Youku Dense Caption: A Large-scale Chinese Video Dense Caption Dataset and Benchmarks

阿里云-计算平台-搜索算法团队

01 Motivation

Video Content Explosion

视频内容呈爆炸式增长，全球视频平台日均上传量达数百万小时，亟需视频理解工具。

中国作为全球最大视频消费市场之一，对本土化视频理解工具需求迫切。

▶ 01

Dataset Limitations

现有大规模视频字幕数据集多以英语为中心，如 MSR- VTT、ActivityNet，中文数据匮乏。

缺乏大规模中文细粒度视频数据集，限制了中文多模态模型开发。

▶ 02

Cultural Bias

现有数据集文化背景单一，缺乏中国文化元素，不利于文化相关视频理解。

中国文化独特性要求构建本土化数据集，以支持文化相关模型训练。

▶ 03

数据集	类型	语言	# 视频数	# 标注数	片段平均长度	领域
MSR-VTT[1]	视频描述	英文	10k	200k	14.8s	开放域
MSVD[2]	视频描述	英文	2k	85k	7s	开放域
VATEX[4]	视频描述	中英	41k	826k	10s	开放域
Youku-mPLUG[16]	视频描述	中文	80k	80k	54.2s	开放域
ActivityNet Captions[5]	密集视频描述	英文	20k	100k	36s	开放域
YouCook2[3]	密集视频描述	英文	2k	15k	7.7s	烹饪
ViTT[7]	密集视频描述	英文	8k	12k	-	开放域
Youku Dense Caption	密集视频描述	中文	31k	311k	8.1s	开放域

02 Key Contributions

Scale & Coverage

Youku Dense Caption 数据集包含 **31k** 个视频，**311K** 条字幕，总时长超 **748** 小时，是最大中文细粒度视频数据集。

覆盖 **11** 个粗粒度、**84** 个细粒度类别，涵盖生活、体育、文化等多领域。

Model Validation

使用 ViCLIP、Qwen- VL 等前沿模型验证**数据集有效性**，显著提升模型性能。

实验证明数据集在**跨领域迁移**、**模型训练**方面具有重要价值。

Benchmark Tasks

提供**检索**、**定位**、**生成**三大任务基准，涵盖部分相关视频检索、时间定位、字幕生成等子任务。

为模型评估提供全面、标准化测试平台，推动多模态模型发展。

03 Data Construction

Youku Platform

数据源自优酷平台，中国领先视频分享网站，拥有海量中文视频资源。

筛选高质量视频，排除分屏、无意义内容，确保数据纯净性。

Annotation Guidelines

精心设计的详细标注标准，对**主体、主体的描述、动作、场景**都有具体的标注要求。

对标注的**文字内容**也有严格的质量要求。

Annotation Process

人工标注视频片段，平均每个片段时长 **8.1 秒**，标注精细度高。

标注团队由**1个leader和10个**标注员组成，共耗时**2个月**，验收标准达到**97%**以上准确率。

A Annotation Guidelines

Objective The goal is to create detailed and coherent descriptions for the main scenes presented in the provided video, including specific start and end timestamps for each description. The final annotation files are released in a video format.

Description Structure Each description must adhere to the following structural components:

(1) Subject: Use general identifiers to describe subjects, avoiding specific names. Acceptable terms include "man", "woman", "young man", "boy", "child", "girl", "doll", "elderly person", and "groom". Avoid proper nouns or distinct identifiers, such as "Yang Mi", "Cangnan Mountain Immortal", or "gangster". (2) Description of the Subject: Include adjectives that pertain to the subject's color, size, state, facial expressions, and emotions. (3) Description of Actions (Optional): Describe the actions performed by the subject, indicating their nature and purpose. (4) Description of the Scene (Optional): Specify the type of location, such as "school", "plaza", "playground", or "shopping mall", while avoiding specific site identifiers like "Mingzhu elementary school" or "Beijing city".

Annotation Requirements (1) Each description sentence must contain a minimum of ten words and consist of at least three sentences for a video. (2) Descriptions should provide specific details for each component (subject, action, and scene) and articulate events occurring in the video while maintaining grammatical and semantic coherence. (3) Common referential terms (e.g., "he", "she") may be used to reference previously introduced subjects. (4) A coherent narrative that encapsulates the primary content of the video should be produced, along with descriptions of noteworthy but non-primary scenes. (5) Annotations for overlapping video timestamps may be permissible. If multiple scenes are present in a single video, they should be described separately.

