Conformal Predictors for Efficient Video Text Spotting

Omar Ben Tanfous
omar.tanfous@terminal-industries.com
Sankha Subhra Mukherjee
rick.mukherjee@terminal-industries.com
Neil M. Robertson
neil@terminal-industries.com

Terminal Industries
River House, High Street
Belfast, UK

Abstract

Video text spotting can be achieved by first detecting and recognizing text instances at the frame level, then tracking these texts across video frames. A major challenge with this approach is the inconsistency of text recognition across frames, where an image text spotter may misrecognize text in some frames while correctly recognizing it in others. We argue that text recognition accuracy is correlated with the model's prediction uncertainty. To improve prediction accuracy in videos, we use a statistically rigorous approach to estimate the uncertainty of predicted texts and leverage tracking information to produce more confident and accurate results. To this end, we employ a pretrained state-of-the-art text spotter and apply the conformal prediction framework to estimate its prediction uncertainties. A second challenge addressed in this work is the inconsistent quality of predicted bounding boxes, which impacts text tracking performance. As a solution, we use a second conformal prediction approach that applies corrections to the predicted bounding boxes, providing guarantees with a predefined probability of success. Extensive experiments on four public datasets demonstrate consistent improvements over state-of-the-art methods, achieving IDF1 gains of **1.6%** on DSText, **1.0%** on ArTVideo, and **2.3%** on BOVText. While being computationally lightweight, the proposed method integrates seamlessly into any pretrained matching-based video text spotter, enhancing overall performance. The code is available at https://github.com/OmarBent/CVTS.

1 Introduction

Video text spotting (VTS), the task of detecting, recognizing, and tracking text instances across video frames, is a key component of numerous real-world applications, including video content understanding, autonomous systems [30], and video retrieval [8]. While image text spotting (ITS), which involves text detection and recognition in static images, has been extensively explored with impressive results [13, 29], progress in VTS has lagged behind due to the added complexity of temporal tracking and the limited availability of annotated video datasets.

Existing end-to-end approaches to VTS jointly train a model to detect, recognize, and track text across frames [16, 25, 27, 31]. While these methods have shown promise, they

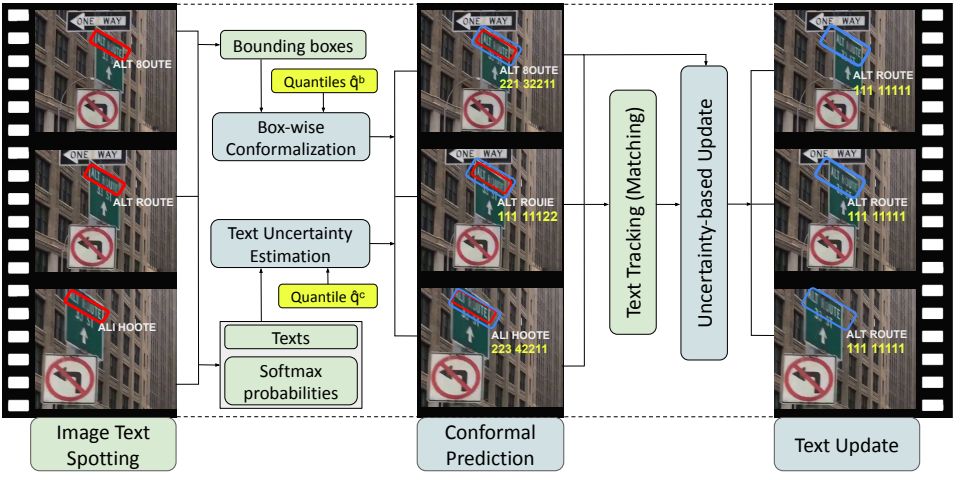

Figure 1: Overview of our **C**onformal **V**ideo **T**ext **S**potter (**CVTS**) framework. A pretrained SOTA text spotter is used to detect and recognize text in input video frames. CVTS proceeds as follows: (1) A box-wise conformal predictor calibrates the detected bounding boxes; (2) The recognized texts and their associated softmax probabilities are passed to a text conformal predictor, which estimates per-character uncertainty scores (shown in yellow in the images). After text matching via a pretrained tracker, an uncertainty-aware algorithm updates the text across tracked instances. Green boxes represent components of the pretrained VTS; blue boxes indicate the different modules of CVTS. Yellow boxes refer to two quantities computed offline during conformal calibration phases.

