Supplementary Material for ‘A Bilingual, Open World Video Text Dataset and End-to-end Video Text Spotter with Transformer’

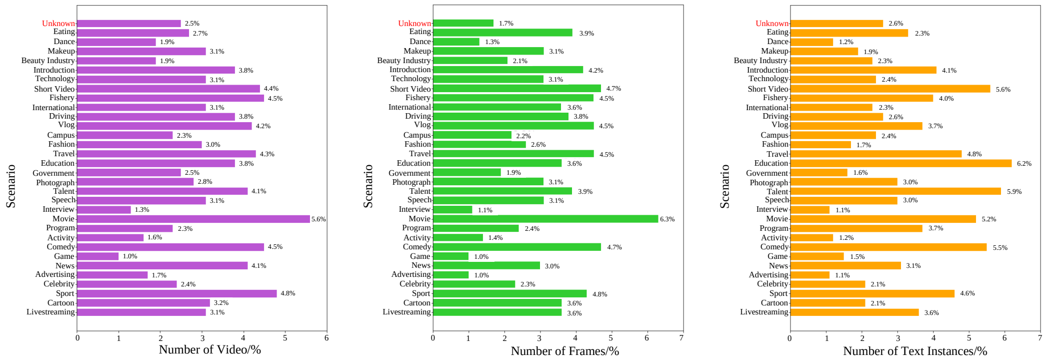
A Appendix

A.1 Data Distribution for BOVText

The data distribution for 32 open scenarios. We present the 31 open scenarios and an "Unknown" scenarios distribution on BOVText in three levels, *i.e.,* video, video frames, and text instances, as shown in Figure. 1 and Table 1. BOVText spans various video domains with these scenarios in the existing datasets (*e.g.,* Driving for RoadText-1k [22], Vlog(supermarket, shopping street, indoor), Travel(hotel, railway station) for LSVTD [3]) and more open domains that are not yet supported (*e.g.,* Game(PUBG mobile, Honor of Kings...), Sport(NBA, world cup...), News). We expect that the open-world scenarios can provide a comprehensive effectiveness evaluation for advanced video text models in different scenarios.

Scenarios	Video	Video Frames	Text Instances	Scenarios	Video	Video Frames	Text Instances
Cartoon	61(3.2%)	60,395(3.6%)	123,191(2.1%)	Sport	91(4.8%)	72,469(4.3%)	266,996(4.6%)
Vlog	81(4.2%)	76,056(4.5%)	214,910(3.7%)	News Report	79(4.1%)	48,868(3.0%)	178,000(3.1%)
Driving	72(3.8%)	61,656(3.7%)	151,994(2.6%)	Celebrity	46(2.4%)	39,440(2.3%)	121,235(2.1%)
Advertising	33(1.7%)	29,329(1.0%)	91,090(1.0%)	Technology	59(3.1%)	52,072(3.1%)	140,172(2.4%)
Activity	31(1.6%)	23,585(1.4%)	67,879(1.2%)	Program	43(2.3%)	40,784(2.4%)	214,561(3.7%)
Comedy	86(4.5%)	79,404(4.7%)	317,865(5.5%)	Game	8(1.0%)	12,565(1.0%)	84,106(1.5%)
Interview	24(1.3%)	18,229(1.1%)	63,616(1.1%)	Livestreaming	59(3.1%)	60,494(3.6%)	211,569(3.6%)
Government	47(2.5%)	32,283(1.9%)	93,874(1.6%)	Speech	59(3.1%)	52,465(3.1%)	175,119(3.0%)
Travel	83(4.3%)	75,266(4.5%)	280,446(4.8%)	Movie	107(5.6%)	105,949(6.3%)	299,760(5.2%)
Campus	44(2.3%)	37,556(2.2%)	139,760(2.4%)	Photograph	53(2.8%)	52,771(3.1%)	173,832(3.0%)
International	60(3.1%)	60,486(3.6%)	132,117(2.3%)	Education	73(3.8%)	60,315(3.6%)	360,774(6.2%)
Short Video	84(4.4%)	79,148(4.7%)	326,930(5.6%)	Dance	37(1.9%)	22,264(1.3%)	71,740(1.2%)
Makeup	60(3.1%)	52,449(3.1%)	111,814(1.9%)	Fishery	86(4.5%)	79,750(4.7%)	230,085(4.0%)
Talent	79(4.1%)	66,038(3.9%)	339,382(5.9%)	Fashion	57(3.0%)	42,868(2.6%)	98,942(1.7%)
Beauty Industry	37(1.9%)	35,851(2.1%)	132,025(2.3%)	Introduction	73(3.8%)	42,086(4.2%)	236,721(4.1%)
Eating	52(2.7%)	65,609(3.9%)	191,035(3.3%)	Unknown	48(2.5%)	29,212(1.7%)	150,721(2.6%)