04 Dataset Statistics

Size & Language

与 MSR- VTT、ActivityNet、YouCook2 等数据集对比，Youku Dense Caption 在规模和中文数据量上优势明显。

弥补中文视频数据集空白，为**中文多模态研究**提供有力支持。

Key Features

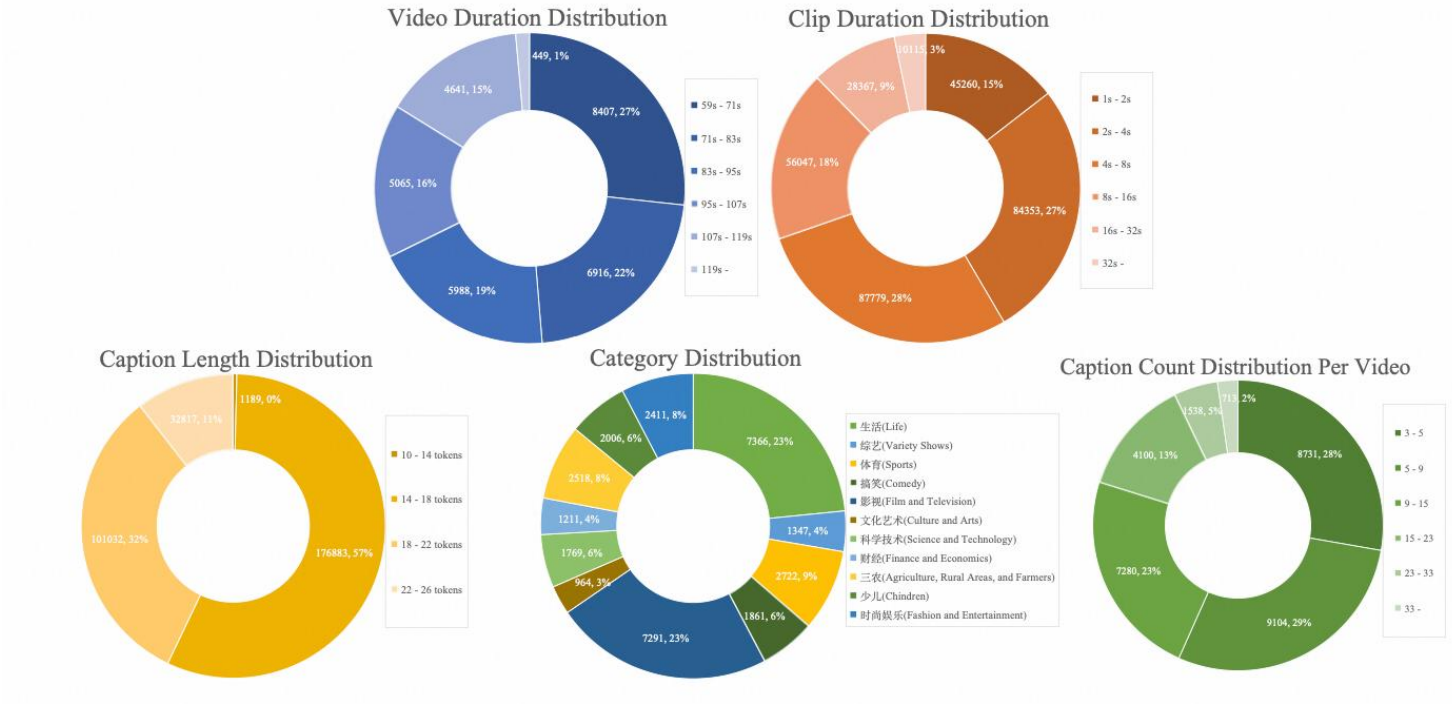
视频时长均衡，多集中在 **1- 2 分钟**，适合多种任务需求。

字幕短小精悍，平均 **17.9 字符**，便于模型理解和生成。

Cultural Relevance

包含**中国文化特色元素**，如新年祝福、茶艺展示等，提升文化相关性。

为文化相关视频理解任务提供丰富语料，推动文化研究。



05 Cultural Uniqueness

Vid: 1089493207 TS: 0.16-6.59	Vid: 1272141709 TS: 0.00-18.34	Vid: 1182715104 TS: 29.48-31.46	Vid: 1255985397 TS: 60.20-100.87	Vid: 1108754691 TS: 26.69-30.59	Vid: 1480996201 TS: 37.04-38.68
戴着厨师帽的男人在摆满月饼前忙活着 A man wearing a chef's hat is busy preparing a full moon cake.	舞台上一个穿着红色衣服的男人正在给大家拜年 On stage, a man wearing red clothes is greeting everyone on New Year's Eve.	中年女子慌忙的从兜里掏出一个红包 The middle-aged woman hurriedly took out a red envelope from her pocket.	穿着黄色衣服的光头男子在表演相声 A bald man wearing yellow clothes is performing cross-talk.	一个身穿灰色衣服的男人正在手拿剪刀剪粽子 A man in gray is cutting Zongzi with scissors.	几个大人手里拿着新年摆件在恭贺大家 A few adults are holding New Year decorations in their hands to congratulate everyone.

Video ID 1473367532						
Timestamps	0.00-4.00	4.00-13.73	13.73-19.86	19.86-30.40	30.40-58.52	63.99-86.79
