ICDAR 2025 Competition on Historical Map Text Detection, Recognition, and Linking

Yijun Lin[1⋆], Solenn Tual[3,4⋆], Zekun Li[1⋆], Leeje Jang[1⋆], Yao-Yi Chiang[1], Jerod Weinman[2], Joseph Chazalon[4], Edwin Carlinet[4], Julien Perret[3,5], Nathalie Abadie[3], Bertrand Duménieu[5], Ta-Chien Chan[6], Hsiung-Ming Liao[6], Wen-Rong Su[6], Mengjie Zou[7†], Tianhao Dai[7†], Rémi Petitpierre[7†], Beatrice Vaienti[7†], Frederic Kaplan[7†], Isabella di Lenardo[7†], Youngmin Baek[8,9†], Michael Hentschel[9†], Yu Nakagome[9†], Ichimura Shuta[9†], Jeongtae Lee[8†], and Chankyu Choi[8†]

⋆ *denotes equal contribution* † *denotes competitions winners*

[1] University of Minnesota - Twin Cities, United States
[2] Grinnell College, United States
[3] Univ. Gustave Eiffel, ENSG, IGN, LASTIG, France
[4] EPITA, France
[5] CRH, EHESS, France
[6] Center for GIS, RCHSS, Academia Sinica, Taiwan
[7] EPFL, Swiss Federal Institute of Technology in Lausanne, Switzerland
[8] Naver Cloud, Republic of Korea
[9] LINE WORKS, Japan

{lin00786, li002666, jang0124, yaoyi}@umn.edu, jerod@acm.org, {solenn.tual, julien.perret, nathalie-f.abadie}@ign.fr, {joseph.chazalon, edwin1.carlinet}@epita.fr, bertrand.dumenieu@ehess.fr, {dachien, veevee, spidersu}@gate.sinica.edu.tw, {mengjie.zou, tianhao.dai, remi.petitpierre, beatrice.vaienti, frederic.kaplan, isabella.dilenardo}@epfl.ch, {youngmin.baek, jeongtae.lee, chankyu.choi}@navercorp.com, {hentschel.michael, y.nakagome, ichimura.shuta}@line-works.com

Abstract. Historical maps are valuable for history, social sciences, and linguistics but pose challenges for automatic transcription. This competition edition continues to address detection, recognition, and linking of text in historical maps, with new features: expanded French Land Registers data, a new Taiwanese dataset with Chinese characters, synthetic training data, and improved linking evaluation metrics. Seven teams participated with over 25 submissions across four tasks and three datasets. While detection performance is strong, recognition and linking remain difficult, though improvements were seen with Bézier curve line fitting and enhanced linking pipelines. All resources are publicly available on Zenodo (https://zenodo.org/communities/icdar-maptext).

Keywords: Hierarchical text detection · Text recognition · Text linking · Historical maps.

Appears in *Proceedings of the 19th International Conference on Document Analysis and Recognition (ICDAR 2025).* Xu-Cheng Yin, Dimosthenis Karatzas, and Daniel Lopresti (Eds.). Springer Nature LNCS 16027. 2025.

1 Introduction: Motivation and Challenges

This report highlights the second edition of the competition on Historical Map Text Detection, Recognition and Linking ("MapText"), which is part of the ICDAR 2025 competition series. Previous competition results reinforced the original motivations by emphasizing how challenging it is to extract text from maps, and especially historical ones [21]. Indeed, historical maps not only constitute underexplored information for historians, economists, and other social scientists; they also are at the edge of what modern transcription systems can process. This is due to the unique combination of challenges these sources exhibit, originating both from their historical and multimodal symbolic natures, among which we can highlight the following (more figures can be found in supplementary material):

- **Faded or damaged content**: Many historical maps suffer from deterioration over time, which can include faded ink, physical damage, or altered content, making them difficult to interpret.
- **Complex hierarchical structures**: The layered organization of maps, including territorial boundaries, landmarks, and labels, often follows a complex hierarchical pattern which is reflected on textual content (Figure 1).
- **Overlapping and dense elements**: Historical maps frequently contain overlapping symbols, text with irregular spacing, and dense clusters of information, presented in various orientations (Figure 2).
- **Contextual interpretation**: A deep understanding of the historical and geographical context is often necessary to correctly interpret the map's content, especially when dealing with ambiguous or incomplete data (Figure 3).
- **Evolving lexicon**: Place names may be replaced, spelling conventions deprecated, or terms altered over time, limiting the ability to rely on existing lexicons to recognize all geographical entities.
- **Exotic fonts and handwritten content**: Many historical maps feature fonts that are no longer in use, and some include handwritten annotations.

Despite having been the subject of many studies [6], the problem of text detection and recognition on maps remains largely unsolved. Thanks to the progresses of modern ML techniques, recent works managed to push the boundaries of what is possible [37,20,33,2,22], in a way similar to the recent achievements in scene text reading [34,13,12,36,11,28].

Some key challenges in map text reading like curved text or rotated text with complex backgrounds are also present in scene text reading, and were the focus of recent robust reading competitions [7,41]. Grouping words into semantic elements, another key challenge in map text reading, was the focus of the ICDAR 2023 Competition on Hierarchical Text Detection and Recognition [27]. However, the unique combination of challenges in map text reading was not addressed until the first edition of the MapText competition in ICDAR 2024, which successfully attracted eight different teams who made over 40 submissions on the four tasks. Although encouraging, the performance attained in the first edition of the competition was not yet satisfactory. This 2025 edition continues to focus on the tasks of text detection, recognition, and word phrase linking,

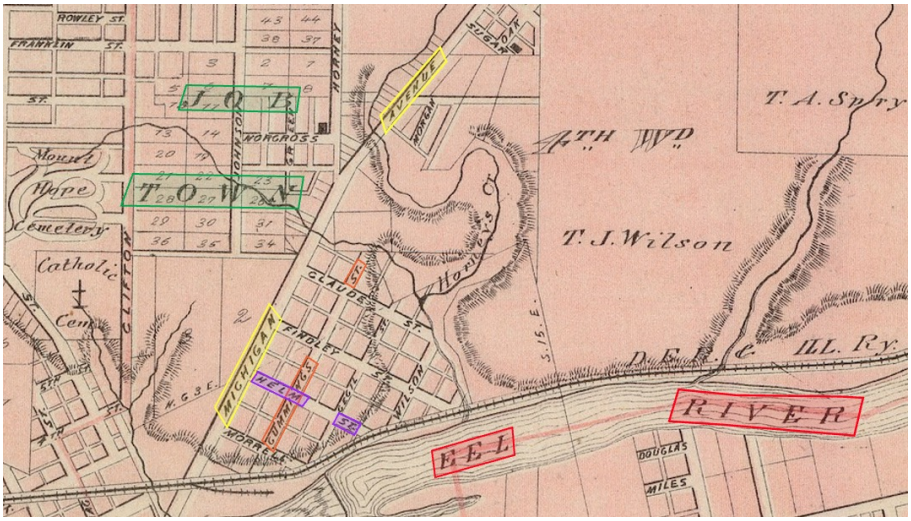


Fig. 1: Complex hierarchical structures: Word detection and linking examples are annotated in colored polygons. The polygons in the same color are linked. The green boxes ("JOB TOWN") with widely spaced characters overlap other words. The yellow boxes are very far apart from each other ("MICHIGAN AVENUE"), but the word "MICHIGAN" is close to and mixed with other street names ("HELM ST." and "GUMMINGS ST."), which are also overlapped each other. Image credit: Rumsey Collection [3] Image 0019.050 (Plan of Logansport, Cass Co., *Indiana*, 1876).

which are essential for extracting meaningful information from historical maps, and led to further advances in the state-of-the-art in map text reading. This edition introduces the following important new features:

1. **Extended French Land Registers dataset**, with more than twice the number of image tiles, extracted from new map sheets.
2. A **new focused subset with traditional Chinese characters** contributed by the GIS Center at Academia Sinica, a new partner of the competition.
3. **Synthetic data** for each dataset, enabling large pre-training strategies.
4. **Improved linking indicators** and overall evaluation.

The competition is hosted on the RRC platform, which standardizes the competition results and enables post-competition submissions.[10] All code, data, and artifacts are released under permissive licenses on archival repositories, to ensure reproducible research and foster further advances in the field.[11] Authors of the two winning approaches were invited to co-author this report.

The rest of this report is structured as follows: Section 2 describes the tasks of text detection, recognition, and linking, the evaluation metrics used for each

[10] https://rrc.cvc.uab.es/?ch=32
[11] https://zenodo.org/communities/icdar-maptext

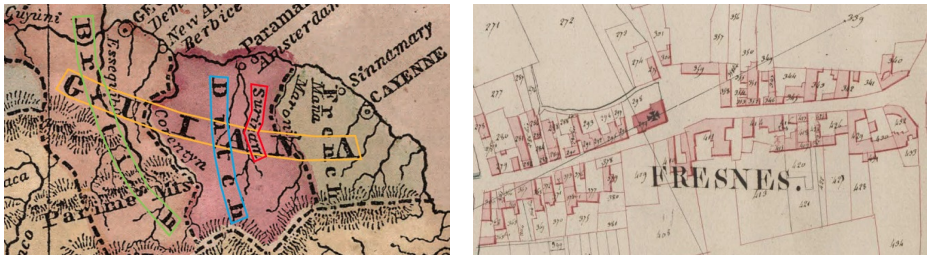
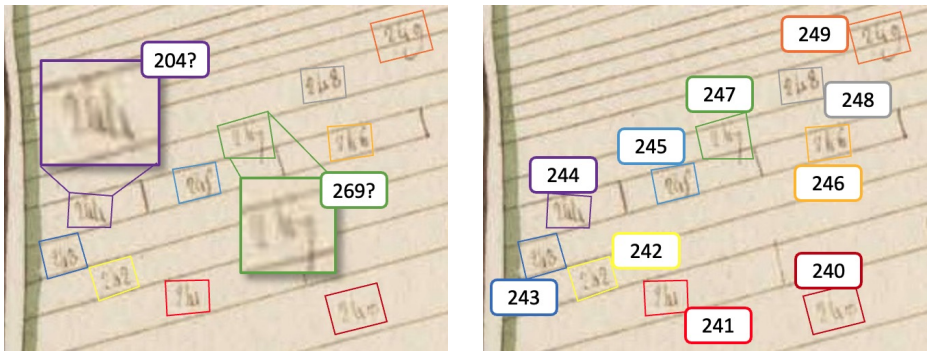


Fig. 2: Overlapping and dense elements: In addition to strong curvature and rotation, large text often overlaps with other elements, and small text can be very dense.



Unordered transcription Numerically-ordered transcription

Fig. 3: Contextual interpretation: Excerpt of French land register map with handwritten text. Even for a human, unordered transcription is challenging (left), while a numerically-ordered transcription is more straightforward (right).

task, and the final ranking strategy. Section 3 presents the datasets used in the competition, including the *David Rumsey Historical Map Collection*, the expanded *French Land Registers*, and the new *Taiwanese map dataset* contributed by the GIS Center at Academia Sinica. Section 4 details the competition protocol and summarizes each participant's method. Finally, Section 5 analyzes the competition results, while Section 6 offers our conclusions.

2 Tasks, Evaluation, and Ranking

This edition of the competition features the same tasks as the previous edition, organized around the detection, recognition, and linking of text on historical maps. For practical reasons, we evaluate the tasks in a combined manner, which leads to the actual four tasks summarized in Table 1.

Despite a desire to compare results between the 2024 and 2025 editions, we found the need to improve the evaluation metrics for each task, particularly

Table 1: Overview of the tasks and the constituent terms of the harmonic mean used for competition ranking. Each task is evaluated on three datasets. All involve detection (which includes segmentation); Tasks 2 and 4 add linking, and Tasks 3 and 4 add recognition. Abbreviations: *Detection with segmentation* — P = precision, R = recall, T = tightness, *Recognition* — C = character accuracy, *Linking* — P_L = link precision, R_L = link recall.

Task Name	Task Metrics					
	Detect		Segment	Recognize	Link	
1: Word Detection	P	R	T			
2: Phrase Detection (Word Grouping)	P	R	T		P_L	R_L
3: Word Detection and Recognition	P	R	T	C		
4: Phrase Detection and Recognition	P	R	T	C	P_L	R_L

linking. The 2024 edition used the Panoptic Quality (PQ) metric [16], modeled after the 2023 HierText competition [27]. However, we discovered that measuring PQ on the union of word regions did not adequately capture the word links desired, owing to the less frequent and much shorter phrases and diverse layouts in maps (typically one or two words and varying distances between words). Indeed, the 2024 winner predicted no links at all, excelling at PQ solely due to the underlying strength in word detection and recognition. For this reason, we transition to explicitly measuring precision and recall of the links between words.

Although metrics were updated, this edition used the evaluation protocol from the previous edition. The evaluation and analysis code is publicly available at https://github.com/icdar-maptext. The remainder of this section briefly reviews these, with further details in the supplementary material.

Task 1: Word Detection

This task requires detecting individual words on map images, i.e., generating bounding polygons that enclose text instances at the word level. To begin, all detected word regions and ground truth elements are optimally matched with a minimum IoU requirement (0.5). This correspondence results in a set of detection "true positives" [21]:

$$\mathrm{TP}_{\mathrm{Det}} \subset \{(g, d) \in G \times D \mid \mathrm{IoU}\,(g, d) > 0.5\}\,, \tag{1}$$

where G is the set of ground truth regions and D is the set of detected regions.