Table 1: **The Data Distribution for 32 Open Scenarios**. In green refers to these scenarios only supported by BOVText. "%" denotes the percentage of each scenario data for whole data.



(a) Video Distribution of Different Scenario (b) Frames Distribution of Different Scenario (c) Text Distribution of Different Scenario

Figure 1: **The Data Distribution for 32 Open Scenarios**. (a) Video distribution for different scenarios. (b) Video frames distribution for different scenarios. (c) Text instance distribution for different scenarios.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

	Languages			Text Category			
	English	Chinese	Alphanumeric	Caption Text	Scene Text	Title	Others
Video	842(45%)	1521(82%)	1023(55%)	1523(82%)	1336(72%)	222(12%)	148(8%)
Video Frames	671,462(40%)	1,449,824(86%)	775,021(46%)	1,327,382(79%)	930,359(55%)	167,011(10%)	116,914(7%)
Text Instances	2,785,643(38%)	5,308,766(72%)	3,150,256(43%)	3,296,101(39%)	3,981,574(52%)	291,690(4%)	364,613(5%)

Table 2: **Statistics of text language and category in BOVText.** In green refers to these text category only supported by BOVText. "%" denotes the percentage of each scenario data for whole data.

The data distribution for multilingual text and text categories. As shown in Figure. 2, BOVText provides two language text annotation and for text categories annotation (*i.e.,* caption, title, scene text, or others). BOVText, as the first benchmark for support various text categories annotation, can provide a novel perspective and in favor of other video-and-language tasks, more details in A.6 Link to Other Video-and-Language tasks and video understanding demo video in the supplementary material. Besides, BOVText, as the first large-scale, multilingual video text benchmark dataset, can promote video text multilingual spotting in the community.

A.2 BOVText Metrics

The proposed BOVText includes four tasks: (1) Video Text Detection; (2) Video Text Recognition; (3) Video Text Tracking. (4) End to End Text Spotting in Videos. $MOTP$ (Multiple Object Tracking Precision) [1], $MOTA$ (Multiple Object Tracking Accuracy) and IDF_1 [5, 23] as the three important metrics are used to evaluate task3 (text tracking) and task4 (text spotting) for BOVText. Following the previous works [9, 22], BOVText evaluates text tracking methods in video and compare their performance with the MOTA and MOTP, which are given by:

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

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

$$MOTA = 1 - \frac{\sum_t(m_t + fp_t + mme_t)}{\sum_t g_t} \,, \tag{2}$$

where m_t, fp_t and mme_t are the number of misses, false positives, and mismatches, respectively. g_t is the number of objects present at time t. It shows the tracker's performance at detecting objects and keeping their trajectories, independent of the precision of the location. ID_{F1} is the ratio of correctly identified detections over the average number of ground-truth and computed detections. And the metric is more reasonable to evaluate ID switches in some cases. We evaluate the metrics in BOVText by:

$$ID_{tp} = \sum_h \sum_t m(h, o, \triangle_t, \triangle_s) \,, \tag{3}$$

$$ID_{F1} = \frac{2ID_{tp}}{2ID_{tp} + ID_{fp} + ID_{fn}} \,, \tag{4}$$

where ID_{tp}, ID_{fp} and ID_{fn} refer to true positive, false positive and false negative of matching ID. Besides, the ID metric [5] also includes MT (Mostly Tracked) Number of objects tracked for at least 80 percent of lifespan, ML (Mostly Lost) Number of objects tracked less than 20 percent of lifespan. $\triangle_t$ and $\triangle_s$ refer to time matching and space location matching, respectively.

