

---

# MMVText: A Large-Scale, Multidimensional Multilingual Dataset for Video Text Spotting

---

Anonymous Author(s)

Affiliation

email

## Abstract

1 Video text spotting is crucial for numerous real application scenarios, but most  
2 existing video text reading benchmarks are challenging to evaluate the performance  
3 of advanced deep learning algorithms due to the limited amount of training data  
4 and tedious scenarios. To address this issue, we introduce a new large-scale bench-  
5 mark dataset named Multidimensional Multilingual Video Text (MMVText), the  
6 first large-scale and multilingual benchmark for video text spotting in a variety of  
7 scenarios. There are mainly three features for MMVText. Firstly, we provide **510**  
8 videos with more than **1,000,000** frame images, four times larger than the existing  
9 largest dataset for text in videos. Secondly, our dataset covers 30 open categories  
10 with a wide selection of various scenarios, *e.g.*, *life vlog*, *sports news*, *automatic*  
11 *drive*, *cartoon*, *etc.* Besides, caption text and scene text are separately tagged for the  
12 two different representational meanings in the video. The former represents more  
13 theme information, and the latter is the scene information. Thirdly, the MMVText  
14 provides multilingual text annotation to promote multiple cultures live and commu-  
15 nication. In the end, a comprehensive experimental result and analysis concerning  
16 text detection, recognition, tracking, and end-to-end spotting on MMVText are pro-  
17 vided. We also discuss the potentials of using MMVText for other video-and-text  
18 research. The dataset and code can be found at [github.com/weijiawu/MMVText](https://github.com/weijiawu/MMVText).

19 

## 1 Introduction

20 Text spotting [23, 13] has received increasing attention due to its numerous applications in computer  
21 vision, *e.g.*, document analysis, image-based translation, image retrieval [35, 28], etc. With the advent  
22 of deep learning and abundance in digital data, reading text from images has made extraordinary  
23 progress in recent years with a lot of great public datasets [9, 14, 6] and algorithms [45, 55, 24, 22]. By  
24 contrast, video text spotting almost remains at a standstill for the lack of large-scale multidimensional  
25 practical datasets, which limited numerous applications of video text, *e.g.*, video understanding [40],  
26 video retrieval [8], video text translation, and license plate recognition [1], etc.

27 Most existing algorithms [55, 45, 20] in text detection and recognition deal with only static frames.  
28 Therefore, one intuitive drawback of these approaches is that they do not necessarily work well  
29 in the video domain, while at the same time they do not take advantage of the extra information  
30 present in the video (*e.g.*, tracking already detected regions). Moreover, the quality of the image  
31 is generally worse than static images, due to motion blur and out of focus issues, while video  
32 compression might create further artefacts. Due to these interferences, methods designed for still  
33 images, may fail to obtain reliable detection and recognition results when applied to a video frame.  
34 Most importantly, these methods based on image-level can not obtain text tracking information in  
35 video. However, spatio-temporal information in video is vital for a number of real-world applications.  
36 For example, video understanding and video caption translation all require temporal text information  
37 in sequential frames. There have been a few previous works [50, 48] in the community for attempting

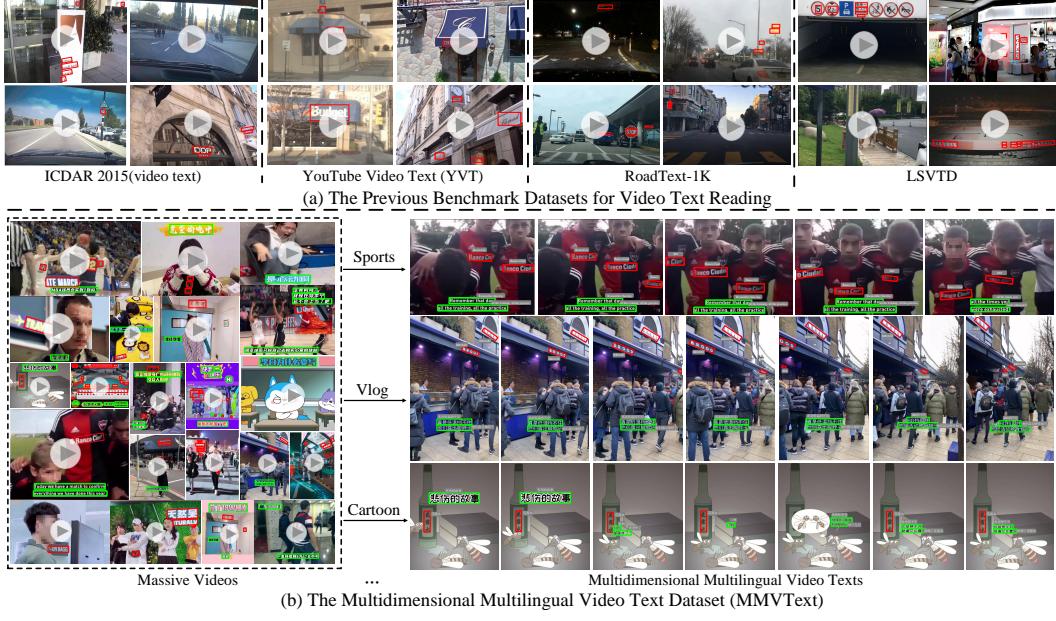


Figure 1: **Example Sequences and Annotations.** Unlike the previous benchmarks, our MMVText contains a wide variety of scenarios and multi-languages. The caption text and scene text are separately tagged for the two different representational meanings.

38 to develop text reading in videos, and there is a handful of datasets [30, 15] that support the research.  
 39 ICDAR2015 (Text in Videos) [14], as one of the common datasets, was introduced during the ICDAR  
 40 Robust Reading Competition in 2015 and mainly includes a training set of 25 videos (13,450 frames  
 41 in total) and a test set of 24 videos (14,374 frames in total). The videos were categorized into  
 42 seven scenarios: walking outdoors, searching for a shop in a shopping street, browsing products in  
 43 a supermarket, etc. YouTube Video Text (YVT) [30] dataset harvested from YouTube, contains 30  
 44 videos (13,500 frames in total), 15 for training, and 15 for testing. The text content in the dataset  
 45 can be divided into two categories, overlay text (*e.g.*, captions, songs title, logos) and scene text (*e.g.*,  
 46 street signs, business signs, words on shirt). RoadText-1K [31] are sampled from BDD100K [53],  
 47 includes 700 videos (210,000 frames) for training and 300 videos for testing. The texts in the  
 48 dataset are all obtained from driving videos and match for driver assistance and self-driving systems.  
 49 LSVTD [5] includes 100 text videos, 13 indoor (*e.g.*, bookstore, shopping mall) and 9 outdoor (*e.g.*,  
 50 highway, city road) scenarios. The existing video text benchmarks are limited by the amount of  
 51 training data (less than 300k frames) and tedium data scenarios, as shown in Figure. 1 (a). There are  
 52 only a few outdoor scene text videos with 13k frames in ICDAR2015 (video text). Similar situation  
 53 for YVT, RoadText-1k and LSVTD, the training set is limited and the dataset scenarios are tedious.  
 54 This makes it difficult to evaluate the effectiveness of more advanced deep learning models.

55 To address this issue, our work intends to contribute a large-scale, multidimensional multilingual  
 56 benchmark dataset (MMVText) to the community for developing and testing video text reading  
 57 systems that can fare in a realistic setting. Our dataset has several advantages. Firstly, the large  
 58 training set (*i.e.*, 1,010,848 video frames) enables the development of deep design specific for video  
 59 text spotting. Secondly, MMVText is a multilingual multidimensional dataset. Abundant videos  
 60 in various scenarios (*e.g.*, driving, street view, news reports, cartoon) are provided for representing  
 61 real-world scenarios, as shown in Figure. 1 (b). Thirdly, caption and scene text are separately tagged  
 62 for the two different representational meanings in the video. This is in favor of other tasks, such as  
 63 video understanding and video retrieval. The main contributions of this work are three folds:

- 64 • We propose a large-scale, multidimensional, and multilingual video text spotting benchmark  
 65 named MMVText. The proposed dataset span various video scenarios, text types, multi-stage  
 66 tasks and is four times the existing largest dataset.
- 67 • Caption text and scene text are separately tagged for the two different representational  
 68 meanings in the video. Based on the previous works [35, 16], this favors other tasks  
 69 theoretically, such as video understanding, video retrieval, and video text translation.