As before, we capture the word detection precision P and recall R of the detected regions as well as the tightness T (average IoU among these true positive regions):

$$P \triangleq \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}|} \qquad R \triangleq \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FN}|} \qquad T \triangleq \frac{1}{|\mathrm{TP}|} \sum_{(g,d) \in \mathrm{TP}} \mathrm{IoU}\,(g, d)\,.$$

The new competition metric is the harmonic mean between these three elements, rather than PQ, the product of the F measure (a harmonic mean between P and R) and T. Although PQ was in the range 0–1, it was an un-normalized geometric mean. The new metric based solely on the harmonic mean has the benefit of equally balancing all the included terms, and it does not skew closer to zero as more terms are included.

Note that tightness T is not included in the IGN French data set evaluation for competition due to high annotator variability on short cursive words.

Task 2: Phrase Detection

This task requires words to be detected (with their polygon boundaries), as in Task 1, and then also further grouped into constituent ordered lists as phrases. Word order determines the links (edges), which are now directly evaluated.

The detected word regions and ground truth regions are optimally matched as in Task 1. In addition to calculating the same word-level statistics (P, R, and T), we evaluated the links between the detected words.

True positive links are where 1) the two endpoint words are both considered true positive matches, 2) the constituent *ground truth* words have a direct link (adjacent within the same group), and 3) the constituent *detected* words have a direct link (adjacent within the same group). Unmatched ground truth links are considered false negatives, and unmatched predicted links are considered false positives. From these, we calculate the link precision P_L and link recall R_L.

The competition metric is the harmonic mean among the five elements: word detection precision P, recall R, tightness T, link precision P_L, and recall R_L.

Task 3: Word Detection and Recognition

This task requires word-level text detection and recognition results, e.g., generating a set of word bounding polygons and their corresponding transcriptions.

As in Task 1, detected word regions d are matched optimally to ground truths g; while only the IoU threshold criterion is required for matching, such valid correspondences are subsequently biased to prefer agreement between text strings [38]. The character accuracy is measured as the average complementary normalized edit distance between the corresponding elements (denoted as the set of true positives, TP) [21]:

$$C \triangleq 1 - \frac{1}{|\mathrm{TP}|} \sum_{(g,d)\in \mathrm{TP}} \mathrm{NED}\,(g,d)\,. \tag{2}$$

The competition metric is thus the harmonic mean between word detection precision P, recall R, tightness T, and character accuracy C.

Note in particular that this differs from the 2024 competition, where an exact string match was required for detections and ground truth to be considered for correspondence (in addition to the IoU threshold).

Task 4: Phrase Detection and Recognition

This task requires detection and recognition at the phrase level (as in Task 3). Submissions must also group words (polygons and transcriptions) into phrases as an ordered list (as in Task 2).

Correspondences between predicted and ground truth words are found as in Task 3. Links are calculated as in Task 2. The competition metric is the harmonic mean among all measurements: word detection precision P, recall R, tightness T, and character accuracy C, as well as the link precision P_L and recall R_L.

Note this also differs from the 2024 competition in several respects. As with Task 2, matches are no longer made at the group level but at the word level, and links are now explicitly evaluated.

3 Datasets

The competition comprises three data sources with human annotations: a set of maps selected from the David Rumsey Historical Map Collection (unchanged in this edition), a series of French land registry maps (extended in this edition), and a new selection of Chinese historical topographic maps provided by the GIS Center at Academia Sinica, Taiwan. This section provides an overview of these distinct data sets. The train, validation, and test sets of the 2024 edition, along with the associated ground-truth (except for the test set) are archived at Zenodo[12] [23,24,4,5]. Table 2 summarizes the changes made to datasets between 2024 and 2025 editions.

Although transcription languages differ, the three datasets share the same annotation format to make methods immediately compatible with each of them. In particular, tiles with a shape of 2000×2000 pixels are extracted from each map sheet, and they form the input for detection, recognition, and linking systems.

David Rumsey Historical Map Collection For the 2025 edition, this dataset is left unchanged from the previous edition [21]:

> This archive hosts an extensive set of over 126,000 maps accessible online [3]. The catalog spans maps from the 16th to the 21st century, encompassing regions from every continent, the Pacific, the Arctic, and the entirety of the World. From this rich assortment, we select **936** representative maps for human annotation using style clustering. The sampled maps cover 80 regions and 183 distinct publication years (from 1623 to 2012). ...
>
> The total number of annotated cropped tiles is **940**. We split the map tiles into 200 for training, 40 for validation, and 700 for testing. Four maps have multiple tiles in training; all test tiles are from distinct maps, and the splits are also disjoint. (pp. 369–370, emphasis added)

[12] A dedicated community collects competition artifacts across editions: https://zenodo.org/communities/icdar-maptext

Table 2: Datasets statistics.

	2024			2025		
	Train	Validation	Test	Train	Validation	Test
	Rumsey — *unchanged except for synthetic tiles*					
Tiles	200	40	700	200	40	700
Map Sheets	196	40	700	196	40	700
Words	34 518	5544	128 457	34 518	5544	128 457
Word Groups	21 205	3502	78 582	21 205	3502	78 582
Synthetic tiles	0	0	0	35 000	0	0
	French Land Registers — *extended*					
Tiles	80	15	50	228	25	144
Map Sheets	37	9	49	78	12	77
Words	8096	1801	7346	25 564	2725	15 708
Word Groups	7449	1661	6814	23 542	2413	14 284
Synthetic tiles	0	0	0	18 073	0	0
	Chinese historical topographic maps — *new*					
Tiles	0	0	0	1478	166	1120
Map Sheets	0	0	0	160	30	112
Words	0	0	0	13 153	5007	10 096
Word Groups	0	0	0	0	0	0
Synthetic tiles	0	0	0	45 000	0	0

To supplement the actual data, this edition features the release of a synthetic map image dataset created by Lin et al. [22]. The method leverages OpenStreetMap[13] and QGIS API [18] to render location names and extract background pixels from real historical maps to synthesize map images. The synthetic dataset also provides the linking labels for words in the location phrases. Because the location names are randomly chosen to render on the images, there are no direct spatial relationships between the phrases. The synthetic images have various sizes ranging from 600×600 to 1440×1440 pixels.

French Land Registry Maps The 2025 edition offers a substantial increase in the number of sheets and tiles, but otherwise has the same features as the previous edition [21]:

> To complement the broad selection of maps from the previous collection—which covers a diversity of scales, styles, geographical regions, and historical periods—we also provide another subset covering a very narrow region and time. This second subset is composed of 19th century land registers from approximately 50 [now 60] French cities and towns, at a very large scale. (In a cartographic context, a "large scale" map means it covers a relatively small area in great detail.) These cadastral plans

[13] https://www.openstreetmap.org/

> contain an important quantity of parcel numbers, now-forgotten place names, and many other local details (in French) relevant to the accurate delineation of parcels in order to identify owners and compute taxes. The entire online collection consists of over 800 map sheets [to which several extra maps from other regions and cities have been added [1]].
>
> For the competition, map sheets were selected according to two criteria: we used stratification to ensure a good representation of the different map types (first or second surveying campaign, geographic area), and we grouped the maps by the city they represent to avoid any overlap between the training, validation, and test sets in terms of geographic area. ... Contrary to the Rumsey data set, the French land registers images have a lower quality, both in terms of resolution and contrast, and the text is mostly handwritten. However, it contains fewer groups and many words represent numbers. (p. 370)

In addition to the large increase in the number of available tiles sampled following the same stratification and grouping policy as the previous edition, we also provided participants with synthetic data matching the style of this dataset. Using a similar method as for the Rumsey dataset, we generated synthetic tiles via the QGIS API [18]. However, for this dataset, plot numbers and place names are chosen from actual French land registry maps and rendered using a style similar to the original maps. Synthetic tiles have a 2000×2000 pixels resolution. To maximize the chances that participants achieve great results, we sampled regions from the same geographic area as the test set to integrate contemporary place names in learned language models. Furthermore, because the area used for the test set has massively urbanized since the original maps' creation, we also sampled rural areas with similar topological properties to help participant methods capture such structures. The code for this generator is publicly available.[14]

Chinese Historical Topographic Maps New to the 2025 edition, we provide a collection of topographic maps related to Taiwan, published during the first half of the 20th century (1900–1960). Map scales range between 1:20,000 and 1:50,000. These maps contain detailed records of the terrain, topography, and rich geographic names (in traditional Chinese), which have significant historical and research value, not only documenting Taiwan's natural and cultural landscape during the first half of the 20th century, but also serving as crucial foundational data for Geographic Information System (GIS) research.

The collection includes both monochrome and multi-color printed maps, with a total of approximately 1,205 sheets (around 60cm×45cm, 6,800×5,300 pixels). We selected 311 sheets, about 25% of the total, for manual annotation. The selection criterion is based on the even distribution across Taiwan, with priority given to locations that have more place names and text on the map, mostly in the western part of Taiwan. The annotations include boundary points and transcriptions for place names (or single characters if they are far apart) and

[14] French land register map generator: https://doi.org/10.5281/zenodo.15498600

each text's writing direction (i.e., horizontal or vertical). We use map sheets from the year 1904, 1905, and 1921 for training, 1924 for validation, and 1950s for testing to examine models' style transfer capabilities.

In addition, we provide a synthetic map image dataset in traditional Chinese, similar to the English dataset [22], but containing horizontal and vertical writing directions. More details can be found in the supplementary material.

Data Format We reused the data format specifications from the previous edition of [21], which proved to be both robust and simple to use. The data format supports two levels of nesting for word regions, which is sufficient for the three datasets considered. The data format also supports flagging illegible or truncated content for accurate evaluation. Please refer to the supplementary material for details.

4 Competition Protocol and Participants

The competition remains on the RRC platform (https://rrc.cvc.uab.es) to standardize evaluations and track further progress. Training and validation sets were released on December 10, 2025, and the test set was released on March 1, 2025. The results submission deadline was April 20, 2025, giving participants just over four months to train their models. Competition rules allowed the use of any open data set, so long as it was disjoint from the test set.

The remainder of the section lists the seven primary teams of participants. Detailed descriptions of their methods are provided in supplementary material.

MapText Strong Pipeline *Y. Xie, C. Xu, J. Zhang, P. Chen, W. Wang, Y. He, P. Li, Y. Meng, L. Gao — Bilibili Inc. and QUST (China)*

This team participates in Task 1 and Task 3 for Rumsey and TWH map data sets, for which they achieved the leading detection performance. For the Rumsey data set, the authors employed DNTextSpotter [32], a novel denoising training method based on DeepSolo. For TWH data set, they utilized DeepSolo [40]. Data augmentation techniques, including cropping, scaling, and adjustments to saturation and contrast, were applied. Pre-training was conducted using public datasets such as TextOCR [35], TotalText [8], ICDAR15 [12], MLT17 [29]. Post-processing methods were also adopted.

Self-Sequencer *M. Zou, T. Dai, R. Petitpierre, B. Vaienti, F. Kaplan, I. di Lenardo — EPFL, Swiss Federal Institute of Technology in Lausanne (Switzerland)*

Winner Task 2 (Rumsey), Task 4 (Rumsey)

This team participates in all four tasks on the Rumsey map dataset. For word detection and recognition, they use DeepSolo [40] with a postprocessing step inspired by Non-Maximum Suppression. For text linking, they propose a four-step approach: (1) neighbor sampling, (2) self-sequencing, (3) graph

assignment, and (4) ordering. In step (2), they introduce Self-Sequencer, a trainable Transformer-based model that iteratively produces ordered local sequences from the input query segment and candidate neighbors [46]. Training uses a mix of real and synthetic datasets: ICDAR MapText, MapKuratorHuman [15], SynthMap [22], and Paris and Jerusalem Maps Text Dataset [10].

CREPE + BezierCurve *Y. Baek, M. Hentschel, Y. Nakagome, S. Ichimura, J. Tae Lee, C. Choi — NAVER (Republic of South Korea) and LINE WORKS (Japan)*

Winner Task 2 (Rumsey), Task 4 (Rumsey)

This team participates in the four tasks on the Rumsey map data set. The team extends the DONUT [14] architecture to perform coordinate regression as in the CREPE [30] approach, using eight control points based on Bezier curves as proposed in ABC-Net [26]. This results in an end-to-end model that performs text detection, recognition, and linking tasks without any post-processing. The model is pretrained on the ArT dataset [7] for text reading and fine-tuned on the Rumsey MapText training dataset.

YOLOv8_ViTAE_PolygonDetector *B. Rajesh, D. Raj Sekar, L. Babu Kuna, Venkatesh — Indian Institute of Information Technology, Sri City (India)*

This team participated in all the tasks on three data sets. The team proposes YOLOv8-ViTAE-Polygon that extends the YOLOv8 architecture [17,25] to predict polygon control points instead of traditional bounding boxes, enabling precise detection of irregularly shaped objects. The model includes a Transformer-based Polygon Decoder to output polygon coordinates for object boundaries. Training uses a combined dataset with Binary Cross-Entropy and Dice Loss for better boundary accuracy.

Word-Level Text Detection Using Multi-Stage Preprocessing and PaddleOCR *A. Prajapati, A. Chakraborty, M. Javed, D. Doermann — IIIT Allahabad (India) and University at Buffalo (USA)*

This team participates in Task 1 and Task 3 for the Rumsey map data set. They use the PaddleOCR framework [19] with polygonal detection enabled to extract precise multi-vertex polygons for each text instance. The system also applies a suite of enhancement techniques to each image, including adaptive histogram equalization, noise reduction, and edge-based filtering, to generate several preprocessed versions, then performs OCR and selects the method yielding the highest text detection count and average confidence.

PolyTextTR *S. Baltaci, R. Baena, F. Meng, M. Aubry — ENPC (France)*

This team participated in all the tasks on three data sets. The team proposes a modified DINO-DETR [42] that predicts bounding polygons for text spotting. For Latin text (Rumsey & IGN), they pretrain the model using scene text datasets (TextOCR[35], TotalText[8], ICDAR15[12], MLT17 [44]). For

Chinese text, there is no pretraining phrase. The approach also adopts test-time augmentations (patch-wise inference) and post-processing.

