This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-031-70546-5_19. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

Counting the Corner Cases: Revisiting Robust Reading Challenge Data Sets, Evaluation Protocols, and Metrics

Jerod Weinman[1][0000−0002−2247−8174], Amelia Gómez Grabowska[2], and Dimosthenis Karatzas[2,3][0000−0001−8762−4454]

[1] Grinnell College, Grinnell, Iowa, USA
jerod@acm.org
[2] Universitat Autònoma de Barcelona
[3] Computer Vision Center, Barcelona, Spain

Abstract. For two decades, robust reading challenges (RRCs) have driven and measured progress of text recognition systems in new and difficult domains. Such standardized benchmarks benefit the field by allowing participants and observers to systematically track steady performance improvements as interest in the problem continues to grow. To better understand their impacts and create opportunities for further improvements, this work empirically analyzes three important aspects of several challenges from the last decade: data sets, evaluation protocols, and competition metrics. First, we explore implications of certain annotation protocols. Second, we identify limitations in existing evaluation protocols that cause even the ground truth annotations to receive less than perfect scores. To remedy this, we propose evaluation protocol updates that boost both recall and precision. Accounting for these corner cases causes almost no changes to current rankings; however, such cases may become more prominent and important to consider as challenges focus on increasingly complex reading tasks. Finally, inspired by the recent HierText challenge's use of Panoptic Quality (PQ), we explore the impact of this simple, parameter-free tightness-aware metric on six prior challenges, and we propose a new variant—Panoptic Character Quality (PCQ)—for simultaneously measuring character-level accuracy and word detection tightness. We find PQ-based metrics have a greater re-ranking impact on detection-only tasks, but predict end-to-end rankings slightly better than F-score. In sum, our empirical analysis and associated code should allow future challenge designers to make better-informed choices.

Keywords: Scene text reading · Evaluation protocols · Optimization.

1 Introduction

Since 2003, ICDAR has regularly hosted robust reading competitions "to establish some common benchmark datasets, and gain a clear understanding of the current state of the art" [25, p. 1]. The early competitions focused primarily on scene text detection and cropped word recognition. Challenges standardizing end-to-end reading tasks appeared in 2015 and continue to attract interest [13].

Appears in *Proceedings of the 18th International Conference on Document Analysis and Recognition (ICDAR 2024)*. Elisa Barney Smith, Marcus Liwicki, and Liangrui Peng (Eds.). Springer Nature LNCS. 2024.

Other competitions with increasingly specific or more challenging tasks have followed [11,29,32,12,36,34,38,4,10,24,20]. The online portal hosting many of these challenges has evaluated over 90,000 submissions since 2011, indicating the broad impact of this framework over the years [3]. Whereas many performance challenges remain to be addressed in the coming decade, we believe it is a good time to take stock of the frameworks, so that future competitions can continue to provide meaningful insights about progress in the field of robust reading.

This work carefully examines several aspects of the challenge pipeline, from the data sets and annotations themselves to the evaluation protocols that lay the groundwork for assessment. In addition, we explore how a recent family of metrics may broaden understanding of system performance. Although the overall impacts of the findings might be thought of as relatively small for prior challenges, we conclude that a variety of special cases—amounting to hundreds in most data sets—remain important to consider and could increase in importance as new challenges are created to drive progress.

The remainder of the work is organized as follows. In Section 2, we elaborate on the history, tasks, evaluations, and connections among the several families of robust reading challenges. Section 3 describes the data used in the study and the overarching methodology of the analyses to follow, while Section 4 begins to examine the competition data sets in ways that motivate several questions about the evaluation protocol addressed in Section 5. Section 6 completes the pipeline by exploring a family of metrics that unify recognition and segmentation quality at both the word and character level.

2 Background and Related Work

Organized benchmark analyses of well-zoned page reading OCR systems predate robust reading challenges [30,31]. However, due to the relative sparsity of words in scenes (compared to pages) and the difficulty of defining analogous "zones," it is unsurprising that evaluations for robust reading systems share many traits with object detection, segmentation, and recognition benchmarks.

Due to the nature of the task, we can understand a distinction between two stages of the performance measurement. First, the predicted detections must somehow be put in correspondence with the ground truth. Depending on the assumptions or goals of the system, these correspondences may be one-to-one, many-to-one, and/or one-to-many. We call this first stage the *evaluation protocol*. With the correspondences in place, their quality may subsequently be measured by some *metric(s)* such accuracy, precision, recall, etc.

2.1 Evaluation Protocols

Lucas *et al.* [25] defined an evaluation protocol for end-to-end systems, as well as the initial detection stage. The RRC protocols have evolved somewhat over time, but three key elements have ossified: greedy correspondence search, cascaded (sequential) assessment of detection and recognition, and the handling of "don't

care" regions to be ignored in the evaluation. In the following, we describe how each element has mutually evolved in both object detection and the RRCs.

Correspondence Lucas *et al.* [25] proposed "soft" versions of precision and recall, which measured the average best match score between a detection and the set of ground truth rectangles (for soft precision) and between a ground truth rectangle and the set of detections (for soft recall). The match score was defined as "the area of intersection divided by the area of the minimum bounding box containing both rectangles" [25, p. 3], similar to the intersection-over-union (IoU) criterion (or Jaccard index) used in most systems today. This procedure somewhat blurs the line between correspondence and metric. Because there were no entries in the end-to-end contest (and 2005 was detection-only [26]), the stated evaluation protocol—"the rectangles must have a match score...of greater than 0.5, and the word text must match exactly" (p. 4)—remains somewhat ambiguous because no explicit process for matching rectangles was stated.

Alternatively, by 2005 the Pascal Visual Object Challenge (VOC) [8] explicitly matched object rectangles first, then determined how or whether the match was to be assessed (*i.e.*, depending on whether it was a "difficult" ground truth example or had already been matched). Only afterward was the set of corresponding rectangles judged by some metric. Because the average precision measure was used, rectangles were one-to-one paired by a greedy algorithm in which the most confidently scored detections are matched to ground truth rectangles first [9], a strategy shared by the COCO object detection challenge [21,6].

Later RRCs adopted similar matching strategies; i.e., *fully greedy* accepts satisfying matches as they are found for each image, as shown below.

Listing 1. Fully greedy annotation/detection correspondence matching.

```
for g in G: # Unordered ground truth items
  for d in D: # Unordered detections
    if unmatched(g) and unmatched(d) and # matchable?
          not ignore(g) and not ignore(d):
        if IoU(g,d) > threshold: # geometric criterion
          match(g,d)
```

In the *ground truth greedy* variant, the best detection is chosen for each ground truth item, with the inner loop modified to ensure the corresponding detection for each $g \in G$ is $\hat{d}(g) \triangleq \arg\max_{d \in D} \mathrm{IoU}(g, d)$ if it satisfies the minimum IoU threshold. A *symmetric greedy* variant adds the further constraint that all matches $\left(g, \hat{d}(g)\right)$ also have $g = \arg\max_{g' \in G} \mathrm{IoU}(g', \hat{d}(g))$.

This fully greedy strategy was sensibly sufficient for early RRC detection tasks where metrics were based purely on counts, rather than quality, and few valid regions overlapped. Satisfying correspondences were essentially optimal. Subsequent strategies choosing maximizing matches suggest a shift toward an implicit optimization problem. In this work, we fully formalize this optimization and measure the detrimental impact of such greedy approximations.

Sequential Assessment In end-to-end challenges, the text recognition evaluation has become a second stage in a cascaded, sequential process, with geometric verification established first, as for detection. That is, in challenges measuring word-level accuracy, correspondences established by geometric criteria (*i.e.*, IoU > 0.5) are irrevocably fixed, as in Listing 1; word accuracy is only subsequently assessed inside the last `if` conditional *after* the `match`. If valid annotations happen to overlap (see Figure 1), meeting the geometric criterion, a mismatch among the proper corresponding detections and the underlying ground truth may result in an inaccurate assessment. Although the geometries align, the texts may not, causing severe scoring penalties (observed by Baek *et al.* [1]). This work measures the impact of the cascaded evaluation and alternatives that are both consistent with the original protocol laid out by Simon *et al.* as well as commensurate with the formal maximization framework we introduce in Section 5.

Handling "Don't Cares" Often there exist identifiable image regions containing text that is to be omitted from the evaluation, *i.e.*, because it is illegible, in a non-target language, etc. In the VOC [8], a detection was counted as either a true positive or false positive once the IoU overlap criterion was met, but only if the corresponding ground truth rectangle was not considered difficult (that is, it was *not* a "don't care") [9]. For COCO, detections were allowed to correspond to "don't care" annotations (and subsequently discounted from the metrics), but only after attempts were made to find a match with a *not* "don't care" annotation failed [6]. In the more recent Panoptic Segmentation (PS) task [16], which unifies semantic and instance segmentation, predictions are not matched to "don't care" regions (unlabeled "void" regions or difficult to separately label groups of instances), but any unmatched regions that sufficiently overlap "don't care" regions do not count as false positives. Importantly, PS is distinct in that regions are mutually disjoint by definition.

As can be seen in Listing 1, the RRCs have taken a different approach by identifying ignorable detections beforehand, similarly based on their overlap with "don't care" regions. Any such detections will not be matched, *even if* they are viable match candidates with valid annotations. Thus, as in PS, although such detections will not be counted as false positives, unlike VOC or COCO neither can they become true positives. We likewise investigate the impact of this prefiltering detection protocol in comparison to alternatives like COCO's or PS's.

Table 1 summarizes the RRCs and their matching strategies. These challenges are firmly rooted in the *word*: 1) Ground truth annotations are made at the "word" level; in scripts without space-separated words, such as Chinese, the layout geometry (*i.e.*, "lines") largely determines the annotation granularity [28,5,34]. 2) Evaluation protocols assume/enforce a one-to-one matching between annotation regions and detections. 3) Metrics largely assess how well methods can recover (detect) and recognize these words. Recent work has examined the limitations of RRC protocols, particularly with respect to the granularity of the one-to-one requirement [35,2,7,18,19,1]. **Our goal is not to counter these important developments in evaluations, but to clarify that—*within* the**

space of these assumptions—applying greater nuance to the approach increases evaluation fairness and completeness. Some ideas presented here may also transfer to other evaluation frameworks.

