Employing Glyphic Information for Chinese Event Extraction with Vision-Language Model

Xiaoyi Bao[1], **Jinghang Gu**[1*], **Zhongqing Wang**[2*], **Mingjie Qiang**[2], **and Chu-ren Huang**[1]
[1]The Hong Kong Polytechnic University, Hong Kong, China
[2]Natural Language Processing Lab, Soochow University, Suzhou, China
p2213545413@outlook.com
wangzq@suda.edu.cn,qiangminjie27@gmail.com
{jinghang.gu,churen.huang}@polyu.edu.hk

Abstract

As a complex task that requires rich information input, features from various aspects have been utilized in event extraction. However, most of the previous works ignored the value of glyph, which could contain enriched semantic information and can not be fully expressed by the pre-trained embedding in hieroglyphic languages like Chinese. We argue that, compared with combining the sophisticated textual features, glyphic information from visual modality could provide us with extra and straight semantic information in extracting events. Motivated by this, we propose a glyphic multi-modal Chinese event extraction model with hieroglyphic images to capture the intra- and inter-character morphological structure from the sequence. Extensive experiments build a new state-of-the-art performance in the ACE2005 Chinese and KBP Eval 2017 dataset, which underscores the effectiveness of our proposed glyphic event extraction model, and more importantly, the glyphic feature can be obtained at nearly zero cost. Code and data can be found at https://github.com/HoraceXIaoyiBao/GlyphicVLM-for-ChineseEE.

1 Introduction

Event extraction aims to extract events from the sentence, each of which consists of four types of elements: a *trigger* and multiple *arguments* are exist as the raw spans in the input text, an *event type* or *role type* are assigned to corresponding trigger and argument as a result of classification. The example in Figure 1 contains an event record: an *Meet* event triggered by "讲话"(speech), the corresponding *Person* argument is "总理"(prime minister) and "民众"(people), and *Place* argument is "受灾山区"(disaster-stricken mountainous area).

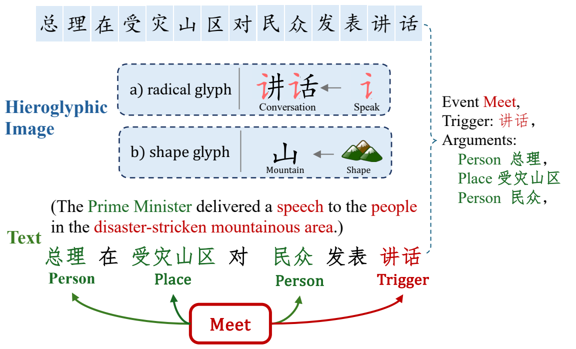

Figure 1: Example of the glyphic information in Chinese Event Extraction.

Recent studies on event extraction have incorporated a variety of features, such as textual elements (Lu et al., 2021; Liu et al., 2023), extra annotations (Lin et al., 2020; Yang et al., 2023b), and multi-modal components (Li et al., 2023a; Nguyen et al., 2023). Nevertheless, research on Chinese event extraction remains sparse. The majority of these studies tend to directly implement English event extraction techniques on Chinese datasets (Lin et al., 2020; Cui et al., 2024), while only a limited number of works have tailored their methodologies based on the inherent traits of the Chinese language (Lin et al., 2020; Xu et al., 2020; Liu et al., 2021).

Despite their effectiveness, previous works with sophisticated feature have encountered high annotation costs and narrow application scopes, making them less than optimal for Chinese event extraction. In this study, we shift our attention to a long-existing yet often neglected feature: glyphs. Chinese, as a hieroglyphic language, embeds substantial information within the glyphs of its characters, which is pivotal for Chinese event extraction. To illustrate, the radical glyph plays a critical role in communicating the semantic essence of the trigger phrase "讲话" (speech). The radical "讠", signifying speech, is present in both characters,

* Jinghang Gu and Zhongqing Wang are the corresponding authors

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1068–1080
November 12-16, 2024

emphasizing its connection to the act of speaking. Furthermore, the shape of the first character “山” in “受灾山区” directly evolved from the actual silhouette of a mountain (a shape glyph). This intuitive glyphic representation facilitates the straightforward extraction and classification of “受灾山区” as a *Place* argument.

However, it is challenging to incorporate glyphic information into Chinese event extraction tasks. This difficulty arises because it is unclear how glyphs impact event triggers or arguments along with their connections, and we also lack effective methods for incorporating glyphs into downstream tasks such as event extraction. The straightforward adoption of the efforts in pretraining from previous works (Yin et al., 2016; Sun et al., 2021) are not applicable since their ways of splitting sequence into characters and radicals to align with the tokenized sequence are hard to capture the semantic connection across words in the sentence, which could be the crucial for downstream tasks.

In this study, we utilize glyphic images at sentence-level as an alternative to radical or character information for capturing glyph details. As illustrated in Figure 2, we transform the character sequence of a sentence directly into a glyphic image with active visual emphasises and leverage this image for Chinese event extraction. This approach is distinct from splitting into radicals or characters, providing a comprehensive representation of the sentence, enabling the model to perceive glyphic features through a high-level visual perspective, enhancing the extraction process. Furthermore, we adopt a Vision-Language Model (VLM) integrated with two modality alignment methods to decipher the interplay between the input sentence and the glyphic image. This integration enables the model to bridge the gap between character sequence and glyphic image, and learn the interaction between them.

The detailed evaluation shows that our proposed model significantly advances the state-of-the-art performance on several benchmarks, indicating that the glyphic information can be obtained to enhance Chinese event extraction at nearly zero cost.