MapTextSpotter *J. Li, C. Xu, C. Shi, Y. Chen, W. Cao — Qingdao University of Science and Technology (China)*

This team participates in Task 1 and Task 3 on the Rumsey dataset, employing the same approach as used in the 2024 competition. The description of the methodology is provided in [21].

Baseline TESTR Finetuned + Heuristic MST *MapText Competition Organizers*

We adopt an existing text spotting model, TESTR [43], to detect and recognize text instances on maps. The model is built upon Deformable DETR [45] and uses dual decoders for text-box control point regression and character recognition, respectively. We finetune the pretrained model weights to generate baseline results for three data sets. The model was pretrained on SynthText and multiple human-annotated scene image data sets and finetuned on the Rumsey train, IGN train, TWH train data sets. Linking was achieved with a heuristic minimum spanning tree metric [31] with edge weights based on distance and similarity.

5 Results and Discussion

This section reports the quantitative results for each task of the competition on each data set. Linking tasks (2 and 4) are not evaluated on the Chinese Historical Topographic Maps data set, as the ground truth does not contain any links. The HMean for the French Land Register does not include the Tightness for any task, as explained in Section 2.

5.1 Results for Task 1: Word Detection

Table 3 presents results for Task 1 on each data set. The retrained version of last year's winning method, "MapText Detection Strong Pipeline", is the best performing method on the Rumsey and Taiwanese data sets. We also note a very close performance of the "Self-Sequencer" approach on the Rumsey data set, which is a strong contender for the top position. It is interesting to note that, for this task, DETR-based methods are the top performing ones, and our finetuned TESTR baseline is a strong contender as well, especially for the French Land Register data set.

The results indicate that isolated word detection can be solved with good performance using general object detection methods, with the help of proper fine-tuning. The overall performance is still far from being perfect, as the best methods are close to a 90% HMean score, and it is also the indication that architectural or training improvements are still possible. However, this must be balanced with the fact that the tightness of the ground truth is not perfect and

Table 3: Results for Task 1 (isolated word detection). Values expressed in percentage. For all metrics, higher is better.

	Rank	Method name	HMean Det.	Tightness Det.	Precision Det.	Recall Det.
Rumsey	1	MapText Strong Pipeline	90.2	83.8	95.9	91.8
	2	Self-Sequencer	88.9	86.1	91.5	89.1
	3	(MapText'24 ...Strong Pipeline	88.7	82.7	94.2	89.9
	4	Baseline TESTR Finetuned	88.5	86.3	89.1	90.0
	5	PolyTextTR	86.7	82.5	90.3	87.8
	6	MapTextSpotter	84.9	81.4	92.6	81.5
	7	CREPE + BezierCurve	81.9	73.6	87.1	86.5
	8	YOLOv8-ViTAE-Polygon	60.4	74.4	54.1	56.3
	9	Word-Level T...g and PaddleOCR	54.2	73.9	59.5	40.0
French	1	Baseline TESTR Finetuned	80.5	69.7	80.7	80.2
	2	PolyTextTR	75.2	70.3	76.5	73.9
	3	YOLOv8-ViTAE-Polygon	67.0	68.2	63.5	71.0
Taiwanese	1	MapText Strong Pipeline	91.1	87.5	96.5	89.7
	2	Baseline TESTR Finetuned	79.4	88.2	71.0	80.8
	3	PolyTextTR	72.7	82.4	65.5	72.3
	4	YOLOv8-ViTAE-Polygon	33.8	84.9	15.4	82.8

Table 4: Results for Task 2 (grouped word detection). Values expressed in percentage. For all metrics, higher is better.

	Rank	Method name	HMean Link	Recall Link	Precision Det.	Tightness Det.	Precision Det.	Recall Det.
Rumsey	1	Self-Sequencer	78.4	61.6	72.6	86.1	91.5	89.1
	2	CREPE + BezierCurve	78.4	71.7	75.7	73.6	87.1	86.5
	3	Baseline TES...+ Heuristic MST	55.3	51.1	27.1	86.3	89.1	90.0
	4	(Baseline MapText'24) DS-LP	42.1	16.6	55.3	71.6	71.8	78.9
FR	1	Baseline TES...+ Heuristic MST	10.2	27.8	3.0	69.7	80.7	80.2

exhibits a lot of variance, which makes it difficult to reach a perfect score. This underlines the fact that detection evaluation, *per se*, is not sufficient to assess the quality of a method in the context of the competition.

5.2 Results for Task 2: Phrase Detection (Word Grouping)

Table 4 presents results for Task 2. This is the part of the competition that features the greatest change compared to last year, with significant progress in linking methods. It is worth mentioning that the two best methods on the Rumsey data set are reaching the exact same level of performance, with a HMean score of 78.4%, despite using different approaches. The "CREPE + BezierCurve" method relies on the CREPE [30] decoder to produce structured output, while the "Self-Sequencer" method uses a new link prediction method [46] which relies on a multi-step process to progressively aggregate words into groups. Finally, the provided baseline, which relies on a simple heuristic based on a minimal spanning tree, enables assessing the relative difficulty between the Rumsey and

Table 5: Results for Task 3 (word detection and recognition). Values expressed in percentage. For all metrics, higher is better.

	Rank	Method name	HMean	Char. Acc.	Det. Tightness	Det. Precision	Det. Recall
Rumsey	1	MapText Strong Pipeline	91.1	94.0	83.7	95.9	91.8
	2	Self-Sequencer	90.3	94.9	86.1	91.5	89.1
	3	Baseline TESTR Finetuned	89.5	92.9	86.3	89.1	90.0
	4	(Baseline Ma...Strong Pipeline	89.3	94.0	83.3	96.2	85.0
	5	(Baseline MapText '24) MapTest	87.4	89.5	81.8	90.5	88.2
	6	CREPE + BezierCurve	84.9	95.5	73.6	87.1	86.5
	7	(Baseline Ma... MapTextSpotter	84.5	83.5	81.4	92.6	81.5
	8	(Baseline MapText'24) DS-LP	77.6	90.8	71.6	71.8	78.9
	9	(Baseline Ma...ESTR Checkpoint	74.6	82.1	79.5	71.9	66.9
	10	Word-Level T...g and PaddleOCR	59.5	83.7	73.9	59.5	40.0
	11	YOLOv8_ViTAE_PolygonDetector	27.4	78.0	76.0	61.7	9.6
FR	1	Baseline TESTR Finetuned	83.1	88.9	69.7	80.7	80.2
	2	YOLOv8_ViTAE_PolygonDetector	16.6	58.2	66.7	66.0	6.8
TW	1	MapText Strong Pipeline	83.3	66.2	87.5	96.5	89.7
	2	Baseline TESTR Finetuned	78.4	75.6	88.2	71.0	80.8
	3	YOLOv8_ViTAE_PolygonDetector	54.6	66.2	85.3	35.9	53.7

Table 6: Results for Task 4 (joint grouped word detection and recognition). Values expressed in percentage. For all metrics, higher is better.

	Rank	Method name	HMean	Char. Acc.	Link Recall	Link Precision	Det. Tightness	Det. Precision	Det. Recall
Rumsey	1	CREPE + BezierCurve	80.8	95.5	71.7	75.7	73.6	87.1	86.5
	2	Self-Sequencer	80.8	94.9	61.7	72.6	86.1	91.5	89.1
	3	Baseline TES...+ Heuristic MST	59.3	92.9	51.1	27.1	86.3	89.1	90.0
	4	(Baseline MapText'24) DS-LP	46.2	90.8	16.6	55.3	71.6	71.8	78.9
FR	1	Baseline TES...+ Heuristic MST	12.5	88.9	28.1	3.1	69.7	80.7	80.2

the French Land Register data sets, for which no valid submission was received for this task.

5.3 Results for Task 3: Word Detection and Recognition

Table 5 presents results for Task 3. As for Task 1, leading approaches are all based on DETR architectures, and the two best methods use the more recent DeepSolo [40] variant to achieve accurate transcription. The TESTR [43] architecture, as illustrated by the performance of the provided baseline, is a strong contender for this task as well, but the performance margin of the "MapText Strong Pipeline" on the Taiwanese data set is significant. This may be because TESTR for traditional Chinese (Taiwanese data set) is finetuned from the pretrained model of Latin script, indicating the necessity of a suitably pretrained model for various languages.

5.4 Results for Task 4: Phrase Detection and Recognition

Table 6 presents results for Task 4. Finally, the general task of the competition also features a significant improvement over last year, with the **"CREPE +**

BezierCurve" and the **"Self-Sequencer"** methods reaching a HMean score of 80.8% on the Rumsey data set, and being recognized as **joint winners of this competition.** Their high linking performance was instrumental in achieving this result, which was made possible by a very solid word detection and recognition performance on the first step, though the "CREPE + BezierCurve" method was not a leading one in pure detection. Such results underscore the capabilities of attention-based architectures regarding structured output generation.

6 Conclusion and Final Ranking

This competition addresses the challenge of robust reading in historical maps, which is especially difficult due to the distant nature of the words that need to be grouped and the unpredictable relative directions between them. These unique challenges highlight the originality of this competition and its contribution to advancing the field. This edition provided more data sets and a better evaluation framework, resulting in a better evaluation of the linking problem. We extend our gratitude to all participants for their innovative contributions and efforts. Their work significantly advances the state of the art in the linking task, which was the main objective of this competition. The results demonstrate substantial progress across all tasks. For Tasks 1 and 3 (Word Detection and Word Detection and Recognition), DeepSolo-based methods dominated, with the "MapText Detection Strong Pipeline" achieving the best performance on the Rumsey and Taiwanese datasets. Task 2 (Phrase Detection) saw significant improvements in linking methods, with the "CREPE + BezierCurve" and "Self-Sequencer" methods achieving identical top scores on the Rumsey dataset. Finally, **Task 4 (Phrase Detection and Recognition) showcased the joint winners, "CREPE + BezierCurve" from the NAVER and LINE WORKS team, and "Self-Sequencer" from the EPFL team, both achieving an HMean score of 80.8% on the Rumsey dataset, demonstrating the potential of attention mechanisms for structured outputs.** We congratulate the winners and all participants for their remarkable achievements. We hope this competition and its results will inspire further innovation in robust reading and linking tasks. We encourage the community to reuse the public materials (datasets and evaluation code) and explore the supplementary material containing qualitative results, available via our Zenodo community at https://zenodo.org/communities/icdar-maptext. We also remind that the submission platform is still open to post-competition submissions at https://rrc.cvc.uab.es/?ch=32.

Acknowledgments. The authors thank David Rumsey for his generous support for the competition. This work is partially supported by the French Ministry of the Armed Forces - Defense Innovation Agency (AID). Digitized French land registers are provided by the Archives of the French *departement* of Val-de-Marne (AD94).

References

1. Archives Départementales du Val de Marne: Cadastre napoléonien. https://archives.valdemarne.fr/recherches/archives-en-ligne/cadastre-napoleonien
2. Can, Y.S., Erdem Kabadayi, M.: Text detection and recognition by using CNNs in the Austro-Hungarian historical military mapping survey. In: The 6th international workshop on historical document imaging and processing. pp. 25–30 (2021)
3. Cartography Associates: David Rumsey map collection. https://www.davidrumsey.com
4. Chazalon, J., Tual, S., Abadie, N., Duménieu, B., Perret, J., Weinman, J.: IGN test data for ICDAR'24 MapText competition (Mar 2024). https://doi.org/10.5281/zenodo.10732281
5. Chazalon, J., Tual, S., Abadie, N., Duménieu, B., Perret, J., Weinman, J.: IGN Train and Validation Data for ICDAR'24 MapText Competition (Apr 2024). https://doi.org/10.5281/zenodo.10987299
6. Chiang, Y.Y., Leyk, S., Knoblock, C.A.: A survey of digital map processing techniques. ACM Computing Surveys (CSUR) **47**(1), 1–44 (2014)
7. Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., et al.: ICDAR2019 robust reading challenge on arbitrary-shaped text - RRC-ArT. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1571–1576. IEEE (2019)
8. Ch'ng, C.K., Chan, C.S., Liu, C.: Total-Text: Toward orientation robustness in scene text detection. International Journal on Document Analysis and Recognition (IJDAR) **23**, 31–52 (2020). https://doi.org/10.1007/s10032-019-00334-z
9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747
10. Dai, T., Johnson, K., Petitpierre, R., Vaienti, B., di Lenardo, I.: Paris and Jerusalem Maps Text Dataset (2025), https://doi.org/10.5281/zenodo.14982662
11. Gomez, R., Shi, B., Gomez, L., Numann, L., Veit, A., Matas, J., Belongie, S., Karatzas, D.: ICDAR2017 robust reading challenge on COCO-text. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1435–1443. IEEE (2017)
12. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: ICDAR 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1156–1160. IEEE (2015)
13. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Gomez i Bigorda, L., Robles Mestre, S., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: ICDAR 2013 robust reading competition. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1484–1493. IEEE (2013)
14. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: OCR-free document understanding transformer. In: Computer Vision – ECCV 2022. pp. 498–517. Springer Nature Switzerland, Cham (2022)
15. Kim, J., Li, Z., Lin, Y., Namgung, M., Jang, L., Chiang, Y.Y.: The mapKurator system: A complete pipeline for extracting and linking text from historical maps. In: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems. SIGSPATIAL '23, Association for Computing Ma-