Annotations	晶莹剔透的糖制品看着让人垂涎欲滴 The crystal-crysted sugar products look like they're salivating.	穿着白色大褂带着白色帽子的大爷站在中间 He's standing in the middle of a big white coat with a white hat.	一块块带着文字和花纹的月饼被制作出来 A mooncake with words and strips was made.	面团经过两个人的揉搓然后放进精致的模具中加工 The face passes through the twigs of two men and is processed in fine molds.	月饼经过专业烤箱一段时间的高温烘焙逐渐成型 The mooncakes have evolved through a period of hot baking in a professional oven.	男士拿起月饼咬了一口满嘴流露出芳香 The man took the mooncake and bit it with his mouth full of fragrance.

Traditional Festivals

中秋节月饼、春节传统习俗等文化元素在数据集中得到体现。

通过标注和数据展示，突出**中国文化特色**，吸引国际关注。

Cultural Terms

图 3 展示标注中包含的文化术语，如**“红包”**等，凸显文化细节。

为多模态模型训练提供文化背景知识，提升**文化理解能力**。

Significance

解决现有数据集**文化偏差**问题，为多模态模型提供本土化训练数据。

推动中国文化在全球多模态研究领域的传播和应用。

06 Benchmark Tasks

Partially Relevant Video Retrieval (PRVR)

算法 3.1 跨视频 PRVR 基准设置

输入： 数据集 $D = \{(t_i, v_i), (t_j, v_i), \cdots, (t_k, v_j)\}$，其中 t_i 为片段级标注。
输出： 基准 $B = \{(t_i, V_i), \cdots, (t_j, V_j)\}$，其中 V_i 为视频 ID 集合。
for 每个标注文本 t_i **do**
    $\text{embed}(t_i) = \text{Text Encoder}(t_i)$。
    创建与 t_i 的嵌入相似度 > 0.9 的所有标注文本集合 T_i：