often struggle to optimize each sub-task effectively. Alternatively, tracking-by-detection methods leverage high-performing ITS models and focus on developing dedicated tracking mechanisms. A notable example is GoMatching [10], which employs the state-of-the-art DeepSolo model [29] for ITS and introduces a robust tracking framework via long- and short-term associations (LST-Matcher). GoMatching significantly outperforms end-to-end methods on four public benchmarks.

Despite these advances, a major challenge persists: inconsistent text recognition across frames. Even state-of-the-art ITS may correctly identify text in some frames while failing in others due to partial occlusion, motion blur, or lighting variations. Such inconsistencies degrade the overall performance of VTS systems. We address this issue by proposing a novel uncertainty-aware framework that leverages both tracking information and ITS model confidence to correct misrecognized text. Our key insight is that recognition reliability is closely correlated with prediction uncertainty. To exploit this, we integrate conformal prediction (CP) [23], a statistically rigorous framework for uncertainty estimation, into the VTS pipeline to estimate per-character confidence scores. These uncertainty estimates enable us to identify the most reliable predictions and use them to correct less confident ones, thereby enhancing the robustness of text recognition in videos.

Specifically, we extend the pretrained GoMatching model by incorporating a text conformal predictor. Using the output of its underlying ITS model (DeepSolo), we compute

character-level uncertainty scores. Based on the tracking associations provided by the LST-Matcher, we then apply a principled algorithm that updates each character in a tracked instance with its most confident prediction. To handle cases where text instances have missing or misaligned characters across frames, we align tracked predictions to a reference using the robust Needleman–Wunsch algorithm [21]. This uncertainty-guided update mechanism ensures consistent and accurate recognition, even under challenging conditions such as partial occlusion, motion blur, or poor lighting.

Furthermore, we hypothesize that text tracking quality is influenced by the precision of detected bounding boxes. To this end, we introduce a second CP module [1] for bounding box refinement. This module applies corrections to the predicted bounding boxes, providing statistical guarantees with respect to a predefined success probability. The result is improved localization accuracy and, consequently, more stable text tracking across frames.

Finally, we enhance the LST-Matcher module by introducing a pre-matching step that performs text comparison based on a thresholded Levenshtein distance [15], thereby improving the reliability of text matching across frames.

Figure 1 illustrates the main components of the proposed **C**onformal **V**ideo **T**ext **S**potter (**CVTS**) framework, which incorporates two conformal predictors: one for text recognition and one for bounding box refinement. The approach introduces negligible computational overhead and can be easily integrated into any pretrained VTS model, yielding substantial performance gains. Our main contributions are as follows.

- We propose a novel method for selecting the most confident text recognition from tracked text instances using conformal prediction (CP) for uncertainty estimation. To the best of our knowledge, this is the first application of CP to text recognition tasks.
- We introduce a second CP-based module to refine predicted text bounding boxes, with statistical guarantees of correctness with respect to a predefined coverage level. Empirically, this refinement significantly improves text tracking performance.
- We further enhance the pretrained LST-Matcher tracker by incorporating a distance-based pre-matching step, which improves association accuracy during inference.
- We validate the effectiveness of our approach through extensive experiments on four public benchmarks, consistently outperforming state-of-the-art methods, particularly the GoMatching baseline.

2 Related Work

2.1 Conformal Prediction in Computer Vision

Conformal Prediction (CP) [4, 22] is a powerful technique that provides reliable uncertainty estimates for predictions, with guarantees under the only assumption that data are i.i.d. First proposed by Gammerman *et al.* [9], CP was later extended to *inductive* (or *split*) CP [23], improving computational efficiency by using a pretrained model and a held-out calibration set, making it suitable for real-time applications.