2 Related Works

In this section, we introduce two related topics: event extraction and applications of glyphic information.

2.1 Event Extraction

Event extraction works have indeed leveraged features from diverse perspectives, from the original contextual features (Chen et al., 2015; Wang et al., 2019; Sha et al., 2016; Cui et al., 2020; Lin et al., 2020) to the features from extra annotations or modalities(Lin et al., 2020; Yang et al., 2023b; Li et al., 2023b,a; Nguyen et al., 2023). Recent trends have shifted towards harnessing the power of large language models to generate the structure of events (Lu et al., 2021; Liu et al., 2023; Yang et al., 2023b).

Although various works have contributed to event extraction, few have tailored their methods specifically to the unique characteristics of the Chinese language (Chen and Ji, 2009; Li and Zhou, 2012; Li et al., 2012; Ding et al., 2019). These prior studies often relied on hand-crafted features and patterns, which limited their compatibility with modern deep learning networks. Recent works with neural networks have shown great advance on the basis of raw inputs. For instance, Xu et al. (2020) addressed the issue of overlapping roles, while Shen et al. (2020) introduced hierarchical event features. Separately, Lin et al. (2018) approached event detection on a character-by-character basis, utilizing a hybrid representation for each character.

Previous studies have typically approached event extraction without fully considering the unique glyphic features inherent in hieroglyphic languages like Chinese. However, in our study, we innovate by manipulating the glyphic characteristics of Chinese characters using vision-language models. To the best of our knowledge, this marks the first instance where methods have been designed specifically with the glyphic attributes of hieroglyphic languages in mind for event extraction.

2.2 Applications of Glyphic Information

Given the routine nature of characters, there is a growing trend to interpret glyphic features through embeddings. Initial efforts focused on capturing glyphs by decomposing characters into radicals (Shi et al., 2015; Yin et al., 2016; Sun et al., 2021). More recent studies have taken a more direct approach, training embeddings by viewing each characters as images (Aoki et al., 2020; Yang et al., 2023a). This method allows glyph information to be naturally learned through image mod-

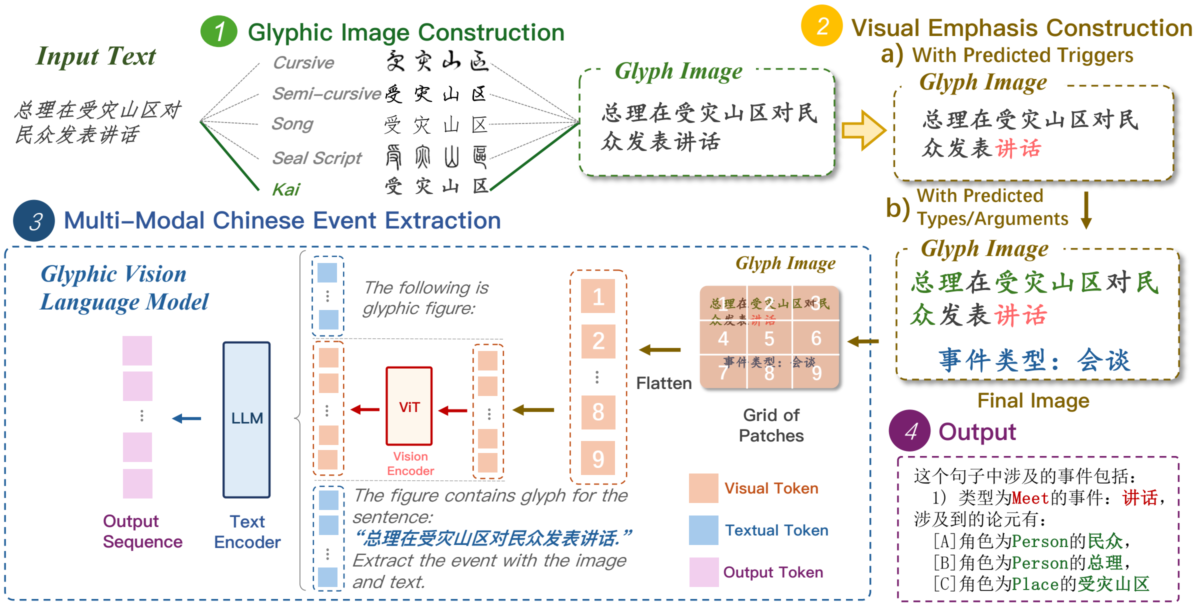

Figure 2: The illustration of our proposed method.

eling. However, there is still a significant gap between training these embeddings and their application in specific downstream tasks. As a result, only a handful of studies have successfully leveraged glyphic information to enhance their downstream task performance (Zhang et al., 2023).

Different from previous studies, we introduce an innovative approach that manipulates the glyphic characteristics of Chinese characters at the sentence-level with the vision-language models specifically tailored for Chinese event extraction. Our method stands out as the first to utilize glyphic features directly in a downstream task, rather than solely relying on pre-training or splitting them.

3 Chinese Event Extraction via Glyphic Vision-Language Model

In this study, we utilize a Glyphic Vision-Language Model specifically designed for Chinese event extraction. As shown in Figure 2, our approach involves several key steps. Firstly, we convert the input sentence into a glyphic image using a visual emphasis construction method. Secondly, we employ a vision-language Model to learn the interactions between the input sentence and the glyphic image. Finally, we generate the event structure based on two fusion strategies. In the below of these section, we will discuss these issues one by one.

