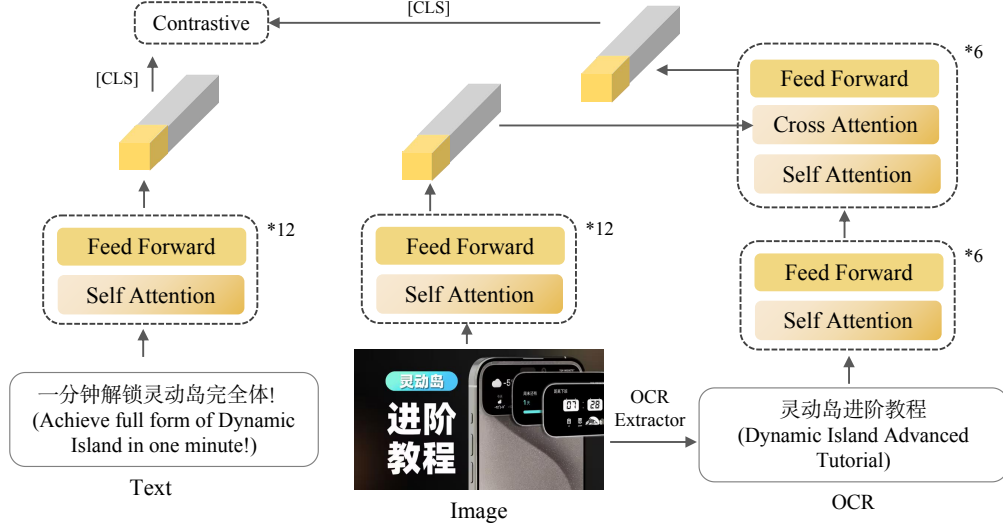
Figure 6: An explicit OCR fusion scheme, which is denoted as ALBEF-CLIP in Tab. 3 and Tab. 5

Table 5
PNR metrics for different OCR texts combinations. $< S_T, S_F >$ stands for that in Eq. 2, $i \in S_T$ and $j \in S_F$.

Model	$< S_T, S_T >$ (11.71%)	$< S_F, S_F >$ (46.51%)	$< S_T, S_F >$ (41.78%)	All (100.00%)
QA-CLIP$_{ViT-B/16}$	3.203	2.722	2.975	2.877
ALBEF-CLIP$_{ViT-B/16}$	**3.375**	2.689	3.051	2.906
UniCLIP$_{ViT-B/16}$	3.331	**2.904**	**3.194**	**3.069**

further generalise our ideas.

5.5. Cover Text Capability Assessment

Since the training process relies on the cover text modality, for a fair comparison, we implement an explicit OCR text fusion scheme, which is denoted as ALBEF-CLIP, as shown in 6. Compared to CLIP, we replace ViT with an ALBEF structure, where the cover image and the cover text go through their respective encoders before passing through a 6-layer Attention-based fusion structure. For the case of missing OCR texts, the prompt word is employed. The results of ALBEF-CLIP are shown in Tab. 3. If the cover text is introduced, but with ALBEF-CLIP rather than our proposal to exploit this modality, the metrics are much lower than UniCLIP in all aspects. We hypothesize that the reason for this is that UniCLIP guides the semantic training of ViT and handles the modality missing problem more consistently, reducing information confusion.

To further evaluate the cover text capability, we categorize the data in CBVS-20K into two main categories according to the presence or absence of cover texts, which are denoted as S_T and S_F, respectively. Tab. 5 demonstrates the PNR metrics for different combinations. Compared to the scheme without cover texts (QA-CLIP), ALBEF-CLIP significantly improves the matching ability for covers with cover texts, increasing the PNR from 3.203 to 3.375. However, for covers without cover texts, the scheme degrades the performance, which may be due to semantic confusions brought about by prompt words.

In comparison, UniCLIP is basically comparable to ALBEF-CLIP for matching between covers with cover texts. This is in line with expectations, as we discard cover texts modalities in our inference, however the very close results are a good indication of the promise of UniCLIP. Besides, UniCLIP

performs much better for both other cases than QA-CLIP and ALBEF-CLIP. For the matching between covers without cover texts, which is most likely to happen, UniCLIP's PNR exceeds that of the fusion scheme by 8.00% and has a lower inference cost. For the hybrid cases, UniCLIP achieves the PNR of 3.194. This suggests that UniCLIP is able to overcome the modality missing problem to some extent and handle cover images with or without cover texts uniformly. Thanks to this, UniCLIP shows the best performance on the full range of data.

