VKIE: The Application of Key Information Extraction on Video Text

Siyu An[1]*, **Ye Liu**[2]*, **Haoyuan Peng**[3] **and Di Yin**[1]
[1]Tencent YoutuLab [2]Nvidia [3]Learnable.ai
{siyuan, endymecyyin}@tencent.com
liuyebug@126.com, haoyuan.peng@learnable.ai

Abstract

Extracting structured information from videos is critical for numerous downstream applications in the industry. In this paper, we define a significant task of extracting hierarchical key information from visual texts on videos. To fulfill this task, we decouple it into four subtasks and introduce two implementation solutions called PipVKIE and UniVKIE. PipVKIE sequentially completes the four subtasks in continuous stages, while UniVKIE is improved by unifying all the subtasks into one backbone. Both PipVKIE and UniVKIE leverage multimodal information from vision, text, and coordinates for feature representation. Extensive experiments on one well-defined dataset demonstrate that our solutions can achieve remarkable performance and efficient inference speed.

1 Introduction

Extracting information from video text is an essential task for many industrial video applications, i.e., video retrieval(Radha, 2016), video recommendation(Yang et al., 2007), video indexing(Yang et al., 2011), etc. Visual text embedded in videos usually carries rich semantic descriptions about the video contents, and this information gives a high-level index for content-based video indexing and browsing.

Conventional methods utilize OCR (Liao et al., 2018; Tian et al., 2016; Zhou et al., 2017) to extract visual texts from videos frames and employ text classification techniques (Le et al., 2018; Li et al., 2020) to categorize the extracted content. However, these methods suffer from two significant shortcomings: 1) Visual texts are typically coarse-grained at the segment level, and are unable to capture fine-grained information at the entity level, which is critical for downstream tasks. 2) Traditional methods have not fully utilized the fusion of features from different modalities.

*These authors contributed equally to this work.

Subtitle:	I achieved everything with the national team, as I always dreamed of. I achieved everything in my career, at Barcelona
Person Info.	Lionel Messi, 2022 World Cup Winner
Person Name	Lionel Messi
Person Identity	2022 World Cup Winner
Entity linking	(2022 World Cup Winner, Lionel Messi)

Figure 1: An example of hierarchical key information extracted by VKIE in a video frame (CGTN Sports Scene, 2023).

Therefore, in our work, we introduce a novel industrial task for extracting key information from video text and exploring the relationship between entities, which we refer to as VKIE. The task aims to extract valuable hierarchical information from visual texts, explore their relationships, and organize them in structured forms. This approach enables effective management and organization of videos through the use of rich hierarchical tags, which can be utilized to index, organize, and search videos at different levels. Figure 1 provides an example of the hierarchical key information extracted by VKIE, where subtitles are captured at the segment level, and personal information is organized with names and identities at the entity level.

To enhance clarity, we decompose VKIE into four subtasks: text detection and recognition (TDR), box text classification (BTC), entity recognition (ER), and entity linking (EL). While the first subtask, TDR, is typically accomplished using off-the-shelf OCR tools, our work concentrates on the remaining three subtasks of BTC, ER, and EL.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 532–540

Since TDR outputs all boxes with text content and coordinates information, there are massive useless texts, such as scrolling texts and blurred background texts, which could have side effects on downstream tasks. BTC aims to eliminate these useless texts and find valuable categories, such as title, subtitle, and personal information.

Although the BTC method can obtain segment level information, the results are relatively coarse-grained and will limit its deployment to many downstream applications. For example, in video-text retrieval, the query is usually in different forms, such as keywords, phrases, or sentences. In video indexing, a video is required to be stored with hierarchical tags. To address these issues, we designed ER to extract entities from text segments and EL to explore the relations among the entities. With this structured information, videos can be well managed with rich hierarchical information at the entity and segment levels.

In this paper, we present two solutions that have been deployed in our industry system. The first approach, called PipVKIE, involves performing the tasks sequentially, which serves as our baseline method. The second approach, called UniVKIE, achieves better performance and efficiency by more effectively integrating multimodal features.