For Task4 (End to End Text Spotting in Videos), the objective of this task is to recognize words in the video as well as localize them in terms of time and space. We use ID_{F1} to evaluate our BOVText, which focuses on text instance ID tracking and recognition results that be required by many downstream tasks. More specifically,

$$ID_{tp} = \sum_h \sum_t m(h, o, \triangle_t, \triangle_s, \triangle_r) \,, \tag{5}$$

$$ID_{F1} = \frac{2TID_{tp}}{2TID_{tp} + TID_{fp} + TID_{fn}} \,, \tag{6}$$

where $\triangle_t$, $\triangle_s$ and $\triangle_r$ refer to ID matching, space location matching and recognition result matching. h and o denote hypothesis set (*e.g.,* predicted ID I_p , box locations L_p, recognition results R_p) and

Method	Text Detection on ICDAR2015/%			Text Tracking on ICDAR2015(video)/%		
	Precision	Recall	F-score	$\mathrm{ID_{F1}}$/%	MOTA	MOTP
Ours,w/o angle prediction	73.5	65.1	68.5	36.1	16.4	68.7
Ours,w/ angle prediction(L1 loss)	84.2	82.3	83.2	54.7	43.2	74.5
Ours,w/ angle prediction(cosine loss)	86.8	81.7	84.2(+15.7)	57.3(+21.2)	44.1(+27.7)	75.8(+7.1)

Table 3: **Ablation study for angle prediction.** The gaps of at least (+7.1%) improvement after using angle prediction are shown in green.

Query	Text Tracking on ICDAR2015(video)/%			$\mathrm{ShortSide}$	Text Tracking on ICDAR2015(video)/%			
	$\mathrm{ID_{F1}}$/%	MOTA	MOTP		$\mathrm{ID_{F1}}$/%	MOTA	MOTP	FPS
Obejct query	54.2	40.5	76.6	480	52.6	41.5	74.5	16
Obejct + tracking query	57.3	44.1	75.8	640	56.7	43.8	74.5	13
-	-	-	-	800	57.3	44.1	75.8	9

Table 4: **Ablation study for input query and input image size.** Tracking query bring huge improvement with 3.1% $\mathrm{ID_{F1}}$.

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

$$\triangle_t : I_p = I_g, \quad \triangle_s : IoU(L_p, L_g) > 0, \quad \triangle_r : R_p = R_g. \tag{7}$$

The match of h and o is a true positives of text ID (*i.e.,* ID_{tp}) when these conditions (*i.e.,* $\triangle_t$, $\triangle_s$ and $\triangle_r$ are met. Similarly, false positive (*i.e.,* ID_{fp}) and false negative (*i.e.,* ID_{fn}) of text ID can be obtained for ID_{F1} calculation.

A.3 More details for experiments

TransVTSpotter. All experiments are conducted on Tesla V100 GPU. We train the model for 150 epochs and the learning rate drops by a factor of 10 at the 100th epoch.

Baselines concerning the existing methods. Video-based text spotting methods are rare and lack open-source algorithms. Therefore, we adopt various mature image-based techniques to compare and evaluate the efficiency of BOVText. **Detection**. EAST [38] as one of the popular regression-based methods is used to test our BOVText. The method adopts FCNs to predict shrinkable text score maps, rotation angles. For segmentation based methods, we adopt PSENet [30] and DB [14] to evaluate our BOVText. PSENet [30] generates various scales of shrinked text segmentation maps, then gradually expands kernels to generate the final text instance. **Recognition**. Recent methods mainly include two techniques, Connectionist Temporal Classification (CTC) and attention mechanism. In CTC-based methods, CRNN [25] as the representation, which introduced CTC decoder into scene text recognition with BiLSTM to model the feature sequence. In Attention-based methods, RARE [26] normalizes the input text image using the Spatial Transformer Network (STN [7]). Then, it estimates the output character sequence from the identified features with the attention module. **Text Tracking Trajectory Generation**. With text detection and recognition in a static image, we only obtain text localization and recognition information without temporal information, which are insufficient for video spotting evaluation (*e.g.,* ID_{F1},$MOTA$ and $MOTP$). Following the work [31], we link and match text objects in the current frame and several frames by IOU and edit the distance of text. All of the experiments use the same strategy: (1) Training detector and recognizer with BOVText. (2) Matching text objects with corresponding text tracking trajectory id.