2.2 Metrics

As suggested earlier, the prominent metrics for evaluating text detection and end-to-end reading tasks have been derived from early use in information retrieval and later use in object detection. Once the correspondence is completed, matched detections can be considered to belong to the set of true positives, while unmatched detections (ground truths, resp.) belong to false positives (false negatives, resp.). Precision and recall rates capture these proportions, and their harmonic mean, the F-score, combines them.

$$P \triangleq \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}|} \qquad R \triangleq \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FN}|} \qquad F \triangleq \frac{2PR}{P + R}$$

For detection tasks, the set of true positives is a subset satisfying geometric overlap constraints,

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

where τ is the minimum IoU match threshold, typically 0.5. Recognition tasks add a textual boolean predicate $M\,(g, d)$ to constrain to acceptable string matches,

$$\mathrm{TP}_{\mathrm{Rec}} \subseteq \{(g, d) \in G \times D \mid \mathrm{IoU}\,(g, d) > \tau \wedge M\,(g, d)\}\,. \tag{2}$$

In this context, the F-score measures word-level performance. The match set TP is found for each image; however, in calculating the final metrics, the values |TP|, |FP|, and |FN| are typically totals accumulated for an entire set of benchmark images.

Edit distance as a measure of total character-level error has a long history in OCR [31], and the normalized Levenshtein edit distance (ED) between two (short) strings r and s,

$$\mathrm{NED}\,(r, s) \triangleq \mathrm{ED}\,(r, s) \,/\, \max\left(|r|\,, |s|\right), \tag{3}$$

was introduced to cropped word tasks [12,38], precursors of end-to-end challenges. For end-to-end detection and recognition tasks [34,5], the average complementary normalized edit distance (or "1 − NED") is given by

$$\mathrm{CNED} \triangleq \frac{|\mathrm{TP}| - \sum_{(g,d)\in \mathrm{TP}} \mathrm{NED}\,(g, d)}{|\mathrm{TP}| + |\mathrm{FN}| + |\mathrm{FP}|}. \tag{4}$$

The NED is taken to be 1 for a false negative (no corresponding detection) or a false positive (no corresponding ground truth), so that these errors are penalized.

Kirillov *et al.* [16] introduced the Panoptic Quality (PQ) measure, which addressed several concerns raised in Wolf and Jolion [35] by clearly disentangling quality and quantity in an intuitive and parameter-free way. The value

PQ $\in [0, 1]$ combines a measure of segmentation quality T (for tightnesss)—the average IoU among true positive regions—with the well-established F-score as the measure of detection quality:

$$\mathrm{PQ} \triangleq F \times T \qquad = F \times \frac{1}{|\mathrm{TP}|} \sum_{(g,d)\in \mathrm{TP}} \mathrm{IoU}\,(g, d)\,. \tag{5}$$

Long *et al.* adopted PQ for the HierText challenge [23,24]. In later sections, we retrospectively apply the PQ metric to other RRC submissions, which have not traditionally been scored for segmentation quality.

The PQ metric is similar in spirit to the contemporaneous tightness-aware metrics of Liu *et al.* [22], which—like Lucas *et al.* [25]—proposes an average form of recall and precision; in this case modulating IoU values of each correspondence with an inverse recall or precision with respect to box pixel areas.

The existing PQ for end-to-end tasks is a word-level metric. To capture the same sort of character-level performance measured by CNED, we introduce another factor to PQ: the average $1 - \mathrm{NED}$ value over true positives (much like tightness is average IoU). We combine this with PQ to define the *Panoptic Character Quality* PCQ $\in [0, 1]$:

$$\mathrm{PCQ} \triangleq F \times T \times C \qquad = \mathrm{PQ} \times \frac{1}{|\mathrm{TP}|} \sum_{(g,d)\in \mathrm{TP}} (1 - \mathrm{NED}\,(g, d))\,. \tag{6}$$

An un-normalized geometric mean like PQ, the PCQ intuitively combines three elements of interest—localization accuracy (tightness) T, detection quality F, and *character*-level recognition accuracy C—into a single measure.

The PCQ metric and the updated evaluation protocol outlined in Section 5 appear in the MapText RRC [20]. This broader study examines their retrospective impacts on prior RRCs. The supplementary material shows the remarkable stability of panoptic measures with respect to the IoU threshold, which can be reduced to $\tau = 0$ for end-to-end tasks, making PQ effectively parameter free.

3 Data and Methodology

The next section offers an empirical analysis of the competition data sets, their evaluation protocols, and metrics. Here we describe the data and basic methodology supporting these experiments.

The seven challenges with twelve tasks total (five detection and seven end-to-end) are all hosted at the RRC site [3] (see Table 1). Evaluation scripts for FST15, IST15, MLT19, and HierText22 are public, but the others (ArT19, LSVT19, and OOV22) remain privately held by the organizers. With the exception of HierText22, all the evaluations are derivatives of the same original protocol scripts for the RRC site. Although HierText22 followed the same protocols outlined in Section 2.1, the implementation is entirely independent.

The train data splits (annotations and images) and test split images are publicly available, but the ground truth test split annotations remain privately

Table 1. Robust reading challenges examined in this study. See text for details.

Challenge		Year	Det	E2E	Matching	Metric
Focused Scene Text	(FST) [13]	2015		✓	Fully greedy	F-score
Incidental Scene Text	(IST) [13]	2015	✓	✓	Fully greedy	F-score
Multi-lingual Scene Text	(MLT) [28]	2019	✓	✓	Fully greedy	F-score
Arbitrary-Shaped Text	(ArT) [5]	2019	✓	✓	GT greedy	CNED
Large-scale Street View Text	(LSVT) [34]	2019	✓	✓	GT greedy	CNED
Out of Vocabulary	(OOV) [10]	2022		✓	Fully greedy	F-score†
Hierarchical Text	(HeirText) [24]	2023	✓	✓	Symmetric greedy	PQ

† Note: OOV uses an average F-score between in- and out-of-vocabulary sets of words; for comparisons, our analysis focuses on the standard F-score over all words.

held by the organizers. Non-public scripts and data have been made available to us strictly for the purpose of supporting this study.

Users who submit results for evaluation to the RRC server may keep their results private or share them publicly, which allows their ranking and results to appear on the site along with a brief description and optional links to supporting papers or code. Nearly 2,000 methods over the twenty-six competitions have been shared; we downloaded the 458 publicly accessible submissions (as of 11 December 2023) for the twelve tasks examined in Sections 5 and 6.

Each competition uses a slightly different format for its ground truth and submission files. Because the overall task is basically shared among these competitions, we wrote converters that exported the ground truth file for each competition to a common format; we similarly converted each submission for all the tasks to a similar common format.

All the analyses proceed through a single evaluation script that supports the protocol and metric variations described in the next sections. For a given protocol variant, all challenges are thus processed in a unified fashion (with the exception of hooks for task-specific string comparisons).[4]

4 Competition Data Sets

Table 1 summarizes protocols and metrics for the various RRCs. Motivated by the protocols described in Section 2.1, this section investigates properties of the annotations for various scene text data sets, focusing on two primary questions: Q1) *How many valid (not "don't care") ground truth annotations overlap?* If greedy matching simply requires meeting an IoU threshold, the "best" or largest set of correspondences may not be found. Q2) *How many valid ground truth words overlap with "don't care" regions?* If such cases are excluded as correspondence candidates, potentially valid words will be ignored, causing an unrecoverable drop in recall. Figure 1 illustrates meaningful cases for Q1.

[4] Code and data: https://github.com/weinman/rrc-evaluation and DOI:11084/34450.

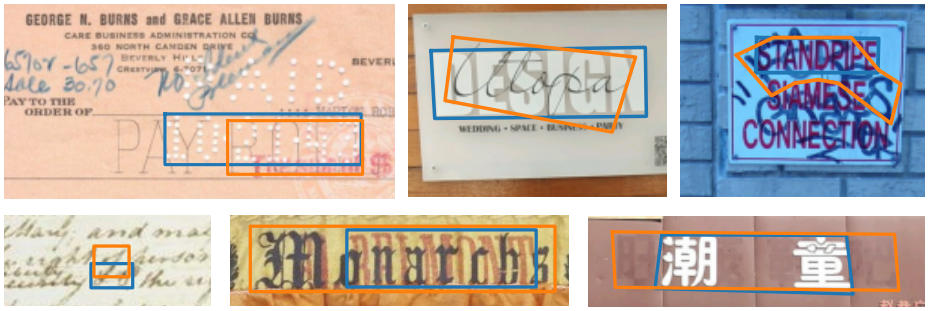

Fig. 1. Overlapping word annotations with IoU > 0.5 in robust reading training data occur for a variety of reasons including design, typography, and incidental.

Table 2. Training data statistics. "Ignores" is the total number of regions marked as a "don't care" for the evaluation. "Ignored Valids" are the valid words that meet the overlap criterion with a "don't care" region. "Matching Valids" are the valid words that can match with a different valid word (their IoU > 0.5).

Train Set	Words	Ignores	Ignored Valids		Matching Valids	
FST15	1,455	698	3	0.40%	2	0.26%
IST15	11,886	7,418	3	0.07%	18	0.40%
COCOText17 [11]	145,862	58,742	1,382	1.59%	4,848	5.56%
MLT17 [29]	86,632	17,814	62	0.09%	74	0.11%
MLT19	111,998	22,562	71	0.08%	78	0.09%
ArT19	62,990	12,899	132	0.26%	108	0.22%
LSVT19	382,606	138,969	262	0.11%	220	0.09%
TextOCR [33]	1,052,354	0	0	0.00%	884	0.08%
IntelOCR [17]	2,353,302	0	0	0.00%	92,258	3.92%
HierText	1,014,142	151,565	582	0.07%	82	0.02%

Table 2 quantitatively addresses these questions. In particular, the "Matching Valids" columns indicate the potential impact of greedy matching (Q1), while the "Ignored Valids" columns indicate the impact of pre-filtering detections to be ignored in evaluation protocols (Q2). While the proportions may yet be small, the raw numbers of such cases have increased as data sets have scaled. Note that the OOV [10] data set is the union of all listed except MLT17, ArT19, and LSVT19. We find that many "matching valids" appear to be due to double annotations, particularly in COCOText (distinct rectangles) and IntelOCR (duplicates).