$$T_i = \{t_j \mid \text{sim}(\text{embed}(t_i), \text{embed}(t_j)) > 0.9\}$$

    从对 (t_j, v_j) 中，若 $t_j \in T_i$，则将 v_j 添加到集合 V_i：

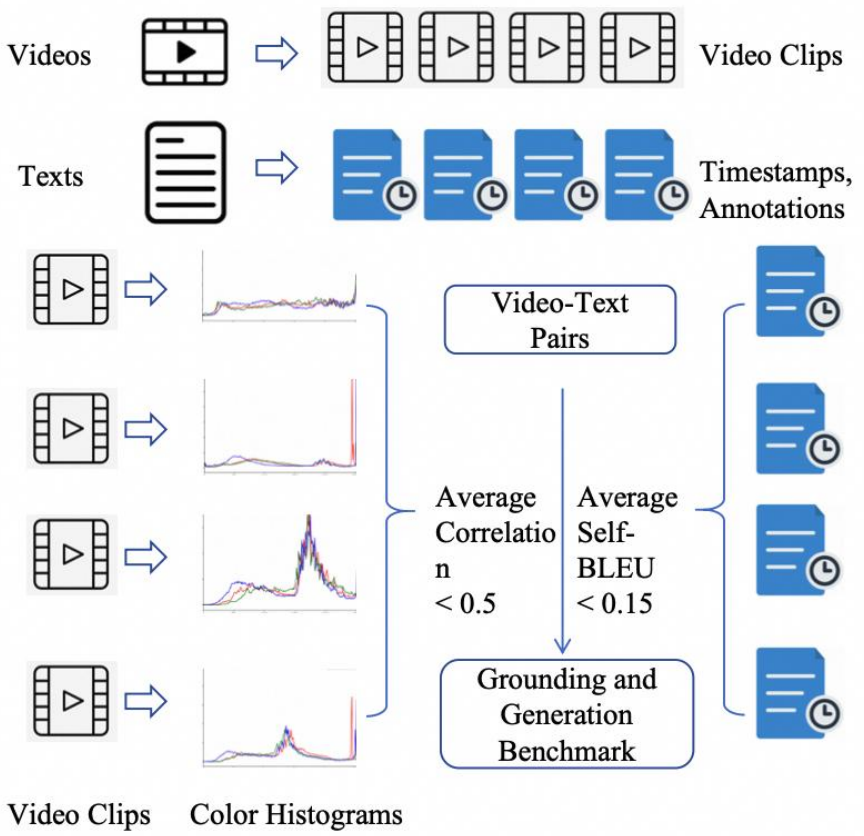
$$V_i = \{v_j \mid (t_j, v_j) \in D \text{ 且 } t_j \in T_i\}$$

end for
for 每一对集合 (V_i, V_j) **do**
    if $\text{len}(V_i) > 1$ 且 $\text{len}(V_j) > 1$ **then**
        计算交并比（IoU）：

$$\text{IoU}(V_i, V_j) = \frac{|V_i \cap V_j|}{|V_i \cup V_j|}$$

        if $\text{IoU}(V_i, V_j) > 0.7$ **then**
            若 $\text{len}(V_i) > \text{len}(V_j)$，则移除 (t_j, V_j)；否则，移除 (t_i, V_i)。
        end if
    end if
end for
输出： 剩余的对 (t_i, V_i)。

Grounding and Generation



06 Benchmark Tasks

Word Comparison

通过算法自动筛除视频**内容相似**、标注**文本相似**的样本。

对比筛除前后的词频变化，无论是极差、标准差都显著降低，词汇使用**分布更加均匀**。

使用筛除后的数据集进行Benchmark任务，实验结果会更具参考价值。

指标	原始标注	处理后标注
词频极差	0.4726	0.4431
词频标准差	0.1928	0.1824

频率排名	筛选前 高频标注词	筛选前 频率	筛选后 高频标注词	处理后 频率
25	一边 (One side)	0.48611	黄色 (Yellow)	0.45717
75	男士 (Gentleman)	0.19027	西装 (Suit)	0.18388
125	手指 (Finger)	0.10943	狮子 (Lion)	0.11035
175	西服 (Suit)	0.07742	一把 (A handful)	0.08013
225	外面 (Outside)	0.06525	一台 (One unit)	0.06621
275	房间内 (Inside the room)	0.05288	美味 (Delicious)	0.05445
325	黯黑 (Dark)	0.04419	牛仔裤 (Jeans)	0.04496
375	毛衣 (Sweater)	0.03774	加入 (Join)	0.03805
425	安静 (Quiet)	0.03356	赛场 (Arena)	0.03444
475	走来走去 (Walk around)	0.02899	长裙 (Long dress)	0.03001
525	桌面上 (On the desk)	0.02641	点头 (Nod)	0.02681
575	墙边 (By the wall)	0.02383	阳光 (Sunshine)	0.02424
625	电动车 (Electric vehicle)	0.02164	愤怒 (Angry)	0.02248
675	撕咬 (Tear with teeth)	0.02023	大厅 (Hall)	0.02063
725	花白 (Grizzled)	0.01881	花纹 (Pattern)	0.01928
775	剪刀 (Scissors)	0.01752	轮胎 (Tire)	0.01836
825	工作服 (Work clothes)	0.01642	笔直 (Straight)	0.01702
875	开着车 (Driving)	0.01546	慢悠悠 (Slowly)	0.01588
925	卡其色 (Khaki)	0.01443	头顶 (Top of head)	0.01475
975	刷子 (Brush)	0.01353	展厅 (Exhibition hall)	0.01403