chinery, New York, NY, USA (2023). https://doi.org/10.1145/3589132.3625579, https://doi.org/10.1145/3589132.3625579
16. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9404–9413 (2019)
17. Krasnov, D.I., Yarishev, S.N., Ryzhova, V.A., Djamiykov, T.S.: Improved YOLOv8 Network for Small Objects Detection. In: 2024 XXXIII International Scientific Conference Electronics (ET). pp. 1–4. IEEE (2024)
18. Lawhead, J.: QGIS python programming cookbook. Packt Publishing Ltd (2017)
19. Li, C., Liu, W., Guo, R., Yin, X., Jiang, K., Du, Y., Du, Y., Zhu, L., Lai, B., Hu, X., et al.: PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system. arXiv preprint arXiv:2206.03001 (2022)
20. Li, Z., Chiang, Y.Y., Tavakkol, S., Shbita, B., Uhl, J.H., Leyk, S., Knoblock, C.A.: An automatic approach for generating rich, linked geo-metadata from historical map images. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3290–3298 (2020)
21. Li, Z., Lin, Y., Chiang, Y.Y., Weinman, J., Tual, S., Chazalon, J., Perret, J., Duménieu, B., Abadie, N.: ICDAR 2024 competition on historical map text detection, recognition, and linking. In: Barney Smith, E.H., Liwicki, M., Peng, L. (eds.) Document Analysis and Recognition - ICDAR 2024. pp. 363–380. Springer Nature Switzerland, Cham (2024)
22. Lin, Y., Chiang, Y.Y.: Hyper-local deformable transformers for text spotting on historical maps. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 5387–5397 (2024). https://doi.org/10.1145/3637528.3671589
23. Lin, Y., Li, Z., Chiang, Y.Y., Weinman, J.: Rumsey Train and Validation Data for ICDAR'24 MapText Competition (Jun 2024). https://doi.org/10.5281/zenodo.11516933
24. Lin, Y., Li, Z., Chiang, Y.Y., Weinman, J.: Rumsey test data for ICDAR'24 MapText competition (Mar 2024). https://doi.org/10.5281/zenodo.10776183
25. Liu, C., Xi, C.: A multi-strategy integrated optimized YOLOv8 algorithm for object detection. In: 2023 IEEE 6th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE). pp. 893–899. IEEE (2023)
26. Liu, Y., Chen, H., Shen, C., He, T., Jin, L., Wang, L.: ABCNet: Real-time scene text spotting with adaptive Bezier-curve network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9809–9818 (2020)
27. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: ICDAR 2023 competition on hierarchical text detection and recognition. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) Document Analysis and Recognition - ICDAR 2023. pp. 483–497 (2023)
28. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., et al.: ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019. In: 2019 International conference on document analysis and recognition (ICDAR). pp. 1582–1587. IEEE (2019)
29. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al.: ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification - RRC-MLT. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 1454–1459. IEEE (2017)

30. Okamoto, Y., Baek, Y., Kim, G., Nakao, R., Kim, D., Yim, M.B., Park, S., Lee, B.: CREPE: Coordinate-aware end-to-end document parser. In: Barney Smith, E.H., Liwicki, M., Peng, L. (eds.) Document Analysis and Recognition - ICDAR 2024. pp. 3–20. Springer Nature Switzerland, Cham (2024)
31. Olson, R., Kim, J., Chiang, Y.Y.: Automatic search of multiword place names on historical maps. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data. p. 9–12. GeoSearch '24 (2024). https://doi.org/10.1145/3681769.3698577
32. Qiao, Q., Xie, Y., Gao, J., Wu, T., Huang, S., Fan, J., Cao, Z., Wang, Z., Zhang, Y.: DNTextSpotter: Arbitrary-shaped scene text spotting via improved denoising training. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 10134–10143. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3680981
33. Schlegel, I.: Automated extraction of labels from large-scale historical maps. AGILE: GIScience Series **2**, 12 (2021)
34. Shahab, A., Shafait, F., Dengel, A.: ICDAR 2011 robust reading competition challenge 2: Reading text in scene images. In: 2011 international conference on document analysis and recognition. pp. 1491–1496. IEEE (2011)
35. Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., Hassner, T.: TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8802–8812 (2021)
36. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: COCO-Text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
37. Weinman, J., Chen, Z., Gafford, B., Gifford, N., Lamsal, A., Niehus-Staab, L.: Deep neural networks for text detection and recognition in historical maps. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 902–909. IEEE (2019)
38. Weinman, J., Gómez Grabowska, A., Karatzas, D.: Counting the corner cases: Revisiting robust reading challenge data sets, evaluation protocols, and metrics. In: Barney Smith, E.H., Liwicki, M., Peng, L. (eds.) Document Analysis and Recognition - ICDAR 2024. pp. 324–342. Springer Nature Switzerland, Cham (2024)
39. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: CCNet: Extracting high quality monolingual datasets from web crawl data. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4003–4012. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.494/
40. Ye, M., Zhang, J., Zhao, S., Liu, J., Liu, T., Du, B., Tao, D.: DeepSolo: Let transformer decoder with explicit points solo for text spotting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19348–19357 (2023)
41. Yu, W., Liu, M., Chen, M., Lu, N., Wen, Y., Liu, Y., Karatzas, D., Bai, X.: ICDAR 2023 competition on reading the seal title. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) Document Analysis and Recognition - ICDAR 2023. pp. 522–535. Springer Nature Switzerland, Cham (2023)
42. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In:

The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=3mRwyG5one
43. Zhang, X., Su, Y., Tripathi, S., Tu, Z.: Text spotting transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9519–9528 (June 2022)
44. Zheng, T., Chen, Z., Huang, B., Zhang, W., Jiang, Y.G.: MRN: Multiplexed routing network for incremental multilingual text recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18644–18653 (2023)
45. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=gZ9hCDWe6ke
46. Zou, M., Dai, T., Petitpierre, R., Vaienti, B., Kaplan, F., Lenardo, I.d.: Recognizing and sequencing multi-word texts in maps using an attentive pointer. https://doi.org/10.21203/rs.3.rs-6330456/v1, https://www.researchsquare.com/article/rs-6330456/v1, ISSN: 2693-5015

Supplementary Material

This supplementary material contains additional information about the ICDAR'25 MapText competition, including:

- details about the production of the data sets (Section A),
- details about the evaluation protocol (Section B),
- graphical submission rankings (Section C),
- sample results from the different submissions (Section D),
- additional figures from the introduction to illustrate the challenges of the tasks (Section E),
- method descriptions provided by the participants (Section F).

A Data Set Details

A.1 Synthetic Map Images

The synthetic map image generation process follows the methodology described in PALETTE [22]. This approach enables the creation of high-quality synthetic map images with realistic text placements, background textures, and cartographic layouts to mimic historical maps.

While the original method focuses on Latin-based scripts, we extend the approach to support the generation of synthetic maps in Traditional Chinese. To support Traditional Chinese, we curated a comprehensive lexicon of place names and phrases from multiple sources. The primary source is a Taiwanese historical place name gazetteer, collected from the Taiwan History and Culture Map System.[15] To expand linguistic coverage and phrase diversity, we also utilized the CC-100 dataset [9,39], a large-scale corpus extracted from web crawl data.[16] Text sampled from these sources are rendered onto synthetic map backgrounds using fonts that replicate the calligraphic and ancient styles of historical Chinese maps. We also adjust layout parameters such as text orientations (enabling both horizontal and vertical writing directions) and spacing sizes to account for the unique properties of historical maps.

A.2 Annotation Protocol for Taiwan Historical Maps

All the map annotations are performed by human annotators. For one map image, the annotators were instructed to perform text detection, recognition, and linking annotations concurrently. The documentation of annotation procedure and instructions (written in Chinese) is publicly available.[17]

[15] https://thcts.sinica.edu.tw/
[16] https://data.statmt.org/cc-100/
[17] https://drive.google.com/file/d/1I6B9TG14hIs8FWWvEwCS_Bynr0s6bo5S/view

B Evaluation Protocol

As with the 2024 competition, evaluation begins with an optimized one-to-one matching between ground truth elements and detected elements [21]. Here we use the term "elements" rather than "words" because the Chinese-language data set may include multiple characters together in a single annotation region. This section lays out evaluation protocol details and further specifies metric calculations.

Official evaluation code is available at https://github.com/icdar-maptext/evaluation.

B.1 Correspondence Optimization

Inspired by previous work analyzing the necessity of robust and optimized correspondences [38], the 2025 edition uses the same basic strategy as in 2024: a full weighted bipartite matching algorithm determines the maximal set of true positives TP from among valid candidates [21]:

$$\mathrm{TP} \subset \{(g,d) \in G \times D \mid \mathrm{IoU}\,(g,d) > 0.5\}\,, \tag{S1}$$

where G is the set of ground truth regions and D is the set of detected regions. The algorithm ensures that when there are multiple correspondence candidates, the true positive assignments maximize the total IoU score [38].

Formally, given a scoring function $\psi : G \times D \to \mathbb{R}$, bipartite linear sum assignment finds the $\mathbf{X} \in \mathbb{Z}_2^{|G|\times|D|}$ with entries x_{gd} maximizing the sum

$$\sum_{(g,d)\in G\times D} \psi\,(g,d)\,x_{gd} \tag{S2}$$

with constraints $\sum_{g\in G} x_{gd} \leq 1$ and $\sum_{d\in D} x_{gd} \leq 1$ for all $d \in D$ and $g \in G$, respectively [38]. Each $x_{gd} = 1$ in the matrix represents a correspondence.

Ground truth regions are marked as a "don't care" are handled as in the 2024 edition (details below).

B.2 Task-Specific Evaluations

This section details specifics of the evaluation protocol for each task. While all tasks utilize the same basic optimization framework denoted by Eq. (S2), match weights $\psi\,(g,d)$ may differ across tasks.

We refer throughout to each ground truth element $g \in G$ as a "word,", even though elements from the Taiwanese maps may contain a close group of typographically aligned Chinese characters corresponding to more than one "word."

Tasks 1 and 2: Word Detection and Linking For Task 1, the competition metric is the harmonic mean among word precision P, recall R, and tightness T. We therefore desire correspondences that maximize true positives for P and R (the size of TP in S1), as well as IoU for T. Thus, the match score function for Task 1 remains unchanged from the 2024 edition [21]:

$$\psi(g,d) = \begin{cases} \text{IoU}(g,d) & \text{if IoU}(g,d) > 0.5 \wedge \neg I(g) \\ \epsilon & \text{if IoU}(g,d) > 0.5 \wedge I(g) \\ -1 & \text{otherwise,} \end{cases} \tag{S3}$$

where predicate $I(g)$ represents whether g is a "don't care" word to be ignored. The very small value of $\epsilon > 0$ allows detections to be matched with "don't care" ground truth items (and later discounted) while still preferring matches with valid ground truths. For scoring with the competition metrics,

$$\text{TP} = \{(g,d) \in G \times D \mid x_{gd} = 1 \wedge \neg I(g)\}. \tag{S4}$$

Competition metric of Task 2 includes link precision and recall. While a match algorithm sensitive to these factors (such as a graph edit distance) could theoretically offer evaluation performance improvements, we choose to use the same correspondences for stability in method comparison across tasks as well as simplicity for the evaluation code.

Task 3 and 4: Word Detection, Recognition, and Linking Unlike the prior edition, true positives for word recognition do not require matching transcriptions. Instead, the competition metric includes true positive character accuracy as a measure of recognition quality (i.e., using CER rather than WER analysis). To determine the true positive set, we also use the recognition accuracy to promote string match quality among correspondences. In the bipartite graph, the edge weight $\psi(g,d)$ between a ground truth group g and detected group d is given by [21]

$$\psi(g,d) = \begin{cases} \text{IoU}(g,d)\,(1 - \text{NED}(g,d)) & \text{if IoU}(g,d) > 0.5 \wedge \neg I(g) \\ \epsilon & \text{if IoU}(g,d) > 0.5 \wedge I(g) \\ -1 & \text{otherwise,} \end{cases} \tag{S5}$$

B.3 Link Evaluation

In order to measure how well the methods perform at identifying individual links among words, the competition also assesses the entries in the adjacency matrix at the individual link level among words. The graph from which such an adjacency matrix is constructed consists of nodes representing the matched words (e.g., IoU > 0.50) and any remaining unmatched ground truth or predicted words. The links are given by the ordering of the word polygons constituting each phrase.

Samples of the various linking scenarios and their classifications are depicted in Figure S1. With these adjacency matrix classifications, we can report the overall precision, recall, and F-score over the matrices' entries.

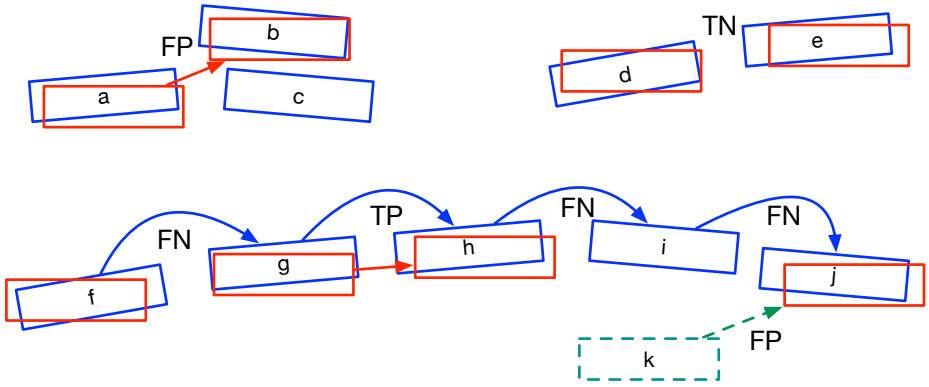


Fig. S1: A graph structure for a link-level analysis of word grouping and reading order. Blue boxes represent ground truth and red boxes represent matched predictions. Two ground truth boxes (c and i) are unmatched. The dashed green box represents an unmatched prediction (k). Blue arrows represent ground truth phrases with reading order, and the red arrows correspond to the predicted groupings and reading order. The classification of each link is given; only one of the several true negatives is explicitly denoted.