The next section carefully quantifies the extent of these issues by verifying a stark reality: using the ground truth annotations as a submission fails to achieve a perfect score using the existing evaluation protocols. Greedy, sequential matching finds a satisfying but incomplete set of correspondences, and pre-filtering detections causes false negatives.

Separately, most RRC annotation protocols allow "don't care" regions to contain multiple words, or even multiple lines of text. In some circumstances it is

difficult or impossible to annotate "don't care" regions at the word level, which then necessitates a many-to-one matching scheme when handling predictions that are to be ignored in the evaluation. Among the public submissions to these challenges, a significant number (5–20%) of ground truth "don't care" elements have multiple predictions overlapping them. (See the supplementary material for quantitative and qualitative analysis of the extent and impact of many-to-one matching.) In particular, forcing a one-to-one matching generally tends to negatively impact the precision of only the systems with highest recall. Ordinarily, the many extra unmatched predictions overlapping "don't cares" would not count as false positives that reduce precision. However, in competition contexts where it *is* possible to give word-level "don't care" annotations (*e.g.*, multilingual [28], printed versus handwritten [11], or words truncated by cropping [20]), requiring one-to-one matching with word-level "don't cares" may be preferable. Such protocols would avoid skewing the precision metric by failing to penalize multiple predictions that overlap such regions.

5 Evaluation Protocols

In this section we propose remedies for the issues and concerns with the evaluation protocols detailed above. Specifically, by framing the correspondence stage as an instance of the linear sum assignment problem, we can easily give a complete, optimal metric value for a given set of predictions.

5.1 Greedy versus Optimal Correspondence Matching

As described in Section 2.1, RRC challenges use matching strategies with varying levels of greediness. Section 4 showed that many benchmark data sets have quite a number of annotations with IoU overlaps above the commonly-used match threshold. As a result, the standard evaluation protocol fails to give perfect scores to the ground truth (see Table 3).

With a fixed set of detections D and ground truth annotations G, maximizing the number of true positives will optimize the F score. To give the best possible evaluation, the protocol should therefore find a correspondence that maximizes the number of matches for each image.

Given a score matrix $\mathbf{\Psi} \in \mathbb{R}^{|G|\times|D|}$ with entries ψ_{gd}, bipartite linear sum assignment finds the binary-valued matrix $\mathbf{X} \in \mathbb{Z}_2^{|G|\times|D|}$ with entries x_{gd} maximizing the sum

$$\sum_{(g,d)\in G\times D} \psi_{gd} x_{gd} \tag{7}$$

with (bipartite) 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. This maximization problem has many polynomial time algorithms [27] and runs quickly in modern software packages such as SciPy.

To maximize the matches for detection tasks, we use the match score

$$\psi(g,d) = \begin{cases} +1 & \text{if IoU}(g,d) > \tau \\ -1 & \text{otherwise,} \end{cases} \tag{8}$$

taking $\mathrm{TP}_{\mathrm{Det}} = \{(g,d) \mid x_{gd} = 1\}$. Setting the last case to some negative value, rather than zero, prevents the optimizer from setting $x_{gd} = 1$ without penalty, ensuring the matches of $\mathbf{X}$ satisfy the geometric constraint.

As described in Section 2.1, current evaluation protocols for end-to-end tasks first establish correspondences and then filter these to the final set of true positives by requiring candidates' transcriptions match. Thus, we would take $\mathrm{TP}_{\mathrm{Rec}} = \{(g,d) \mid x_{gd} = 1 \wedge M(g,d)\}$ where M is the string match predicate. Note that in this case $\mathrm{TP}_{\mathrm{Rec}} \subseteq \mathrm{TP}_{\mathrm{Det}}$.

5.2 Sequential versus Joint Matching for End-to-End Tasks

Section 2.1 observed that in the end-to-end task identifying candidate correspondences by geometry alone may lead to sub-optimal performance evaluations. Rather than cascade the string constraint verification, we can effortlessly include it within the optimal framework described above so that all valid geometries are considered. To this end, the match score function is

$$\psi(g,d) = \begin{cases} +1 & \text{if } \mathrm{IoU}(g,d) > \tau \wedge M(g,d) \\ -1 & \text{otherwise.} \end{cases} \tag{9}$$

Thus, it is no longer necessary to include the match constraint in a secondary filtering step and we may directly take $\mathrm{TP}_{\mathrm{Rec}} = \{(g,d) \mid x_{gd} = 1\}$. As mentioned above, geometric consistency alone with Equation (8) may not produce the best correspondences for recognition, so in this case $\mathrm{TP}_{\mathrm{Rec}} \not\subseteq \mathrm{TP}_{\mathrm{Det}}$ in general.

5.3 Pre- versus Post-filtering "Don't Care" Predictions

As shown in Table 2, a small but measurable number of valid words overlap with ignore regions. Eliminating such valid words from match candidacy causes an inflation in false negatives. We can remedy this by following the precedent of object recognition challenges, which only discount any such ignorable predictions *after* the matching stage. Within the joint optimization framework we have

$$\psi(g,d) = \begin{cases} +1 & \text{if } \mathrm{IoU}(g,d) > \tau \wedge V(g) \\ -1 & \text{otherwise,} \end{cases} \tag{10}$$

where V indicates the ground truth annotation is valid (*not* a "don't care.") Any unmatched detections that meet the overlap threshold with "don't care" ground truth regions are subsequently excluded from the false positive tally.

For end-to-end tasks with the F-score metric, we insert the match predicate $M(g,d)$ to the positive case of Equation (10) as we did for Equation (9).

Tasks using the CNED metric (4) do not use the string match constraint. Importantly, we cannot maximize only the true positives, but must also minimize total NED. To accomplish this we take

$$\psi(g,d) = \begin{cases} +1 - \mathrm{NED}(g,d) & \text{if } \mathrm{IoU}(g,d) > \tau \wedge V(g) \\ -1 & \text{otherwise.} \end{cases} \tag{11}$$

5.4 Empirical Analysis

Here we quantitatively examine the impact of the protocols and variants proposed. Table 3 reports the results of using the ground truth annotations as prediction input to the evaluation protocol. When the input is reversed, the

Table 3. Evaluation false negatives using ground truth as prediction submissions.

				ArT		MLT		LSVT		OOV	HierText	
Rev.	Opt.	Joint	Post	Det	E2E	Det	E2E	Det	E2E	E2E	Det	E2E
				306	306	125	135	258	258	226	28	28
✓				306	306	125	141	258	258	448	28	28
✓	✓			306	306	125	137	258	258	230	28	28
✓	✓	✓		-	306	-	135	-	258	226	-	28
✓			✓	0	0	2	14	1	1	273	0	0
✓	✓		✓	0	0	2	5	1	1	7	0	0
✓	✓	✓	✓	-	0	-	2	-	1	2	-	0

false negatives increase for two tasks due to the greedy matching—doubling for OOV; changing to optimal matching nearly restores the prior performance. Post-filtering the ignores, even with a greedy matcher, makes the biggest difference in performance. The OOV false negative count remains stubbornly high without the optimal matcher, which also reduces errors for the MLT E2E task. Finally, addressing all of the issues, in the bottom row of the table, minimizes the number of false negatives. Tellingly, columns for IST and FST (not shown) in Table 3 are all zeros, suggesting that protocol issues were not readily apparent in early competitions and data sets. (We note that the remaining non-zeros in the bottom row are due to ground truth annotations that have an un-matchable area of zero; IoU is undefined.)

For competition entries, Figure 2 shows the changes in false negative counts relative to the original RRC evaluation protocol. The sharp initial increases for some E2E entries highlight the dangers of sequential constraint verification (geometry then text); systems that produce multiple predictions overlapping a ground truth element are more likely to result in a correspondence with incorrect text. Joint constraint processing maintains or reduces false negatives in all cases. As expected, allowing all predictions to be match candidates further reduces false negatives.

There is almost no change in the count of false positives among all challenges and entries (four differ by one or two each). The mean (max, resp.) reduction in false negatives is 65 (201) for detection tasks and 96 (817) for end-to-end. Among all tasks, the increase in F-score is +0.034% (+0.27% max). Fully 10% of the submissions have an F-score difference with an adjacently ranked system less than that average. (See the supplementary material for a more detailed visual analysis of the net positive improvements on precision, recall, and F-score.)

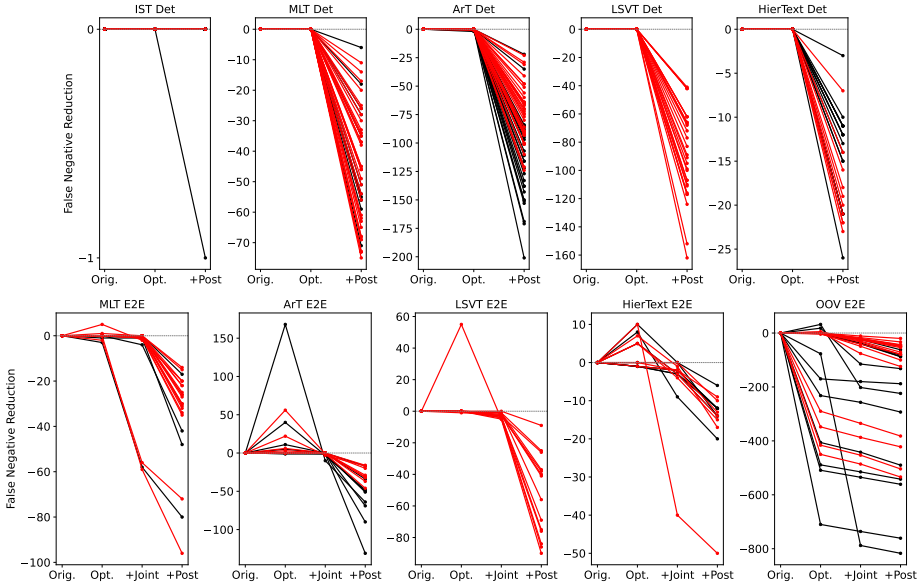

Fig. 2. Changes in raw false negative count from the original protocol for public RRC submissions. Red is competition entries; black are post-competition.