- 70 • We evaluate the current state-of-the-art techniques for scene text detection, recognition, text  
71 tracking, and end-to-end video text spotting. Besides, a thorough analysis of performance  
72 on this dataset is provided.

73 **2 Related Work**

74 **2.1 End-to-End Text Spotting**

75 For image-level text spotting, various methods [17, 10, 24] based on deep learning have been proposed  
76 and have improved the performance considerably. Li et al. [17] proposed the first end-to-end trainable  
77 scene text spotting method. The method successfully uses a RoI Pooling [32] to joint detection  
78 and recognition features via a two-stage framework. Liao et al. [24] propose a Mask TextSpotter  
79 which subtly refines Mask R-CNN and uses character-level supervision to detect and recognize  
80 characters simultaneously. However, these methods based on the static image can not obtain temporal  
81 information in the video, which is essential for some downstream tasks such as video understanding.  
82 Compared to text spotting in a static image, video text spotting methods are rare. Yin et al. [52]  
83 provides a detailed survey, summarizes text detection, tracking and recognition methods in video  
84 and their challenges. Wang et al. [46] introduced an end-to-end video text recognition method.  
85 Multi-frame text tracking is employed through associations of texts in the current frame and several  
86 previous frames to obtain final results. Cheng et al. [5] propose a video text spotting framework by  
87 only recognizing the localized text one-time. Nguyen et al. [30] improves detection and recognition  
88 performance by temporal redundancy and linearly interpolate to recover missing detection results.  
89 Rong et al. [34] tracked video text using tracking-by-detection. An MSER detector was used  
90 to locate scene text character (STC), which was used as a constraint to optimize the trajectory  
91 search. To promote text spotting in the video, we attempt to establish a standardized evaluation and  
92 benchmark (MMVText), covering various open scenarios and multilingual text annotation.

93 **2.2 Text Spotting Datasets for Static Images**

94 The various and practical benchmark datasets [14, 41, 15, 6] contribute to the huge success of  
95 scene text detection and recognition at the image level. ICDAR2015 [14] was provided from the  
96 ICDAR2015 Robust Reading Competition, which is commonly used for oriented scene text detection  
97 and spotting. Google glasses capture these images without taking care of position, so text in the  
98 scene can be in arbitrary orientations. ICDAR2017MLT [29] is a large-scale multilingual text dataset,  
99 which is composed of complete scene images which come from 9 languages, and text regions in this  
100 dataset can be in arbitrary orientations, so it is more diverse and challenging. ICDAR2013 [15] is  
101 a dataset proposed in the ICDAR 2013 Robust Reading Competition, which focuses on horizontal  
102 text detection and recognition in natural images. The COCO-Text dataset [41] is currently the largest  
103 dataset for scene text detection and recognition. It contains 50,000+ images for training and testing.  
104 The COCO-Text dataset is very challenging since the text in this dataset is in arbitrary orientation.

105 **2.3 Text Spotting Datasets for Videos**

106 The development of video text spotting is limited in recent years due to the lack of efficient data  
107 sets. ICDAR 2015 Video [15] consists of 28 videos lasting from 10 seconds to 1 minute in indoors  
108 or outdoors scenarios. Limited videos (*i.e.*, 13 videos) used for training and 15 for testing. Minetto  
109 Dataset [27] consists of 5 videos in outdoor scenes. The frame size is 640 x 480 and all videos  
110 are used for testing. YVT [30] contains 30 videos, 15 for training and 15 for testing. Different  
111 from the above two datasets, it contains web videos except for scene videos. USTB-VidTEXT [50]  
112 with only five videos mostly contain born-digital text (captions and subtitles) sourced from Youtube.  
113 RoadText-1K provides a driving videos dataset with 1000 videos. The 10-second long video clips in  
114 the dataset are sampled from BDD100K [53]. As shown in Table. 1, the existing datasets contain a  
115 limited training set and tedium video scenarios. To promote the development of video text spotting  
116 and extension of application based on video text, we create a large scale, multidimensional and  
117 multilingual dataset, and attempt to provide a more reasonable metric.



Figure 2: **Distributions of MMVText.** (a) Chinese caption and English scene text. (b) Only Chinese caption. (c) Multilingual caption and English scene text. (d) The benchmark dataset covers a wide and open range of life scenes (30 categories) with multilingual texts. Caption text (blue box) and scene text (red box) are distinguished in MMVText, which is favorable for downstream tasks.

### 118 3 MMVText Benchmark

119 This section firstly introduces the collection and annotation of MMVText and provides a comprehensive  
120 analysis and comparison. And then, the related tasks and corresponding metrics are described.  
121 Finally, we discuss the link to application scenarios and potential impacts.

#### 122 3.1 Data Collection and Annotation

123 **Data Collection.** To obtain abundant and various text videos, we first start by acquiring a large list  
124 of text videos class using *KuaiShou*<sup>1</sup> - an online resource that contains billions of videos with various  
125 scene text from cartoon movies to human relation. Then, we choose 30 live video categories, *i.e.*, ,  
126 *E-commerce*, *Game*, *Home*, *Fashion*, and *Technology*, as shown in Figure. 2 (d). With each raw video  
127 category, we first choose the video clips with text, then make two rounds of screening to remove  
128 the ordinary videos. As a result, we obtain 512 videos with 1,010,848 video frames, as shown in  
129 Table 1. Finally, to fair evaluation, we divide the dataset into two parts: the training set with 641,049  
130 frames from 331 videos, and the testing set with 369,799 frames from 179 videos. As shown in  
131 Figure 2 (a), different from the existing data sets, which only focus on one type of video text and the  
132 video scene is limited, our dataset not only care about scene text spotting in the real world, but also  
133 focus on caption texts in the video. For the most part, caption text represents more global information  
134 than scene text, which is quite favorable for some downstream tasks, *e.g.*, *video understanding*, *video*  
135 *caption translation*. Therefore, the MMVText can cover a wider and open range of life scenes, and  
136 contains various text with a more comprehensive description of the video.

137 **Data Annotation.** We invite a professional annotation team to label each video text with four kinds of  
138 description information: the bounding box describing the location information, judging the tracking  
139 identification (ID) of the same text instance, identifying the content of the text information, and  
140 distinguishing the category label of the caption or scene text. To save the annotation cost, we first  
141 sample the videos, annotate each sampled video frame at an instance level, and then transform the  
142 annotation information from the sampled video frame to the unlabeled video frame by interpolation.  
143 Finally, we invite an audit team to carry out another round of annotation checks, and re-label part  
144 video frames with unqualified annotation. *For video sampling*, we use uniform sampling with a  
145 sampling frequency of 7 to sample all the videos in the dataset, and obtain the sampled video frame  
146 set. *For sampling video frame annotation*, each text instance is labeled in the same quadrilateral way  
147 as in the ICDAR 2015 incidental text dataset [56]. In addition, the text instance also will be marked  
148 with two description information: the category of the caption or scene, and the recognition content.  
149 After the spatial location, content, and category of the video text are determined, the annotator will

<sup>1</sup><https://www.kuaishou.com/en>

Table 1: **Statistical Comparison.** Comparisons between MMVText and existing datasets for caption and scene text in videos. *D*, *R*, *T*, and *S* denotes the Detection, Recognition, Tracking, and Spotting respectively. ‘**Incidental**’ denotes indoor and outdoor scenarios in daily life (*e.g.*, *walking outdoors*, *driving*). ‘**Open**’ refers to any scenarios, *e.g.*, *Game*, *Sport(NBA)*.