In summary, our contributions are as follows:

(1) We define a new task in the industry to extract key information from video texts. By this means, structured information could be effectively extracted and well managed at hierarchical levels.

(2) We introduce and compare two deployed solutions based on the framework includes TDR, BTC, ER, and EL. Experiments show our solutions can achieve remarkable performance and efficient inference speed.

(3) To make up the lack of datasets, we construct a well-defined dataset to provide comprehensive evaluations and promote this industrial task.

2 Approaches

2.1 PipVKIE

The PipVKIE solution fulfills three subtasks of BTC, ER, and EL in a sequential pipeline and processes a single visual box at a time. In this process, BTC acts as a filter, selecting only the valuable text segments. After BTC performs, ER is carried out only on the segments selected by BTC. Similarly, when performing EL, only the entities extracted by ER are inputted, while other irrelevant information is filtered out.

BTC In our design, the objective of BTC is to categorize the text segments that appear on the OCR boxes into different classes, such as titles and subtitles. As illustrated in Fig.2, in PipVKIE, BTC takes the visual and textual features as input and outputs the corresponding class label. Specifically, for visual modality, in contrast to conventional approaches that usually use the classical VGG (Simonyan and Zisserman, 2014) or ResNet-based (He et al., 2016) network, we construct a shallow neural network as the backbone. In fact, we observe that texts differ in low-level features of colors and fonts, thus the above-mentioned deeper networks are abandoned as high-level semantic information is extracted. Consequently, transformer (Vaswani et al., 2017) is selected as the backbone of textual extraction.

The fusion of multimodal features is a critical step in obtaining the multimodal representation of one box. The process of visual and positional modalities is shown below:

$$\boldsymbol{h}_{vb} = \mathrm{Trans}(\mathrm{ROIAlign}(\boldsymbol{h}_{vf}, \boldsymbol{h}_p), \boldsymbol{h}_{vf}) \quad (1)$$

where $\boldsymbol{h}_{vf}$ is visual embedding of frame directly obtained by CNN (Krizhevsky et al., 2012), and $\boldsymbol{h}_p$ is positional embedding of box obtained by coordinates respectively. Firstly, ROIAlign (He et al., 2017) is utilized to extract visual box embedding conditioned on $\boldsymbol{h}_p$ and $\boldsymbol{h}_{vf}$. Then, we take the transformer (Vaswani et al., 2017) to learn the implicit relation between a box and its corresponding frame, which denoted as $\boldsymbol{h}_{vb}$. The visual box embedding $\boldsymbol{h}_{vb}$ and the textual box embedding $\boldsymbol{h}_{tb}$, which is obtained by applying the transformer encoder on text, are simply concatenated to obtain the final multimodal vector representation $\boldsymbol{h}_b$. Subsequently, we perform softmax classification by multiplying $\boldsymbol{h}_b$ with trainable weight parameters.

ER Contrary to commonly known NER in flat text (Lample et al., 2016), the goal of ER in VKIE is to identify entities from a single video frame. In this context, factors such as the entity's position and background features can significantly influence the recognition process. In PipVKIE, what we need to accomplish at this stage is the extraction of entities from the valuable text segments selected in the previous step. We obtain the hidden representation of text tokens by transformer encoder, and then predict their tags with the BIO2 tagging schema (Sang and Veenstra, 1999).