For ArT19 and LSVT19, which rank by the CNED metric, optimal processing with the NED-based match score (11) uniformly improves entries' CNED by an average (max, resp.) of 0.09% (0.32%).

Importantly, we note that the only rank change resulting from these modifications is in OOV positions 10 and 11, though we reiterate the overall F-score analyzed here was not the competition metric.

In summary, it is not enough to simply find satisfying correspondences in general, particularly with a one-to-one matching scheme. Even non-overlapping ground truth annotations may be close enough to allow for multiple satisfying correspondences. Framing the correspondence task as an optimization problem for a specific metric allows for more complete performance assessments.

6 Panoptic Metrics

With a robust evaluation protocol that addresses correspondence matching more completely, we turn our attention to the competition metrics, which can justifiably vary to emphasize different aspects of performance. In contexts where cropped word matches are displayed as search results [15] or erased for privacy [37], it may make sense to reward the "segmentation quality", or the tightness of the match, as done by the Panoptic Quality (PQ) metric (5).

For the evaluation protocol to find correspondences, simply maximizing true positives (as for F) will not optimize PQ, which also considers the IoU-based

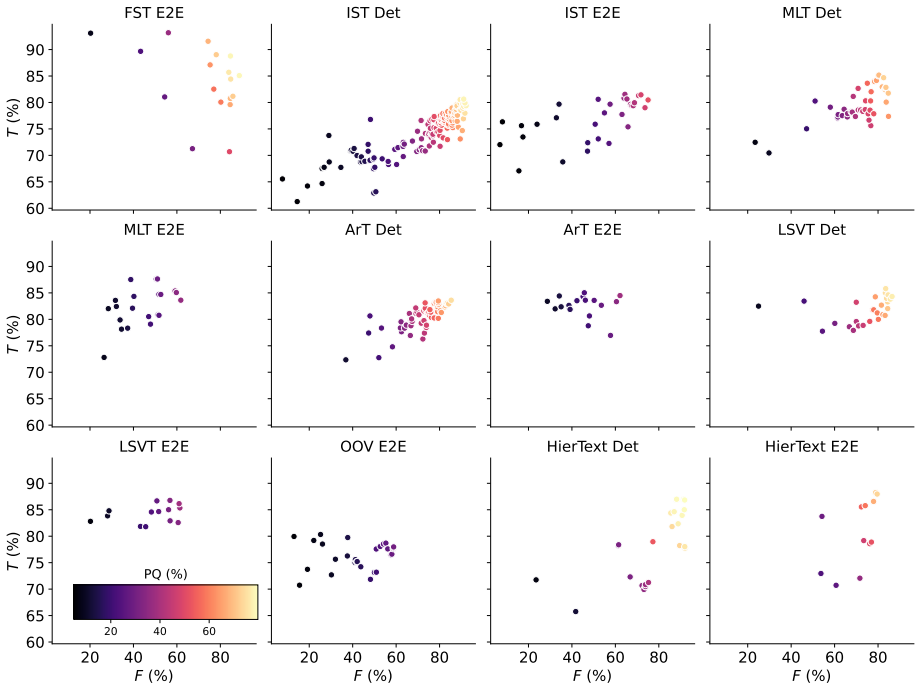

Fig. 3. Traditional F-score versus tightness T for RRC submissions. Color is PQ.

tightness. To improve the PQ, we use the score

$$\psi(g,d) = \begin{cases} \text{IoU}(g,d) & \text{if IoU}(g,d) > \tau \wedge V(g) \\ -1 & \text{otherwise,} \end{cases} \tag{12}$$

which maximizes the total IoU among matches for tightness T, but which also retains strong true positive scores for F due to a preference for including as many matches as possible, up to the bipartite constraint. As before, we add the constraint $M(g,d)$ to the positive case for end-to-end tasks.

Whereas PCQ contains sums among factors (as opposed to being a sum of products), there is no clear way to use the linear sum assignment model (7) to directly optimize correspondences for the PCQ metric. However, since IoU and $1 - \text{NED}$—both in $[0,1]$—are each important terms among the sums, it seems natural to use their product (also in $[0,1]$) in the correspondence match function:

$$\psi(g,d) = \begin{cases} \text{IoU}(g,d)\,(1 - \text{NED}(g,d)) & \text{if IoU}(g,d) > \tau \wedge V(g) \\ -1 & \text{otherwise.} \end{cases} \tag{13}$$

Using Equation (12) with the fully optimal protocol uniformly raises entries' PQ in all tasks by an average (max, resp.) of 0.04% (0.75%) over the original protocols; using IoU in the match score contributes to improve PQ uniformly

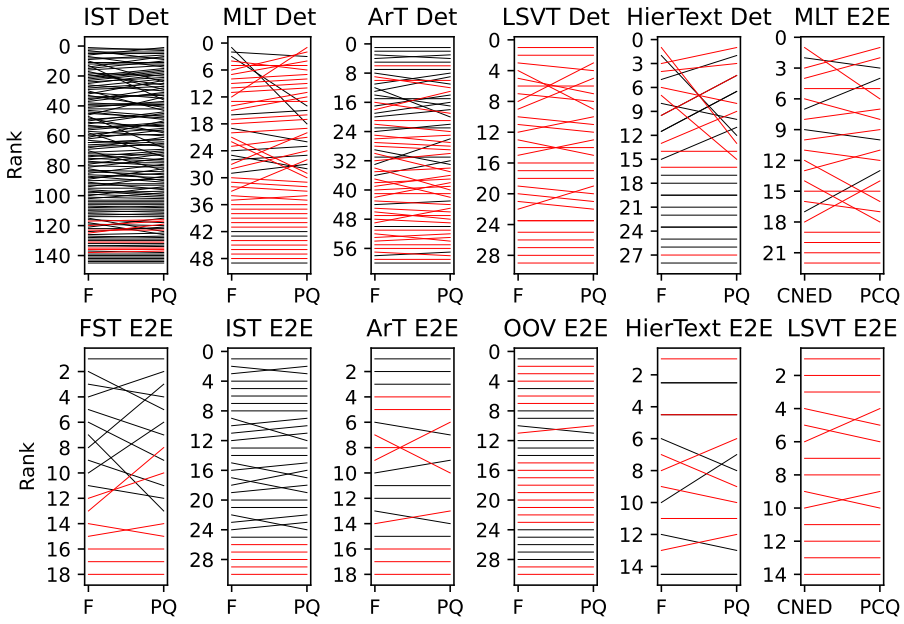

Fig. 4. Rank changes from F-score to PQ metric or CNED to PCQ (at right, for MLT and LSVT). Red is competition entries; black are post-competition.

by 0.01% (0.18%) over the binary match score (10). The proposed IoU·NED match score (13) yields an average (max, resp.) improvement in PCQ of 0.07% (0.88%) compared to the original protocols, 0.01% (0.12%) over optimal F-score matching (10), 0.003% (0.08%) over the NED-only matcher (11), and 0.01% (0.10%) over the IoU-only matcher (12).

More interesting than the effects of protocol on metric values, perhaps, is the effect of metrics on rankings. Before HierText, methods were not scored on any tightness-sensitive metric since FST13 [14] used DetEval [35] for detection. Figure 3 demonstrates that for roughly the same F-score, there can be a wide range of tightness values T. Intuitively, we might prefer higher tightness when all else is equal. Tightness and F-score are only mildly correlated, even for end-to-end tasks where one might expect higher tightness to improve recognition. While better tightness should reduce reliance on robustness, we find the robustness of the recognition modules varies somewhat independently of tightness among these entries. However, the Pearson correlations are higher than previously reported [24]: $\rho\left(T_{\text{Det}}, F_{\text{Det}}\right) = 0.5307$, $\rho\left(T_{\text{Det}}, F_{\text{E2E}}\right) = 0.4661$, and $\rho\left(T_{\text{E2E}}, F_{\text{E2E}}\right) = 0.2895$.

Figure 4 illustrates the ranking changes for entries in the various competitions. (The supplemental material provides concrete examples.) Because the text match constraint strongly influences results, ranks for E2E competitions tend to be more stable. As can be seen in Figure 3, the F-score leaders for MLT and LSVT have a wide spread in tightness T, which results in a shuffling when

Panoptic Measures are used, whether PQ or PCQ. HeirText already reported and ranked with PQ; Figure 4 confirms the rankings would have been quite different with F-score, where the highest F values differ in T by around 10%.

The PQ metric is also a slightly better predictor of end-to-end performance than F-score, as evidenced by the Spearman rank correlations: $r_s\left(F_{\text{Det}}, F_{\text{E2E}}\right) = 0.8594$ while $r_s\left(\text{PQ}_{\text{Det}}, \text{PQ}_{\text{E2E}}\right) = 0.8874$.

7 Conclusions

The data sets for robust reading challenges have grown in scale and complexity over time. Cases such as overlapping words that must be independently detected and recognized may once have been an oddity, but they now form a substantial presence in the data. Newer challenges might even put greater focus on such cases. The changes we have proposed here do not fundamentally alter the extant evaluation protocols; instead, they aim to put each entry in its best light. Indeed, we have shown that each change—optimizing the correspondences, satisfying geometric and text constraints jointly (not sequentially), and filtering predictions that overlap "don't care" ground truth regions *after* correspondence matching—are all necessary for even the ground truth to achieve a perfect score.

Our goal is not to cast doubt on prior results—virtually no ranks on the competition metrics are altered. Instead, we have reformulated the evaluation protocol to be more fair to entries and allow for new ways of thinking about solving the correspondence problem endemic to such competitions.

In addition, we have applied tightness-aware metrics to prior challenge submissions and found wide differences among otherwise similarly-performing systems. With optimization, we conclude the metrics to be stable for end-to-end tasks, allowing the IoU threshold to be eliminated (see supplementary material).