Figure 3: Example of fonts which interpret glyphic information with different writing styles.

3.1 Glyphic Image Construction

We first illustrate the construction process of the glyphic image. This process can be divided into two stages. The first stage is **sequence image construction**, which focuses on capturing the internal morphological structure of characters and their interactions. To interpret the glyphic information in the sequence, different fonts will be selected. Once the sequence image is built, we further refine it with **visual emphasis** based on the intrinsic characteristics of event extraction. This refinement is to actively helps the image better adapt to downstream tasks, as discussed in the next subsection.

Given a Chinese sentence, each character is transformed to an image of size $p \times q$ with a specific font, such as Song (宋体) and Seal Script (篆书) as shown in Figure 3, each with their unique writing styles, are instrumental in interpreting the specific meanings of glyph. We also have traditional Chinese included for a richer array of graphic content to explore the best fonts of capturing the semantic information and enable the model to amalgamate pictographic data. Then, a sentence containing N characters is constructed in a image of size $K \times K$, composing of the characters' pixel maps that are concatenated sequentially or start another new line with a common modern Chinese writing order: from left to right and starting a new line below current one.

3.2 Visual Emphasis Construction

As the glyph information is delivered to the model passively and waits for the model to dig into them by itself, we further consider actively directing the model to focus on specific parts of the event from the image. Specifically, we designed two parts of visual emphasises as follows:

Trigger Emphasis

Using a concept akin to tag embedding, Trigger Emphasis visually distinguishes the event trigger from the surrounding plain text. This visual cue guides the model to focus on the corresponding part of the image. Since actual triggers are not provided in advance, we first train a generative model[1] using only the trigger annotations from the ground truth data. This trained model then predicts the triggers for each sample. As shown in Figure 4(a), the predicted triggers are highlighted in red on the glyph image, serving as an active reminder for the model.

Type and Argument Emphasis

In addition, based on the hypothesis of shared glyphs as introduced in Figure 1, we incorporate the glyphs representing the event type and corresponding arguments into the image. This is done to further emphasize the event context. As shown in Figure 4(b), these types and arguments are predicted using a similar approach to the trigger prediction mentioned earlier. Specifically, the predicted trigger from the previous emphasis step is concatenated with the input text to infer the corresponding arguments and types. In the glyphic image, the predicted types are translated (translation can be found in Appendix A) and printed in blue, while the arguments are printed in green. This visual representation helps the model better understand and extract events from the text.

[1]LLaMA-3-8B-Instruct, https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

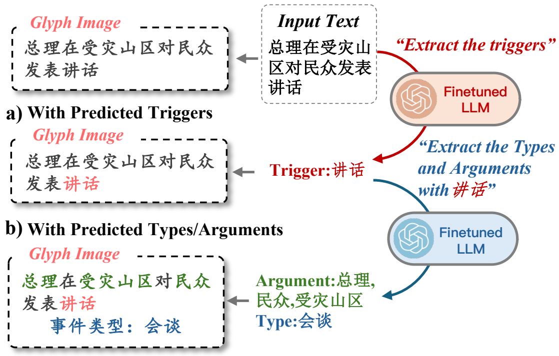

Figure 4: Process of visual emphasis construction.

3.3 Vision Encoder with Sequence Order Alignment

Given a glyph image with Visual Emphasis, we use Vision Transformer (ViT) as the image encoder to learn the visual representation. ViT is crafted to distill high-level visual features from unprocessed images, attaining excellent results compared to state-of-the-art convolutional networks. Besides, to align with the textual writing order, we employ a **Sequence Order Alignment** method that simulates the reading order of the glyph image.

Specifically, the input image is divided into a grid of patches, and each patch is then embedded into a visual token. As shown in Figure 5 a), the grid is then flatten into a sequence that follow the order of human reading and align with the textual tokens in the review inputted into the LLM. The patch in the upper right corner (marked as 1 in Figure 5 a)) of the image will be placed in the start of the flattened sequence when inputted into the transformer, followed by the patch on its right. Once reach the end of a line, the next patch would be the rightmost patch in the line below (marked as 4).

With the alignment method, the visual tokens are in the same order with the textual tokens, which then augmented with positional encodings before being fed into the Transformer. Then the encoded image representations x_v can be obtained from image I.

3.4 Text Encoder with Fusion Instruction

As Large Language Models (LLMs) has shown great capability in understanding the semantic information, we employ the LLM as our text encoder also the modality fusioner.

We specifically design the instructions for fu-

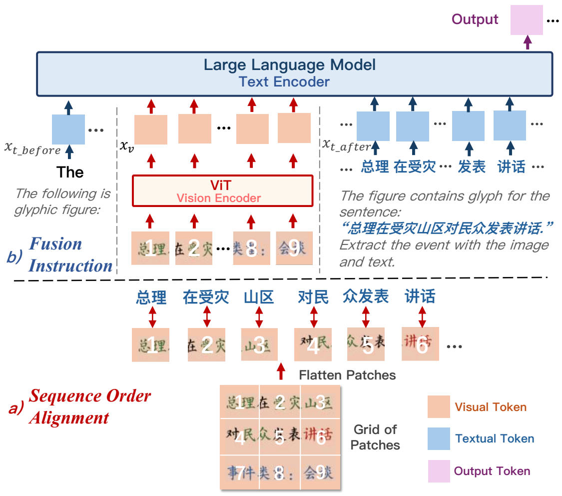

Figure 5: The illustration of our fusion strategies.

sion in natural language, responsible for guiding the VLM to fuse visual input. The fusion instruction are designed as shown in Figure 5 b), which include a guiding instruction at both before and after the visual tokens, along with the specific text for extracting.