A.4 Ablation study for TransVTSpotter

Angle Prediction. The angle prediction and corresponding loss are the main contributions for TransVTSpotter. As shown in Table 3, we conduct three experiments to test the effectiveness of the angle prediction and corresponding cosine loss. Without using angle prediction, model with upright-bounding box(two points) results show a dissatisfactory performance(*i.e.,* 68.5% f-score for detection and 36.1% $\mathrm{ID_{F1}}$ for tracking). Compared with using angle prediction, there is around 20% performance gap. Besides, using cosine loss present a better performance than the counterpart of L1 loss(84.2% v.s 83.2% for text detection F-score)

Tracking Query Tracking query is important for our framework, which uses the knowledge of previously detected objects to obtain a set of tracking boxes. As shown in Table 4, for only object

Method	ICDAR2013 [9](video,D)/%			ICDAR2013 [21](image,D)/%			ICDAR2015 [39](image,D)/%			FPS
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	
Wang *et al.* [33]	71.9	58.6	62.6	-	-	-	-	-	-	-
Wu *et al.* [34]	63.0	68.0	65.0	-	-	-	-	-	-	-
Yu *et al.* [37]	**82.4**	56.4	66.9	-	-	-	-	-	-	-
Wei *et al.* [6]	75.5	64.1	69.3	-	-	-	-	-	-	9.6
TextBoxes [13]	-	-	-	88.0	83.0	85.0	-	-	-	1.5
SegLink [24]	-	-	-	87.7	83.0	85.3	73.1	76.8	75.0	8.9
Mask Textspotter [16]	-	-	-	88.6	**95.0**	**91.7**	**91.6**	81.0	86.0	4.8
CharNet [35]	-	-	-	-	-	-	88.3	**91.1**	**89.7**	-
TransVTSpotter(ours)	73.4	**69.4**	**71.1**	**91.4**	89.6	91.0	89.8	84.7	87.2	**13.0**

Table 5: **Experiments for TransVTSpotter on ICDAR2015(image) [9], ICDAR2013 [9] and ICDAR2013(video) [9].** 'D','R', 'T' and 'S' denotes the Detection, Recognition, Tracking, Spotting, respectively. 'video' or 'image' denote the image or video level dataset.

Method	ICDAR2015(video,T)/%			ICDAR2015(video,S)/%			Minetto [19](video,T)/%			YVT [21](video,T)/%		
	MOTA	MOTP	$\mathrm{ID_{F1}}$	MOTA	MOTP	$\mathrm{ID_{F1}}$	MOTA	MOTP	$\mathrm{ID_{F1}}$	MOTA	MOTP	$\mathrm{ID_{F1}}$
USTB_TexVideo [8]	7.4	70.7	25.9	15.6	68.5	28.2	-	-	-	-	-	-
StradVision-1 [8]	7.9	70.2	25.8	8.9	70.2	31.9	-	-	-	-	-	-
USTB_TexVideo II-2 [8]	12.3	71.8	21.9	13.2	66.6	21.3	-	-	-	-	-	-
AJOU [8, 10]	16.4	72.7	36.1	-	-	-	-	-	-	-	-	-
AGD with AGD [37]	-	-	-	-	-	-	75.6	74.7	-	-	-	-
AGD with EAGD [37]	-	-	-	-	-	-	81.3	75.7	-	-	-	-
Wei *et al.* [6]	-	-	-	-	-	-	83.5	76.8	-	-	-	-
Free [4]	43.2	**76.8**	**57.9**	52.9	74.8	**61.8**	-	-	-	**54.0**	**78.0**	-
TransVTSpotter(ours)	**44.1**	75.8	57.3	**53.2**	**74.9**	61.5	**84.1**	**77.6**	**74.7**	53.9	75.9	**64.5**