Overall, a primary weakness of the underlying formulation remains granularity; when one-to-one matching undergirds protocols, nearly perfect and disastrously imperfect detections or recognitions may both end up with similar scores. However, the strength may be its simplicity and intuitiveness. The formulation requires no extra parameters and asks systems to produce the output of direct end-use: tightly framed indexed words or even phrases. With ever more robust and accurate models, it may become increasingly appropriate to focus on word-level evaluations to the exclusion of the character level.

We hope that designers of future challenges in this family will consider adopting participant-friendly protocols or might be further inspired by the general optimization framework and the link between final metric and correspondence matching it exposes.

Acknowledgments. Thanks to reviewers for constructive feedback. Joseph Chazalon revealed the CNED form in Eq. (4). We thank the many organizers of all the competitions over the decades. We especially thank Sergi Robles for his role in developing and maintaining the RRC server that has made this centralized study possible. Finally, we thank all the researchers whose enthusiasm for the task continues to further progress.

References

1. Baek, Y., Nam, D., Park, S., Lee, J., Shin, S., Baek, J., Lee, C.Y., Lee, H.: CLEval: Character-level evaluation for text detection and recognition tasks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2404–2412 (2020). https://doi.org/10.1109/CVPRW50498.2020.00290
2. Calarasanu, S., Fabrizio, J., Dubuisson, S.: What is a good evaluation protocol for text localization systems? Concerns, arguments, comparisons and solutions. Image and Vision Computing **46**, 1–17 (2016). https://doi.org/https://doi.org/10.1016/j.imavis.2015.12.001
3. Center, C.V.: Robust reading competition, https://rrc.cvc.uab.es
4. Cheng, Z., Lu, J., Zou, B., Zhou, S., Wu, F.: ICDAR 2021 competition on scene video text spotting. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) Document Analysis and Recognition – ICDAR 2021. pp. 650–662. Springer International Publishing, Cham (2021)
5. Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., Liu, J., Karatzas, D., Chan, C.S., Jin, L.: ICDAR2019 robust reading challenge on arbitrary-shaped text - RRC-ArT. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1571–1576 (2019). https://doi.org/10.1109/ICDAR.2019.00252
6. COCO - Common Objects in Context: COCO API (2014), https://github.com/cocodataset/cocoapi
7. Dangla, A., Puybareau, E., Tochon, G., Fabrizio, J.: A first step toward a fair comparison of evaluation protocols for text detection algorithms. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 345–350. IEEE (2018)
8. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision **88**(2), 303–338 (Jun 2010)
9. Everingham, M., Winn, J.: The PASCAL visual object classes challenge 2012 (VOC2012) development kit (May 2011), http://host.robots.ox.ac.uk/pascal/VOC/voc2012/VOCdevkit_18-May-2011.tar
10. Garcia-Bordils, S., Mafla, A., Biten, A.F., Nuriel, O., Aberdam, A., Mazor, S., Litman, R., Karatzas, D.: Out-of-vocabulary challenge report. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) Computer Vision – ECCV 2022 Workshops. pp. 359–375. Springer Nature Switzerland, Cham (2023)
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. 01, pp. 1435–1443 (2017). https://doi.org/10.1109/ICDAR.2017.234
12. Iwamura, M., Morimoto, N., Tainaka, K., Bazazian, D., Gomez, L., Karatzas, D.: ICDAR2017 robust reading challenge on omnidirectional video. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 1448–1453 (2017). https://doi.org/10.1109/ICDAR.2017.236
13. Karatzas, D., Gomez Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., Shafait, F., Uchida, S., Valveny, E.: ICDAR 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1156–1160 (2015). https://doi.org/10.1109/ICDAR.2015.7333942

14. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Gomez i Bigorda, L., Mestre, S.R., Mas, J., Mota, D.F., Almazàn, 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 (2013). https://doi.org/10.1109/ICDAR.2013.221
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. pp. 1–4 (2023)
16. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9396–9405 (2019). https://doi.org/10.1109/CVPR.2019.00963
17. Krylov, I., Nosov, S., Sovrasov, V.: Open images v5 text annotation and yet another mask text spotter (2021) arXiv:2106.12326 [cs.CV]
18. Lee, C.Y., Baek, Y., Lee, H.: TedEval: A fair evaluation metric for scene text detectors. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 7, pp. 14–17. IEEE (2019)
19. Lee, H.S., Yoon, Y., Jang, P.H., Choi, C.: PopEval: A character-level approach to end-to-end evaluation compatible with word-level benchmark dataset. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1207–1213 (2019). https://doi.org/10.1109/ICDAR.2019.00195
20. Li, Z., Lin, Y., Chiang, Y.Y., Weinman, J., Chazalon, J., Tual, S., Perret, J., Duménieu, B., Abadie, N.: ICDAR 2024 competition on historical map text detection, recognition, and linking. In: 18th International Conference on Document Analysis and Recognition (ICDAR 2024) (2024)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
22. Liu, Y., Jin, L., Xie, Z., Luo, C., Zhang, S., Xie, L.: Tightness-aware evaluation protocol for scene text detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9604–9612. IEEE (2019)
23. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1049–1059 (2022)
24. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: ICDAR 2023 competition on hierarchical text detection and recognition (2023) arXiv:2305.09750 [cs.CV]
25. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R.: ICDAR 2003 robust reading competitions. In: Proc. Intl. Conf. on Document Analysis and Recognition. vol. 2, pp. 682–687 (2003). https://doi.org/10.1109/ICDAR.2003.1227749
26. Lucas, S.M.: Text locating competition results. In: Proc. Intl. Conf. on Document Analysis and Recognition. pp. 80–85 (2005). https://doi.org/10.1109/ICDAR.2005.231
27. Mulmuley, K., Vazirani, U.V., Vazirani, V.V.: Matching is as easy as matrix inversion. In: Proceedings of the nineteenth annual ACM Symposium on Theory of Computing. pp. 345–354 (1987)
28. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., Ogier, J.M.: 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 (2019). https://doi.org/10.1109/ICDAR.2019.00254
29. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., Khlif, W., Luqman, M.M., Burie, J.C., Liu, C.l., Ogier, J.M.: 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. 01, pp. 1454–1459 (2017). https://doi.org/10.1109/ICDAR.2017.237
30. Rice, S.V., Kanai, J., Nartker, T.A.: A report on the accuracy of OCR devices. Tech. rep., University of Nevada, Information Science Research Institute (1992)
31. Rice, S.V.: Measuring the accuracy of page-reading systems. Ph.D. thesis, University of Nevada, Las Vegas (1996)
32. Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Belongie, S., Lu, S., Bai, X.: ICDAR2017 competition on reading chinese text in the wild (RCTW-17). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 1429–1434 (2017). https://doi.org/10.1109/ICDAR.2017.233
33. 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: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8798–8808 (2021). https://doi.org/10.1109/CVPR46437.2021.00869
34. Sun, Y., Ni, Z., Chng, C.K., Liu, Y., Luo, C., Ng, C.C., Han, J., Ding, E., Liu, J., Karatzas, D., Chan, C.S., Jin, L.: ICDAR 2019 competition on large-scale street view text with partial labeling - RRC-LSVT. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1557–1562 (2019). https://doi.org/10.1109/ICDAR.2019.00250
35. Wolf, C., Jolion, J.M.: Object count/area graphs for the evaluation of object detection and segmentation algorithms. International Journal of Document Analysis and Recognition (IJDAR) **8**(4), 280–296 (2006)
36. Zayene, O., Hennebert, J., Ingold, R., BenAmara, N.E.: ICDAR2017 competition on Arabic text detection and recognition in multi-resolution video frames. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 1460–1465 (2017). https://doi.org/10.1109/ICDAR.2017.238
37. Zdenek, J., Nakayama, H.: Erasing scene text with weak supervision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2238–2246 (March 2020)
38. Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., Wang, L., Wang, D., Liao, M., Yang, M., Bai, X., Shi, B., Karatzas, D., Lu, S., Jawahar, C.V.: ICDAR 2019 robust reading challenge on reading chinese text on signboard. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1577–1581 (2019). https://doi.org/10.1109/ICDAR.2019.00253

Supplementary Material

Section A contains additional visualizations of the effects of protocol changes proposed in the main paper. Section B provides examples that illustrate the changes in rank with the panoptic metrics. Section C explores the effects of changing the evaluation protocol to allow only one-to-one matching with "don't care" regions. Finally, Section D demonstrates the stability of metrics (particularly end-to-end PQ) with respect to the IoU threshold under the proposed optimization framework.

A Evaluation Protocol

This section contains additional visualizations of the effects of protocol changes described in the main paper.

Figure S1 plots a subset of the data from Figure 2 in the main paper; omitting the sequential variant more visibly demonstrates the effect of joint optimal processing relative to the original protocol.

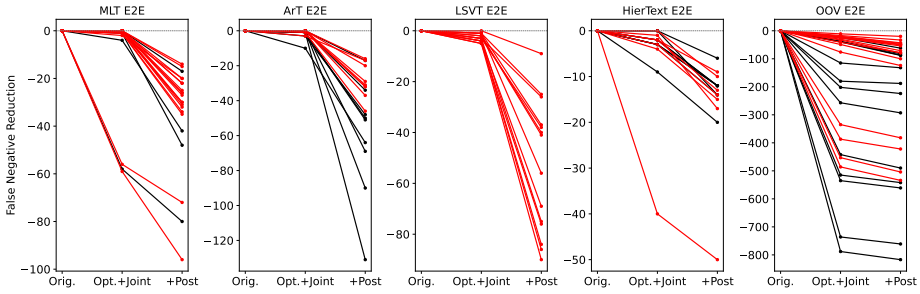

Fig. S1. Changes in raw false negative count from the original protocol for public RRC submissions. Red is competition entries; black are post-competition.

Figures S2 and S3 show the net difference for each submission with all proposed protocol changes: i) correspondences optimizing true positives, ii) jointly (rather than sequentially) satisfying geometric and textual constraints for end-to-end tasks, and iii) matching all predictions before filtering out remaining unmatched predictions that overlap with "don't care" regions. From these changes we calculate the improvements in the main paper—ΔF-score average (max, resp.) +0.034% (+0.27% max).