Figure S2 shows the adjacency matrix resulting from Figure S1. With one true positive (g→e), two false positives (a→b and k→j), and three false negatives (f→g, h→i, and i→j), the resulting link level scores are $P_L = \frac{1}{1+2}$ and $R_L = \frac{1}{1+3}$.

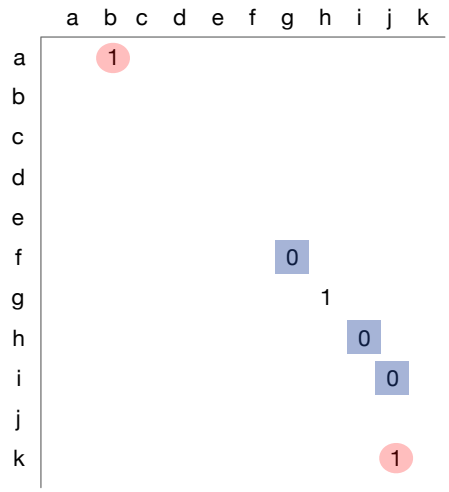


Fig. S2: The adjacency matrix for the graph in Figure S1. Red circles indicate false positives; blue squares indicate false negatives. Unspecified entries are zeros.

C Graphical Submission Ranking

This section features ranking summaries as visual bar plots for fast comparison:

- results for **Task 1** are shown in Figure S3,
- results for **Task 2** are shown in Figure S4,
- results for **Task 3** are shown in Figure S5,
- results for **Task 4** are shown in Figure S6.

Because the Chinese Historical Topographic Maps (TWH) data set does not feature work groups, Tasks 2 and 4 are not applicable to it. For the French Land Register (IGN) data set, the final HMean indicator does not take tightness into account to limit the effect of word region annotation variance in the ground truth.

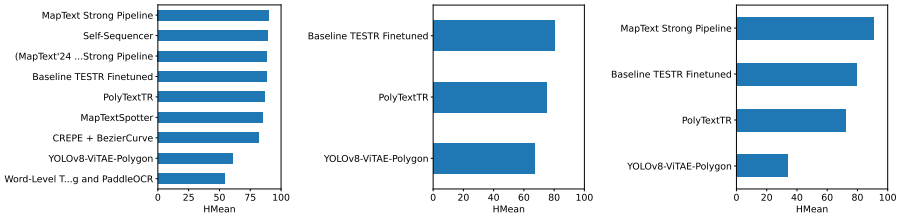


Fig. S3: Final ranking overview for Task 1 on the Rumsey *(left)*, French Land Register *(center)*, and Chinese Historical Topographic Maps *(right)* data sets. Methods are sorted by descending HMean (%).

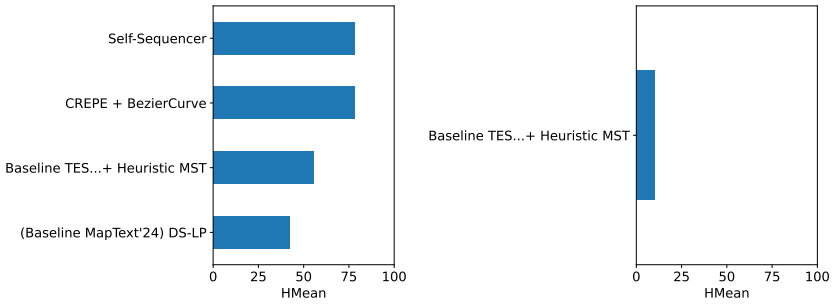


Fig. S4: Final ranking overview for Task 2 on the Rumsey *(left)* and French Land Register *(right)* data sets. Methods are sorted by descending HMean (%).

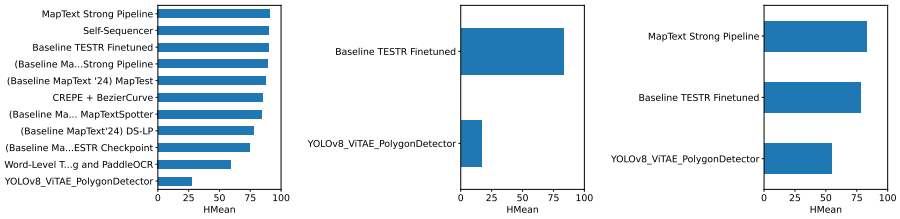


Fig. S5: Final ranking overview for Task 3 on the Rumsey *(left)*, French Land Register *(center)*, and Chinese Historical Topographic Maps *(right)* data sets. Methods are sorted by descending HMean (%).

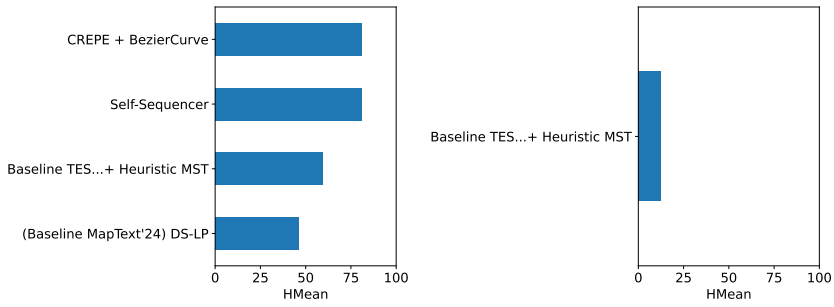


Fig. S6: Final ranking overview for Task 4 on the Rumsey *(left)* and French Land Register *(right)* data sets. Methods are sorted by descending HMean (%), taking into account detection quality and transcription quality (character accuracy). Compared to other tasks, words are grouped and must be detected and recognized.

D Example Results

This section provides illustrations of select example predictions from each system on every task and data set. Additional examples are permanently archived at https://zenodo.org/communities/icdar-maptext.

Every image contains a comparison of the raw predictions of each submission with the ground truth. Each figure provides examples of map tiles that are easiest, hardest, and randomly selected. Because the Chinese Historical Topographic Maps dataset does not provide word grouping annotations, no examples for Tasks 2 and 4 are provided for this data set. For each task, the selection of easy and hard images is based on the submissions' mean performance (HMean) on the task:

Task 1: Isolated words
- Rumsey: Figure S7
- French Land Register: Figure S8
- Chinese Historical Topographic Maps: Figure S9

Task 2: Word groups
- Rumsey: Figure S10
- French Land Register: Figure S11

Task 3: Isolated words transcription
- Rumsey: Figure S12
- French Land Register: Figure S13
- Chinese Historical Topographic Maps: Figure S14

Task 4: Word groups transcription
- Rumsey: Figure S15
- French Land Register: Figure S16

Refer the main paper and Appendix B for formal definitions of the metrics.

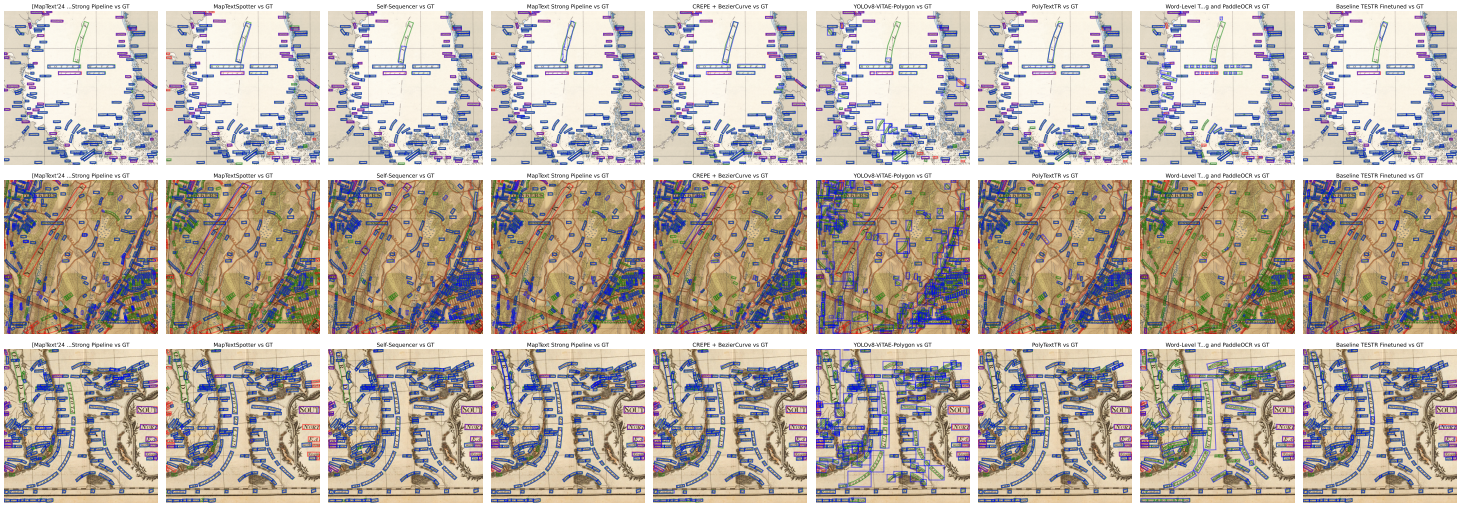


Fig. S7: Example results for Task 1 (word detection) on the Rumsey data set. Blue regions are predictions for the entitled submission. Green indicates valid ground truth words, while red indicates cropped or ignored words. Top–Bottom: Easiest, hardest, random.

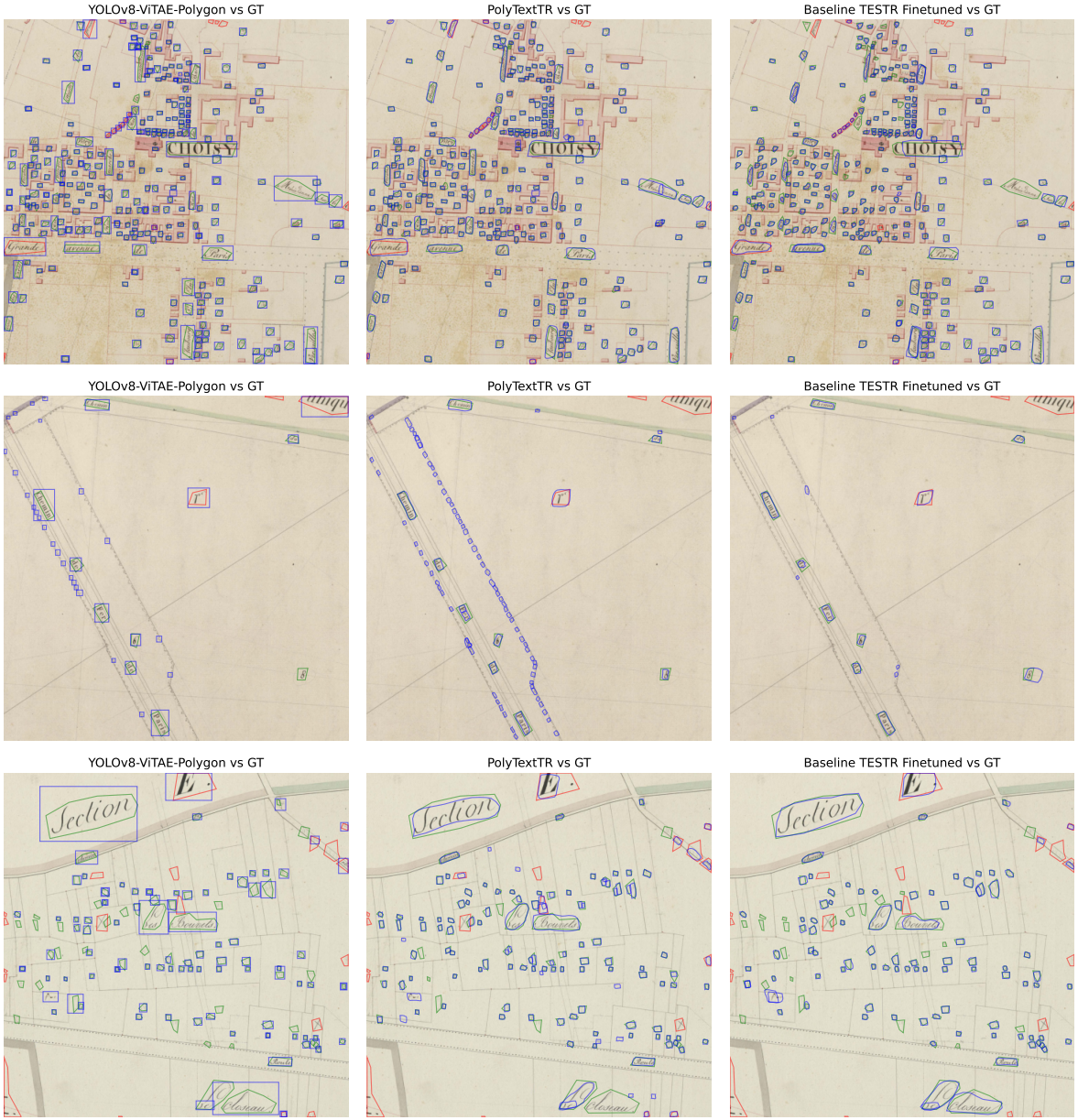


Fig. S8: Example results for Task 1 (word detection) on the French Land Register data set. Blue regions are predictions for the entitled submission. Green indicates valid ground truth words, while red indicates cropped or ignored words. Top–Bottom: Easiest, hardest, random.