Dataset	Category	MLingual	Scenario	Videos	Frames	Texts	Task
AcTiV-D [54]	Caption	-	News video	8	1,843	5,133	D
UCAS-STLData [4]	Caption	-	Teleplay	3	57,070	41,195	D
USTB-VidTEXT [50]	Caption	-	Web video	5	27,670	41,932	D&S
YVT [30]	Both	-	Incidental	30	13,500	16,620	D&R&T&S
ICDAR 2015 VT [56]	Scene	-	Incidental	51	27,824	143,588	D&R&T&S
LSVT [5]	Scene	✓	Incidental	100	66,700	569,300	D&R&T&S
RoadText-1K [31]	Scene	-	Driving	1000	300,000	1,280,613	D&R&T&S
MMVText (ours)	Both	✓	Open	510	<b>1,010,848</b>	<b>4,513,525</b>	D&R&T&S

150 determine the tracking ID by browsing the length of the same video text in the continuous sampling  
 151 video frames. We also invited other text-related people to conduct two rounds of cross-checking  
 152 to ensure the annotation quality. *For interpolation on unlabeled video frames*, each text instance  
 153 is marked with tracking ID and recognition content, so we can judge whether the texts in adjacent  
 154 sampling frames are the same text with the same ID. For the same text instance, we first determine  
 155 whether the text annotation of the sampled video frame is the starting and end frame of the text  
 156 instance. If not, we look forward and backward for the starting and end position of the text instance  
 157 and label it. Then we use the linear interpolation way to calculate the position of the text object in  
 158 the middle of the unmarked video frame, and give tracking ID, recognition content, and category.  
 159 *For check and re-label bad cases*, the linear interpolation shows a dissatisfied performance in some  
 160 cases, *e.g.*, *the new text appears on starting frame, text suddenly disappears on ending frame*, which  
 161 are difficult to capture. Therefore, we invite an audit team to carry out another round of annotation  
 162 checks. Around 150,000 video frames with unqualified annotation from 1,010,848 video frames are  
 163 selected to refine, taking 20 men in three weeks. As a labor-intensive job, the whole labeling process  
 164 takes 30 men in two months, *i.e.*, 20,160 man-hours, to complete about 200,000 sampled video frame  
 165 annotations.

### 166 3.2 Dataset Analysis

167 **Statistic Comparison.** The qualitative and statistic comparison between the established MMVText  
 168 and other datasets are visualized in Figure. 1, and summarized in Table. 1. *Category* denotes the  
 169 category of the text type in the corresponding dataset. *MLingual* denotes whether the dataset contains  
 170 multiple language texts. *Scenario* denotes the scene range of the video. *Videos*, *Frames*, *Texts*  
 171 represents the number of videos, video frames, video texts in the dataset, respectively. *Task* denotes  
 172 which tasks the dataset supports. **Caption Text and Scene Text.** For comprehensive evaluation and  
 173 research, we not only expand the scale of the dataset (*i.e.*, the number of videos, video frame, and  
 174 video text), and label the spatial quadrilateral position, recognition content, and tracking ID, but also  
 175 additionally collect and annotate the category of caption or scene for each text instance. As shown  
 176 in Figure. 2 (a), in a video, different types of text instances may exist simultaneously, and they are  
 177 helpful to understand videos synergistically. Concretely, caption text can directly show the dialogue  
 178 between people in video scenes and represent the time or topic of the video scenes, scene text can  
 179 unambiguously define the object and can identify important localization and road paths in video  
 180 scenes. Besides, nowadays, caption text frequently exists in all kinds of life scenarios video. Even  
 181 for some videos, without any scene texts, there is a lot of caption text, as shown in Figure. 2 (b). To  
 182 favor downstream tasks (*e.g.*, video text translation, video understanding, and video retrieval), we  
 183 also provide multilingual text annotations, as shown in Figure. 2 (c).

184 To provide the community with unified text-level quantitative descriptions, and facilitate controlled  
 185 evaluation for different approaches, we will compare our dataset with caption or scene text datasets  
 186 from four aspects, *i.e.*, text description, video scene, dataset size, and supported tasks. *For text*  
 187 *description attribute* (*i.e.*, *Category*, *MLingual*), our MMVText contains both types (caption and  
 188 scene) of video text and has multi-language features, which obviously has more extensive description  
 189 ability than caption or scene text dataset. *For video scene attribute* (*i.e.*, *Scenario*), the caption  
 190 text datasets choose videos with certain professional purposes (*e.g.*, news reports, TV dramas, and

191 documentaries), which shows that the scenes they cover are relatively limited. And the existing  
 192 scene text datasets often choose some video scenes captured by mobile shooting, and the number of  
 193 collectors is small, the range of captured scenes is also limited. However, the videos in our dataset  
 194 are from videos uploaded voluntarily by all kinds of users. Therefore, the proposed MMVText  
 195 covers various scenarios, but it also brings significant challenges to researchers. *For the size of*  
 196 *the dataset(i.e., Videos, Frames, Texts)*, we can find that our MMVText has advantages over the  
 197 superimposed caption text dataset and the scene text dataset in the indicators of videos, frames, and  
 198 texts. The number of videos in RoadText-1K is more than ours (1,000 vs. 510), but the number of  
 199 frames in RoadText-1K is far less than ours (300,000 vs. 1,010,848), which imply that the average  
 200 video length of RoadText-1K is much shorter than ours (300 vs. 1,982). *For the supported tasks*,  
 201 the proposed MMVText supports four common video text tasks: detection, recognition, video text  
 202 tracking, end to end video text spotting. The focus and application scenarios of each task is entirely  
 203 different. For example, detection task used in the static image focus on localization performance,  
 204 paving the way for recognition task, which apply to license plate recognition. End to end video text  
 205 spotting task focuses on recognition and tracking performance, which apply to video understanding  
 206 and video retrieval. In conclusion, the high efficiency of MMVText for evaluating advanced deep  
 207 learning methods is very favorable for promoting various text spotting applications in real life.

### 208 3.3 MMVText Tasks and Metrics

209 Standardized benchmark metrics are crucial as same as the dataset for the majority of computer vision  
 210 applications, and we attempt to provide a reasonable evaluation for video text spotting methods. The  
 211 proposed MMVText includes four tasks: (1) Text Detection. (2) Text Recognition. (3) Video Text  
 212 Tracking, aimed at describing text location information in continuous frames. (4) End to End Text  
 213 Spotting in Videos, to understand text and track multiple frames.

214 Following ICDAR2015 [56], the evaluation protocols [44] are used for text detection and recognition  
 215 task. For video text tracking task, the existing video text tracking datasets such as ICDAR2015  
 216 (video) [15] and RoadText-1k [31] all adopted the MOT metrics (*i.e.*, Multiple Object Tracking  
 217 Accuracy (*MOTA*) and Multiple Object Tracking Precision (*MOTP*)). However, there are two  
 218 sets of measures for Multiple Object Tracking: the MOT metrics (*MOTA, MOTP*) [2] and ID  
 219 metrics (*ID<sub>F1</sub>*) [18, 33]. The CVPR19 MOTChallenge evaluation framework [7] presents that  
 220 different measures serve different purposes. *Event-based* measures like MOT help pinpoint the  
 221 source of some errors, and are thereby informative for the designer of certain system components.  
 222 *Identity-based* measure(*ID<sub>F1</sub>*) is more favorable for evaluating how well computed identities  
 223 conform to true identities. Therefore, except for using the standard measures(*MOTA, MOTP*),  
 224 *Identity-based* measures(*ID<sub>F1</sub>*), as a new metric is adopted firstly for video text spotting task.  
 225 *ID<sub>F1</sub>* is the ratio of correctly identified detections over the average number of ground-truth and  
 226 computed detections. And the metric is more reasonable to evaluate ID switches in some cases. We  
 227 evaluate the metrics in MMVText by:

$$ID_{F1} = \frac{2ID_{tp}}{2ID_{tp} + ID_{fp} + ID_{fn}}, \quad (1)$$

228 where *ID<sub>tp</sub>*, *ID<sub>fp</sub>* and *ID<sub>fn</sub>* refer to true positive, false positive and false negative of matching ID.  
 229 Besides, the ID metric [7] also includes *MT* (Mostly Tracked) Number of objects tracked for at least  
 230 80 percent of lifespan, *ML* (Mostly Lost) Number of objects tracked less than 20 percent of lifespan.

231 For Task4 (End to End Text Spotting in Videos), the objective of this task is to recognize words in the  
 232 video as well as localize them in terms of time and space. And we argue that the final recognition  
 233 result is more important than text localization in videos. Thus, we modify the *ID<sub>F1</sub>* to *TID<sub>F1</sub>*, which  
 234 focuses on text instance ID tracking and recognition results that be required by many downstream  
 235 tasks. More specifically,

$$TID_{tp} = \sum_h \sum_t m(h, o, \Delta_t, \Delta_s, \Delta_r), \quad (2)$$

$$TID_{F1} = \frac{2TID_{tp}}{2TID_{tp} + TID_{fp} + TID_{fn}}, \quad (3)$$

236 where  $\Delta_t$ ,  $\Delta_s$  and  $\Delta_r$  refer to ID matching, space location matching and recognition result matching.  
 237  $h$  and  $o$  denote hypothesis set (*e.g.*, predicted ID  $I_p$ , box locations  $L_p$ , recognition results  $R_p$ ) and

238 ground truth set with (ID  $I_g$ , box locations  $L_g$ , recognition ground true  $R_g$ ). And the three matching  
239 can be obtained by:

$$\Delta_t : I_p = I_g, \quad \Delta_s : IoU(L_p, L_g) > 0, \quad \Delta_r : R_p = R_g. \quad (4)$$

240 The match of  $h$  and  $o$  is a true positives of text ID (*i.e.*,  $TID_{tp}$ ) when these conditions (*i.e.*,  $\Delta_t$ ,  $\Delta_s$   
241 and  $\Delta_r$  are met. Similarly, false positive (*i.e.*,  $TID_{fp}$ ) and false negative (*i.e.*,  $TID_{fn}$ ) of text ID can  
242 be obtained for  $TID_{F1}$  calculation. More details concerning metrics in the supplementary material.