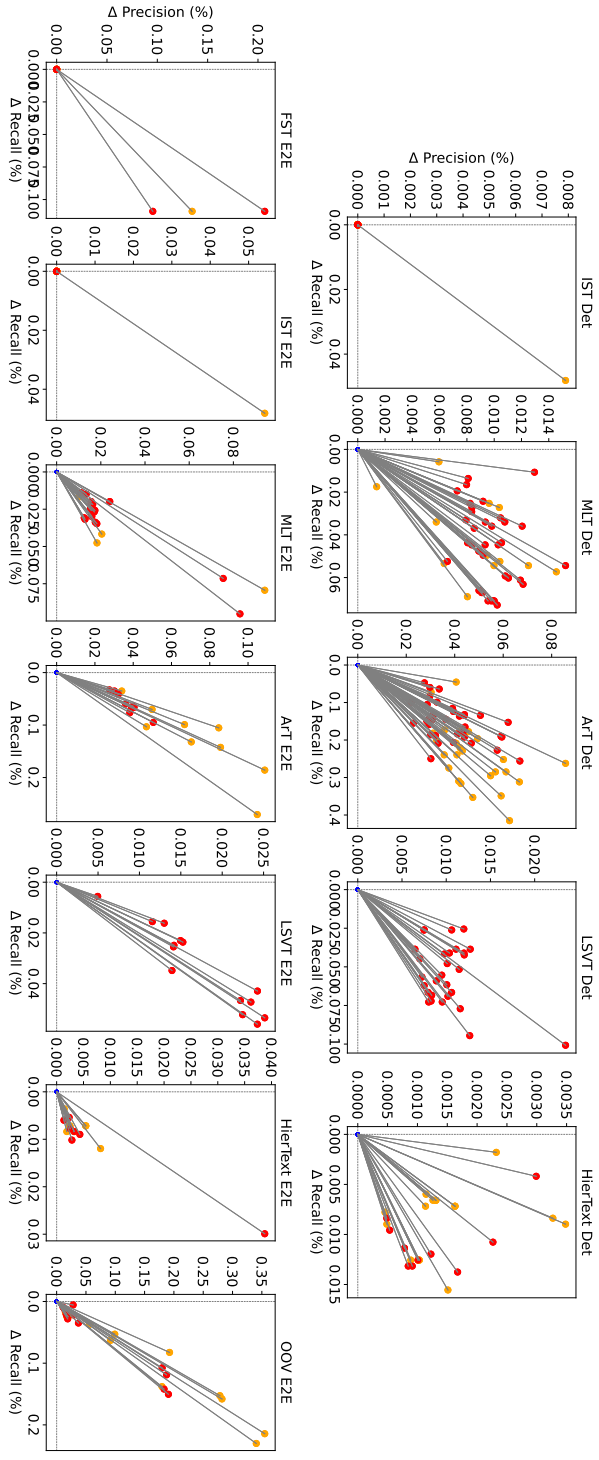

Fig. S2. Net changes to precision P and recall R for submissions with all protocol changes described in the main paper. Red indicates competition entries; others are later submissions.

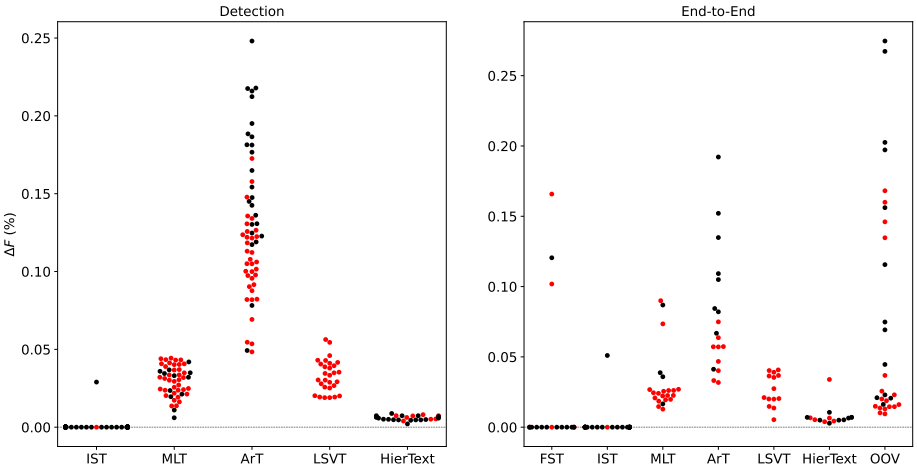

Fig. S3. Net changes to F-score for submissions with all protocol changes described in the main paper. Red indicates competition entries; others are later submissions.

B Panoptic Metric

Figure S4 provides indicative example detections from the Focused Scene Text end-to-end task (where submissions and test ground truth are both publicly available). These examples illustrate how otherwise similarly performing systems can diverge in rank once tightness becomes a factor in the metric. The systems with rank seven (blue) and eight (red) end up at ranks thirteen and three after accounting for tightness; the blue boxes are visibly poorer than the red boxes. Table S1 provides the accompanying metric values underlying the rank changes.

Table S1. Overall performance statistics (on entire FST E2E data set) accompanying the systems in Figure S4.

Rank (F)	F	T	PQ	Rank (PQ)
7	84.2041	70.6937	59.5270	13
2	85.7778	81.1673	69.6235	5
4	84.6501	88.7825	75.1545	2
8	83.8895	85.7028	71.8957	3

C Handling "Don't Cares"

The annotation protocol of most RRCs allows "don't care" regions to contain multiple words or even multiple lines of text. The evaluation protocol therefore

Fig. S4. Cropped example correct detections from FST E2E illustrating differences in tightness (IoU) that cause re-ranking under the PQ metric.

allows many-to-one matching of predicted detections against such "don't care regions." This section explores potential effects of changing the evaluation protocol to allow only one-to-one matching with "don't care" regions.

First, Table S2 quantifies the extent of many-to-one matchings for "don't care" regions among all public submissions to the RRCs.

Table S2. Relative frequency (%) of number of matches to "ignore" regions, among those with any matches at all. (Accumulations over all public challenge submissions.)

Challenge	Det		E2E	
	Two	More	Two	More
FST15	–	–	2.49	0.00
IST15	4.54	1.03	4.26	0.84
ArT19	6.77	3.35	6.83	3.45
MLT19	9.88	5.04	9.65	4.53
LSVT19	6.08	3.35	4.17	1.15
OOV22	–	–	12.87	8.14
HierText23	6.77	4.14	8.18	5.39

To accommodate the exploration of one-to-one matching with "don't care" regions, we alter the match score slightly:

$$\psi(g,d) = \begin{cases} +1 & \text{if IoU}(g,d) > \tau \wedge V(g) \\ \epsilon & \text{if IoU}(g,d) > \tau \wedge \neg V(g) \\ -1 & \text{otherwise,} \end{cases} \tag{14}$$

where V indicates the ground truth annotation is valid (*not* a "don't care.") In this framework, using a small positive value ϵ for matches to "don't care" annotations allows such correspondences, while leaving a preference for matches with valid words, which improve performance metrics. Such one-to-one matches are discounted (neither true positives, false positives, nor false negatives), but remaining unmatched predictions are counted as false positives, even if they overlap with a "don't care" region.

Figures S5 and S6 illustrate the impact of *changing* the evaluation protocol to allow only one-to-one matches with "don't care" regions of the ground truth. Detection tasks are more heavily impacted than end-to-end tasks. The impacts are not uniformly distributed; the primary paper observes that high-recall submissions have the most degraded precision values. This is likely because under current protocols, such systems can be rewarded with the potential for an improved recall without harm to precision since most of the additional predicted detections will be filtered by the *a priori* "don't care" filtering.

D IoU Match Threshold τ and Metric Stability

As in the established competition protocols, the main paper utilizes an IoU match threshold of $\tau = 0.5$, but evaluations have used other values (e.g., $\tau = 0.75$ for COCOText).

This section examines the stability of the various metrics across different values $0 \leq \tau < 1$.

The main paper introduces the optimization framework for the matching stage of the evaluation protocol. That framework allows us to find the best overall correspondence, as measured by total IoU. Because the panoptic quality (PQ) metrics also incorporate the IoU tightness of the matches, it may make sense to lower the allowable threshold for a match, even down to zero.

Requiring only a positive IoU ($\tau = 0$) allows PQ-based metrics to become essentially parameter-free with an increasing penalty for decreasing tightness quality. For detection tasks, exceptionally low values of τ may not have experimental validity, but end-to-end tasks requiring matching strings still produce strongly grounded correspondences.

F-score can only increase as the match threshold τ decreases, because it allows more detections to be considered true positives (thus reducing both false positives and false negatives). Similarly, the average tightness T decreases as more matches with lower IoU are included. Whether higher F or lower T dominates in PQ depends on "where" (τ value) any increase in recall occurs.