6. Conclusion

As the title of the paper indicates, one of the most significant contributions of this work is to establish the first large-scale cover-text benchmark for Chinese short video search scenarios, which provides short video covers and real user queries. we release the largest publicly available Chinese video cover-video title dataset to fill in the lack of cover data for short video search scenarios. We further build a manual fine-labeling video cover-user query benchmark test for short video search domain.

Based on this, we further propose UniCLIP, which integrates the semantic information of cover-texts without increasing the inference cost, is uniform with and without cover text, and has the advantage of online deployment. UniCLIP are proposed to unify cover texts to guide contrastive learning, where the image classification task and the image-text matching task are performed in an OCR-free manner. We are the first to integrate the semantics of cover text into CLIP. UniCLIP has demonstrated significant performance gains and has been deployed in our online system. With more than 700 million daily active users worldwide for short video products represented by TikTok, our benchmark and model have great potential for application. We believe in the ability of CBVS-5M/10M to expand the domain of large-scale Chinese image-text training. Besides, we are pleasantly surprised to observe the model-agnostic potential of UniCLIP.

One the one hand, UniCLIP is language-agnostic and can be generalised to other languages, but relies on the establishment of large-scale open-source cover-text datasets in other languages. We expect subsequent work to build data benchmarks for more languages and apply UniCLIP. On the other hand, there is significant room for exploration remains in balancing the video search domain with the generalized domain. We look forward to the extension of CBVS to downstream tasks such as title generation, as well as inspiration from UniCLIP for performing multi-modal fusion in the CLIP framework.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[5] M. Hendriksen, M. Bleeker, S. Vakulenko, N. van Noord, E. Kuiper, M. de Rijke, Extending clip for category-to-image retrieval in e-commerce, in: European Conference on Information Retrieval, Springer, 2022, pp. 289–303.

[6] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, International Journal of Computer Vision 130 (2022) 2337–2348.

[7] A. Yang, J. Pan, J. Lin, R. Men, Y. Zhang, J. Zhou, C. Zhou, Chinese clip: Contrastive vision-language pretraining in chinese, arXiv preprint arXiv:2211.01335 (2022).

[8] C. Xie, H. Cai, J. Li, F. Kong, X. Wu, J. Song, H. Morimitsu, L. Yao, D. Wang, X. Zhang, et al., Ccmb: A large-scale chinese cross-modal benchmark, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4219–4227.

[9] J. Gu, X. Meng, G. Lu, L. Hou, N. Minzhe, X. Liang, L. Yao, R. Huang, W. Zhang, X. Jiang, et al., Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark, Advances in Neural Information Processing Systems 35 (2022) 26418–26431.

[10] N. Spolaôr, H. D. Lee, W. S. R. Takaki, L. A. Ensina, C. S. R. Coy, F. C. Wu, A systematic review on content-based video retrieval, Engineering Applications of Artificial Intelligence 90 (2020) 103557.

[11] M. Wray, H. Doughty, D. Damen, On semantic similarity in video retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3650–3660.

[12] Z. Zhang, Y. Chen, Z. Ma, Z. Qi, C. Yuan, B. Li, Y. Shan, W. Hu, Create: A benchmark for chinese short video retrieval and title generation, arXiv preprint arXiv:2203.16763 (2022).

[13] L. Nie, L. Qu, D. Meng, M. Zhang, Q. Tian, A. D. Bimbo, Search-oriented micro-video captioning, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3234–3243.

[14] H. Xu, Q. Ye, X. Wu, M. Yan, Y. Miao, J. Ye, G. Xu, A. Hu, Y. Shi, G. Xu, et al., Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks, arXiv preprint arXiv:2306.04362 (2023).

[15] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, S. Park, Ocr-free document understanding transformer, in: European Conference on Computer Vision, Springer, 2022, pp. 498–517.