CP has found applications in computer vision, particularly in two main areas: (i) classification tasks and (ii) structured prediction tasks, such as object detection. For classification, CP produces *prediction sets* rather than single-label predictions. These prediction sets can

be used to provide a formal confidence measure, improving uncertainty quantification, especially in the presence of class imbalance and distribution shift [6, 11, 14, 19, 23]. For instance, in the case of semantic segmentation, CP has been extended to derive pixel-wise confidence sets using post-hoc calibration methods [20], particularly to address distribution shifts such as label imbalance and class prevalence drift.

CP is also gaining traction in structured prediction tasks like object detection [7], where it provides spatial guarantees by calibrating bounding boxes, making it particularly valuable in safety-critical applications such as railway signaling [1].

Moreover, CP has been successfully applied in medical imaging, where model confidence is vital for critical decisions such as cancer detection. CP can improve the trustworthiness of diagnostic models, ensuring that predictions are not only accurate but also accompanied by uncertainty quantification [18, 19, 24]. Despite its success in these domains, CP has not yet been applied to textual tasks. Our work is the first to introduce it to VTS by means of both text and bounding-box conformalizations.

2.2 Video Text Spotting

VTS methods typically fall into two categories: end-to-end systems and tracking-by-detection frameworks. End-to-end models like TransDETR [27] treat detection, recognition, and tracking as a unified sequence modeling task. GloTSFormer [25] improves tracking accuracy by modeling global associations using Transformers and Wasserstein distance, while LOGO [16] incorporates language priors and glyph perception to enhance recognition in noisy conditions. Despite their elegance, these unified models often struggle to balance performance across sub-tasks. Tracking-by-detection methods offer a more modular approach by leveraging powerful ITS. GoMatching [10], for instance, uses DeepSolo [29] for frame-wise spotting and introduces the LST-Matcher to associate text across time using long- and short-term cues. This approach achieves SOTA performance, but remains sensitive to recognition inconsistencies caused by motion blur, occlusion, or lighting variation. Other recent efforts have explored generalization and temporal coherence. VimTS [17] unifies ITS and VTS to improve cross-domain performance, aided by a large synthetic dataset. VLSpotter [31] enhances recognition stability with super-resolution and semantic reasoning modules. While these methods address tracking robustness, they rarely quantify prediction confidence or exploit it for correction. In this work, we introduce an uncertainty-aware framework that leverages CP to refine recognition and bounding boxes produced by GoMatching [10], improving consistency and robustness across challenging video conditions.

3 Preliminaries

We begin by reviewing the necessary background on CP, focusing on two specific settings relevant to our work: classification and bounding-box detection. For a comprehensive overview of the general framework, we refer the reader to [3].

3.1 Conformal Prediction for Classification

Conformal prediction [3, 22], a.k.a conformal inference, is a simple, model-agnostic framework for constructing prediction sets with rigorous statistical guarantees. Let $\hat{f}_c$ denote a trained classifier that outputs softmax scores, i.e., $\hat{f}_c(x) \in [0,1]^K$ for K classes. To construct

prediction sets, we use a separate calibration dataset consisting of n i.i.d. image-label pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, disjoint from the training data. Given a test input X_{test}, the objective is to generate a prediction set $C(X_{\text{test}}) \subset \{1, \ldots, K\}$ such that

$$1 - \alpha \leq \mathbb{P}(Y_{\text{test}} \in C(X_{\text{test}})) \leq 1 - \alpha + \frac{1}{n+1}, \tag{1}$$

where Y_{test} is the true label and $\alpha \in [0,1]$ is a user-specified erro rate. This inequality guarantees *marginal coverage* [3], ensuring that the prediction set contains the correct label with high probability over the randomness of the calibration and test samples.

The prediction set $C(X_{\text{test}})$ is constructed from $\hat{f}_c$ and the calibration data via the following procedure. (1) Compute the nonconformity score for each calibration point as $s_i = 1 - \hat{f}_c(X_i)_{Y_i}$, where $\hat{f}_c(X_i)_{Y_i}$ is the softmax score assigned to the true class. (2) Estimate the empirical quantile $\hat{q}_c$ as the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of the nonconformity scores, where $\lceil \cdot \rceil$ is the ceiling function, yielding a conservative estimate of the $1-\alpha$ quantile. (3) Define the prediction set for a test point X_{test} as

$$C(X_{\text{test}}) = \left\{ y \in \{1, \ldots, K\} : \hat{f}_c(X_{\text{test}})_y \geq 1 - \hat{q}_c \right\}, \tag{2}$$

which includes all classes whose softmax scores exceed the threshold $1 - \hat{q}_c$.