When provided with a image and text, the LLM processes the vision encoder's output as visual tokens x_v and the tokenized text as language tokens x_{t_before} and , x_{t_after}. These tokens are subsequently merged to create the input sequence x, specifically:

$$x = [x_{t\_before},\ x_v,\ x_{t\_after}] \tag{1}$$

Given the fused token sequence $x = x_1, ..., x_{|x|}$ as input, the model outputs the linearized representation $y = y_1, ..., y_{|y|}$. The decoder predicts the output sequence token-by-token. At the i-th step of generation, the decoder predicts the i-th token y_i in the linearized form, and decoder state h_i^d as:

$$y_i, h_i^d = ([h_1^d, ..., h_{i-1}^d], y_{i-1}) \tag{2}$$

The conditional probability of the whole output sequence $p(y|x)$ is progressively combined by the probability of each step $p(y_i|y_{<i}, x)$:

$$p(y|x) = \prod_{i=1}^{|y|} p(y_i|y_{<i}, x) \tag{3}$$

where $y_{<i} = y_1 ... y_{i-1}$, and $p(y_i|y_{<i}, x)$ are the probabilities over target vocabulary V.

The objective functions is to maximize the output target sequence X_T probability given the review sentence X_O. Therefore, we optimize the negative log-likelihood loss function:

$$\mathcal{L} = -\frac{1}{|\tau|} \sum_{(X_O, X_T) \in \tau} \log p(X_T|X_O; \theta) \tag{4}$$

where θ is the model parameters, and (X_O, X_T) is a (*sentence, target*) pair in training set τ, then

$$\begin{aligned} &\log p(X_T|X_O; \theta) = \\ &= \sum_{i=1}^{n} \log p(x_T^i | x_T^1, x_T^2, ... x_T^{i-1}, X_O; \theta) \end{aligned} \tag{5}$$

where $p(x_T^i | x_T^1, x_T^2, ... x_T^{i-1}, X_O; \theta)$ is calculated by the decoder.

4 Experiment

In this section, we introduce the datasets used for evaluation and the baseline methods employed for comparison. We then report the experimental results conducted from different perspectives, and analyze the effectiveness of the proposed model with different factors.

4.1 Dataset and Experiment Setting

In this study, we use ACE2005 Chinese (**ACE05**) (Walker et al., 2006) for Event Extraction and TAC KBP 2017 Event Nugget Detection Evaluation (**KBP17**) datasets for Event Detection. For these two dataset, we follow the splittin setting from ONEIE (Lin et al., 2020) and Lin et al. (2018) respectively.

For our Vision-Language Model, we employ the pre-trained weight InternLM-XComposer2-VL(Dong et al., 2024) and LoRA fine-tune the LLM adapter parameters. We tune the parameters of our models by grid searching on the validation dataset and average the 5 runs as the final result. The LoRA alpha is set to 128 and LoRA rank is set to 64. The model parameters are optimized by Adam (Kingma and Ba, 2015), with a learning rate of 5e-5. The batch size is set to 1 with a cut-off length of 4096 and image size of 490 × 490. The glyph is interpreted with traditional Chinese and Song (宋体). The LoRA adapter would be merged with the original parameters and freeze during the inference process. Our experiments are carried out with two Nvidia RTX A6000 GPUs.

We use the same criteria as (Zhang et al., 2019; Wadden et al., 2019) for evaluation. A **Trigger** is correctly identified (Tri-I) if its offsets match a ground truth trigger. It is correctly classified (Tri-C) if its event type also matches the ground truth

Method	ACE05						KBP17					
	Tri-I			Tri-C			Tri-I			Tri-C		
	P.	R.	F1.	P.	R.	F1.	P.	R.	F1.	P.	R.	F1.
FBRNN(Char)	0.613	0.456	0.523	0.575	0.428	0.491	0.579	0.369	0.451	0.517	0.329	0.402
DMCNN(Char)	0.601	0.616	0.609	0.571	0.585	0.578	0.536	0.499	0.517	0.501	0.465	0.482
C-BiLSTM*	0.656	0.667	0.661	0.600	0.609	0.604	-	-	-	-	-	-
FBRNN(Word)	0.641	0.637	0.639	0.599	0.596	0.597	0.651	0.468	0.545	0.601	0.432	0.502
DMCNN(Word)	0.666	0.636	0.651	0.616	0.588	0.602	0.604	0.516	0.556	0.548	0.468	0.505
HNN*	**0.742**	0.631	0.682	**0.771**	0.531	0.630	-	-	-	-	-	-
Rich-C*	0.622	0.719	0.667	0.589	0.681	0.632	-	-	-	-	-	-
NPN*	0.648	**0.738**	0.690	0.609	**0.693**	0.648	0.643	0.531	0.582	0.576	0.476	0.521
TLNN	0.651	0.716	0.681	0.606	0.680	0.639	0.622	0.563	0.591	0.572	0.501	0.534
ONEIE*	-	-	-	-	-	0.656	-	-	-	-	-	-
DEGREE	0.647	0.709	0.676	0.613	0.681	0.645	0.624	0.559	0.589	0.577	0.502	0.535
LLaMA-3	0.724	0.682	0.702	0.676	0.641	0.658	0.652	0.578	0.612	0.609	0.512	0.556
ChatGLM-3	0.560	0.453	0.501	0.491	0.401	0.441	0.439	0.377	0.409	0.485	0.436	0.459
Ours	0.741	0.708	**0.724**	0.695	0.664	**0.679**	**0.683**	**0.596**	**0.636**	**0.638**	**0.531**	**0.581**

Table 1: Comparison with baselines in Event Detection, * indicates the results adapted from the original paper.