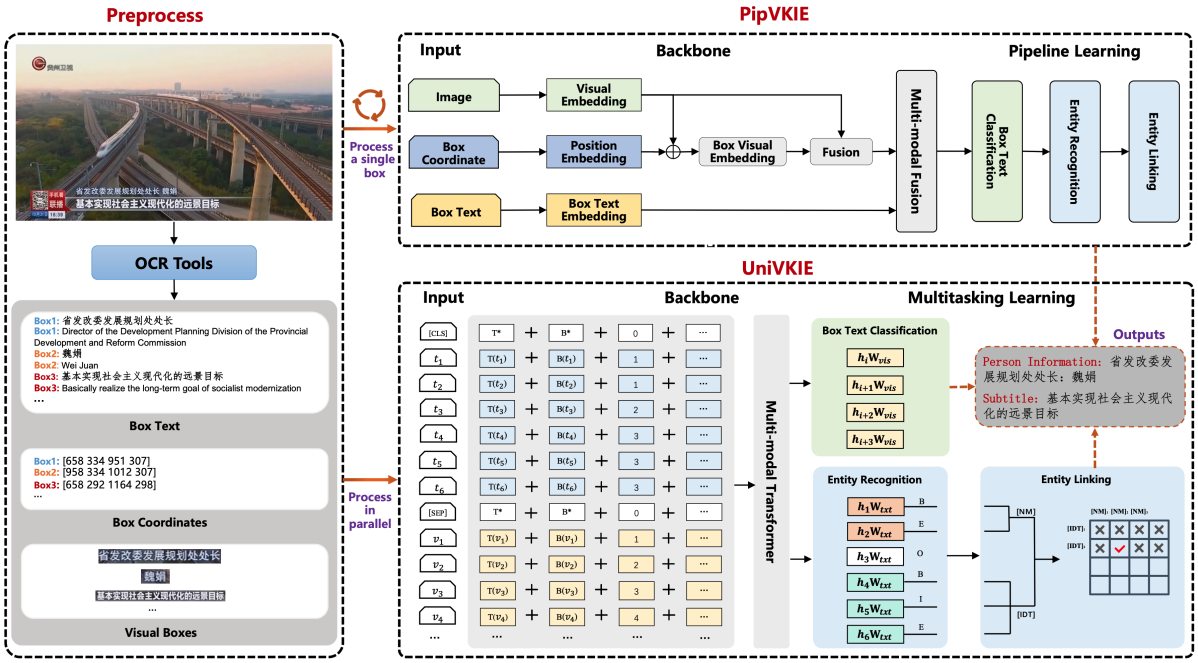

Figure 2: The overall architecture of two deployed solutions PipVKIE and UniVKIE.

EL EL aims to explore the relations between the extracted entities in each frame. Specifically, let $\boldsymbol{h}_p^{\mathrm{N}}$ denote the hidden representation respect to p-th entity of the category *Name*, $\boldsymbol{h}_q^{\mathrm{I}}$ denote the hidden representation respect to q-th entity of the category *Identity*, the representation of each entity is generated by the average pooling of text tokens. Subsequently, in each frame, we build the matrix D as inputs for the classifier. The element of D is described in Eq.2, where $D(p,q)$ represents the vector concatenated with the hidden representations of the entity pair $[\boldsymbol{h}_p^{\mathrm{N}}, \boldsymbol{h}_q^{\mathrm{I}}]$.

$$D(p,q) = [\boldsymbol{h}_p^{\mathrm{N}}, \boldsymbol{h}_q^{\mathrm{I}}] \tag{2}$$

2.2 UniVKIE

Although PipVKIE is effective in practice, we have identified several problems with it: 1) PipVKIE does not effectively utilize the layout relationships between different boxes within the same frame. 2) The three tasks (BTC, ER, and EL) are trained separately and cannot benefit from each other. 3) Processing only one box at one time during inference is not efficient enough. To tackle the challenges posed by PipVKIE, we propose UniVKIE, a unified model that processes all boxes of each frame in parallel. UniVKIE leverages a shared multimodal backbone and employs a multitask learning approach. Fig.2 provides an overview of our model's architecture.