### 243 3.4 Methods

244 Text detection and recognition in the static image have made tremendous progress, and abundant  
245 great work [45, 55, 36] be proposed. By contrast, the counterparts in video text reading are rare and  
246 lack quality open-source algorithms. Therefore, we adopt various mature techniques in the static  
247 image to better evaluate the efficiency of MMVText.

248 **Detection.** The deep learning-based text detection methods can be roughly divided into two categories:  
249 regression-based method and segmentation-based method. EAST [55] as one of the popular  
250 regression-based methods is used to test our MMVtext. The method adopts FCNs to predict shrink-  
251 able text score maps, rotation angles and perform per-pixel regression, followed by a post-processing  
252 NMS. For segmentation based methods, we adopt PSENet [45] and DB [20] to evaluate our MMVtext.  
253 PSENet [45] generates various scales of shrunked text segmentation maps, then gradually expands  
254 kernels to generate the final instance segmentation map. Similarly, DB [20] utilizes the shrunked  
255 text segmentation maps and differentiable binarization to detect text instances. **Recognition.** Recent  
256 methods mainly use two techniques to train the scene text recognition model, namely Connectionist  
257 Temporal Classification (CTC) and attention mechanism. In CTC-based methods, CRNN [36] as  
258 the representation, which introduced CTC decoder into scene text recognition with a Bidirectional  
259 Long Short-Term Memory (BiLSTM) to model the feature sequence. In Attention-based methods,  
260 RARE [37] firstly normalizes the input text image using the Spatial Transformer Network (STN [12]),  
261 then utilizes CNN to extract feature and captures the contextual information within a sequence of  
262 characters. Finally, it estimates the output character sequence from the identified features with the  
263 attention module.

264 **Text Tracking Trajectory Generation.** With text detection and recognition in a static image, we  
265 only obtain text localization and recognition information without temporal information, which are  
266 insufficient for video spotting evaluation (*e.g.*,  $TID_{F1}$ ,  $MOTA$  and  $MOTP$ ). The work [46] based  
267 on multi-frame tracking provides a method to track text instances temporally based on attributes of  
268 the text objects in multiple frames. Following the work [46], we link and match text objects in the  
269 current frame and several frames by IOU and edit distance of text.

## 270 4 Experimental

271 In this section, we conduct experiments on our MMVText to demonstrate the effectiveness of the  
272 proposed benchmark. Note that we denote Ground Truth of ID tracking in all the experiments, Mostly  
273 Tracked and Mostly Lost as ‘GT’, ‘MT’ and ‘ML’, respectively.

### 274 4.1 Implementation Details

275 All of the experiments use the same strategy: (1) Training detector and recognizer with MMVText.  
276 (2) Matching text objects with corresponding text tracking trajectory id. **Detection:** without pretrained  
277 model, we train detectors directly with training set (*i.e.*, 641,049 frame images) of MMVText.  
278 **Recognition:** the network is pre-trained on the *chinese ocr*<sup>2</sup> and MJSynth [11], and further fine-tuned  
279 on our MMVText. All of our experiments are conducted on 8 V100 GPUs. PSENet [45], EAST [55]  
280 and DB [20] are adopted as the base detectors because of their popularity. CRNN [36] and RARE [37]  
281 as the base text recognizers to evaluate our MMVText. In the PSENet, EAST, DB, CRNN and RARE  
282 experiments, all settings follow the original reports.

<sup>2</sup>[https://github.com/YCG09/chinese\\_ocr](https://github.com/YCG09/chinese_ocr)

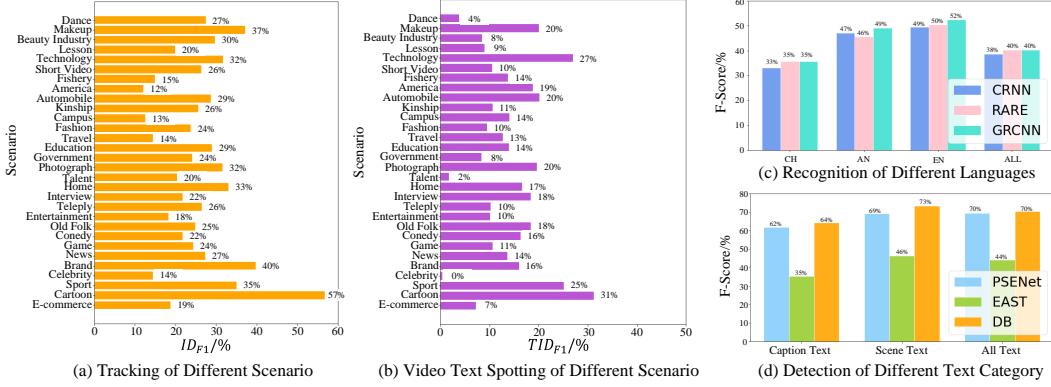


Figure 3: **Attribute Experiments of MMVText.** (a) Tracking performance (*i.e.*,  $ID_{F1}$ ) with EAST [55] in different scenarios. (b) End to end video text spotting performance (*i.e.*,  $TID_{F1}$ ) with PSENet [45] and CRNN [36] in different scenarios. (c) Recognition accuracy of different models in different languages. (d) Detection performance of different model in caption or scene text. ‘CH’, ‘AN’, ‘EN’ and ‘ALL’ refer to ‘Chinese Characters’, ‘Alphanumeric Characters’, ‘English’ and ‘All Characters’, respectively.

## 283 4.2 Attribute Experiments Analysis

284 **Text Tracking in Different Scenarios.** Figure. 3 (a) gives the tracking performance  $ID_{F1}$  of  
285 EAST [55] in different scenarios of MMVText. The model achieves the best performance with a  
286  $ID_{F1}$  of 57% in cartoon videos, since the conspicuous text instances and simple background are  
287 designed in cartoon videos. By comparison, several scene categories obtain extremely dissatisfied  
288 performance due to complex background and various text appearance, such as *Campus* and *Travel*.

289 **New Scenarios, New Challenge for Text Spotting.** Figure. 3 (b) gives the end-to-end performance  
290  $TID_{F1}$  using PSENet [45] as the detector and CRNN [36] as the recognizer in different scenarios  
291 of MMVText. Similar to tracking performance using EAST [55], the end-to-end video spotting  
292 performance shows the best performance with a  $TID_{F1}$  of 31% in scenario of *Cartoon*. Outdoor  
293 scenes video (e.g., *Campus*, *Travel*, *Celebrity*) usually have a more complex environment and unsteady  
294 camera movement. On the contrary, some easy video types such as cartoons hardly contain complex  
295 scene text, and the camera movement extremely stable by manual handling.

296 **Multilingual Recognition, New Challenge.** As shown in Figure. 3 (c), the text recognition results  
297 for different languages are provided. In summary, the alphanumeric recognition result (about 47%)  
298 is better than the Chinese recognition result (about 35%), regardless of the models. The final  
299 results (about 40%) for all characters are satisfactory, can not meet the requirement of the application.  
300 Unlike English, Chinese contains thousands of characters(3,856 Chinese characters v.s. 26 English  
301 characters on MMVText), which are difficult to recognize. As shown in Figure. 4, we argue that  
302 many clear image patches with Chinese text still bring recognition errors, since the classification of  
303 such many characters is difficult for the network to learn. By comparison, most English errors are  
304 caused by blur and out-of-focus.