Method	Arg-I			Arg-C		
	P.	R.	F1.	P.	R.	F1.
C-BiLSTM *	0.530	0.522	0.526	0.473	0.466	0.469
Rich-C*	0.436	0.573	0.495	0.392	0.516	0.446
ONEIE*	-	-	-	-	-	0.520
JMCEE*	**0.663**	0.452	0.537	0.537	0.467	0.500
LLaMA-3	0.562	0.578	0.569	0.533	0.526	0.529
Ours	0.581	**0.601**	**0.590**	**0.547**	**0.562**	**0.554**

Table 2: Comparison with baselines in Argument Extraction in ACE05-CN.

trigger. An **Argument** is correctly identified (Arg-I) if its offsets and event type match a ground truth argument mention. It is correctly classified (Arg-C) if its role label also matches the ground truth argument mention.

4.2 Main Results

In Table 1 and Table 2, we present a comprehensive comparison of our proposed model with various state-of-the-art baselines. These baselines include character-feature, word-feature models, feature-enriched models as well as large language models.

Character-feature methods, such as C-BiLSTM (Zeng et al., 2016), FRCNN (Ghaeini et al., 2016), DMCNN (Chen et al., 2015), solving Chinese Event Detection in a character-level sequential labeling paradigm. On the other hand, word-feature methods segment sentence into words, such as HNN (Feng et al., 2016) and word-based FRCNN and DMCNN. Feature-enriched models have extra information inputted then the previous two, include Rich-C (Chen and Ng, 2012), NPN (Lin et al., 2018), TLNN (Ding et al., 2019), JMCEE (Xu et al., 2020). We also deploy English methods on Chinese, include: ONEIE (Lin et al., 2020) and DEGREE (Hsu et al., 2022). The newly released LLaMA-3-8B (AI@Meta, 2024) and ChatGLM-3-6B (Zeng et al., 2023) are also included as our LLM baseline.

As shown in Table 1 and Table 2, we find that word-feature methods outperform character-feature methods, revealing that words could better represent the semantic information in Chinese event extraction then characters. In addition, the methods integrate hybrid features surpass the single feature methods, showing us the value of employing lavish features for the complex task such as event extraction.

Moreover, our proposed model exhibits significant improvements over all prior studies ($p < 0.05$), demonstrating the efficacy of our visual glyphic information when applied with large language models for Chinese event extraction. To the best of our knowledge, this is the first attempt to leverage glyphic information in visual modality and sequence formation in event extraction.

4.3 Contribution of Glyphic Information

After analyzing the overall performance, a natural question arises: *How much does the glyphic feature contribute to it*? To investigate this, we gradually incorporate various glyphic information into LLM, starting from the sequence image up to the visual emphasises. We use "Basic" in Table 3 to refer to the removing of visual modality, relying solely on textual features.

Method	ACE05		KBP17
	Tri-C	**Arg-C**	**Tri-C**
Basic	0.633	0.523	0.547
+Sequence Image	0.661	0.539	0.567
+Trigger Emphasis	0.671	0.545	0.572
+Argument Emphasis	0.662	0.548	0.568
+Type Emphasis	0.665	0.542	0.564
Ours	**0.679**	**0.554**	**0.581**

Table 3: Results of the contribution of the glyphic feature, measured by F1-score.

As depicted in Table 3, when using only textual features, the performance of VLM is notably low, underscoring the necessity of enriched features to achieve SOTA results in complex tasks like event extraction. Significantly improved performance is observed when the Sequence Image is included in the input, highlighting the superiority of glyphic information in capturing semantic details for event extraction. Furthermore, all the visual emphasises contribute positively to event extraction, demonstrating the effectiveness of active visual reminders that guide the model to focus on specific image components. Among these emphasises, Trigger Emphasis outperforms the others. Additionally, our proposed model, which combines both active and passive methods of incorporating visual glyphic features, achieves the best performance and showcases the value of glyphs in event extraction.

We subsequently add cases study in Appendix B to make a more intuitive illustration of the effects of the glyphic information.

5 Analysis and Discussion

In this section, we give some analysis and discussion to show the effectiveness of proposed glyphic vision-language model.

5.1 Comparison of Glyph Rationales

Different fonts represent different rationales towards the glyph as well as the formations (simplified and traditional). Thus we first analysis the impact of the rationales in Table 4 by replacing the characters in the image with various fonts.

From Table 4, we observe that the traditional Chinese characters outperform the simplified ones, which is expected since traditional characters contain more radicals. This feature not only extends the pool of shared radicals between characters, but also provides us with more semantic information

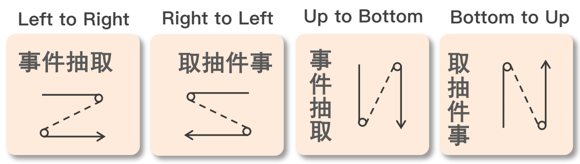

Figure 6: Illustration of different orders of writing.

behind the characters, such as 讲话” (speech, simplified) and 講話” (speech, traditional): the traditional one contains one more radical of “口” (mouth), indicating that speech is an action from the mouth. In terms of the fonts, Song (宋体) surpasses the other fonts. This may be due to the fewer adhesions between radicals within a character in Song, making it easier for the visual encoder to distinguish them and establish connections between the shared radicals across different characters.