Table 6: **Experiments for TransVTSpotter on ICDAR2015(video), Minetto and YVT.** 'D','R', 'T' and 'S' denotes the Detection, Recognition, Tracking, Spotting, respectively. 'video' or 'image' denote the image or video level dataset.

query, with only learned object query is input as decoder query, and associating the generated detection to track. This solution achieves a not bad performance by $54.2\ \mathrm{ID_{F1}}$. But, with tracking query, the performance present further improvement($3.1\%\ \mathrm{ID_{F1}}$)

Accuracy vs. Speed We analyze the inference speed of TransVTSpotter. The time cost is measured using a single Tesla V100 GPU. Table 4 shows the effect of input image size. With input image size increasing, the model present a better $\mathrm{ID_{F1}}$ performance from 52.6 to 67.3. When the short-side of the input image is by 800 pixels, the speech and performance all present relatively satisfactory results, so we set it as the default setting in the experiment.

A.5 Comparison with State-of-The-Arts for TransVTSpotter

As shown in Table 5, we evaluate our TransSpotter on ICDAR2015(image) [9], ICDAR2013 [9] and ICDAR2013(video) [9] for detection task. The proposed TransSpotter presents a great performance with 71.1% F-score on video level dataset(ICDAR2013 Text in video), at least 0.8% improvement than the previous methods. Although for image-based datasets(*e.g.,* ICDAR2013(image) and ICDAR2015(imgae), TransSpotter also give competitive results with 91.0% and 87.2%, respectively. As shown in Table 6, we evaluate TransSpotter on ICDAR2015(video) [39], Minetto(video) [19] and YVT(video) [21] for tracking and video text spotting task. Minetto consists of 5 videos in outdoor scenes. The frame size is 640 x 480 and all videos are used for test when the model training on ICDAR2015(video). Our TransVTSpotter obtains 44.1% and 84.1% for tracking task(MOTA) on ICDAR2015(video) and Minetto, at least 0.6% improvement than the previous models.

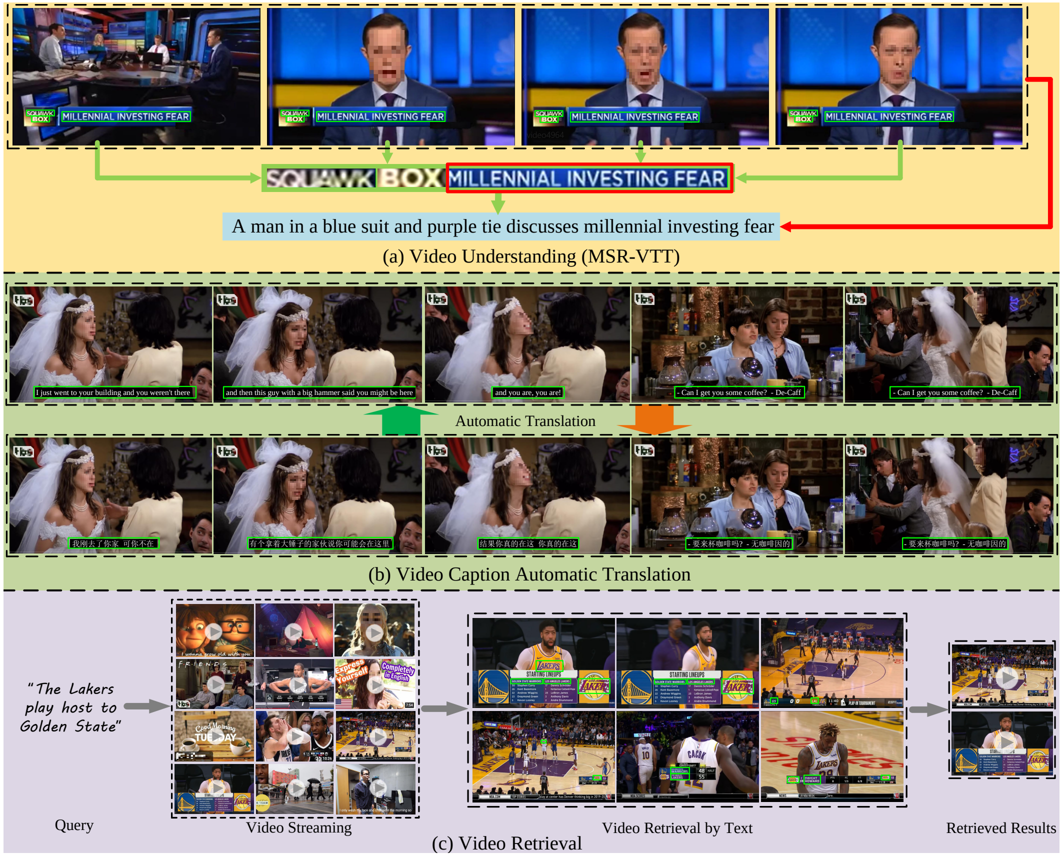
A.6 Link to Other Video-and-Language Applications