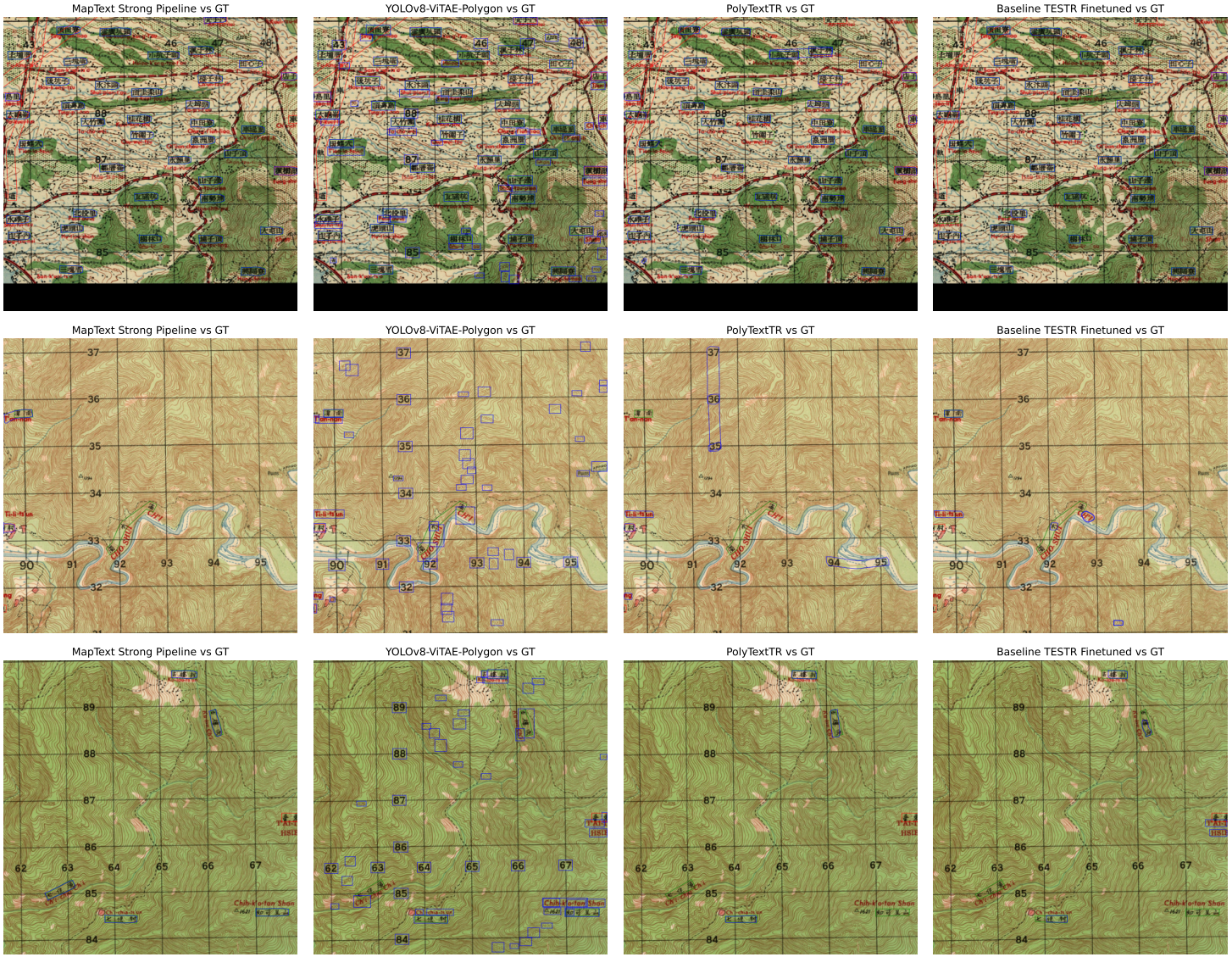


Fig. S9: Example results for Task 1 (word detection) on the Chinese Historical Topographic Maps (TWH) data set. Blue regions are predictions for the entitled submission. Green indicates valid ground truth words, while red indicates cropped or ignored words. Top–Bottom: Easiest, hardest, random.

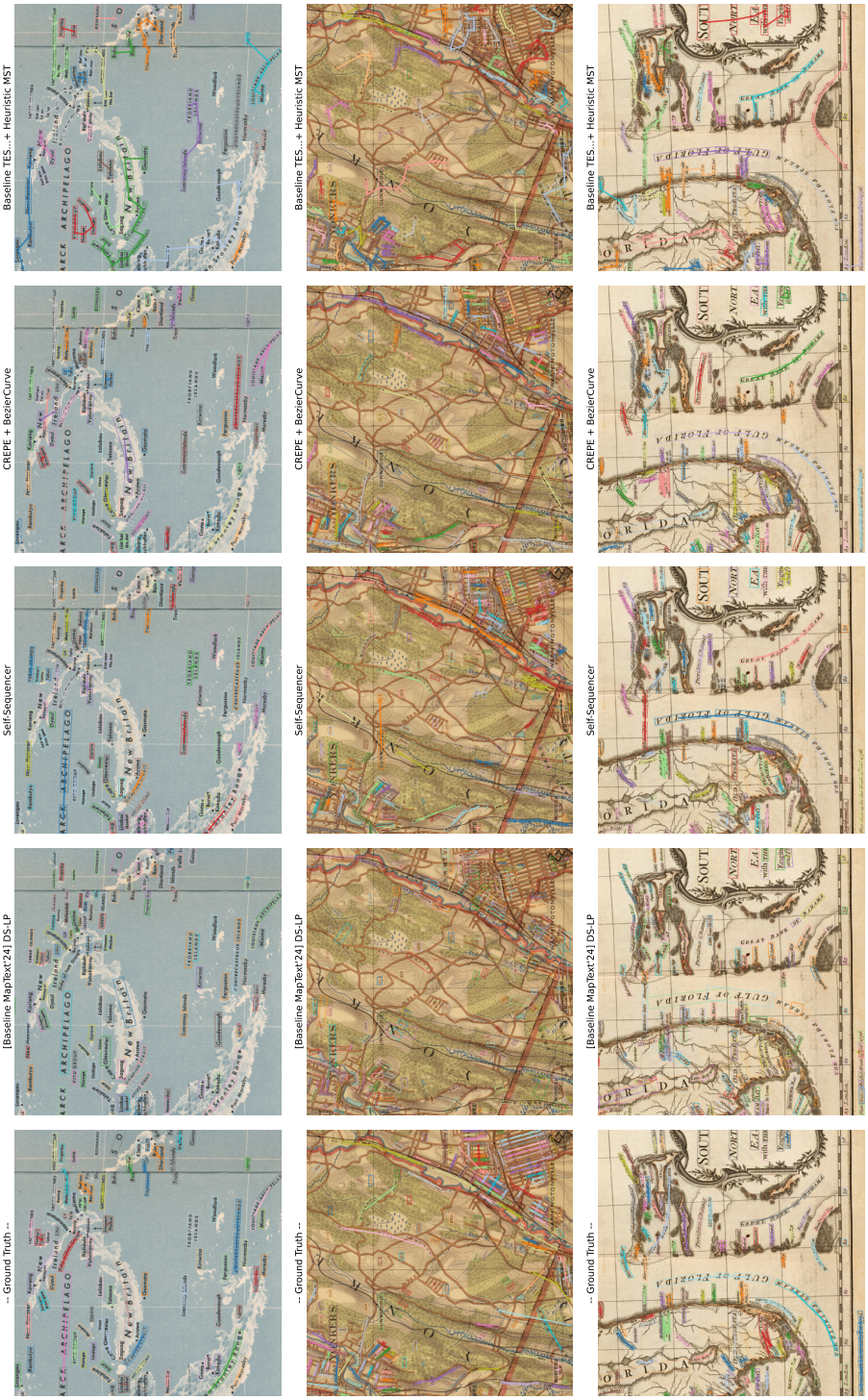


Fig. S10: Example results for Task 2 (phrase detection/word grouping) on the Rumsey data set. Phrase groups have the same color with links drawn between successive group members. Top–Bottom: Easiest, hardest, random.

Fig. S11: Example results for Task 2 (phrase detection/word grouping) on the French Land Register data set. Phrase groups have the same color with links drawn between successive group members. Top–Bottom: Easiest, hardest, random.

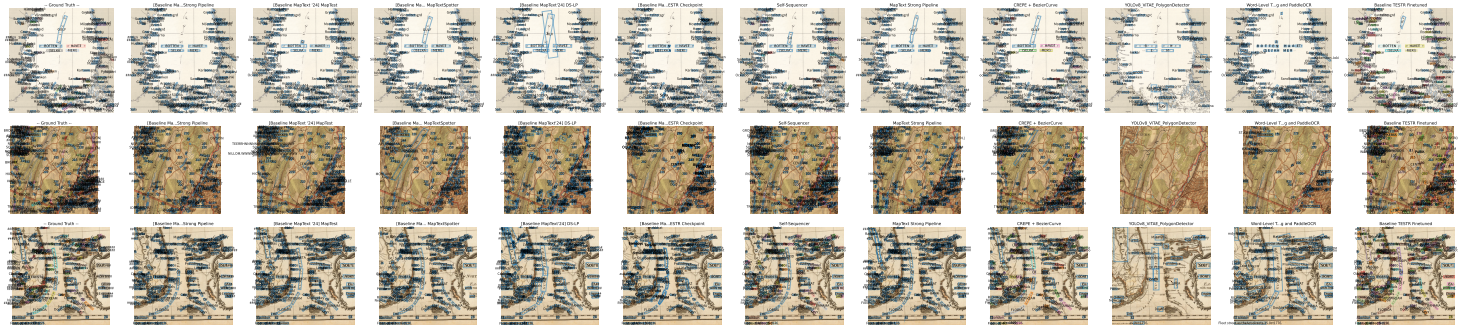


Fig. S12: Example results for Task 3 (word detection and recognition) on the Rumsey data set. Phrase groups have the same color with overlaid transcription. (Ground truth transcriptions including "###" indicate an ignored word.) Top–Bottom: Easiest, hardest, random.

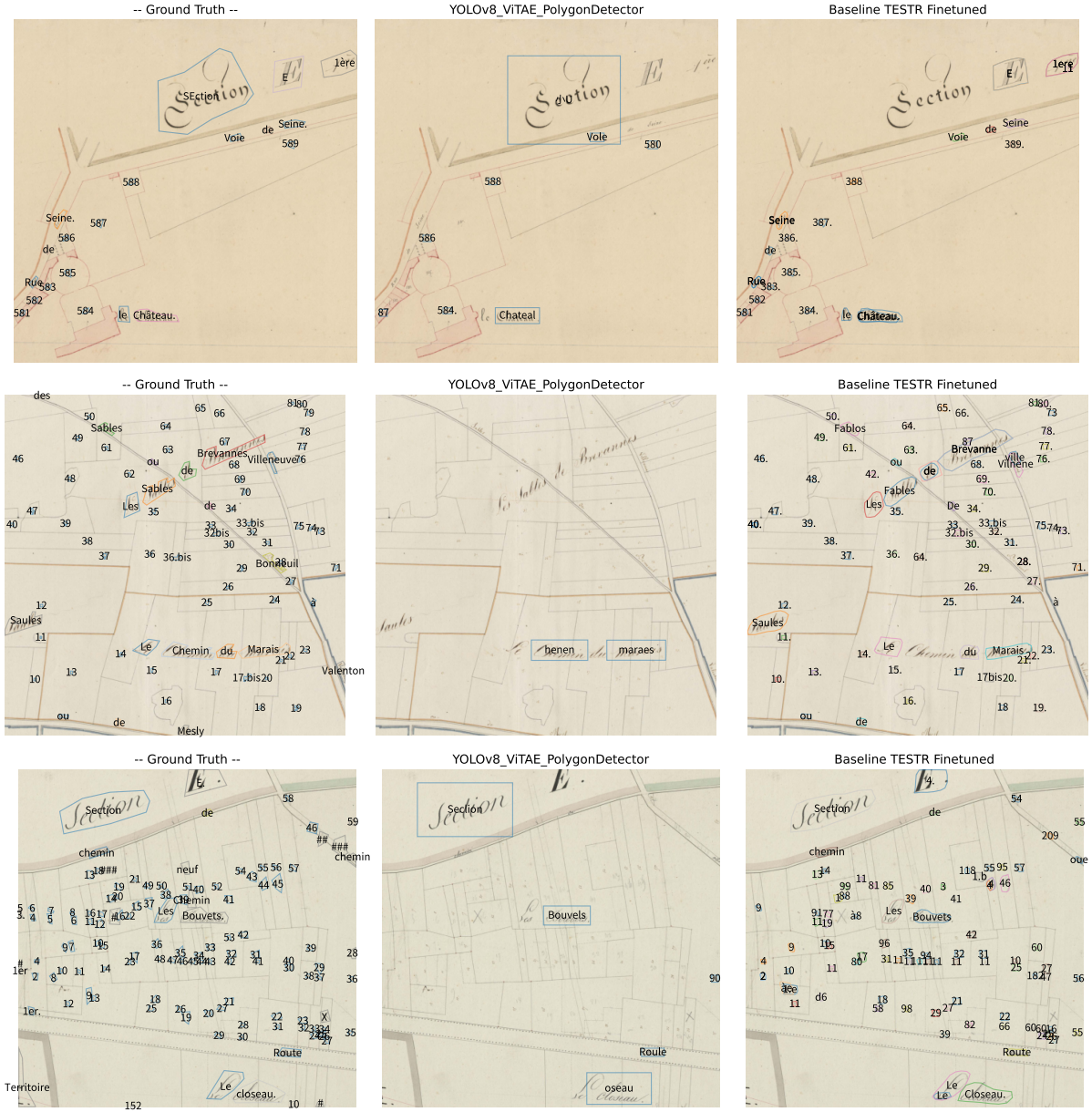


Fig. S13: Example results for Task 3 (word detection and recognition) on the French land register data set. Phrase groups have the same color with overlaid transcription. Top–Bottom: Easiest, hardest, random.