5.2 Impact of Order Alignment

We evaluate the effect of order alignment in the image by inputting our glyphic information with different orders, also examining if human writing habits influence the visual encoder's capture.

Particularly, besides from the order shown in Figure 4 that writing the sentence from left to right, we also include the writing habits where the sentence are wrote: 1) from top to bottom (Classic Chinese); 2) from bottom to top; 3) from right to left (Arabic, Hebrew) as shown in Figure 6.

Based on the findings presented in Table 5, it is evident that different writing orders demonstrate comparable performance, suggesting that the model's understanding of the sequence is not significantly influenced by human writing habits. Notably, the left-to-right writing order yields better results compared to other orders. We attribute this improvement to the utilization of **Sequence Order Alignment** in the visual encoder, as depicted in Figure 5 a). In this approach, the patch situated in the upper right corner of the image is positioned at the beginning of the flattened sequence before being fed into the transformer. Subsequently, the patch to its right follows in a sequential manner, ensuring that the linearized sequence aligns with the order of the textual input.

5.3 Impact of Sequence Image

We subsequently compare different ways of incorporate glyph information, especially compared with the previous way of splitting into charac-

Font	Formation	Illustration	ACE05		KBP17
			Tri-C	Arg-C	Tri-C
Song(宋体)	Simplified	发表讲话	0.673	0.545	0.569
Semi-cursive(行书)		发表讲话	0.665	0.539	0.566
Cursive(草书)		发表讲话	0.662	0.534	0.558
Song(宋体) - Ours	Traditional	發表講話	**0.679**	0.554	**0.581**
Semi-Cursive(行书)		發表講話	0.677	**0.556**	0.576
Cursive(草书)		發表講話	0.672	0.547	0.572
Seal Script(篆体)		發表講話	0.664	0.540	0.563

Table 4: Result of different fonts and formations, measured by F1-score.

Orders	ACE05		KBP17
	Tri-C	Arg-C	Tri-C
Top to Bottom	0.677	0.551	0.578
Bottom to Top	0.675	0.548	**0.581**
Right to Left	0.673	0.552	0.576
Left to Right (Ours)	**0.679**	**0.554**	**0.581**

Table 5: Comparison with different human writing orders, measured by F1-score.

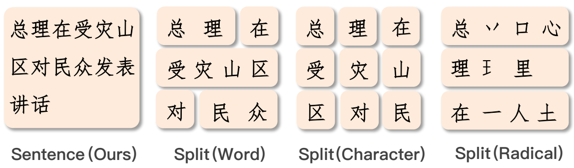

Figure 7: Illustration of different organizations.

Organization	Manner	ACE05		KBP17
		Tri-C	Arg-C	Tri-C
Split (Word)	Splitting	0.658	0.536	0.564
Split (Character)		0.652	0.531	0.556
Split (Radical)		0.639	0.527	0.553
Sentence (Ours)	Serial	**0.661**	**0.539**	**0.567**

Table 6: Comparison with different sentence organizations, measured by F1-score.

ters (Aoki et al., 2020; Yang et al., 2023a) or radicals (Lyu et al., 2021). Concretely, besides from the organization shown in Figure 4 that writing the characters follow the sentence order, we also include organizations as shown in Figure 7 where the sentences are: 1) split into words; 2) split into characters; 3) split into radicals.

As shown in Table 6, we first find the sentence manner outperform the splitting, indicating that, the sequence image can better help the model capturing the correlations over sentence in downstream tasks. Among the three splits, the splitting by character falls behind the word-based, we believe this due to basic meaning unit of Chinese is word instead of character (and this is why it needs segmentation). The splitting by radical does not surpass the splitting by characters, this may due to their shapes have already been covered by characters, leading to no improvement in glyph.

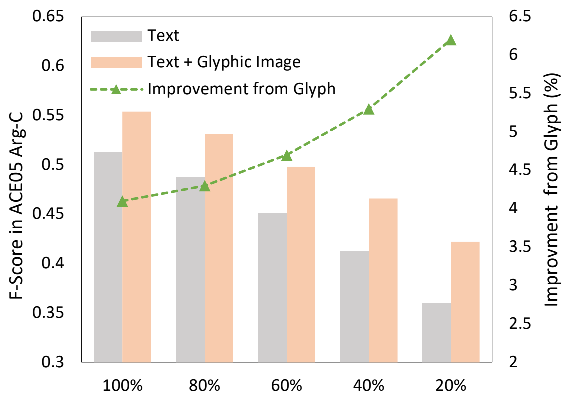

Figure 8: Improvement of data efficiency from glyph.

5.4 Analysis of Data Efficiency

Compared with textual features, one of the advantages of glyphic feature is that there are large amount of shared radicals, making it easier to build semantic connection across characters with a small size of training data. We thus investigate how the glyph improves the data efficiency of our model by comparing with using textual modality solely under limited training data in Figure 8.

From the figure, we find that the more training data, the higher performance our proposed model can reach. Moreover, the advantage of the performance brought by the glyphic information increases under limited data size, showing the superiority of glyphic information in low resource situation where a pool of shared features can be easily build compared with relying on textual modality solely.

