CIKM MMSR '24

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

Xiangshuo Qiao[1,†],Xianxin Li[1,†],Xiaozhe Qu[1],JieZhang [1,*],Yang Liu[1],YuLuo [1], Cihang Jin[1] andJinMa [2]

1 Tencent PCG, Beijing, China
2 University of Science and Technology of China, Hefei, Anhui, China
∗Corresponding author.
†These authors contributed equally.

Background

1. Why do we need video covers?

2. However, there is a significant difference between images for pre-training and video covers.

Open Domain Images

Short Video Cover Images

- 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. However, there is a phenomenon of missing video covers, and existing models have not taken this issue into consideration.

Overview of Our Work

1. 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**.

2. We build a **manual fine-labeling image-text benchmark** test for Chinese short video search scenarios, containing real user queries from browser logs.

3. 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.

dataset, code and checkpoints are available at
https://github.com/QQBrowserVideoSearch/CBVS-UniCLIP

Dataset Construction ：

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 ***Relevance Level***: *2*	***OCR text***: 小鹏G6 买它! Go for the Xpeng G6! ***Relevance Level***: 1	***OCR text***: 尼古拉·特斯拉究竟有多强? Just how brilliant was Nikola Tesla? ***Relevance Level***: 0	***OCR text***: 无 Null ***Relevance Level***: *2*	***OCR text***: 无 Null ***Relevance Level***: 1	***OCR text***: 鸡胸肉黄瓜丁 Diced Chicken Breast with Cucumber ***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!

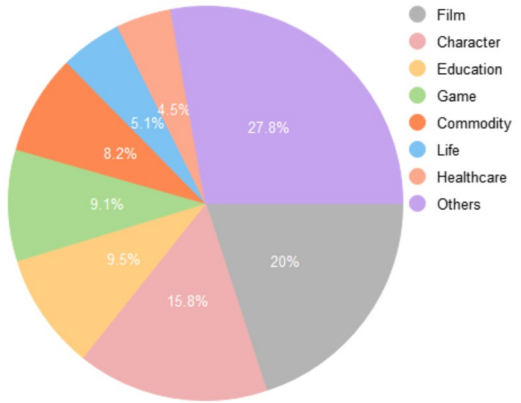
OCR text
故宫保和殿
The Hall of Preserving Harmony in the Forbidden City

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	✓

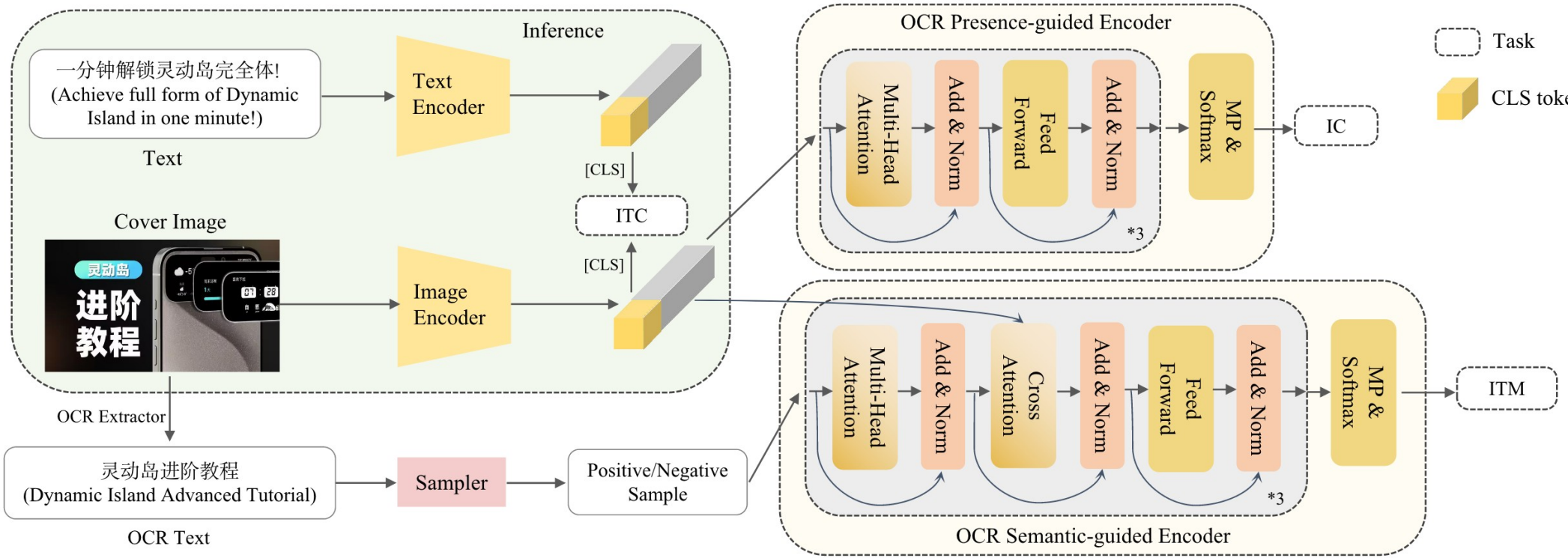
Top:
Presentation of **CBVS-20K data**.