2.2.1 Multimodal Backbone

Similar to the model structure defined (Li et al., 2021; Xu et al., 2020b,a; Hong et al., 2022), we utilize a shared multimodal backbone for the three tasks. Given a frame of video, we firstly apply OCR to obtain text recognition results which could be described as a set of 2-tuples including M text segments and box coordinates. Then, we concatenate these M text segments from top left to bottom right into one text with length N. In this concatenated text, let $v_i \in \{v_1, v_2, ...v_M\}$ denote the i-th visual token with respect to i-th box and $t_j \in \{t_1, t_2, ...t_N\}$ denote the j-th token of text. Then we add `[CLS]`, `[SEP]` and pad the sequence to fixed length L. The input sequence is established as the format in Eq.3.

$$\begin{aligned} S = \{&\texttt{[CLS]}, t_1, \ldots, t_N, \texttt{[SEP]}, \\ &v_1, \ldots, v_M, \texttt{[PAD]}, \ldots\} \end{aligned} \tag{3}$$

UniVKIE benefits from this structure in two aspects: 1) visual token and text token can interact with each other, thus the feature representation is reinforced by multimodal fusion. 2) the relations between boxes are explored to fully extract layout information. 3) all boxes in each frame are processed in parallel in these concatenated form.

2.2.2 Multitask Learning

While the models corresponding to the three subtasks are trained separately in PipVKIE, UniVKIE unifies these subtasks and employs a multitask learning approach(Vandenhende et al., 2020) to jointly train the model. As illustrated in Fig.2, UniVKIE takes the embeddings of M text segments defined in Equation 3 as input to the BTC branch, which outputs the categories of M boxes. The ER branch takes the N tokens in the text concatenated by all box texts as input to identify the entities, which are then passed to the EL branch to explore their relationships.

By summing the losses of the three subtasks, we calculate the final loss as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{BTC} + \beta\mathcal{L}_{ER} + (1-\alpha-\beta)\mathcal{L}_{EL} \quad (4)$$

where $\mathcal{L}_{BTC}$, $\mathcal{L}_{ER}$, and $\mathcal{L}_{EL}$ is the loss of BTC, ER and EL respectly, α and β are hyperparameters to make trade-offs.

3 Experiments

3.1 Experimental Setup

Dataset To promote the new task, we have created a real-world dataset consisting of 115 hours of videos collected from 88 different sources. In preprocess, we uniformly sampled 23,896 frames from these videos and obtained over 123k visual boxes with text segments and coordinates by an off-the-shelf OCR tool. Afterwards, the dataset was carefully annotated and strictly checked by 8 professional annotators. Further details about the dataset are shown in Table 5.

Metrics and Implementation Details We evaluate the performance of BTC, ER, and EL by Precision (P), Recall (R), F1-score, and Accuracy (Acc). To ensure the reliability of our results, we conducted ten runs with distinct random seeds for each setting and report the average results obtained from these runs. Details of the hyperparameters settings for PipVKIE and UniVKIE are presented in Table 6 and Table 7 respectively.

3.2 Experimental Results

3.2.1 BTC

The upper part of Table 1 presents the performance of BTC. To evaluate how modality contributes to performance, we also take unimodal methods for comparison. This includes two text backbones, BERT (Devlin et al., 2018) and xlm-RoBERTa (Conneau et al., 2019), as well as ResNet-50 (He et al., 2016), which serves as a visual backbone. Our results show that PipVKIE and UniVKIE outperform unimodal methods, with UniVKIE performing better than PipVKIE. This demonstrates the superiority of utilizing multimodal information and the unifying strategy.

3.2.2 ER

In PipVKIE, subtasks are completed in sequential stages, which means that errors can accumulate in the downstream task ER after BTC. To isolate the accumulated error, we evaluated the performance of ER by replacing the prediction of BTC with the ground truth. The performance of ER is shown in the bottom left of Table 1, where PipVKIE* represents the results obtained by using ground truth input instead of predicted input. Our observations show that the performance of PipVKIE* with ground truth input is better than that with predicted input, indicating that errors accumulate in downstream tasks. Furthermore, UniVKIE achieves better results than PipVKIE, demonstrating that unifying is a better strategy.

3.2.3 EL

The performance of EL is shown in the bottom right of Table 1. Similar to ER, we compared the performance of PipVKIE and UniVKIE when feeding them with either the ground truth entity boundaries or predicted hidden representations. Our observations show that the performance is slightly lower when using the predictions of ER. In real-world applications where errors can accumulate, UniVKIE achieves better results than PipVKIE, which demonstrates its superiority.