6 Conclusion

In this study, we move our sight to the sentence-level glyphic information in Chinese event extraction and introduce a Glyphic Vision-Language Model along with active visual emphasizes and modalities alignments. By leveraging the long-existing yet often overlooked feature of glyphs, our proposed VLM achieves SOTA performance in several benchmarks without the need for complex and costly annotation of additional features.

Furthermore, our results validate that the conventional approaches of incorporating extra features during pre-training may not align with the specific requirements of downstream tasks. Instead, task-specific methods should be designed to effectively inject and utilize these additional features.

Limitations

The limitations of our work can be stated from two perspectives. Firstly, besides the glyph, there is another feature whose effect on downstream tasks is not yet known: Pinyin. In future research, further exploration of the impact of Pinyin could provide valuable insights.

Secondly, our focus has been primarily on utilizing glyph in a single hieroglyphic language. While we have achieved promising results in this language, it is important to acknowledge that the performance of our approach in other hieroglyphic languages remains unknown. Extending our investigation to multiple hieroglyphic languages would allow us to gain a more comprehensive understanding of the generalizability and effectiveness of our methodology.

Acknowledgments

We would like to thank Prof. Zhongqing Wang for his helpful advice and discussion during this work. Also, we would like to thank the anonymous reviewers for their excellent feedback. This work is supported by The Hong Kong Polytechnic University Projects (#P0048932, #P0051089).

References

AI@Meta. 2024. Llama 3 model card.

Takumi Aoki, Shunsuke Kitada, and Hitoshi Iyatomi. 2020. Text classification through glyph-aware disentangled character embedding and semantic subcharacter augmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 1–7, Suzhou, China. Association for Computational Linguistics.

Chen Chen and Vincent Ng. 2012. Joint modeling for Chinese event extraction with rich linguistic features. In *Proceedings of COLING 2012*, pages 529–544, Mumbai, India. The COLING 2012 Organizing Committee.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Zheng Chen and Heng Ji. 2009. Language specific issue and feature exploration in Chinese event extraction. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 209–212, Boulder, Colorado. Association for Computational Linguistics.

Shi-Yao Cui, Bo-Wen Yu, Xin Cong, Ting-Wen Liu, Qing-Feng Tan, and Jin-Qiao Shi. 2024. Label-aware chinese event detection with heterogeneous graph attention network. *Journal of Computer Science and Technology*, 39(1):227–242.

Shiyao Cui, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Xuebin Wang, and Jinqiao Shi. 2020. Edge-enhanced graph convolution networks for event detection with syntactic relation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2329–2339, Online. Association for Computational Linguistics.

Ning Ding, Ziran Li, Zhiyuan Liu, Haitao Zheng, and Zibo Lin. 2019. Event detection with trigger-aware lattice neural network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 347–356, Hong Kong, China. Association for Computational Linguistics.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 66–71, Berlin, Germany. Association for Computational Linguistics.

Reza Ghaeini, Xiaoli Fern, Liang Huang, and Prasad Tadepalli. 2016. Event nugget detection with forward-backward recurrent neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 369–373, Berlin, Germany. Association for Computational Linguistics.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jiaqi Li, Chuanyi Zhang, Miaozeng Du, Dehai Min, Yongrui Chen, and Guilin Qi. 2023a. Three stream based multi-level event contrastive learning for text-video event extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1666–1676, Singapore. Association for Computational Linguistics.

Peifeng Li and Guodong Zhou. 2012. Employing morphological structures and sememes for Chinese event extraction. In *Proceedings of COLING 2012*, pages 1619–1634, Mumbai, India. The COLING 2012 Organizing Committee.

Peifeng Li, Guodong Zhou, Qiaoming Zhu, and Libin Hou. 2012. Employing compositional semantics and discourse consistency in Chinese event extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1006–1016, Jeju Island, Korea. Association for Computational Linguistics.

Ruiqi Li, Patrik Haslum, and Leyang Cui. 2023b. EDeR: Towards understanding dependency relations between events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14969–14983, Singapore. Association for Computational Linguistics.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2018. Nugget proposal networks for Chinese event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1565–1574, Melbourne, Australia. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Jian Liu, Dianbo Sui, Kang Liu, Haoyan Liu, and Zhe Zhao. 2023. Learning with partial annotations for event detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 508–523, Toronto, Canada. Association for Computational Linguistics.

Jiangwei Liu, Jingshu Zhang, Xiaohong Huang, and Liangyu Min. 2021. Syntactic-gcn bert based chinese event extraction. *Preprint*, arXiv:2112.09939.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Boer Lyu, Lu Chen, and Kai Yu. 2021. Glyph enhanced Chinese character pre-training for lexical sememe prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4549–4555, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Khanh Duy Nguyen, Zixuan Zhang, Reece Suchocki, Sha Li, Martha Palmer, Susan Windisch Brown, Jiawei Han, and Heng Ji. 2023. RESIN-EDITOR: A schema-guided hierarchical event graph visualizer and editor. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 365–372, Singapore. Association for Computational Linguistics.

Lei Sha, Jing Liu, Chin-Yew Lin, Sujian Li, Baobao Chang, and Zhifang Sui. 2016. RBPB: Regularization-based pattern balancing method for event extraction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1224–1234, Berlin, Germany. Association for Computational Linguistics.

Shirong Shen, Guilin Qi, Zhen Li, Sheng Bi, and Lusheng Wang. 2020. Hierarchical Chinese legal event extraction via pedal attention mechanism. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 100–113, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. 2015. Radical embedding: Delving deeper to Chinese radicals. In *Proceedings* *of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 594–598, Beijing, China. Association for Computational Linguistics.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2065–2075, Online. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. Linguistic Data Consortium.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019. HMEAE: Hierarchical modular event argument extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5777–5783, Hong Kong, China. Association for Computational Linguistics.