Bottom:
Presentation of **CBVS-5M/10M data.**



Distribution of **categories** of **user queries** in CBVS-20K.

Model Construction ： UniCLIP



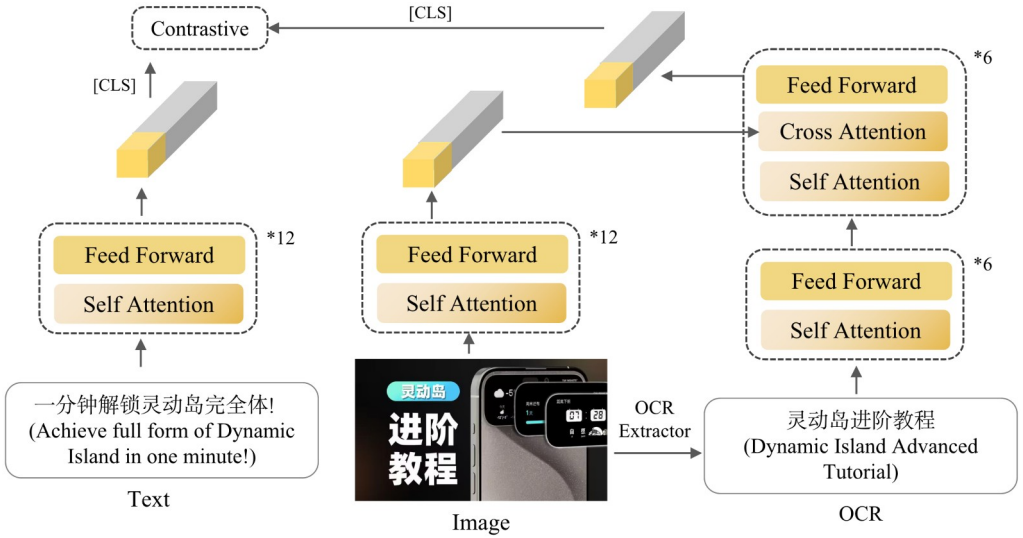
Model structure of UniCLIP. When the model performs inference, only the green area works.

ITC stands for ”Image-Text Contrastive” IC stands for ”Image Classification” , and ITM stands for ”Image-Text Matching” . L_{ITC} and L_{ITM} are computed in the same way as ALBEF. 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 LI_{TM} are used for guidance.

Experiment

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**



- An explicit OCR fusion scheme, which is denoted as **ALBEF-CLIP**

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

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**

- Results of **ablation study** of UniCLIP

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**

- PNR metrics for **different OCR texts combinations**

Summary

1. We establish the **first large-scale** **c**over-text **b**enchmark for Chinese short **v**ideo **s**earch 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
2. 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
3. We believe CBVS could further facilitate advanced research in short video search scenarios

https://github.com/QQBrowserVideoSearch/CBVS-UniCLIP