In Table 1 UniVKIE outperforms PipVKIE in major metrics. We identified that this is primarily due to the efficient fusion of different modalities and the elimination of error accumulation caused by the pipeline method. Another factor is that the subtasks within PipVKIE operate independently and could not benefit from each other.

3.2.4 Ablation Study

We design a series of ablation experiments to verify the contributions of each component in our solutions. We evaluate the effectiveness of modalities by eliminating one or some of them in UniVKIE, as illustrated in Table 2. While text modal is necessary for ER and EL, we notice a manifest performance degradation in BTC after removing textual infor-

BTC Task													
Methods	*Title*			*Person Info*			*Subtitle*			*Misc*			Avg
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc
BERT	86.32	83.58	84.93	92.30	87.46	89.81	91.63	88.21	89.89	85.35	82.84	84.08	86.00
xlm-RoBERTa	89.17	86.91	88.02	92.85	89.95	91.38	83.19	85.31	84.24	85.97	89.00	87.46	87.87
ResNet-50	73.43	66.56	69.84	84.51	79.38	81.86	79.24	78.51	78.87	85.80	76.99	81.16	77.57
PipVKIE	95.10	92.19	93.62	95.58	89.48	92.43	95.28	91.17	93.18	95.71	98.45	97.06	95.57
UniVKIE	84.37	86.42	85.38	98.90	98.77	98.83	90.36	98.74	94.36	99.53	97.25	98.37	**97.22**†

ER Task										EL Task	
Methods	*Name*			*Identity*			Avg			**Methods**	Avg
	P	R	F1	P	R	F1	P	R	F1		Acc
PipVKIE*	92.92	92.08	92.50	74.43	77.30	75.84	84.71	85.69	85.19	PipVKIE*	81.51
PipVKIE	92.78	88.45	90.56	73.99	75.46	74.72	84.32	82.82	83.56	PipVKIE	69.33
UniVKIE*	-	-	-	-	-	-	-	-	-	UniVKIE*	79.96
UniVKIE	97.39	97.81	97.60	90.38	91.30	90.84	94.26	94.91	**94.58**†	UniVKIE	**71.61**†

Table 1: Experimental results of BTC, ER, and EL. * indicates the results obtained by replacing the prediction of the upstream task with ground truth. - indicates the meaningless results, since for UniVKIE, ER does not rely on BTC in the pipeline. † indicates that UniVKIE performs better with p-value < 0.05 based on paired t-test.

Modals		**BTC**	**ER**	**EL**
Visual	Text	Acc	F1	Acc
✓		65.99	-	-
	✓	95.30	90.42	61.59
✓	✓	97.22	94.58	71.61

Table 2: Modality ablation study of UniVKIE. - indicates the meaningless results, as the text modal cannot be omitted in ER and EL.

Loss			**BTC**	**ER**	**EL**
$\mathcal{L}_{BTC}$	$\mathcal{L}_{ER}$	$\mathcal{L}_{EL}$	Acc	F1	Acc
✓	✓	✓	97.22	94.58	71.61
	✓	✓	-	93.46	73.29
✓	✓		97.05	94.19	-
✓			97.84	-	-

Table 3: Loss ablation study of UniVKIE. - indicates the meaningless results as the task-specific loss is necessary for the corresponding subtask.

mation, this confirms that the text modality plays a dominant role in our task. In addition, UniVKIE with multimodal information achieves the best results in all comparisons. To explore the reason, even for identical text, the visual features such as its location and background in a frame can affect the identification of segment categories, entities, and relationships. For example, subtitles are often located at the bottom of the image and have a special background color. Similarly, related names and identities often appear in visually adjacent positions within a frame of video.