07 Experiment Highlights

Models	Language	text2video			video2text		
		Top 1 Acc	Top 5 Acc	Top 10 Acc	Top 1 Acc	Top 5 Acc	Top 10 Acc
ViCLIP (Wang et al., 2024b)	English	26.78	47.53	56.33	56.38	76.95	84.39
InternVideo2 (Wang et al., 2024c)	English	26.56	47.46	57.17	47.87	70.21	79.07
ViCLIP	Chinese	0.91	4.71	7.92	2.12	8.86	12.41
InternVideo2	Chinese	1.31	3.90	6.94	1.77	4.60	7.80
mPLUG-Owl3 (Ye et al., 2024)	Chinese	2.04	6.57	10.55	0.35	2.48	3.19
CLIP Radford et al. (2021)	English	1.05	4.34	8.18	2.12	9.92	13.82
DFN5B-CLIP (Fang et al., 2024)	English	6.61	15.49	24.00	16.31	36.87	47.51
Chinese-CLIP (Yang et al., 2022)	Chinese	**31.78**	**55.49**	**65.72**	**41.13**	**64.53**	**70.92**

Features	Dim	Language	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	mAP@avg
CLIP	512	English	6.38	2.16	11.70	3.24	4.35
ViCLIP	768	English	6.23	2.17	11.40	3.31	4.30
Chinese CLIP	512	Chinese	**13.89**	**4.95**	**20.10**	**5.91**	**7.76**

Models	Vision Encoder	Language Decoder	BLEU-4	METEOR	Rouge-L	CIDEr	Bert Score
InternVideo2-Chat-8B	Internvideo2-1B	Mistral-7B	0.35	6.92	7.25	4.97	58.69
MiniCPM-V-2_6	SigLip-400M	Qwen2-7B	0.57	12.86	12.14	3.66	62.72
InternVL2-8B	Intern-ViT-6B	InternLM2.5-7B	1.09	13.62	13.20	9.19	65.34
Qwen2-VL-7B-Instruct	DFN	Qwen2-7B	**1.51**	**13.84**	**15.10**	**19.32**	**66.68**

Retrieval

采用三类模型：**视频检索**（ViCLIP、Intern-Video2）、**视频生成**（mPLUG-Owl3）、**CLIP类**。

English代表直接中文标注翻译成英文。

Chinese-CLIP效果最好说明**原生中文视频检索模型**重要性。

Grounding

采用CLIP类模型抽取feature后通过**Moment-DETR**训练时刻定位模型。

仍然是Chinese-CLIP效果最好。

Generation

后三个模型效果更好，因为采用了视频抽帧多图片表示的训练模式。

07 Experiment Highlights

Ablation Study

Training Set		Testing Set	video2text			text2video		
% Youku-mPLUG	% YDC	Youku-mPLUG	R1	R5	R10	R1	R5	R10
0%	0%	Validation	1.03	2.94	4.78	0.40	1.32	2.24
100%	0%	Validation	2.42	9.86	16.55	1.90	8.36	14.12
0%	13%	Validation	0.46	2.30	4.09	0.34	1.55	3.17
0%	100%	Validation	3.46	11.18	17.30	3.46	11.36	17.07
100%	13%	Validation	2.99	9.91	16.03	2.24	8.24	13.26
100%	100%	Validation	5.88	17.12	24.62	5.47	15.62	23.47

Improving Retrieval Performance

在Youku-mPLUG的验证集进行实验，选择**ViCLIP**模型做微调。

YDC的13%数量级和Youku-mPLUG训练集相同，同数量级微调效果下降是由于**OOD**。

模型效果随着YDC数据量提升**同步提升**。

Training Set		VATEX Val				Generation Benchmark on YDC			
% VATEX	% YDC	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
0%	0%	16.00	22.52	39.01	33.06	0.94	9.39	12.57	15.87
100%	0%	28.16	30.17	48.05	53.12	2.41	14.06	18.36	26.88
50%	50%	27.71	29.81	47.72	51.52	4.71	17.31	25.23	42.41
0%	100%	22.77	26.50	43.74	35.64	4.86	17.50	25.51	43.06
100%	50%	28.42	30.14	47.98	52.82	4.63	17.34	25.28	41.51
100%	100%	**29.57**	**30.49**	**48.53**	**54.06**	**4.91**	**17.54**	**25.68**	**44.32**