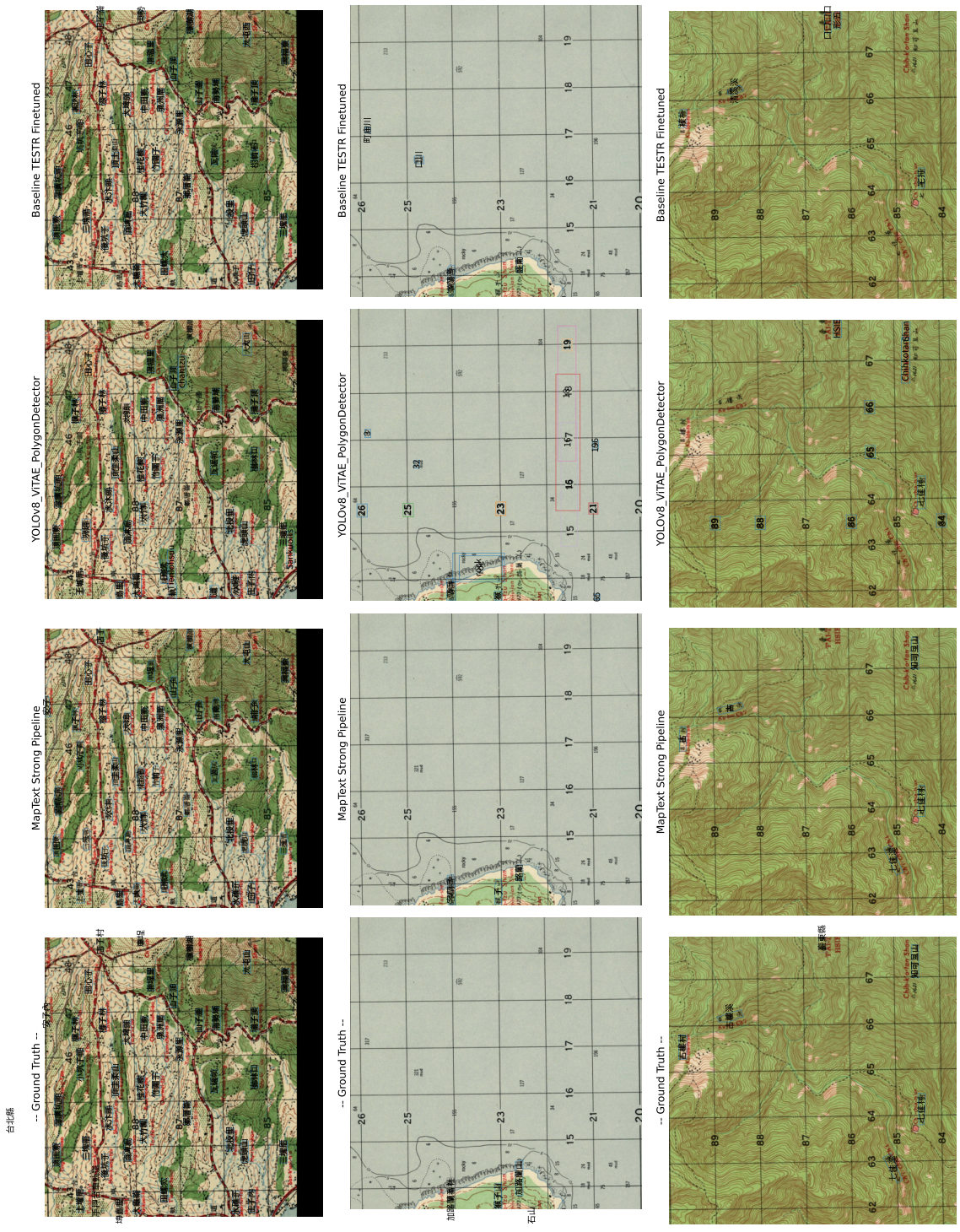


Fig. S14: Example results for Task 3 (word detection and recognition) on the Chinese Historical Topographic Maps (TWH) data set. Phrase groups have the same color with overlaid transcription. Top–Bottom: Easiest, hardest, random.

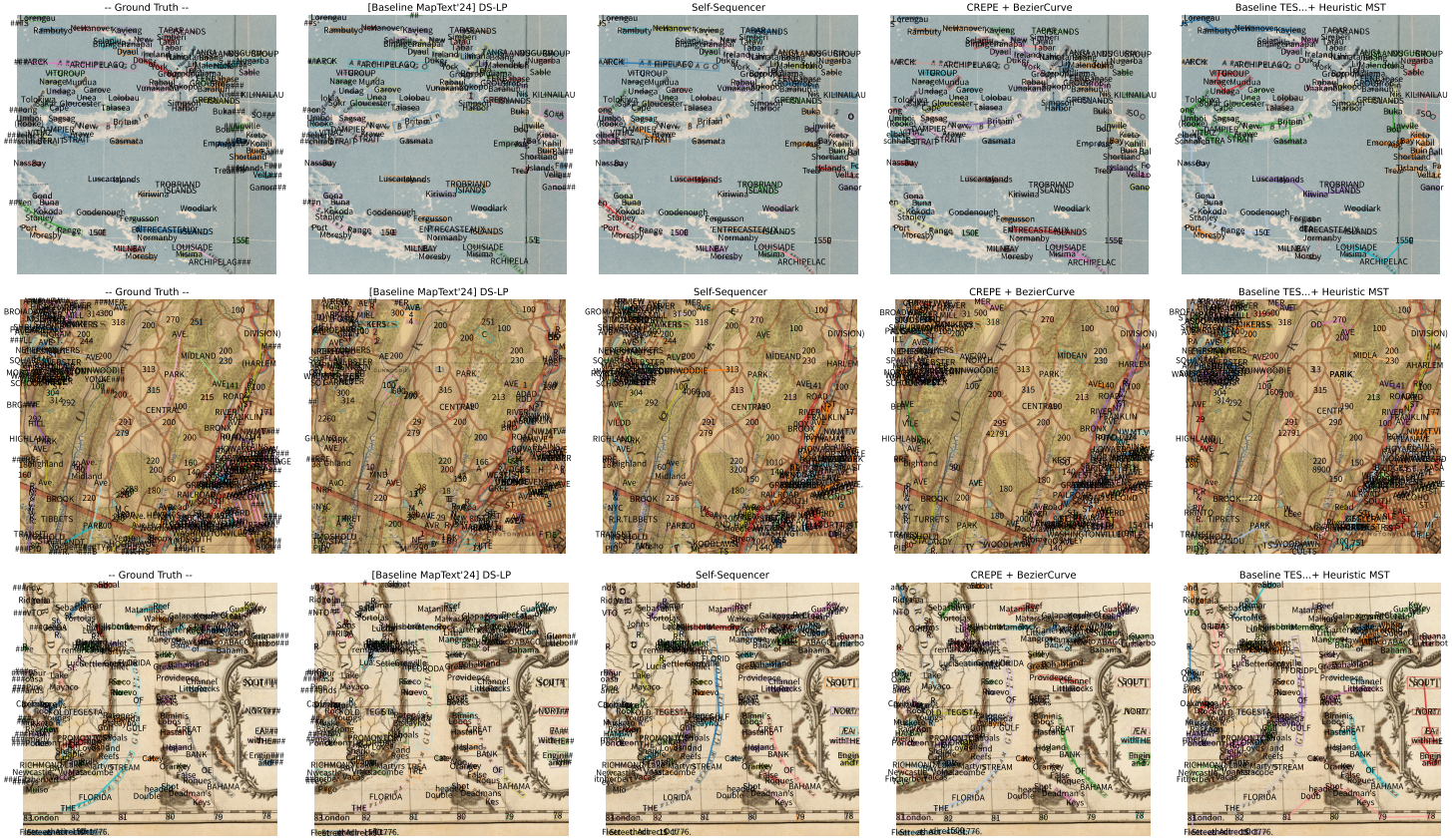


Fig. S15: Example results for Task 4 (phrase detection and recognition) on the Rumsey data set. Phrase groups have the same color with overlaid transcription. (Ground truth transcriptions including "###" indicate an ignored word.) Top–Bottom: Easiest, hardest, random.

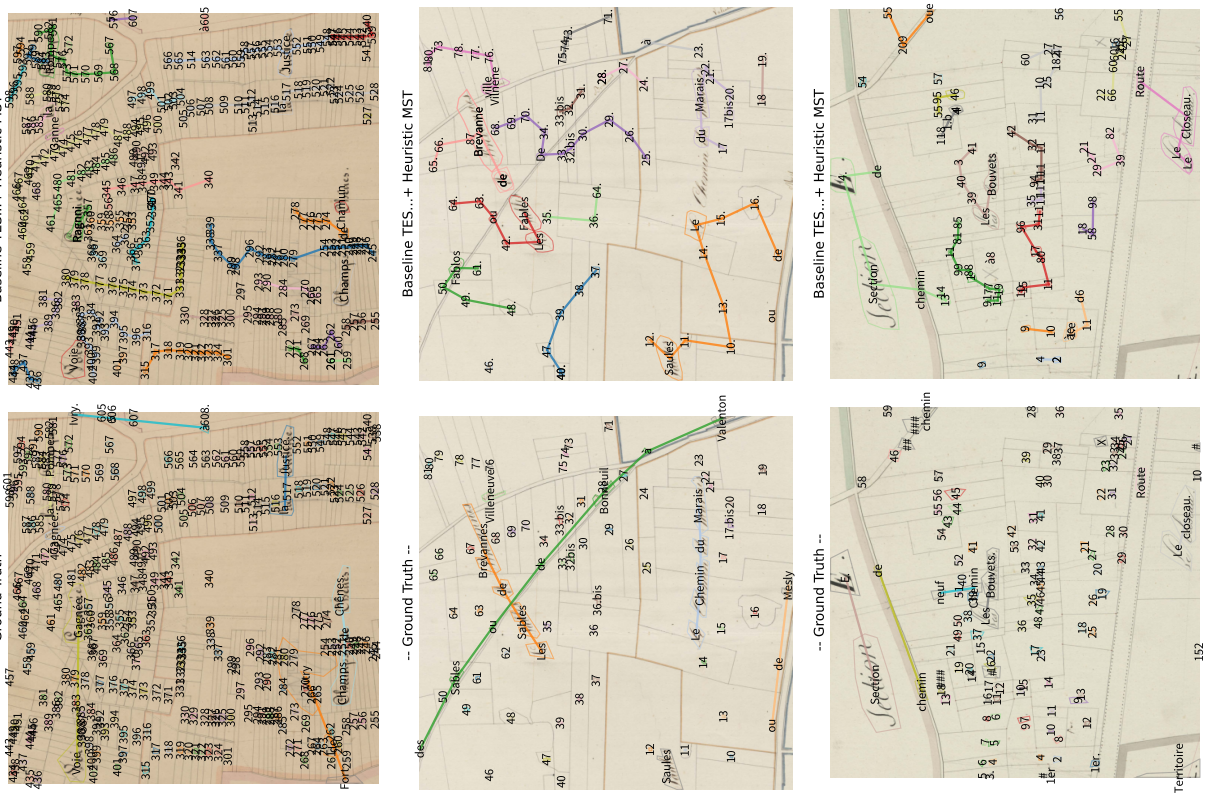


Fig. S16: Example results for Task 4 (phrase detection and recognition) on the French land register data set. Phrase groups have the same color with overlaid transcription. Top–Bottom: Easiest, hardest, random.

E Challenges in Historical Map Analysis

In the introduction of the main report, we discussed the challenges historical map sources exhibit, originating both from their historical and multimodal symbolic natures. We mentioned the following points but omitted some figures for space reasons. We provide them here for completeness.

- **Faded or damaged content**: Many historical maps suffer from deterioration over time, which can include faded ink, physical damage, or altered content, making them difficult to interpret (Figure S17).
- **Complex hierarchical structures**: The layered organization of maps, including territorial boundaries, landmarks, and labels, often follows a complex hierarchical pattern which is reflected on textual content (Figure S18).
- **Overlapping and dense elements**: Historical maps frequently contain overlapping symbols, text with irregular spacing, and dense clusters of information, presented in various orientations (Figure S19).
- **Contextual interpretation**: A deep understanding of the historical and geographical context is often necessary to correctly interpret the map's content, especially when dealing with ambiguous or incomplete data (Figure S20).
- **Evolving lexicon**: Place names may be replaced, spelling conventions deprecated, or terms altered over time, requiring extensive lexicon analysis to recognize all geographical entities (Figure S21).
- **Exotic fonts and handwritten content**: Many historical maps feature fonts that are no longer in use, and some include handwritten annotations (Figure S22).

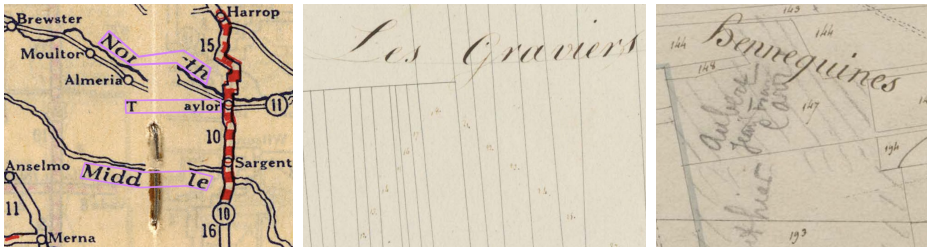


Fig. S17: Faded or damaged content: Paper folding, ink fading or altered content are common challenges in historical documents.