Furthermore, we conduct additional experiments to explore how each task impacts the others, which is shown in Table 3. To explore the impact of BTC on ER and EL, we find that UniVKIE without BTC loss achieves slightly worse results on ER, but obtains improvement on EL. Moreover, by removing the ER loss and the EL loss, we find that the performance is almost steady on BTC. These phenomena indicate that BTC is hardly influenced by the other two tasks. UniVKIE unifies the three tasks into one model and achieves overall balanced performance.

Methods	**Speed**	**Params**
PipVKIE (BTC + ER + EL)	205ms	350M
UniVKIE (BTC + ER + EL)	56ms	106M

Table 4: Efficiency comparison of PipVKIE and UniVKIE.

4 Discussion

4.1 Modality

In the section of the ablation study, we find text modality plays the leading role. Besides, visual information also plays a crucial role in our task. For example, in BTC, box of a specific category often has a particular background color and location, which can serve as complementary features to the text. As the associated names and identities are usually located in associated position in one frame, it is important to consider visual information when performing EL tasks. The experimental results in Table 2 validate this point.

4.2 Efficiency

Table 4 compares the inference speed and resource cost between PipVKIE and UniVKIE. We deploy both models on Tesla V100-SXM2-32GB. By shar-

ing the same multimodal backbone and unifying the three tasks into one model, UniVKIE achieves satisfactory inference speed and costs lower GPU resources. This is mainly attributed to the fact that in the inference of PipVKIE, only the feature in a single box is required, while in UniVKIE, the features of all boxes in a whole frame are inputted, which increases parallelism and thus improves efficiency.

4.3 Deployment Cases

Both PipVKIE and UniVKIE have already been deployed on an AI platform for industrial media, which is a well-designed video understanding platform with comprehensive video processing services. We give three cases of real-world news videos, as shown in Fig.3 . The red boxes illustrate the hierarchical information extracted from the current video frame. In these cases, titles and subtitles are shown at the segment level while personal information is organized at the entity level with name and identity. Therefore, these valuable hierarchical information extracted by VKIE from the visual texts can be used effectively to index, organize, and search videos in real applications. More details about how our application works on the AI platform could be found in the supplementary material A.

5 Conclusion

This paper introduces a novel task in the industry, referred to as VKIE, which aims to extract crucial information from visual texts in videos. To address the task, we decouple VKIE into four subtasks: text detection and recognition, text classification, entity recognition, and relation extraction. Furthermore, we propose two complete solutions utilizing multimodal information: PipVKIE and UniVKIE. PipVKIE performs these three subtasks in different stages, while UniVKIE unifies all of them in one model with higher efficiency and lower resource cost. Experimental results on one well-defined dataset demonstrate that our solutions can achieve remarkable performance and satisfactory resource cost. With VKIE, structured information could be effectively extracted and well organized with rich semantic information. VKIE has been deployed on an industrial AI platform.

Figure 3: Real-world cases on our AI platform, the red boxes illustrate the extraction results of current frame.

Limitations

While VKIE could be easily extended to multilingual tasks, our dataset in practical application centered on Chinese videos. For general use, we are formulating plans to extend the application to multilingual tasks in the future.

Ethics Statement

The authors declare that the data in our work is publicly available and does not involve political and moral sensitivities. Ethical concerns include the usage of the proposed solution for a purpose different from that previously mentioned in the paper, such as video inputs of racism, violence, etc.

References

CGTN Sports Scene. 2023. Messi on kissing the world cup trophy and his regrets at behavior against the netherlands argentina. https://www.youtube.com/watch?v=MhnT_aHkgSQ.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10767–10775.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Hoa T Le, Christophe Cerisara, and Alexandre Denis. 2018. Do convolutional networks need to be deep for text classification? In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Qi Li, Pengfei Li, Kezhi Mao, and Edmond Yat-Man Lo. 2020. Improving convolutional neural network for text classification by recursive data pruning. *Neurocomputing*, 414:143–152.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1912–1920.

Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. 2018. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5909–5918.

N Radha. 2016. Video retrieval using speech and text in video. In *2016 International Conference on Inventive Computation Technologies (ICICT)*, volume 2, pages 1–6. IEEE.