305 **Long Caption Text, New Challenge for Text Detection.** As shown in Figure. 3 (d), we provide  
306 the detection performance comparison for different models in different text categories (*i.e.*, caption  
307 text or scene text) of MMVText. It is obvious that the performance for scene text is better than the  
308 counterpart of caption text, regardless of which detection model. Similar to the long text detection  
309 challenge MSRA-TD500 [51], we argue that the prime reason is that caption texts are all long text (310  
311 average width-height ratio: 6.8 for caption *v.s.* 2.3 for scene text on MMVText), a different case to  
312 detect without any model refinement.

## 312 4.3 Text Detection and Recognition in Images

313 Although text detection and recognition in static images are not the focus in this work, we provide  
314 the corresponding performance for comparison, as shown in Table. 2. For text detection, we adopts  
315 EAST [55], PSENet [45] and DB [20] to evaluate the proposed MMVText. We observe that frame-

Table 2: **Detection and Recognition Performance on MMVText.** Frame level text Detection and Recognition performance of existing models on MMVText. ‘CH’, ‘AN’, ‘EN’ and ‘ALL’ refer to ‘Chinese Characters’, ‘Alphanumeric Characters’, ‘English’ and ‘All Characters’, respectively.

Method	Detection Performance/%				Method	Recognition Performance/%								
	Precision	Recall	F-score			Pretrained				Fine tuned				
						CH	AN	EN	ALL	CH	AN	EN	ALL	
EAST	52.2	38.1	44.1		CRNN	26.0	32.1	36.1	23.2	33.2	47.1	49.5	38.6	
PSENet	74.3	65.2	69.5		RARE	25.2	34.2	37.4	23.5	35.6	45.7	50.4	40.2	
DB	77.2	64.5	70.3		GRCNN	23.1	39.8	40.4	26.7	35.6	49.2	52.4	40.3	



Figure 4: **Bad cases for multilingual text recognition.** The incorrectly recognized characters are marked in red. ‘[ ]’ denotes the prediction of the corresponding character is missing.

316 level text detection and recognition results on MMVText are not unsatisfactory, with lower results than  
 317 these methods report on existing scene text datasets. For example, EAST only obtains an f-score of  
 318 44.1% compared to the F-score of 80.7% on icdar2015 [56]. For text recognition, CRNN [36] based  
 319 on CTC loss, RARE [37] with attention mechanism and GRCNN [43] as the base text recognizers to  
 320 test our MMVText. The text annotation in our MMVText covers two languages (*i.e.*, English and  
 321 Chinese), thus we conduct several experiments for each language. ‘CH’ and ‘AN’ refer to Chinese  
 322 text instances and alphanumeric characters. ‘ALL’ denotes all characters regardless of which language.  
 323 Similar to the detection task, the recognition model only yields about 40% accuracy on our dataset,  
 324 but the same model reports > 90+ on most benchmark datasets [15] for scene text recognition. The  
 325 main reasons have two points: (1) The proposed MMVText is multilingual, and the category number  
 326 of Chinese characters in real-world images is much larger than those of Latin languages. (2) The  
 327 video texts are quite blurred, out-of-focus, and the distribution of characters is relatively smaller than  
 328 the static image counterparts.

#### 329 4.4 Text Tracking and Spotting in Videos

330 **Video Text Tracking.** Table. 3 shows the comparing results of text tracking on MMVText. We  
 331 observe that the overall performances of the used detectors are dissatisfactory on MMVText. Besides,  
 332 the  $IDF_1$  of EAST [55] is lower with 6.7% gap than that of PSENet [45]. The main reason is that  
 333 MMVText contains a mass of long text instances, but regression-based EAST can not deal with  
 334 the long text cases well. The performance of DB is similar to that of PSENet for both all are the  
 335 segmentation-based methods. According to Table. 3, MOTP shows a better performance than  
 336 MOTA. We argue that detectors such as PSENet or DB provide strong detecting capacity, but the  
 337 tracking ability is relatively weak. By comparison,  $IDF_1$  is a comprehensive metric for object ID  
 338 tracking.  $IDF_1$  (31.7%) of DB achieves the best performance of the three detectors, and EAST shows  
 339 the worst performance with a  $IDF_1$  of 23.2%.

Table 3: **Text Tracking Performance on MMVText.** Text tracking trajectory id generation use a method proposed in [46].

Method	MOTP	MOTA	ID <sub>P</sub> /%	ID <sub>R</sub> /%	ID <sub>F1</sub> /%	GT	MT	ML
EAST [55]	0.275	-0.301	23.5	22.9	23.2	48321	20.1%	74.2%
PSENet [45]	0.112	0.334	34.7	26.7	29.9	48321	26.4%	69.1%
DB [20]	0.102	0.438	33.7	29.9	<b>31.7</b>	48321	30.1%	65.1%

Table 4: **End to End Video Text Spotting Performance on MMVText.** Text tracking trajectory id generation use a method proposed in [46].  $TID_P$ ,  $TID_R$ ,  $TID_{F1}$ ,  $MOTP_T$  and  $MOTA_T$  refer to the corresponding metrics with recognition results in Table. 3.

Method								
Detection	Recognition	$TID_P/$ %	$TID_R/$ %	$TID_{F1}/$ %	$MOTA_T$	$MOTP_T$	MT	ML
EAST [55]	CRNN [36]	5.3	5.1	5.2	-0.835	0.173	3.2%	95.0%
	RARE [37]	3.0	3.6	3.2	-1.130	0.173	2.6%	95.4%
PSENet [45]	CRNN [36]	14.7	9.8	11.8	-0.300	0.197	7.8%	88.9%
	RARE [37]	15.2	10.4	12.4	-0.280	0.201	8.0%	87.8%
DB [20]	CRNN [36]	15.6	9.6	11.9	-0.284	0.230	6.9%	89.5%
	RARE [37]	20.1	15.2	<b>17.3</b>	-0.293	0.150	8.7%	82.0%

340 **End to End Text Spotting in Video.** Detection or text tracking tasks are paving the way for the  
 341 recognition task. Table. 4 shows the performance of text spotting in the video. And  $TID_{F1}$  in  
 342 Equation. 10 as an integrated metric to evaluate algorithms in spatial location, content, and temporal  
 343 information three dimensions. Similar to the text tracking performance of EAST, the corresponding  
 344 performance  $TID_{F1}$  using CRNN [36] as the recognizer in video text spotting is still not satisfied  
 345 with a 5.2%  $TID_{F1}$ . The combination of DB [20] and RARE [37] achieves the best performance  
 346 with a 17.3%  $TID_{F1}$  among all the cases, but the performance still is inadequate to meet application  
 347 requirements. MT (Mostly Tracked) and ML (Mostly Lost) as the metrics concerning statistical  
 348 number can be used to evaluate from another aspect. For the combination of DB [20] and RARE [37],  
 349 39650 text tracking trajectories are lost, less than 20 percent of lifespan. By comparison, only 4230  
 350 tracking trajectories are satisfactory, more than 80 percent of lifespan tracked.

#### 351 4.5 Discussion for Real Applications

352 Text spotting in static images has numerous application scenarios: (1) Automatic data entry. SF-  
 353 Express<sup>3</sup> utilizes OCR techniques to accelerate the data entry process. NEBO<sup>4</sup> performs instant  
 354 transcription as the user writes down notes. (2) Autonomous vehicle [26, 25]. Text-embedded  
 355 panels carry important information, e.g., geo-location, current traffic condition, navigation, and etc.  
 356 (3) **Text-based reading comprehension.** TextCaps [38] and text-based VQA [39, 3] show the new  
 357 vision-and-language tasks, which need to recognize text, relate it to its visual context, semantic,  
 358 and visual reasoning between multiple text tokens and visual entities, such as objects. Similarly,  
 359 there are many application demands for video text understanding across various industries and in our  
 360 daily lives. We list the most outstanding ones that significantly impact, improving our productivity  
 361 and life quality. **Firstly**, automatically describing video with natural language [49, 47] can bridge  
 362 video and language. **Secondly**, video text automatic translation<sup>5</sup> can be extremely helpful as people  
 363 travel, and help video-sharing websites<sup>6</sup> to cut down language barriers. **Finally**, text-based video  
 364 retrieval [16, 21] is a irreplaceable business for many companies, such as Google and YouTube. More  
 365 details and analyses for application scenarios concerning MMVText in the supplementary material.

#### 366 4.6 Potential Negative Societal Impacts

367 We argue that there mainly exists slight potential negative societal impacts for personal privacy.  
 368 Although much personal information, e.g., *names*, *identifying information*, *human faces*, have been  
 369 blurred to protect privacy, there still might exist a little risk.