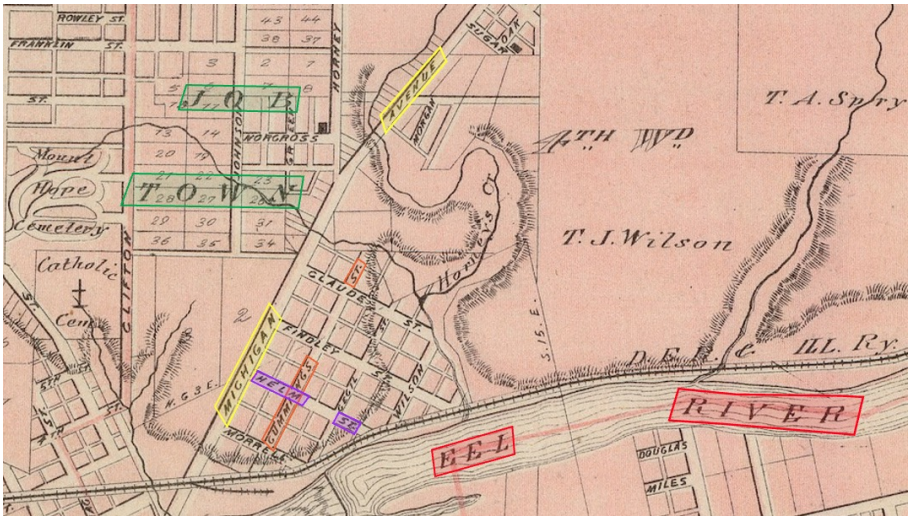


Fig. S18: Complex hierarchical structures: Word detection and linking examples are annotated in colored polygons. The polygons in the same color are linked. The green boxes ("JOB TOWN") with widely spaced characters overlap other words. The yellow boxes are very far apart from each other ("MICHIGAN AVENUE"), but the word "MICHIGAN" is close to and mixed with other street names ("HELM ST." and "GUMMINGS ST."), which are also overlapped each other. Image credit: Rumsey Collection [3] Image 0019.050 (Plan of Logansport, Cass Co., *Indiana*, 1876).

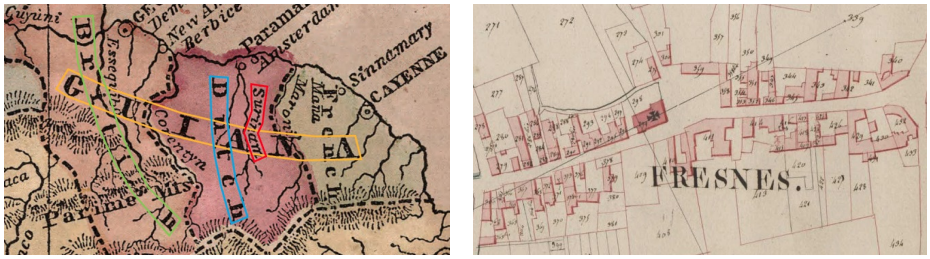
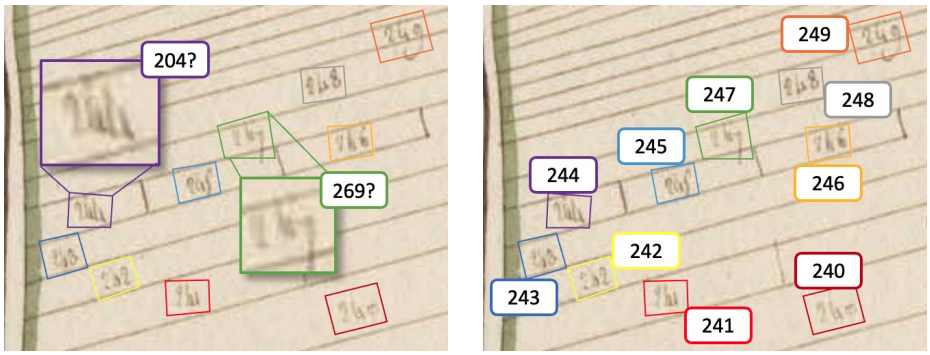


Fig. S19: Overlapping and dense elements: In addition to strong curvature and rotation, large text often overlaps with other elements, and small text can be very dense.



Unordered transcription Numerically-ordered transcription

Fig. S20: Contextual interpretation: Excerpt of French land register map with handwritten text. Even for a human, unordered transcription is challenging (left), while a numerically-ordered transcription is more straightforward (right).

Fig. S21: Evolving lexicon: a simple example of how the city of "Quebec" was spelled differently in the past.

Fig. S22: Exotic fonts and handwritten content: a hand-picked selection of eye-catching elements in the Rumsey collection and the French land registers.

F Detailed description of methods

This section provides a detailed description of the methods used by the participants in the competition, as provided by the participants themselves. We thank them for sharing the details of their methods.

F.1 MapText Strong Pipeline

Authors	Yu Xie, Canhui Xu, Jielei Zhang, Pengyu Chen, Weihang Wang, Yuchen He, Peiyi Li, Yihan Meng, Longwen Gao
Affiliation	Bilibili Inc. and QUST (China)
References	[32,40]
Code/Weights	https://github.com/yyyyyxie/DNTextSpotter

Information sent by the authors:

For the English MapText detection task, we employed DNTextSpotter [32], a novel denoising training method based on DeepSolo [40]. For the Chinese MapText detection task, we utilized DeepSolo. Data augmentation techniques, including cropping, scaling, and adjustments to saturation and contrast, were applied. Pre-training was conducted using available real-world datasets such as TextOCR [35], TotalText [8], ICDAR15 [12], MLT17 [29]. Post-processing methods were also adopted.

F.2 Self-Sequencer

Authors	Mengjie Zou, Tianhao Dai, Remi Petitpierre, Beatrice Vaienti, Frederic Kaplan, Isabella di Lenardo
Affiliation	EPFL, Swiss Federal Institute of Technology in Lausanne (Switzerland)
References	[46,15,10]
Code/Weights	https://github.com/SesamePaste233/ToponymExtractor

Information sent by the authors:

For word detection and recognition, our approach relies on DeepSolo, whose architecture is derived from Detection Transformers (DETR). In short, DeepSolo extracts hierarchical visual features from map images and processes them through an encoder-decoder architecture to detect words as segments bounded by Bézier curves. The model specifically returns four control points of central Bézier curves per word and then uniformly samples query points along these curves to segment, classify, and

delineate each text instance precisely. To resolve duplicate word detections, we implement a postprocessing step inspired by Non-Maximum Suppression. It involves calculating the Fréchet distance between the Bézier curves of potential duplicate word pairs, or "directional synonyms", and merging those below a defined threshold. More details on the model, algorithms, and specific implementation are provided in our separate article [46].

Our text linking methodology consists of four steps: (1) neighbor sampling, (2) self-sequencing, (3) graph assignment, and (4) ordering. In the first step, a word segment is designated as the query, and neighbor segments are used as candidates. For the second step, we introduce Self-Sequencer, a trainable Transformer-based model that iteratively returns ordered local sequences based on the input query segment, and candidate neighbors. Each input text segment is represented only by the control points of its bounding Bézier curves. A Transformer Encoder generates a deep representation from spatial-directional features. Then, an Attentive Pointer module predicts local word sequences based on the concatenated hidden states of candidate word pairs. The aim of the third step, which we call graph assignment, is to aggregate the local link predictions. In this perspective, the links predicted by the Self-Sequencer are used to create a global directional graph. Each strongly connected component of the global graph is considered a distinct linked word set. In the fourth and last step, the order of the words in the sequence is retrieved by applying a consensus ranking algorithm, based on the local sequence order predicted by the Self-Sequencer. More details on the model, algorithms, and specific implementation are provided in our separate article [1].

The model training leverages several real and synthetic datasets: ICDAR MapText [21], MapKuratorHuman [15], SynthMap [22], and Paris and Jerusalem Maps Text Dataset [10].

F.3 CREPE + BezierCurve

Authors	Yu Xie, Jielei Zhang, Ziyue Wang, Yuchen He, Yihan Meng, Weihang Wang, Peiyi Li, Longwen Gao, Qian Qiao
Affiliation	NAVER (Republic of South Korea) and LINE WORKS (Japan)
References	[14,30,26,7]
Code/Weights	Not available.

Information sent by the authors:

We participated in the competition using an image-to-text sequence generation model. The overall architecture follows DONUT [14], which combines a Swin Transformer as the vision encoder and BART as the text decoder. Additionally, inspired by CREPE [30], our model performs

coordinate regression using a 3-layer FFN whenever the decoder outputs an end-of-OCR token (</ocr>).

In this competition, text instances are represented as arbitrarily shaped polygons with a variable number of vertices. To handle this, we adopt the Bezier curve representation as proposed in ABC-Net [26], which enables us to regress a fixed number of control points. Specifically, we predict 8 control points, each with x and y coordinates, resulting in a 16-dimensional output vector for each text instance.

The original CREPE does not generate spatial coordinate tokens within the sequence. However, map datasets often contain multiple occurrences of the same text (e.g., river or mountain) within a single image. To distinguish these instances, we introduce coordinate tokens as initial guidance. The top-left coordinate of each text instance is quantized into 100 bins, and tokens are generated in the format: <ocr.x><ocr.y> text </ocr>. These coordinate tokens serve solely as positional guidance during decoding; the actual text locations are obtained through regression.

For tasks that require linking multiple text segments into a single entity—such as place names—we enclose them with a special token <s_toponym>. For instance, the phrase "Yosemite National Park" is represented as: <s_toponym> <ocr.10><ocr.15>YOSEMITE</ocr> <ocr.8><ocr.20>NATIONAL</ocr> <ocr.21><ocr.20>PARK</ocr> </s_toponym>

This entity grouping strategy enables the model to link related text instances without requiring any post-processing.

For pretraining, we trained the model on the ArT dataset [7] for 10 epochs, focusing exclusively on the text reading task with arbitrary-shaped text. We then fine-tuned the model on the Rumsey dataset for 30 epochs. Input images were resized to 1920×1290, and standard data augmentation techniques—such as rotation, scaling, and basic photometric transformations—were applied during training.

F.4 YOLOv8_ViTAE_PolygonDetector

Authors	Dr. Bulla Rajesh, Dola Raj Sekar, Lokesh Babu Kuna, Venkatesh
Affiliation	Indian Institute of Information Technology, Sri City (India)
References	[17,25]
Code/Weights	Not available.

Information sent by the authors:

Our method, YOLOv8-ViTAE-Polygon, extends the YOLOv8 architecture to predict polygon control points instead of traditional bounding boxes, enabling precise detection of irregularly shaped objects. It incorporates three ViTAE-like transformer layers for feature refinement,

enhancing spatial and contextual understanding. The model includes a Transformer-based Polygon Decoder to output polygon coordinates for object boundaries. Training uses a combined dataset with Binary Cross-Entropy and Dice Loss for better boundary accuracy. Augmentations like mosaic, rotation, and translation improve robustness. The method is suitable for tasks requiring detailed boundary detection, with open-source code and weights provided for reproducibility.

F.5 Word-Level Text Detection Using Multi-Stage Preprocessing and PaddleOCR

Authors	Anand Prajapati, Apurba Chakraborty, Mohammed Javed, David Doermann
Affiliation	IIIT Allahabad (India) and University at Buffalo (USA)
References	[19]
Code/Weights	https://github.com/AnandvPrajapati/Map_Text_Detection_Recognition/tree/main

Information sent by the authors:

Our methodology combines advanced deep learning-based text detection with robust image preprocessing to accurately localize text in challenging historical map images. We use the PaddleOCR framework with polygonal detection enabled, allowing us to extract precise multi-vertex polygons for each text instance, which is essential for handling curved or irregular text layouts typical in historical documents. Each image is preprocessed using adaptive techniques—such as contrast enhancement, denoising, and thresholding—to improve text visibility, and multiple preprocessing variants are evaluated to select the best result. Detected text regions are output as polygons (lists of (x, y) pairs) in the required JSON structure, with each image's predictions grouped accordingly. This approach ensures high-fidelity localization and direct compatibility with the competition's evaluation format, focusing solely on detection accuracy without relying on text recognition content.

F.6 PolyTextTR

Authors	Sonat Baltaci, Raphael Baena, Fei Meng, Mathieu Aubry
Affiliation	ENPC (France)
References	[42]
Code/Weights	Not available.

Information sent by the authors:

Our architecture is a modified DINO-DETR that predicts bounding polygons for text spotting. For pre-training for Latin text (Rumsey & IGN), we utilized scene text datasets (TextOCR [35], TotalText [7], ICDAR15 [12], MLT17 [44]). For Chinese text, no pre-training is done. We also adopted test-time augmentations (patch-wise inference) and post-processing. Our code, architecture, training details, and preprint will be released, soon.

F.7 MapTextSpotter

Authors	Jialiang Li, Canhui Xu, Cao Shi, Yaqi Chen, Wei Cao
Affiliation	Qingdao University of Science and Technology (China)
References	Not available.
Code/Weights	Not available.

Information sent by the authors:

Unlike natural scene text, digitized historical maps have densely distributed text regions, rotated and curved text and widely spaced characters. The text instances have multiple granularities, which hierarchically represent structured geolocation context. To address the new challenges in map text spotting, we have proposed a novel unified network called MapTextSpotter, which jointly explores distinct characteristics in text detection and recognition. Our MapTextSpotter utilized a single decoder with shared queries based on Transformer. The queries are specifically designed spatially and semantically according to text distribution in historical maps. Both point queries and character queries are incorporated and interacted to train the model so as to predict text instance curve Bezier points and character classification in parallel. Notably, densely distributed text instances are often accompanied by smaller fonts. We extract multi-scale visual features with high-resolution detailed convolutional features, which help capture text instances with multiple granularities. Furthermore, with the aid of priori knowledge, Large Language Model is employed to enhance interaction with contextual information to replace the lexicon matching process, which significantly boosts recognition precision. For words highly spaced with complicated text-like noisy distractors, and word phrases divided across multiple lines, we infer that the LLM could alleviate widely space text problems and improve recognition performance by performing instance linkage with prior knowledge.