A key feature of CP is its *adaptiveness*: prediction sets expand when the model is uncertain or the input is intrinsically difficult. As a result, the size of the prediction set naturally reflects model uncertainty—larger sets indicate greater uncertainty. We exploit this property by defining the uncertainty score as $\mathrm{U}(X_{\text{test}}) = |C(X_{\text{test}})|$. However, this adaptiveness can lead to impractically large prediction sets in regions of high uncertainty. To address this, we employ the Regularized Adaptive Prediction Sets (RAPS) method [2], which introduces a regularization term to encourage compact prediction sets while preserving the coverage guarantee from Eq. (1). We refer the reader to Section 2.2 of [2] for a detailed exposition.

3.2 Conformal Prediction for Bounding-Box Detection

We now consider the setting of text detection, where a CP variant modifies the output of a pretrained $\hat{f}_d$ detector by expanding the predicted box margins to ensure coverage guarantees. In this work, we use the method of De Grancey *et al.* [7] which was applied for object detection.

Let $B_k = \{x_{\min}, y_{\min}, x_{\max}, y_{\max}\}$ and $\hat{B}_k = \{\hat{x}_{\min}, \hat{y}_{\min}, \hat{x}_{\max}, \hat{y}_{\max}\}$ denote the coordinates of the ground-truth and predicted bounding boxes, respectively. To compute nonconformity scores, [7] proposes pairing predictions with ground-truth boxes using the Hungarian algorithm and a predefined IoU threshold. Once pairs are established, the procedure is: (1) compute the nonconformity scores as $R = (\hat{x}_{\min} - x_{\min},\ \hat{y}_{\min} - y_{\min},\ \hat{x}_{\max} - x_{\max},\ \hat{y}_{\max} - y_{\max})$; (2) for each coordinate $c \in \{x_{\min}, y_{\min}, x_{\max}, y_{\max}\}$, compute the conformal quantile using the Bonferroni correction [5], where $q^c_{1-\frac{\alpha}{4}}$ is defined as the $\left\lceil (n_{\text{box}}+1)(1-\frac{\alpha}{4}) \right\rceil$-th element of the sorted R^c; and (3) at inference, define the conformalized bounding box for a test input as

$$\hat{C}_\alpha(X_{\text{test}}) = \{\hat{x}_{\min} - q^{x_{\min}},\ \hat{y}_{\min} - q^{y_{\min}},\ \hat{x}_{\max} + q^{x_{\max}},\ \hat{y}_{\max} + q^{y_{\max}}\}. \tag{3}$$

This procedure guarantees that, with high probability, the conformalized bounding box contains the true object, while accounting for uncertainty in the model's localization accuracy. We adopt this approach to conformalize text bounding boxes in our application.

4 Proposed Approach

As illustrated in Figure 1, we propose an uncertainty-aware video text spotting (VTS) framework, **CVTS**, designed to enhance the performance of any pretrained VTS system through a simple yet effective pipeline. In this work, we demonstrate its effectiveness using the state-of-the-art GoMatching framework [10]. CVTS builds upon a pretrained model by integrating frame-level detection via an image-based text spotter, conformal uncertainty quantification through two calibrated predictors, and temporal refinement to improve text tracking and recognition across video frames. Our method operates in two main stages: (1) a *conformal calibration stage*, where quantiles q_c and q_b are computed as described in Section 3, and (2) an *inference stage*, where these quantiles are used to perform uncertainty-aware post-processing of text and bounding box predictions.

Given a sequence of N video frames $\{f_1, f_2, \ldots, f_N\}$, a robust pretrained image-based text spotter (ITS) model $\mathcal{M}$—specifically DeepSolo [29], as used in GoMatching [10]—is applied independently to each frame. For each frame f_n, the model outputs a set of detected texts T_n, their corresponding per-character softmax probabilities P_n, and bounding boxes B_n.