<sup>3</sup><https://www.sf-express.com/cn/sc/>

<sup>4</sup><https://www.myscript.com/nebo/>

<sup>5</sup><https://translate.google.com/intl/en/about/>

<sup>6</sup><https://www.youtube.com/>

370 **5 Conclusion and Future Work**

371 In this paper, we establish a large-scale multidimensional and multilingual dataset for video text  
372 tracking and spotting, termed as MMVText, with four description information, *i.e.*, , bounding box,  
373 tracking ID, recognition content, and text category label. Compare with the existing benchmarks, the  
374 proposed MMVText mainly contains three advantages: large-scale, multidimensional, multilingual.  
375 MMVText spans various video scenarios, text types, and multi-stage tasks, promoting video text  
376 research. We also conduct several experiments on this dataset and shed light on what attributes are  
377 especially difficult for the current task, which cast new insight into the video text tracking, spotting  
378 field. In general, we hope the MMVText would facilitate the advance of video-and-text research.

379 **References**

- 380 [1] Christos-Nikolaos E Anagnostopoulos, Ioannis E Anagnostopoulos, Ioannis D Psoroulas, Vassili  
381 Loumos, and Eleftherios Kayafas. License plate recognition from still images and video  
382 sequences: A survey. *IEEE Transactions on intelligent transportation systems*, 9(3):377–391,  
383 2008.
- 384 [2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the  
385 clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.
- 386 [3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny,  
387 CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings*  
388 *of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019.
- 389 [4] Yuanqiang Cai, Weiqiang Wang, Shao Huang, Jin Ma, and Ke Lu. Spatiotemporal text localiza-  
390 tion for videos. *Multimedia Tools and Applications*, 77(22):29323–29345, 2018.
- 391 [5] Zhanzhan Cheng, Jing Lu, Yi Niu, Shiliang Pu, Fei Wu, and Shuigeng Zhou. You only recognize  
392 once: Towards fast video text spotting. In *ACM International Conference on Multimedia*, pages  
393 855–863, 2019.
- 394 [6] Chee Kheng Ch’ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene  
395 text detection and recognition. In *IEEE International Conference on Document Analysis and*  
396 *Recognition*, pages 935–942, 2017.
- 397 [7] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid,  
398 Stefan Roth, Konrad Schindler, and Laura Leal-Taixe. Cvpr19 tracking and detection challenge:  
399 How crowded can it get? *arXiv preprint arXiv:1906.04567*, 2019.
- 400 [8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang.  
401 Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine*  
402 *Intelligence*, 2021.
- 403 [9] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation  
404 in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages  
405 2315–2324, 2016.
- 406 [10] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end  
407 textspotter with explicit alignment and attention. In *IEEE conference on computer vision and*  
408 *pattern recognition*, pages 5020–5029, 2018.
- 409 [11] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and  
410 artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*,  
411 2014.
- 412 [12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial trans-  
413 former networks. In *Neural Information Processing Systems*, pages 2017–2025, 2015.
- 414 [13] Keechul Jung, Kwang In Kim, and Anil K Jain. Text information extraction in images and  
415 video: a survey. *Pattern recognition*, 37(5):977–997, 2004.

- 416 [14] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew  
 417 Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar,  
 418 Shijian Lu, et al. Icdar 2015 competition on robust reading. In *IEEE International Conference*  
 419 on Document Analysis and Recognition, pages 1156–1160, 2015.
- 420 [15] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Big-  
 421 orda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and  
 422 Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *IEEE International*  
 423 *Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.
- 424 [16] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for  
 425 video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference,*  
 426 *Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020.
- 427 [17] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional  
 428 recurrent neural networks. In *IEEE International Conference on Computer Vision*, pages  
 429 5238–5246, 2017.
- 430 [18] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target  
 431 tracker for crowded scene. In *2009 IEEE conference on computer vision and pattern recognition*,  
 432 pages 2953–2960. IEEE, 2009.
- 433 [19] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetraault, Larry Goldberg, Alejandro Jaimes,  
 434 and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings*  
 435 of the *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.
- 436 [20] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text  
 437 detection with differentiable binarization. In *AAAI Conference on Artificial Intelligence*, pages  
 438 11474–11481, 2020.
- 439 [21] Jingzhou Liu, Wenhui Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu.  
 440 Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF*  
 441 *Conference on Computer Vision and Pattern Recognition*, pages 10900–10910, 2020.
- 442 [22] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text  
 443 spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and*  
 444 *pattern recognition*, pages 5676–5685, 2018.
- 445 [23] Shangbang Long, Xin He, and Cong Yao. Scene text detection and recognition: The deep  
 446 learning era. *International Journal of Computer Vision*, 129(1):161–184, 2021.
- 447 [24] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter:  
 448 An end-to-end trainable neural network for spotting text with arbitrary shapes. In *European*  
 449 *Conference on Computer Vision*, pages 67–83, 2018.
- 450 [25] Abdelhamid Mammeri, Azzedine Boukerche, et al. Mser-based text detection and commu-  
 451 nication algorithm for autonomous vehicles. In *2016 IEEE symposium on computers and*  
 452 *communication (ISCC)*, pages 1218–1223. IEEE, 2016.
- 453 [26] Abdelhamid Mammeri, El-Hebri Khiari, and Azzedine Boukerche. Road-sign text recogni-  
 454 tion architecture for intelligent transportation systems. In *2014 IEEE 80th Vehicular Technology*  
 455 *Conference (VTC2014-Fall)*, pages 1–5. IEEE, 2014.
- 456 [27] Rodrigo Minetto, Nicolas Thome, Matthieu Cord, Neucimar J Leite, and Jorge Stolfi. Snooper-  
 457 track: Text detection and tracking for outdoor videos. In *IEEE International Conference on*  
 458 *Image Processing*, pages 505–508, 2011.
- 459 [28] Anand Mishra, Kartek Alahari, and CV Jawahar. Image retrieval using textual cues. In  
 460 *Proceedings of the IEEE International Conference on Computer Vision*, pages 3040–3047,  
 461 2013.

- 462 [29] Nibal Nayef, Fei Yin, Imen Bizard, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo  
 463 Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading  
 464 challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *IEEE*  
 465 *International Conference on Document Analysis and Recognition*, volume 1, pages 1454–1459,  
 466 2017.
- 467 [30] Phuc Xuan Nguyen, Kai Wang, and Serge Belongie. Video text detection and recognition:  
 468 Dataset and benchmark. In *IEEE winter conference on applications of computer vision*, pages  
 469 776–783, 2014.
- 470 [31] Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and  
 471 CV Jawahar. Roadtext-1k: Text detection & recognition dataset for driving videos. In *IEEE*  
 472 *International Conference on Robotics and Automation*, pages 11074–11080, 2020.
- 473 [32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time  
 474 object detection with region proposal networks. pages 91–99, 2015.
- 475 [33] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance  
 476 measures and a data set for multi-target, multi-camera tracking. In *Workshops of European*  
 477 *conference on computer vision*, pages 17–35, 2016.
- 478 [34] Xuejian Rong, Chucai Yi, Xiaodong Yang, and Yingli Tian. Scene text recognition in multiple  
 479 frames based on text tracking. In *2014 IEEE International Conference on Multimedia and Expo*  
 480 (*ICME*), pages 1–6. IEEE, 2014.
- 481 [35] Georg Schroth, Sebastian Hilsenbeck, Robert Huitl, Florian Schweiger, and Eckehard Steinbach.  
 482 Exploiting text-related features for content-based image retrieval. In *2011 IEEE international*  
 483 *symposium on multimedia*, pages 77–84. IEEE, 2011.
- 484 [36] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-  
 485 based sequence recognition and its application to scene text recognition. *IEEE transactions on*  
 486 *pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.
- 487 [37] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text  
 488 recognition with automatic rectification. In *IEEE conference on computer vision and pattern*  
 489 *recognition*, pages 4168–4176, 2016.
- 490 [38] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset  
 491 for image captioning with reading comprehension. In *European Conference on Computer*  
 492 *Vision*, pages 742–758. Springer, 2020.
- 493 [39] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi  
 494 Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the*  
 495 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.
- 496 [40] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video  
 497 representations using lstms. In *International Conference on Machine Learning*, pages 843–852,  
 498 2015.
- 499 [41] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text:  
 500 Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint*  
 501 *arXiv:1601.07140*, 2016.
- 502 [42] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text  
 503 retrieval via joint text detection and similarity learning. *arXiv preprint arXiv:2104.01552*, 2021.
- 504 [43] Jianfeng Wang and Xiaolin Hu. Gated recurrent convolution neural network for ocr. In *Neural*  
 505 *Information Processing Systems*, pages 334–343, 2017.
- 506 [44] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011*  
 507 *International Conference on Computer Vision*, pages 1457–1464. IEEE, 2011.
- 508 [45] Wenhui Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape  
 509 robust text detection with progressive scale expansion network. In *IEEE conference on computer*  
 510 *vision and pattern recognition*, pages 9336–9345, 2019.