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

Text spotting in static images has numerous application scenarios: (1) Automatic data entry. SF-Express [1] utilizes OCR techniques to accelerate the data entry process. NEBO [2] performs instant transcription as the user writes down notes. (2) Autonomous vehicle [18, 17]. Text-embedded panels

[1] https://www.sf-express.com/cn/sc/

[2] https://www.myscript.com/nebo/


SQUAWK BOX
MILLENNIAL INVESTING FEAR
A man in a blue suit and purple tie discusses millennial investing fear
I just went to your building and you weren't there
and then this guy with a big hammer said you might be here
and you are, you are!
- Can I get you some coffee? - De-Caff
- Can I get you some coffee? - De-Caff
Automatic Translation
我刚去了你家 可你不在
有个拿着大锤子的家伙说你可能会在这里
结果你真的在这 你真的在这
- 要来杯咖啡吗? - 无咖啡因的
- 要来杯咖啡吗? - 无咖啡因的
"The Lakers play host to Golden State"
Query
Video Streaming
Video Retrieval by Text
Retrieved Results

(a) Video Understanding (MSR-VTT)

(b) Video Caption Automatic Translation

(c) Video Retrieval

Figure 2: **The Real Application Tasks Link to BOVText**. (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.

carry important information, *e.g.,* geo-location, current traffic condition, navigation, and etc. (3) Text-based reading comprehension. TextCaps [27] and text-based VQA [28, 2] show the new vision-and-language tasks, which need to recognize text, relate it to its visual context, semantic, and visual reasoning between multiple text tokens and visual entities, such as objects. Similarly, there are many application demands for video text understanding across various industries and in our daily lives. We list the most outstanding ones that significantly impact, improving our productivity and life quality. **Firstly**, automatically describing video with natural language [36, 32] can bridge video and language. **Secondly**, video text automatic translation [3] can be extremely helpful as people travel, and help video-sharing websites [4] to cut down language barriers. **Finally,** text-based video retrieval [11, 15] is an irreplaceable business for many companies, such as Google and YouTube. More details and analyses for application scenarios concerning BOVText in the supplementary material.

Video Understanding. As shown in Figure. 2 (a), the example concerning the task of describing video with natural language is from MSR-VTT [36], and there has been increasing interest in video understanding [32, 12]. However, video description with only visual information is difficult and limited, even for a human. For the annotation of the sample video, *i.e., "A man in a blue suit and purple tie discusses millennial investing fear"*, we can not learn the information of "millennial investing fear" from the visual information in the video. By comparison, caption and scene texts in the video contain accurate information of *"millennial investing fear"*, which can help the model to describe the video better. We argue that the same as general human understanding, videos without captions and audio, is difficult to be properly understood by the model. We hope the release of BOVText can promote efficient video text reading, further enhancing automatic video description.