To estimate uncertainty, we employ two conformal prediction (CP) modules using the precomputed quantiles. The box-level CP module utilizes q_b to adjust the bounding boxes $\{B_1, \ldots, B_N\}$, resulting in more reliable predictions $\{\tilde{B}_1, \ldots, \tilde{B}_N\}$. Simultaneously, the character-level CP module applies q_c to the softmax scores $\{P_1, \ldots, P_N\}$ to compute uncertainty estimates over the detected texts $\{T_1, \ldots, T_N\}$.

Algorithm 1: Conformal Video Text Spotting (CVTS)

Input: Video frames $\{f_1, f_2, \ldots, f_N\}$, pretrained ITS $\mathcal{M}$, q_b, q_c
Output: Updated texts and calibrated bounding boxes
Step 1: Frame-wise Text Inference
for $n = 1$ **to** N **do**
  Apply $\mathcal{M}$ to frame f_n to detect text instances;
  Extract texts T_n, character-level softmax probabilities P_n, and bounding boxes B_n
Step 2: Conformal Calibration
(a) Apply BCP with q_b to $\{B_1, \ldots, B_N\}$ to obtain calibrated boxes $\{\tilde{B}_1, \ldots, \tilde{B}_N\}$, where B_n denotes all bounding boxes detected in frame f_n
(b) Apply TCP with q_c to estimate per-character uncertainty in $\{T_1, \ldots, T_N\}$ based on $\{P_1, \ldots, P_N\}$
Step 3: Cross-frame Text Tracking
Use a pretrained text tracker (LST-Matcher) with our TSF module to associate similar text instances across frames, resulting in a set of tracks $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$
Step 4: Uncertainty-Aware Track Update
foreach *track* $\mathcal{T}_k$ **do**
  Select a reference text $t_{\text{ref}} \in \mathcal{T}_k$ with the lowest average uncertainty;
  Align all other text instances in $\mathcal{T}_k$ to t_{ref} using Needleman–Wunsch algorithm;
  for $i = 1$ **to** $length(t_{ref})$ **do**
    Select the most confident character c_{opt} at position i across aligned instances;
    Update all aligned instances at position i with c_{opt}

Table 1: End-to-end Text Spotting performance on ArTVideo and DSText.

Method	ArTVideo			DSText		
	IDF1	MOTA	MOTP	IDF1	MOTA	MOTP
TransDETR [27]	49.3	2.8	69.7	26.43	-22.63	79.73
LOGO [16]	—	—	—	46.26	12.94	78.78
GoMatching [10]	82.3	75.7	**83.5**	46.27	23.24	80.41
CVTS (Ours)	**83.3**	**76.1**	**83.5**	**47.87**	**25.08**	**80.42**

Table 2: End-to-end Text Spotting performance on BOVText and ICDAR15

Method	BOVText			ICDAR15		
	IDF1	MOTA	MOTP	IDF1	MOTA	MOTP
LOGO [16]	—	—	—	68.07	12.94	73.00
GloTSFormer [25]	—	—	—	74.20	62.50	78.20
GoMatching [10]	62.7	**52.9**	**87.2**	80.11	72.20	78.52
CVTS (Ours)	**65.0**	52.9	**87.2**	**80.92**	**72.54**	**78.54**

Next, we perform temporal refinement by tracking detected text instances across frames using a pretrained LST-Matcher [10], which associates similar detections over time to produce K text tracks $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$. For each track $\mathcal{T}_k$, we identify a reference text $t_{\text{ref}} \in \mathcal{T}_k$ with the lowest average per-character uncertainty. The remaining text instances in the track are then aligned to t_{ref} using the Needleman–Wunsch algorithm [21], which handles variations in text length due to misdetections or occlusions, enabling robust character-level comparison. For each aligned character position i in t_{ref}, the most confident character c_{opt} across aligned instances is selected and propagated to all others at that position, improving consistency and reliability in the final predictions. A detailed overview of the full pipeline is provided in Algorithm 1. For clarity, we refer to the text-level and box-level conformal prediction modules as **TCP** and **BCP**, respectively.