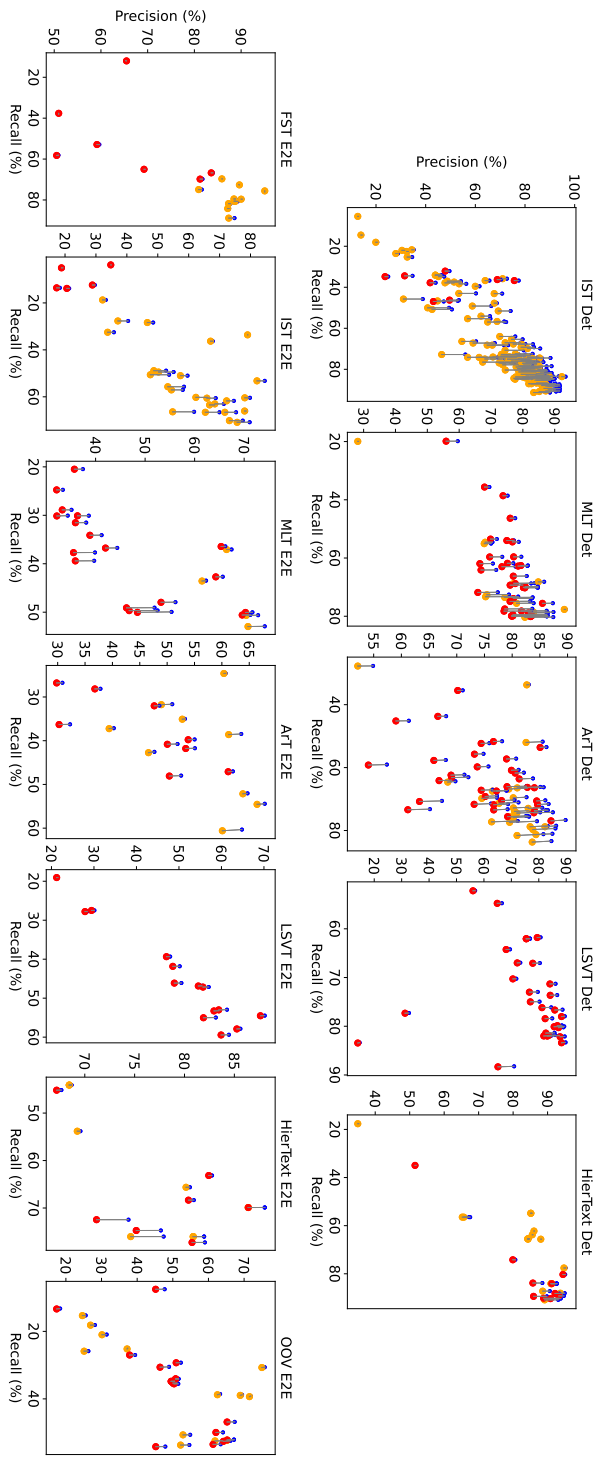

Fig. S5. Requiring one-to-one matches with "don't care" regions: Net changes to precision and recall for submissions with all protocol changes. Red indicates competition entries; others are later submissions.

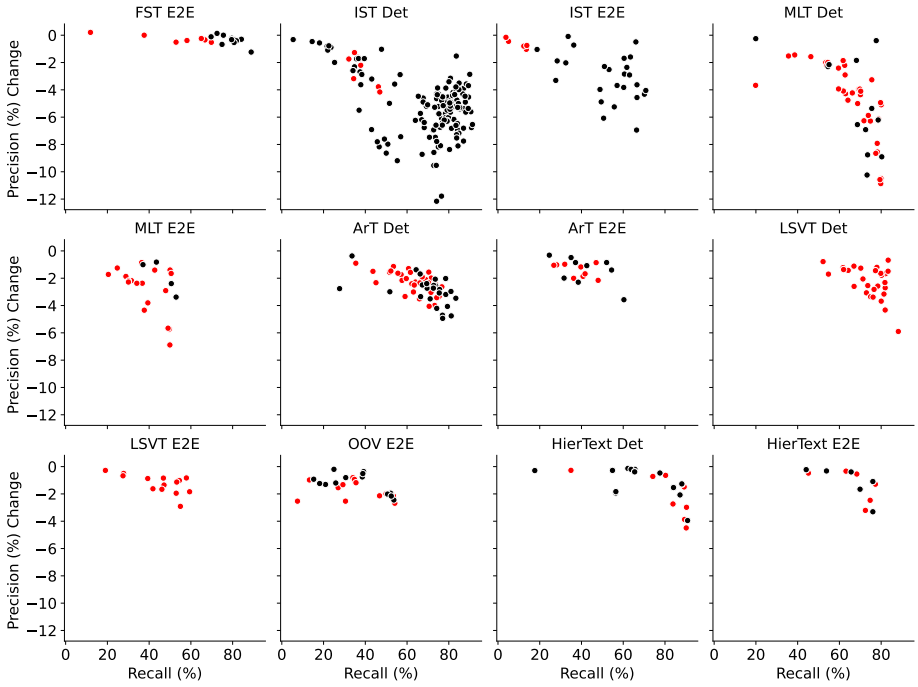

Fig. S6. Requiring one-to-one matches with "don't care" regions: Net changes to precision given recall for submissions with all protocol changes. Red indicates competition entries; others are later submissions.

Figures in the following subsections demonstrate the effect of altering the IoU match threshold on F-score, tightness T, and PQ for both detection and end-to-end tasks.

D.1 Detection Tasks

Several prior works have demonstrated problems with IoU-based one-to-one matching for text detection evaluation (cf. Section 2.1 of the main paper). Specific downstream tasks are likely the best arbiter of any evaluation protocol's validity. Nevertheless, we explore the stability of the metrics under various IoU match thresholds using the IoU-based match score function—Equation (12) in the main paper.

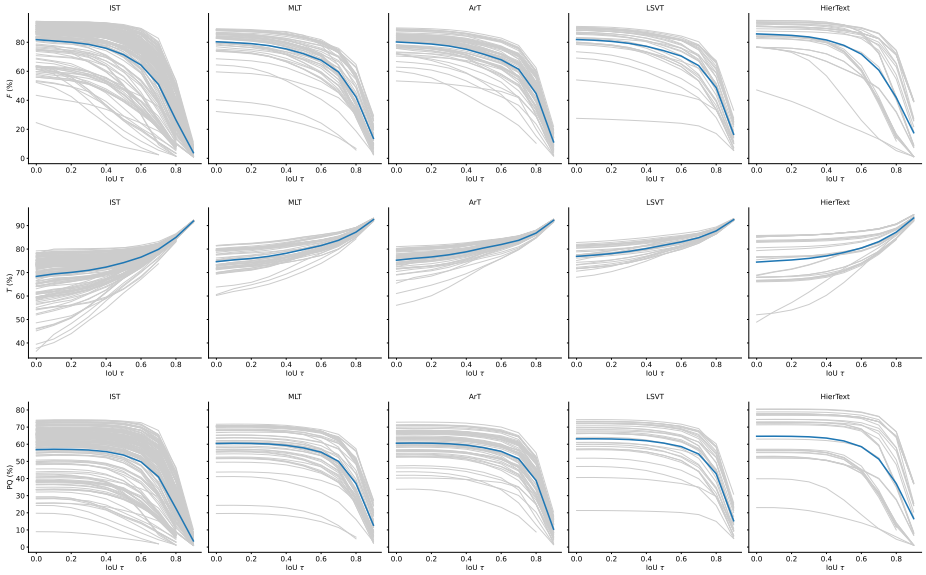

Fig. S7. Changes in F-score (top), tightness T (middle), and panoptic quality PQ (bottom) with varying IoU match threshold τ for detection tasks. Individual submissions in light gray and the average in bold blue.

Figure S7 shows that although F indeed increases while T decreases, the PQ value is mostly stable beyond a certain point, perhaps around $\tau = 0.4$.

In the detection task, it is not obvious whether small values of τ (say 0.1) result in meaningful detections worthy of being called "true positives"; this will depend on the downstream task. However, the lower tightness of such additional matches does end up penalizing the final quality, thus balancing out any increase in F score.

D.2 End-to-end Tasks

Unlike for detections, even small values of τ can still produce meaningful correspondences in the end-to-end task because the recognized string must also match or have low NED. Figure S8 illustrates the metric changes with decreasing τ for end-to-end tasks. As before, the IoU-based match score is used for the F-score, tightness T, and panoptic quality PQ results. The PCQ results are generated using the IoU and NED-based match function—Equation (13) in the main paper.

As suggested above, recognition modules are less likely to produce correctly matching strings below a certain IoU. Whereas decreasing τ progressively raises the F-score for detections, below $\tau = 0.4$ there is very little increase and the metrics are highly stable.

This suggests that the threshold parameter could be eliminated entirely from a PQ-based metric, which rewards both quantity (F-score) and quality (tightness T) of all corresponding string matches.

Whereas the other metrics (PQ, F, and T) remain completely stable below a certain threshold, a slight peak may be observed for the Panoptic Character Quality (PCQ) in the right column of Figure S8. Although additional rectangles are included as matches, the concomitant cost to average tightness and edit distance eventually outweigh the boost to F-score.

Figure S9 captures the total reduction in false negatives by changing the revised protocol described in the main paper from an IoU threshold of $\tau = 0.5$ to $\tau = 0.0$. It indicates the absolute magnitude of the change can be quite substantial, particularly for some submissions whose predictions have an IoU just below the cutoff threshold. Similarly, Figure S10 demonstrates the net effect on the normalized precision and recall metrics.

Figure S11 marks the final changes in F-score and PQ with *tau* changed to 0.0 from 0.5 for both detection and end-to-end tasks.

Because the recall for end-to-end tasks is only sensitive to around $\tau = 0.4$, it may make sense to use the parameter-free version in future competitions. To examine the potential effects on rankings, we extend the rank bump plot from the main paper (Figure 4) to consider not only the change in metric using the updated protocol where $\tau = 0.5$, but also with the additional change to $\tau = 0$.

After the initial change to a panoptic metric, the rankings of end-to-end tasks remain largely consistent in the parameter-free regime, indicating its stability.