[3] https://translate.google.com/intl/en/about/

[4] https://www.youtube.com/

Video Text Automatic Translation. Another practical application is video text automatic translation, as shown in Figure. 2 (b). The application may be unnecessary for several professional teams or classic movies due to the professional translator or huge cost investment. But for international video-sharing websites [5] [6] with millions of users, it isn't easy to apply multilingual caption and scene text in billions of videos. Therefore, efficient translation concerning caption text (e.g., overlap, song title, logos) and scene text (e.g., street signs, business signs, words on shirt) still need further exploration and research. The large-scale and multilingual BOVText contributes various real scenarios for the development of video text automatic translation.

Video Retrieval. Video retrieval with textual cues [20, 29] is also a very important application direction for video-and-text research, as shown in Figure. 2 (c). To the best of my knowledge, video retrieval with text information in the video is still almost a blank field of study and immature application in the industry. The most existing video retrieval methods are stiff combinations of text detection and recognition, invalid for the example with a sentence query. Besides, similar to video understanding, for the query of the sample video, *i.e., "The lakers play host to Golden State"*, we can not obtain the correct related video without scene text or caption information. The missing information needs to recover by understanding the video with key video text information such as *"GOLDEN STATE WARRIORS, LOS ANGELES LAKERS"*. The proposed BOVText with various text types (*e.g.,* caption, song title, logos, street signs, business signs) and annotation can promote the research concerning efficient video retrieval.

A.6.1 Limitations

Although the proposed BOVText supports all video text spotting tasks, *i.e., text detection, recognition, tracking end to end video text spotting*, the potential contributions for other tasks still need mining. For example, as shown in Figure. 2 (c), we do not provide the corresponding annotation (*i.e.,* the query for each video) and metrics concerning video retrieval, but the annotation and metric are easy to obtain due to text spotting annotation already existed. Therefore, there are still many potential contributions for other tasks on BOVText, we want to take these as the future research directions and provide a complete solution method.

A.7 Potential Negative Societal Impacts and Solution

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

How to blur human faces. We blur the human faces in BOVText with four steps. Firstly, human faces in each frame would be detected by *face recognition*[7] - a powerful, simple, and easy-to-use face recognition open source project with complete development documents and application cases. Secondly, with the location box from the previous step, we extract face ROI from the original image. Thirdly, we blur the face ROI with Gaussian Blur operation in OpenCV[8]. Finally, we store the blurred face in the original image and recover to video.

A.8 License and Copyright

The released video dataset includes two parts: $1,494$ videos from *KuaiShou* [9] and 356 videos from *YouTube* [10]. For those videos from *KuaiShou*, we mask the private information such as the human face, which has passed the examination of the legal department and copyright department of KuaiShou corporation. Thus, we own the copyright for these videos. For those videos from *YouTube*, to the best of our knowledge at the time of download, we have exercised caution to download only those videos that were available on YouTube with a Creative Commmons CC-BY (v3.0) License. We don't own the copyright of those videos and provide them for non-commercial research purposes only. All data in our project is open source under CC-by 4.0 license and only be used for research purposes.

[5] https://www.youtube.com/

[6] https://www.kuaishou.com/en

[7] https://github.com/ageitgey/face_recognition

[8] https://www.tutorialspoint.com/opencv/opencv_gaussian_blur.htm

[9] https://www.kuaishou.com/en

[10] https://www.youtube.com/

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[2] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019.

[3] Zhanzhan Cheng, Jing Lu, Yi Niu, Shiliang Pu, Fei Wu, and Shuigeng Zhou. You only recognize once: Towards fast video text spotting. In *ACM International Conference on Multimedia*, pages 855–863, 2019.

[4] Zhanzhan Cheng, Jing Lu, Baorui Zou, Liang Qiao, Yunlu Xu, Shiliang Pu, Yi Niu, Fei Wu, and Shuigeng Zhou. Free: A fast and robust end-to-end video text spotter. *IEEE Transactions on Image Processing*, 30:822–837, 2020.