Text Similarity Filter (TSF). To further enhance the robustness of the LST-Matcher during inference, we introduce a third module: the Text Similarity Filter (TSF). LST-Matcher comprises two components—the Short-Term Matcher (ST-Matcher) and the Long-Term Matcher (LT-Matcher)—each implemented as a one-layer transformer encoder-decoder. The matching process follows a two-stage approach: ST-Matcher first links instances in frame t to trajectories from frame $t-1$; LT-Matcher then attempts to associate any remaining unmatched instances using a memory bank of past trajectories. Both stages rely on threshold-based score filtering, as described in [10]. To improve association quality, we insert TSF as a final verification step before any text is linked to an existing trajectory. TSF computes the Levenshtein distance [15] between candidate texts and their potential matches, requiring the normalized similarity to exceed a threshold θ_t. This additional constraint helps prevent erroneous associations caused by spatially or temporally proximate but semantically unrelated texts, resulting in more accurate and consistent tracking.

Table 3: Text tracking performance on BOVText, ICDAR15, ArTVideo and DSText

Method	BOVText		ICDAR15		ArTVideo		DSText	
	IDF1	MOTA	IDF1	MOTA	IDF1	MOTA	IDF1	MOTA
GoMatching	68.3	68.8	72.87	62.41	76.5	68.2	58.51	48.09
CVTS (BCP)	**68.4**	**69.0**	**73.42**	**62.69**	**76.7**	**68.3**	**58.69**	**48.15**

5 Experiments

5.1 Datasets and Metrics

DSText [28] is a recent dataset targeting dense and small text in video-based scene text reading. It contains 50 training and 50 test videos with word-level annotations in quadrilateral format. DSText presents key challenges: (1) densely packed text, (2) many small-scale instances, and (3) diverse real-world scenes such as *Game* and *Sports*.

ArTVideo [10] is a word-level benchmark focused on arbitrarily-shaped text. It includes 20 videos, with around 30% of text instances exhibiting curved shapes. Straight text uses quadrilateral annotations, while curved text is labeled with polygons.

ICDAR15-Video [12] is a standard benchmark with 25 training and 24 test videos. It emphasizes real-world challenges in scenarios like driving, shopping streets, and supermarkets. Word-level annotations are provided via quadrilateral bounding boxes.

BOVText [26] is a large-scale bilingual VTS dataset, featuring English and Chinese text. It consists of over 1,800 videos sourced from YouTube and KuaiShou, annotated at the text-line level using quadrilateral boxes.

Evaluation Metrics. We adopt three standard metrics to evaluate VTS: IDF1, Multiple Object Tracking Accuracy (MOTA), and Multiple Object Tracking Precision (MOTP).

5.2 Comparison with state-of-the-art

As shown in Tables 1 and 2, our method (CVTS) consistently improves upon the main baseline (GoMatching) across all four datasets, and similarly to GoMatching it outperforms all other SOTA approaches. The improvements in IDF1 and MOTA demonstrate that our uncertainty-aware recognition refinement leads to more consistent and accurate text spotting across frames. Specifically, our method achieves a **1%** improvement in IDF1 on ArTVideo, a **1.6%** improvement on DSText, and a **2.3%** improvement on BOVText. Compared to recent end-to-end methods like TransDETR [27], LOGO [16], and GloTSFormer [25], which often face challenges in balancing recognition accuracy and tracking robustness, our lightweight and modular extension to GoMatching delivers substantial performance gains with negligible computational overhead (see Section 5.3.1 for details).

Table 3 presents text tracking performance. This experiment aims to isolate and evaluate the contribution of our BCP module to text tracking. To this end, we compare GoMatching with a BCP-only variant of our CVTS framework. The results consistently show that applying BCP improves both IDF1 and MOTA scores across all datasets. These gains highlight the positive impact of conformalizing bounding boxes. By reducing the effect of noisy or imprecise detections, BCP leads to more stable and accurate text association over time.

Table 4: Ablation study on ArTVideo and DSText datasets.