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. *arXiv preprint cs/9907006*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. 2016. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision*, pages 56–72. Springer.

Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. 2020. Revisiting multi-task learning in the deep learning era. *arXiv preprint arXiv:2004.13379*, 2(3).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Bo Yang, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Mingjing Li. 2007. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 73–80.

Haojin Yang, Maria Siebert, Patrick Luhne, Harald Sack, and Christoph Meinel. 2011. Lecture video indexing and analysis using video ocr technology. In *2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems*, pages 54–61. IEEE.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560.

A Media AI Platform

The task of key information extraction from visual texts in videos has been deployed on an media AI platform, which is a well-designed video understanding platform with comprehensive video processing services. We uniformly sample key frames from the uploaded video. Then, a OCR engine is used to extract visual boxes and their corresponding coordinates. Afterwards, VKIE completes the three subtasks of BTC, ER and EL, and obtains hierarchical information at the entity and segment levels. Here we present one result for clear viewing as in Fig.4.

B Details of dataset

Table 5 illustrates the concrete categories contained in **BTC**, **ER**, and **EL** in our practice. We collected 88 sources, totaling 115 hours, from publicly available videos, including news programs, variety shows, and other sources. All 88 video sources are split for training, developing, and testing with the ratio 3:1:1. We then extract frames from these videos by taking their average over time. To prevent data leakage, we ensure that frames from the same video are not present in different splits. In **BTC**, we assign the samples to 4 categories including Title, Person Info, Subtitle, and Misc. We further annotate mentions and labels on the samples of Person Info for **ER** as shown in Fig.5 . Finally, **EL** is annotated on the pairs of entities extracted from each frame.

Task	Type	Value
Video	Total Hours	115
	Total Sources	88
	Total Videos	264
	Total Frames	23896
BTC	Categories	Title, Personal Info Subtitle, Misc
	Samples	train/dev/test: 76k/22k/25k
ER	Categories	Name, Identity
	Samples	train/dev/test: 34k/12k/12k
EL	Categories	Matched, Not matched
	Samples	train/dev/test: 18k/6k/6k

Table 5: The basic statistics of our datasets

C Training Hyperparameters

Table 6 illustrates the hyperparameters of the three models corresponding to **BTC**, **ER** and **EL** in PipVKIE. In UniVKIE, we use a shared multimodal backbone and build task-specific branches as in Table 7.

Hyperparameters	Value
BTC	
visual feature extractor	3-layers CNN
textual feature extractor	4-layers transformers
hidden dimension of visual feature	266
hidden dimension of textual feature	768
optimizer	Adam
batch size	48
epochs of training	10
ER	
textual feature extractor	transformer
hidden dimension of textual feature	768
optimizer	AdamW
batch size	16
epochs of training	10
EL	
textual feature extractor	transformer
hidden dimension of textual feature	768
optimizer	AdamW
batch size	16
epochs of training	10

Table 6: Hyperparameters of PipVKIE

Hyperparameters	Value
image channels	3
normalized coordinate size	128
hidden dimension of multimodal feature	768
batch size	32
epochs of training	10
optimizer	AdamW
learning rate	5e-5
hidden layer dropout prob	0.1
number of hidden layers	12
hidden dimension	768
token max length in encoder	128
2d position embedding dimension	1024
1d position embedding dimension	512
vocabulary size	21128
BTC/ER/EL trade-off factors in loss	0.3/0.3/0.4

Table 7: Hyperparameters of UniVKIE

D Integration with LLMs

Recently, large language models(LLMs) have attracted widespread interest. We have noticed this and conducted experiments with LLMs within the VKIE scenario. However, we found these approaches are not sufficiently stable for practical industrial applications. Therefore, we have decided to defer the exploration of integration with LLMs as a future extension of our work, rather than incorporating it into this submission.

Figure 4: A real-world case on the AI platform for clear viewing.

Figure 5: Examples of ER on the annotation platform. The red box indicates the candidate labels of ER.