[5] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixe. Cvpr19 tracking and detection challenge: How crowded can it get? *arXiv preprint arXiv:1906.04567*, 2019.

[6] Wei Feng, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Semantic-aware video text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1695–1705, 2021.

[7] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Neural Information Processing Systems*, pages 2017–2025, 2015.

[8] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *IEEE International Conference on Document Analysis and Recognition*, pages 1156–1160, 2015.

[9] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *IEEE International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.

[10] Hyung Il Koo and Duck Hoon Kim. Scene text detection via connected component clustering and nontext filtering. *IEEE transactions on image processing*, 22(6):2296–2305, 2013.

[11] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020.

[12] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.

[13] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[14] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI Conference on Artificial Intelligence*, pages 11474–11481, 2020.

[15] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910, 2020.

[16] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *European Conference on Computer Vision*, pages 67–83, 2018.

[17] Abdelhamid Mammeri, Azzedine Boukerche, et al. Mser-based text detection and communication algorithm for autonomous vehicles. In *2016 IEEE symposium on computers and communication (ISCC)*, pages 1218–1223. IEEE, 2016.

[18] Abdelhamid Mammeri, El-Hebri Khiari, and Azzedine Boukerche. Road-sign text recognition architecture for intelligent transportation systems. In *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pages 1–5. IEEE, 2014.

[19] Rodrigo Minetto, Nicolas Thome, Matthieu Cord, Neucimar J Leite, and Jorge Stolfi. Snoopertrack: Text detection and tracking for outdoor videos. In *IEEE International Conference on Image Processing*, pages 505–508, 2011.

[20] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3040–3047, 2013.

[21] Phuc Xuan Nguyen, Kai Wang, and Serge Belongie. Video text detection and recognition: Dataset and benchmark. In *IEEE winter conference on applications of computer vision*, pages 776–783, 2014.

[22] Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Roadtext-1k: Text detection & recognition dataset for driving videos. In *IEEE International Conference on Robotics and Automation*, pages 11074–11080, 2020.

[23] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Workshops of European conference on computer vision*, pages 17–35, 2016.

[24] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2550–2558, 2017.

[25] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.

[26] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *IEEE conference on computer vision and pattern recognition*, pages 4168–4176, 2016.

[27] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758. Springer, 2020.

[28] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[29] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. *arXiv preprint arXiv:2104.01552*, 2021.

[30] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *IEEE conference on computer vision and pattern recognition*, pages 9336–9345, 2019.

[31] Xiaobing Wang, Yingying Jiang, Shuli Yang, Xiangyu Zhu, Wei Li, Pei Fu, Hua Wang, and Zhenbo Luo. End-to-end scene text recognition in videos based on multi frame tracking. In *IEEE International Conference on Document Analysis and Recognition*, pages 1255–1260, 2017.

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

[33] Yang Wang, Lan Wang, and Feng Su. A robust approach for scene text detection and tracking in video. In *Pacific Rim Conference on Multimedia*, pages 303–314. Springer, 2018.

[34] Liang Wu, Palaiahnakote Shivakumara, Tong Lu, and Chew Lim Tan. A new technique for multi-oriented scene text line detection and tracking in video. *IEEE Transactions on multimedia*, 17(8):1137–1152, 2015.

[35] Linjie Xing, Zhi Tian, Weilin Huang, and Matthew R Scott. Convolutional character networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9126–9136, 2019.

[36] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[37] Hongyuan Yu, Yan Huang, Lihong Pi, Chengquan Zhang, Xuan Li, and Liang Wang. End-to-end video text detection with online tracking. *Pattern Recognition*, 113:107791, 2021.

[38] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *IEEE conference on computer vision and pattern recognition*, pages 5551–5560, 2017.

[39] Xinyu Zhou, Shuchang Zhou, Cong Yao, Zhimin Cao, and Qi Yin. Icdar 2015 text reading in the wild competition. *arXiv preprint arXiv:1506.03184*, 2015.