Improving Generation Performance

在VATEX和YDC的验证集进行实验，选择**Qwen2-VL-2B**模型做微调。

YDC和VATEX数量级**接近**，交叉训练效果下降明显同样由于**OOD**。

模型效果随着YDC数据量提升**同步提升**。

08 Applications

Video Search Optimization

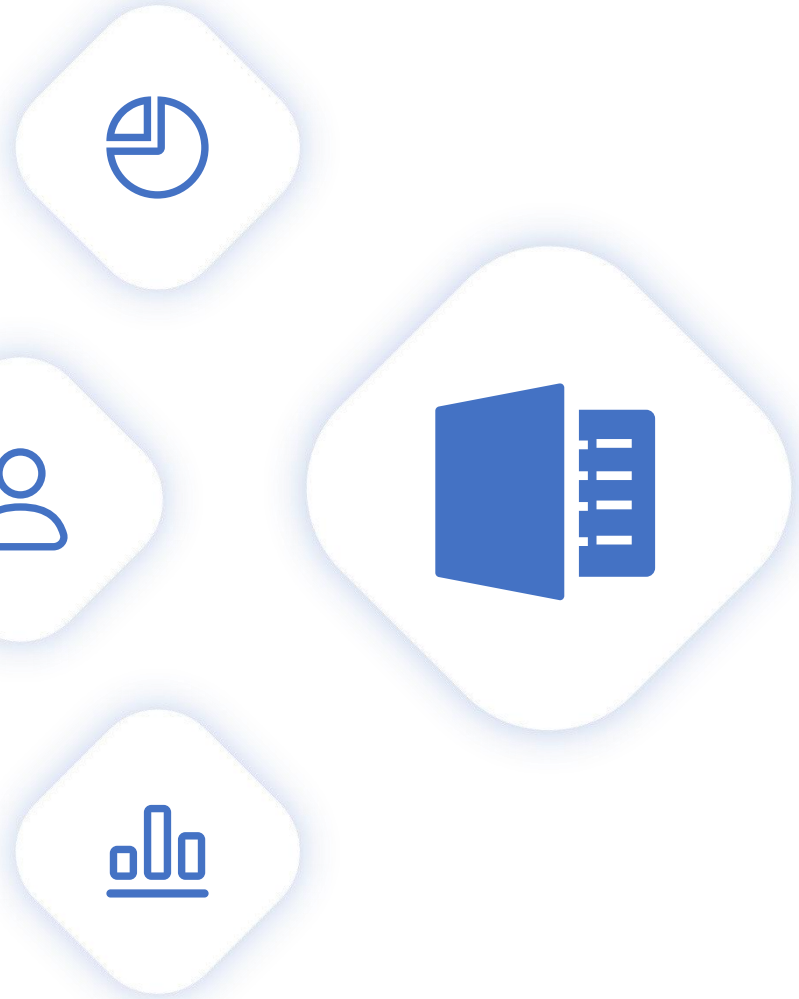
优化视频搜索算法，提高**搜索准确性**和**用户体验**。
为视频平台和搜索引擎提供技术支持，**拓展应用领域**。

Content Moderation

助力视频内容审核，提升**审核效率**和**准确性**。
为视频平台内容管理提供有力工具，保障平台**健康发展**。

Accessibility

支持自动**生成字幕**，提升视频可访问性。
为视障人士和非母语观众提供便利，**拓展应用范围**。



09 Conclusion & Future Work

Gap Filling

Youku Dense Caption 数据集填补了**中文视频 – 细粒度语言**数据集空白，具有重要意义。

高质量标注和丰富文化元素为**多模态研究**提供有力支持。

Future Directions

计划扩展至**长视频领域**，探索多模态大语言模型集成。

研究**跨语言迁移学习**，推动多模态技术全球化发展。

WWW.ALIYUN.COM