- 511 [46] Xiaobing Wang, Yingying Jiang, Shuli Yang, Xiangyu Zhu, Wei Li, Pei Fu, Hua Wang, and  
 512 Zhenbo Luo. End-to-end scene text recognition in videos based on multi frame tracking. In  
 513 *IEEE International Conference on Document Analysis and Recognition*, pages 1255–1260,  
 514 2017.
- 515 [47] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex:  
 516 A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings*  
 517 *of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.
- 518 [48] Liang Wu, Palaiahnakote Shivakumara, Tong Lu, and Chew Lim Tan. A new technique for  
 519 multi-oriented scene text line detection and tracking in video. *IEEE Transactions on multimedia*,  
 520 17(8):1137–1152, 2015.
- 521 [49] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for  
 522 bridging video and language. In *Proceedings of the IEEE conference on computer vision and*  
 523 *pattern recognition*, pages 5288–5296, 2016.
- 524 [50] Chun Yang, Xu-Cheng Yin, Wei-Yi Pei, Shu Tian, Ze-Yu Zuo, Chao Zhu, and Junchi Yan.  
 525 Tracking based multi-orientation scene text detection: A unified framework with dynamic  
 526 programming. *IEEE Transactions on Image Processing*, 26(7):3235–3248, 2017.
- 527 [51] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary  
 528 orientations in natural images. In *2012 IEEE conference on computer vision and pattern*  
 529 *recognition*, pages 1083–1090. IEEE, 2012.
- 530 [52] Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu. Text detection, tracking and  
 531 recognition in video: a comprehensive survey. *IEEE Transactions on Image Processing*,  
 532 25(6):2752–2773, 2016.
- 533 [53] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and  
 534 Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling.  
 535 *arXiv preprint arXiv:1805.04687*, 2018.
- 536 [54] Oussama Zayene, Mathias Seuret, Sameh Masmoudi Touj, Jean Hennebert, Rolf Ingold, and  
 537 Najoua Essoukri Ben Amara. Text detection in arabic news video based on SWT operator and  
 538 convolutional auto-encoders. In *Workshop on Document Analysis Systems*, pages 13–18, 2016.
- 539 [55] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang.  
 540 East: an efficient and accurate scene text detector. In *IEEE conference on computer vision and*  
 541 *pattern recognition*, pages 5551–5560, 2017.
- 542 [56] Xinyu Zhou, Shuchang Zhou, Cong Yao, Zhimin Cao, and Qi Yin. Icdar 2015 text reading in  
 543 the wild competition. *arXiv preprint arXiv:1506.03184*, 2015.

544 **A Appendix**

545 **A.1 MMVText Metrics**

546 The proposed MMVText mainly includes two tasks: (1) Video Text Tracking. (2) End to End Text  
 547 Spotting in Videos. *MOTP* (Multiple Object Tracking Precision) [2], *MOTA* (Multiple Object  
 548 Tracking Accuracy) and *IDF<sub>1</sub>* [7, 33] as the three important metrics are used to evaluate task1 (text  
 549 tracking) for MMVText. Following the previous works [15, 31], MMVText evaluates text tracking  
 550 methods in video and compare their performance with the MOTA and MOTP, which are given by:

$$MOTP = \frac{\sum_{i,t} (1 - d_t^i)}{\sum_t c_t}, \quad (5)$$

551 where  $c_t$  denotes the number of matches found for time  $t$ . For each of these matches, calculate the  
 552 iou  $d_t^i$  between the object  $o^i$  and its corresponding hypothesis. It shows the ability of the tracker to  
 553 estimate precise object positions. MOTA is calculated as follows:

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}, \quad (6)$$

554 where  $m_t$ ,  $fp_t$  and  $mme_t$  are the number of misses, false positives, and mismatches, respectively.  $g_t$   
 555 is the number of objects present at time  $t$ . It shows the tracker's performance at detecting objects and  
 556 keeping their trajectories, independent of the precision of the location. Besides,  $ID_{F1}$  as the new  
 557 metrics for tracking is calculated as follows:

$$ID_{tp} = \sum_h \sum_t m(h, o, \Delta_t, \Delta_s), \quad (7)$$

$$ID_{F1} = \frac{2ID_{tp}}{2ID_{tp} + ID_{fp} + ID_{fn}}, \quad (8)$$

558 where  $\Delta_t$  and  $\Delta_s$  refer to time matching and space location matching, respectively.  $ID_{tp}$ ,  $ID_{fp}$  and  
 559  $ID_{fn}$  refer to true positive, false positive and false negative of matching ID.

560 In Task2 (End to End Text Spotting in Videos), we modify the  $MOTA$ ,  $MOTA$  and  $ID_{F1}$  to  
 561  $MOTA_T$ ,  $MOTA_T$  and  $TID_{F1}$ , the only difference is the matches need to meet the correct of the  
 562 recognition result, since the recognition result is more important than tracking and localization. More  
 563 specifically,  $TID_{F1}$  is calculated as follows:

$$TID_{tp} = \sum_h \sum_t m(h, o, \Delta_t, \Delta_s, \Delta_r), \quad (9)$$

$$TID_{F1} = \frac{2TID_{tp}}{2TID_{tp} + TID_{fp} + TID_{fn}}, \quad (10)$$

564 where  $\Delta_t$ ,  $\Delta_s$  and  $\Delta_r$  refer to time matching, space location matching and recognition result  
 565 matching. And  $h$  and  $o$  denote hypothesis and true text trajectory with recognition result. The match  
 566 of  $h$  and  $o$  is a true positives of text ID (*i.e.*,  $TID_{tp}$ ) when these conditions (*i.e.*,  $\Delta_t$ ,  $\Delta_s$  and  $\Delta_r$ )  
 567 are met. Similarly, false positive (*i.e.*,  $TID_{fp}$ ) and false negative (*i.e.*,  $TID_{fn}$ ) of text ID can be  
 568 obtained for  $TID_{F1}$  calculation.

## 569 A.2 Link to Real Applications

570 In this section, we show that the practicability of the proposed MMVText, not a toy benchmark,  
 571 which can promote other video-and-text application research.

572 **Video Understanding.** As shown in Figure. 5 (a), the example concerning the task of describing  
 573 video with natural language is from MSR-VTT [49], and there has been increasing interest in video  
 574 understanding [47, 19]. However, video description with only visual information is difficult and  
 575 limited, even for a human. For the annotation of the sample video, *i.e.*, "A man in a blue suit  
 576 and purple tie discusses millennial investing fear", we can not learn the information of "millennial  
 577 investing fear" from the visual information in the video. By comparison, caption and scene texts in  
 578 the video contain accurate information of "*millennial investing fear*", which can help the model to  
 579 describe the video better. We argue that the same as general human understanding, videos without  
 580 captions and audio, is difficult to be properly understood by the model. We hope the release of  
 581 MMVText can promote efficient video text reading, further enhancing automatic video description.

582 **Video Text Automatic Translation.** Another practical application is video text automatic translation,  
 583 as shown in Figure. 5 (b). The application may be unnecessary for several professional teams or  
 584 classic movies due to the professional translator or huge cost investment. But for international video-  
 585 sharing websites <sup>7</sup> <sup>8</sup> with millions of users, it isn't easy to apply multilingual caption and scene text  
 586 in billions of videos. Therefore, efficient translation concerning caption text (*e.g.*, overlap, song title,  
 587 logos) and scene text (*e.g.*, street signs, business signs, words on shirt) still need further exploration  
 588 and research. The large-scale and multilingual MMVText contributes various real scenarios for the  
 589 development of video text automatic translation.