Modules	ArTVideo				DSText		
	IDF1	MOTA	MOTP	FPS	IDF1	MOTA	MOTP
CVTS (overall)	**83.3**	**76.1**	83.5	0.63	**47.93**	**25.19**	80.39
CVTS (TSF)	83.2	75.9	83.5	0.65	46.95	24.55	**80.52**
CVTS (BCP)	82.3	75.8	83.5	0.66	46.31	22.57	80.43
CVTS (TCP)	82.4	76.0	83.5	0.66	46.32	23.35	80.41
GoMatching [10]	82.3	75.7	83.5	0.66	46.27	23.24	80.41

Table 5: Ablation study on BOVText and ICDAR15 datasets.

Modules	BOVText			ICDAR15		
	IDF1	MOTA	MOTP	IDF1	MOTA	MOTP
CVTS (overall)	**65.0**	52.9	87.2	78.01	70.34	78.73
CVTS (TSF)	64.8	52.9	87.2	77.06	69.67	**78.76**
CVTS (BCP)	63.0	52.9	87.2	**80.91**	**72.53**	78.54
CVTS (TCP)	62.8	52.9	87.2	80.23	72.13	78.54
GoMatching [10]	62.7	52.9	87.2	80.11	72.20	78.52

5.3 Ablation Studies

We evaluate the effectiveness of each module in the proposed CVTS framework (BCP, TCP and TSF) and compare them against the GoMatching baseline. Results presented in Tables 4 and 5 consistently demonstrate the value of each component. Starting from the baseline, we observe that introducing either TCP or BCP leads to steady performance gains across all datasets. On ArTVideo, DSText, and BOVText, each module individually improves both IDF1 and MOTA, and their combination yields the highest scores, highlighting their complementary benefits. The complete CVTS pipeline (TCP+BCP+TSF) achieves the best overall performance on these datasets. The only exception is ICDAR15, where incorporating TSF results in a slight performance drop compared to using TCP or BCP alone. This can be attributed to the nature of ICDAR15: it contains many short, low-resolution text instances such as license plates and alphanumeric strings that are especially susceptible to character recognition errors. TSF, which enforces strict text similarity constraints, may incorrectly reject legitimate matches due to minor recognition inconsistencies. Despite this, the results affirm that both TCP and BCP provide consistent, dataset-agnostic improvements, while TSF offers strong gains in contexts with longer or more distinctive text content. Overall, CVTS proves to be a flexible and effective enhancement to tracking-based VTS systems.

5.3.1 Runtime Performance Analysis

We evaluated the computational overhead of TCP, BCP, and TSF features on an M1 MacBook Pro with Apple Silicon GPU (MPS). The ablation study on the ArTVideo dataset (884 frames, 20 videos), reported in Table 4 (FPS column), shows minimal performance impact: baseline achieves 0.66 FPS, while adding individual features (BCP-only: 0.66 FPS, TSF-only: 0.65 FPS, TCP-only: 0.66 FPS) maintain nearly identical throughput. Even with all features combined, the system achieves 0.63 FPS, representing only a 4.5% decrease from

baseline. The lightweight computational overhead demonstrates that TCP and BCP corrections, can be cheaply integrated into any video text spotting framework while significantly improving tracking accuracy and text recognition performance.

6 Conclusion

In this work, we introduced a novel framework to enhance Video Text Spotting (VTS) by tackling two key challenges: inconsistent text recognition across frames and inaccurate bounding box predictions. Leveraging the statistically grounded conformal prediction method, we estimate uncertainty scores in pretrained models to improve recognition confidence and accuracy over time. Additionally, we apply a conformal correction mechanism for bounding boxes, consistently enhancing tracking robustness. Our approach is lightweight, modular, and easily integrates into existing VTS pipelines. Extensive experiments on four public datasets validate the effectiveness of each component and show that our method consistently outperforms state-of-the-art techniques. These results highlight the potential of incorporating uncertainty estimation into VTS and open promising directions for future research, particularly in more complex or real-world scenarios.

References

[1] Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In *Conformal and Probabilistic Prediction with Applications*, pages 36–55. PMLR, 2023.

[2] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[3] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[4] Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[5] J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. *Bmj*, 310(6973):170, 1995.

[6] A. Chou et al. Uncertainty quantification with conformal prediction in medical image analysis. *Medical Image Analysis*, 58:101673, 2020.