Nuo Xu, Haihua Xie, and Dongyan Zhao. 2020. A novel joint framework for multiple Chinese events extraction. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 950–961, Haikou, China. Chinese Information Processing Society of China.

Xinmei Yang, Abhishek Arora, Shao-Yu Jheng, and Melissa Dell. 2023a. Quantifying character similarity with vision transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13982–13996, Singapore. Association for Computational Linguistics.

Yuqing Yang, Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023b. An AMR-based link prediction approach for document-level event argument extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12876–12889, Toronto, Canada. Association for Computational Linguistics.

Rongchao Yin, Quan Wang, Peng Li, Rui Li, and Bin Wang. 2016. Multi-granularity Chinese word embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 981–986, Austin, Texas. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Glm-130b: An open bilingual pre-trained model. *Preprint*, arXiv:2210.02414.

Ying Zeng, Honghui Yang, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2016. A convolution bilstm neural network model for chinese event extraction. In *Natural Language Understanding and Intelligent Applications*, pages 275–287, Cham. Springer International Publishing.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint Entity and Event Extraction with Generative Adversarial Imitation Learning. *Data Intelligence*, 1(2):99–120.

Xiaotian Zhang, Yanjun Zheng, Hang Yan, and Xipeng Qiu. 2023. Investigating glyph-phonetic information for Chinese spell checking: What works and what's next? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

A Translation of Event Types

We give the translation of event types in Table 7, which in used for the active visual emphasis.

B Cases Study

We launch case studies from ACE05-CN dataset to make a more intuitive illustration of the effects of the glyphic information in Chinese event extraction. We select samples from each subtasks that are predicted wrongly without glyphic information, but have been correct with it. As demonstrated in Table 8, the correct prediction would be with a ✓ notation.

The first example: without glyph, the model misses the argument "西岸" (west bank) which contains a radical "山" (mountain) whose shape comes from a mountain and clearly expresses the word represent a place. With glyph, our method easily gives a right answer.

The second example: the argument "地方" (place) has a radical "土" (soil) which is a widely shared radical across characters that represent a place such as "场"(field) and "坝" (dam), indicating "地方" is a destination of a transport instead of a start of a organization.

English	Translation	English	Translation
Life	生活	Start-Position	起始位置
Movement	运动	End-Position	结束位置
Transaction	交易	Nominate	提名
Business	业务	Elect	选举
Conflict	冲突	Arrest-Jail	逮捕入狱
Contact	联系	Release-Parole	释放假释
Personnel	人员	Trial-Hearing	审判听证
Justice	审判	Charge-Indict	指控
Be-Born	出生	Sue	起诉
Marry	结婚	Convict	定罪
Divorce	离婚	Sentence	判决
Injure	受伤	Fine	罚款
Die	死亡	Execute	执行
Transport	运输	Extradite	引渡
Transfer-Ownership	所有权转移	Acquit	无罪释放
Transfer-Money	转账	Appeal	上诉
Start-Org	成立组织	Pardon	赦免
Merge-Org	合并组织	Demonstrate	示威
Declare-Bankruptcy	宣布破产	Meet	会面
End-Org	终止组织	Phone-Write	电话写作
Attack	攻击		

Table 7: Translations of the event types

Input	Subtask	w/o Glyph	w Glyph
1 5 名巴勒斯坦伤员将乘直升飞机从约旦西岸飞抵约旦接受治疗。	Argument Identification	伤员, 飞机, 约旦 ✗	伤员, 飞机, 西岸, 约旦✓
后来去了另外一个地方工作，又巧了，附近的一个小镇子自封为" CHICKEN CAPITAL OF THE WORLD"	Argument Classification	Business:Start-Org✗ Destination✓ 地方 ✓	Movement:Transport✓ Destination✓ 地方 ✓
正在日本访问的俄罗斯国防部长塞吉耶夫 29 号表示，北韩很有能削弱他 120 万人部队的部分兵源	Trigger Identification	(Blank) ✗	Movement:Transport✓ 访问✓
北韩最高领导人金正日今天在北韩时间 23 号下午 3 点突然前往平壤百花院迎宾馆和 23 号早上抵达平壤的美国国务卿奥尔布赖特就北韩研发飞弹、反恐怖活动等等阻碍北韩和美国关系正常化的问题进行 3 个小时的会谈。	Trigger Identification	抵达, 前往 ✗	抵达, 前往, 会谈✓
几个小时之前抗议民众冲进议会和国家电视台大楼。	Trigger Classification	Conflict:Attack✗ 冲进✓	Movement:Transport✓ 冲进✓

Table 8: Case study

The third example: the model predicts nothing without glyph and misses the trigger “访问”, which contains a radical “ 讠” (speak) that represents a events. The glyphic information offered to the model gives the right answer.
The fourth example: The trigger “会谈” (conversation) features the radical “ 讠” (speak), which is a commonly used radical in characters related to verbal events. The glyphic information provided to the model leads to the correct answer.
The fifth example: the trigger “冲进” (rush) features the radical “辶” (walk), which is a commonly used radical in characters denoting movement, such as “过”(pass) and “返” (back). This suggests that the term “冲进” is a transport event rather than a conflict or attack.

From the cases shown in Table 8, we can find that, with the extra information form glypy, our method shows significant superiority in improving the performance of Chinese event extraction.