[16] B. Davis, B. Morse, B. Price, C. Tensmeyer, C. Wigington, V. Morariu, End-to-end document recognition and understanding with dessurt, in: European Conference on Computer Vision, Springer, 2022, pp. 280–296.

[17] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C. L. Zitnick, Microsoft coco captions: Data collection and evaluation server. arxiv 2015, arXiv preprint arXiv:1504.00325 (2015).

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International journal of computer vision 123 (2017) 32–73.

[19] X. Zhan, Y. Wu, X. Dong, Y. Wei, M. Lu, Y. Zhang, H. Xu, X. Liang, Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11782–11791.

[20] T. Abdullah, L. Rangarajan, Image-text matching: Methods and challenges, Inventive Systems and Control: Proceedings of ICISC 2021 (2021) 213–222.

[21] J. Chen, H. Hu, H. Wu, Y. Jiang, C. Wang, Learning the best pooling strategy for visual semantic embedding, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15789–15798.

[22] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, Vse++: Improving visual-semantic embeddings with hard negatives, arXiv preprint arXiv:1707.05612 (2017).

[23] L. Qu, M. Liu, D. Cao, L. Nie, Q. Tian, Context-aware multi-view summarization network for image-text matching, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1047–1055.

[24] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, J. Han, Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 12655–12663.

[25] H. Diao, Y. Zhang, L. Ma, H. Lu, Similarity reasoning and filtration for image-text matching, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 1218–1226.

[26] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10921–10930.

[27] Z. Fu, Z. Mao, Y. Song, Y. Zhang, Learning semantic relationship among instances for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15159–15168.

[28] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al., Grounded language-image pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10965–10975.

[29] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, H. Li, Tip-adapter: Training-free clip-adapter for better vision-language modeling, arXiv preprint arXiv:2111.03930 (2021).

[30] N. Mu, A. Kirillov, D. Wagner, S. Xie, Slip: Self-supervision meets language-image pre-training, in: European Conference on Computer Vision, Springer, 2022, pp. 529–544.

[31] Z. Chen, G. Liu, B.-W. Zhang, F. Ye, Q. Yang, L. Wu, Altclip: Altering the language encoder in clip for extended language capabilities, arXiv preprint arXiv:2211.06679 (2022).

[32] J. Zhang, R. Gan, J. Wang, Y. Zhang, L. Zhang, P. Yang, X. Gao, Z. Wu, X. Dong, J. He, et al., Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence, arXiv preprint arXiv:2209.02970 (2022).

[33] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, W. Y. Wang, Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4581–4591.

[34] S. Zhang, Z. Tan, J. Yu, Z. Zhao, K. Kuang, J. Liu, J. Zhou, H. Yang, F. Wu, Poet: Product-oriented video captioner for e-commerce, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1292–1301.

[35] T. Gan, Q. Wang, X. Dong, X. Ren, L. Nie, Q. Guo, Cnvid-3.5 m: Build, filter, and pre-train the large-scale public chinese video-text dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14815–14824.

[36] C. Lei, S. Luo, Y. Liu, W. He, J. Wang, G. Wang, H. Tang, C. Miao, H. Li, Understanding chinese video and language via contrastive multimodal pre-training, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2567–2576.

[37] J. Lin, R. Men, A. Yang, C. Zhou, M. Ding, Y. Zhang, P. Wang, A. Wang, L. Jiang, X. Jia, et al., M6: A chinese multimodal pretrainer, arXiv preprint arXiv:2103.00823 (2021).

[38] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[40] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S. C. H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Advances in neural information processing systems 34 (2021) 9694–9705.

[41] Y. A. Malkov, D. A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, IEEE transactions on pattern analysis and machine intelligence 42 (2018) 824–836.

[42] C. D. Manning, An introduction to information retrieval, Cambridge university press, 2009.

[43] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q. V. Le, Autoaugment: Learning augmentation strategies from data, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 113–123.

[44] B. Shan, W. Yin, Y. Sun, H. Tian, H. Wu, H. Wang, Ernie-vil 2.0: Multi-view contrastive learning for image-text pre-training, arXiv preprint arXiv:2209.15270 (2022).