[7] Florence De Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and David Vigouroux. Object detection with probabilistic guarantees: A conformal prediction approach. In *International Conference on Computer Safety, Reliability, and Security*, pages 316–329. Springer, 2022.

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021.

[9] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

[10] Haibin He, Maoyuan Ye, Jing Zhang, Juhua Liu, Bo Du, and Dacheng Tao. Gomatching: A simple baseline for video text spotting via long and short term matching. *Advances in Neural Information Processing Systems*, 37:25663–25686, 2024.

[11] M. Jones et al. Crop-weed identification with conformal prediction. *Agricultural Applications of Machine Learning*, 12(2):134–145, 2021.

[12] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

[13] Taeho Kil, Seonghyeon Kim, Sukmin Seo, Yoonsik Kim, and Daehee Kim. Towards unified scene text spotting based on sequence generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15223–15232, 2023.

[14] K. Lee et al. Ferrous material detection using conformal prediction. *Materials Science and Engineering*, 124:80–90, 2023.

[15] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. Originally published in Russian in Doklady Akademii Nauk SSSR, 1965.

[16] Hongen Liu, Xu Huang, Yan Sun, Kai Liu, Xiang Zhou, Yu Qiao, et al. Logo: Video text spotting with language collaboration and glyph perception model. *arXiv preprint arXiv:2405.19194*, 2024.

[17] Yuliang Liu, Wenwen Huang, Chuhui Yang, Lianwen Jin, and Yaqiang Zhang. Vimts: A unified video and image text spotter for enhancing the cross-domain generalization. *arXiv preprint arXiv:2404.19652*, 2024.

[18] W. Lu et al. Breast density assessment using conformal prediction. *IEEE Transactions on Medical Imaging*, 40(3):789–799, 2021.

[19] W. Lu et al. Confident skin lesion classification using conformal prediction. *IEEE Transactions on Medical Imaging*, 41(5):1202–1211, 2022.

[20] Luca Mossina, Joseba Dalmau, and Léo Andéol. Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3574–3584, June 2024.

[21] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for homology in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48 (2):443–453, 1970. doi: 10.1016/S0022-2836(70)800327-6. URL https://www.sciencedirect.com/science/article/pii/S0022283670800327.

[22] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

[23] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

[24] P. Taylor et al. Are foundation models good for medical imaging? a conformal prediction approach. *Nature Machine Intelligence*, 5(1):1–9, 2022.

[25] Han Wang, Hui Wang, Shiliang Wang, Wenhao Yang, Shuai Liu, and Wenyu Liu. Glotsformer: Global video text spotting transformer. *arXiv preprint arXiv:2401.03694*, 2024.

[26] Weijia Wu, Debing Zhang, Yuanqiang Cai, Sibo Wang, Jiahong Li, Zhuang Li, Yejun Tang, and Hong Zhou. A bilingual, openworld video text dataset and end-to-end video text spotter with transformer. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[27] Weijia Wu, Yuanqiang Cai, Chunhua Shen, Debing Zhang, Ying Fu, Hong Zhou, and Ping Luo. End-to-end video text spotting with transformer. *International Journal of Computer Vision*, 132(9):4019–4035, 2024.

[28] Weijia Wu, Yiming Zhang, Yefei He, Luoming Zhang, Zhenyu Lou, Hong Zhou, and Xiang Bai. Dstext v2: A comprehensive video text spotting dataset for dense and small text. *Pattern Recognition*, 149:110177, 2024.

[29] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. Deepsolo: Let transformer decoder with explicit points solo for text spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19348–19357, 2023.

[30] Chongsheng Zhang, Yuefeng Tao, Kai Du, Weiping Ding, Bin Wang, Ji Liu, and Wei Wang. Character-level street view text spotting based on deep multisegmentation network for smarter autonomous driving. *IEEE Transactions on Artificial Intelligence*, 3 (2):297–308, 2021.

[31] Xinyan Zu, Zhaoyang Zhang, Yujun Xu, Jianfeng Wang, Jinhui Tang, and Jiebo Zhou. Vlspotter: Towards accurate video text spotting with text-wise semantic reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1904–1912, 2023.