---

<sup>7</sup><https://www.youtube.com/>

<sup>8</sup><https://www.kuaishou.com/en>



Figure 5: **The Real Application Tasks Link to MMVText.** (a) Video Understanding, automatically describing visual content with natural language. (b) Video Caption Translation, extremely helpful for people who travel abroad and video-sharing websites such as YouTube. (c) Video Retrieval, accurate semantic information for text in videos can promote video retrieval.

590     **Video Retrieval.** Video retrieval with textual cues [28, 42] is also a very important application  
 591 direction for video-and-text research, as shown in Figure. 5 (c). To the best of my knowledge,  
 592 video retrieval with text information in the video is still almost a blank field of study and immature  
 593 application in the industry. The most existing video retrieval methods are stiff combinations of text  
 594 detection and recognition, invalid for the example with a sentence query. Besides, similar to video  
 595 understanding, for the query of the sample video, *i.e.*, "The lakers play host to Golden State", we  
 596 can not obtain the correct related video without scene text or caption information. The missing  
 597 information needs to recover by understanding the video with key video text information such as  
 598 "*GOLDEN STATE WARRIOR*, *LOS ANGELES LAKERS*". The proposed MMVText with various  
 599 text types (*e.g.*, caption, song title, logos, street signs, business signs) and annotation can promote the  
 600 research concerning efficient video retrieval.

### 601     A.2.1 Limitations

602     Although the proposed MMVText supports all video text spotting tasks, *i.e.*, *text detection*, *recognition*,  
 603 *tracking* end to end video text spotting, the potential contributions for other tasks still need mining.  
 604 For example, as shown in Figure. 5 (c), we do not provide the corresponding annotation (*i.e.*, the  
 605 query for each video) and metrics concerning video retrieval, but the annotation and metric are easy  
 606 to obtain due to text spotting annotation already existed. Therefore, there are still many potential  
 607 contributions for other tasks on MMVText, we want to take these as the future research directions  
 608 and provide a complete solution method.

Scenarios	Video	Video Frames	Text Instances	Scenarios	Video	Video Frames	Text Instances
Cartoon	10(2.0%)	10,813(1.1%)	23,191(1.8%)	Sport	21(4.1%)	26,138(2.6%)	106,996(2.4%)
Vlog	20(4.0%)	35,136(3.5%)	114,910(2.5%)	News Report	20(4.0%)	23,453(2.3%)	378,000(8.0%)
Driving	20(4.0%)	54,274(5.4%)	151,994(3.4%)	Celebrity	10(2.0%)	11,053(1.1%)	71,235(1.6%)
Home	10(2.0%)	19,870(2.0%)	41,090(1.0%)	Technology	20(4.0%)	30,518(3.0%)	340,172(7.5%)
Old Folk	10(2.0%)	12,475(1.2%)	47,879(1.1%)	Entertainment	20(4.0%)	34,648(3.4%)	214,561(4.8%)
Comedy	20(4.0%)	55,558(5.5%)	217,865(4.8%)	Game	20(4.0%)	43,635(4.3%)	444,106(9.8%)
Interview	5(1.0%)	9,310(0.8%)	23,616(0.6%)	E-commerce	10(2.0%)	25,068(2.5%)	81,569(1.8%)
Government	20(4.0%)	16,603(1.6%)	83,874(1.9%)	Brand	10(2.0%)	13,453(1.3%)	25,119(0.6%)
Travel	20(4.0%)	40,992(4.1%)	120,446(2.7%)	Teleplay	21(4.1%)	54,013(5.3%)	289,760(6.4%)
Campus	21(4.1%)	54,013(5.3%)	289,760(6.4%)	Photograph	20(4.0%)	27,680(2.7%)	123,832(2.7%)
America	19(3.7%)	48,135(4.8%)	132,117(2.9%)	Education	20(4.0%)	28,112(2.8%)	160,774(3.6%)
Short Video	20(4.0%)	68,738(6.8%)	306,930(6.8%)	Dance	20(4.0%)	14,250(1.4%)	71,740(1.6%)
Makeup	20(4.0%)	48,226(4.8%)	111,814(2.5%)	Fishery	19(3.7%)	74,386(7.4%)	230,085(5.1%)
Talent	21(4.1%)	42,246(4.2%)	339,382(7.5%)	Fashion	20(4.0%)	17,849(1.8%)	48,942(1.1%)
Beauty Industry	17(3.3%)	31,918(3.2%)	141,035(3.1%)	Lesson	20(4.0%)	42,086(4.2%)	186,721(4.1%)

Table 5: **The Data Distribution for 30 Open Scenarios.** Red refers to these scenarios only supported by MMVText. "%" denotes the percentage of each scenario data for whole data.

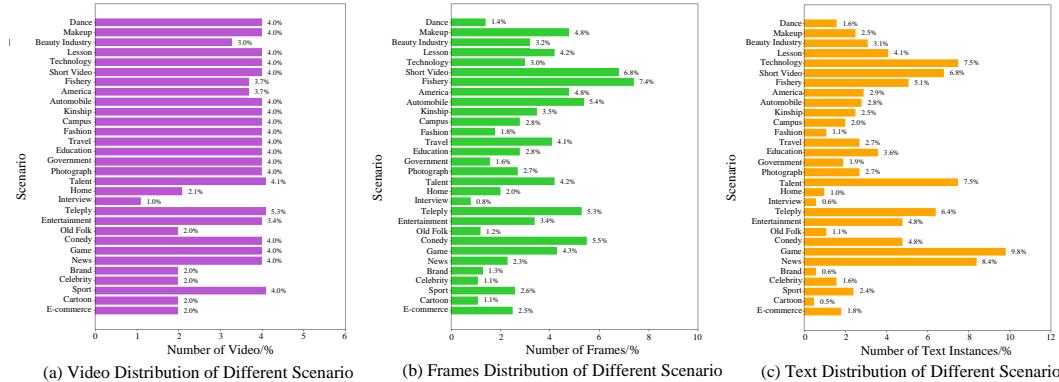


Figure 6: **The Data Distribution for 30 Open Scenarios.** (a) Video distribution for different scenarios. (b) Video frames distribution for different scenarios. (c) Text instance distribution for different scenarios.

### 609 A.3 More statistical information for MMVText.

#### 610 Checklist

- 611 1. For all authors...
  - 612 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
  - 613 (b) Did you describe the limitations of your work? [Yes] We describe the limitations in supplementary material.
  - 614 (c) Did you discuss any potential negative societal impacts of your work? [Yes]
  - 615 (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
- 616 2. If you are including theoretical results...
  - 617 (a) Did you state the full set of assumptions of all theoretical results? [N/A]
  - 618 (b) Did you include complete proofs of all theoretical results? [N/A]
- 619 3. If you ran experiments (e.g. for benchmarks)...
  - 620 (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have provided the URL concerning the coding and the data to promote further research.
  - 621 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
  - 622 (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

	Languages			Text Category	
	English	Chinese	Alphanumeric	Caption Text	Scene Text
Video	GT 141(27.6%)	432(84.7%)	158(31.0%)	389(76.3%)	359(70.4%)
Video Frames	325,739(32.2%)	856,504(85.8%)	367,991(36.4%)	803,382(79.5%)	542,380(53.7%)
Text Instances	1,364,690(30.2%)	3,057,751(67.8%)	1,455,774(32.2%)	2,182,474(48.4%)	2,331,051(51.6%)

Table 6: Statistics of text language and category in MMVText.

- 630 (d) Did you include the total amount of compute and the type of resources used (e.g., type  
 631 of GPUs, internal cluster, or cloud provider)? [Yes]
- 632 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...  
 633 (a) If your work uses existing assets, did you cite the creators? [Yes]  
 634 (b) Did you mention the license of the assets? [Yes]  
 635 (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]  
 636 (d) Did you discuss whether and how consent was obtained from people whose data you're  
 637 using/curating? [N/A]  
 638 (e) Did you discuss whether the data you are using/curating contains personally identifiable  
 639 information or offensive content? [Yes] We have blurred identifiable information or  
 640 offensive content.
- 641 5. If you used crowdsourcing or conducted research with human subjects...  
 642 (a) Did you include the full text of instructions given to participants and screenshots, if  
 643 applicable? [Yes]  
 644 (b) Did you describe any potential participant risks, with links to Institutional Review  
 645 Board (IRB) approvals, if applicable? [N/A]  
 646 (c) Did you include the estimated hourly wage paid to participants and the total amount  
 647 spent on participant compensation? [Yes] We have paid salary to the related participants.