Taken together, the increases of Figure S11 combined with the stability in rankings suggest that eliminating the threshold parameter may increase valuable metric sensitivity for end-to-end methods

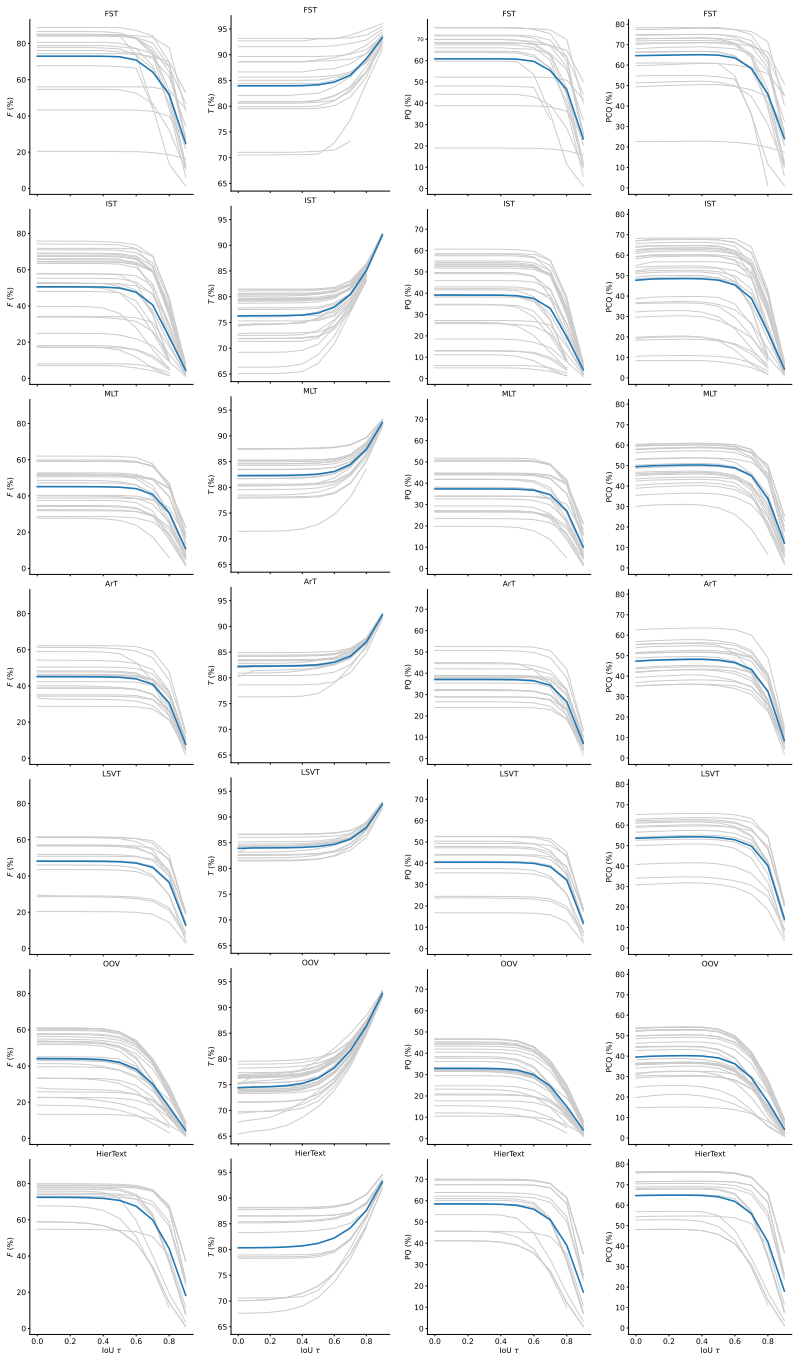

Fig. S8. Changes (left to right) in F-score, tightness T, panoptic quality PQ, and panoptic character quality PCQ with varying IoU match threshold τ for end-to-end tasks. Individual submissions in light gray and the average in bold blue.

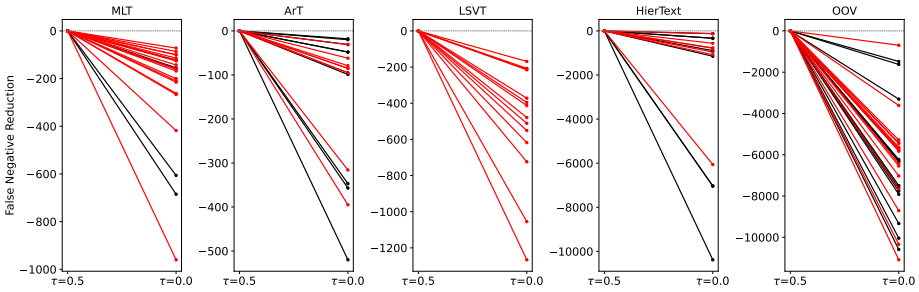

Fig. S9. Changes in raw false negative count from the revised protocol for public RRC submissions to a parameter-free version ($\tau = 0.0$) on E2E tasks. Red is competition entries; black are post-competition.

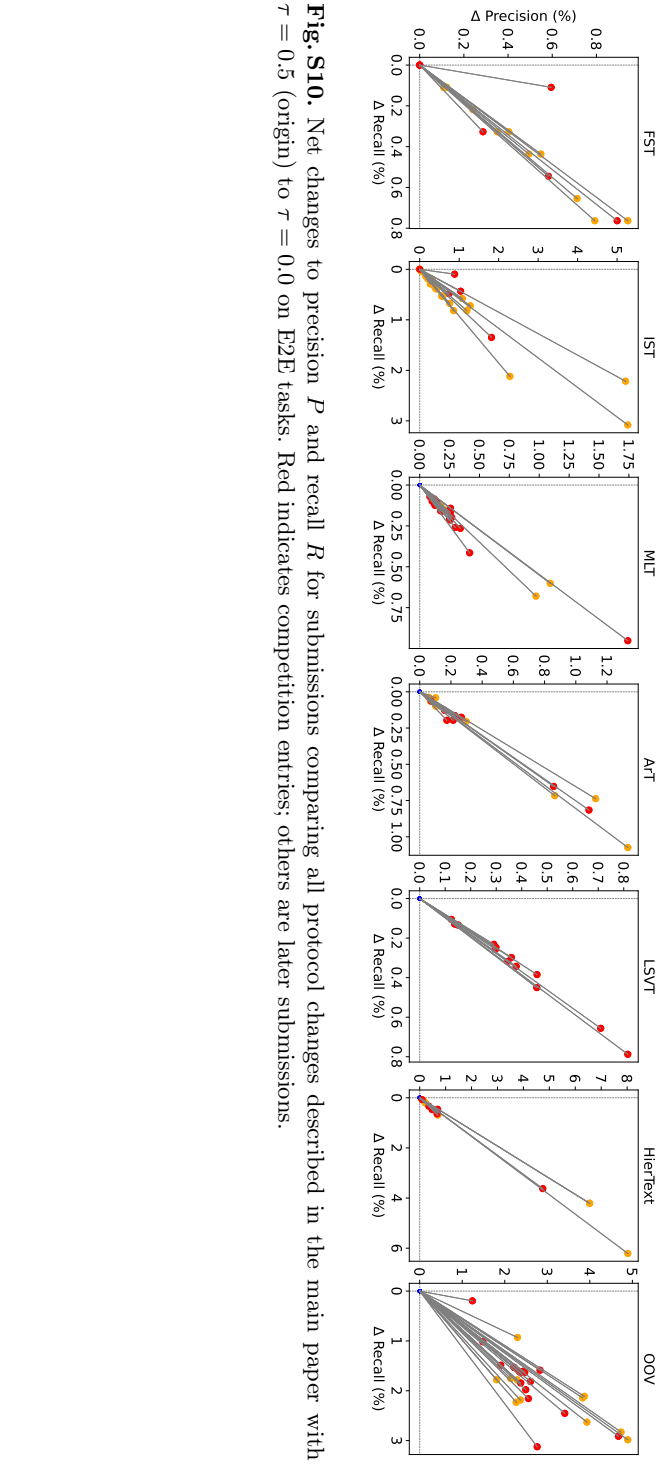

Fig. S10. Net changes to precision P and recall R for submissions comparing all protocol changes described in the main paper with $\tau = 0.5$ (origin) to $\tau = 0.0$ on E2E tasks. Red indicates competition entries; others are later submissions.

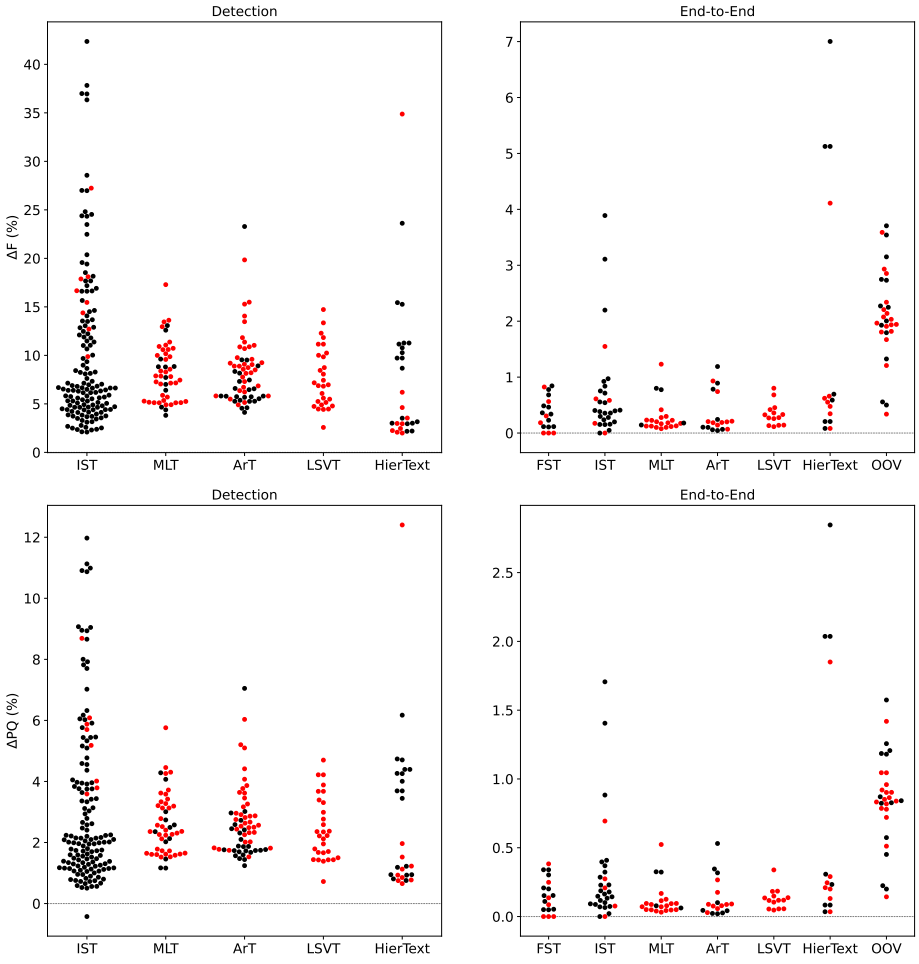

Fig. S11. Net changes to F-score and PQ for submissions comparing all protocol changes described in the main paper with $\tau = 0.5$ (y origin) to $\tau = 0.0$. Red indicates competition entries; others are later submissions.

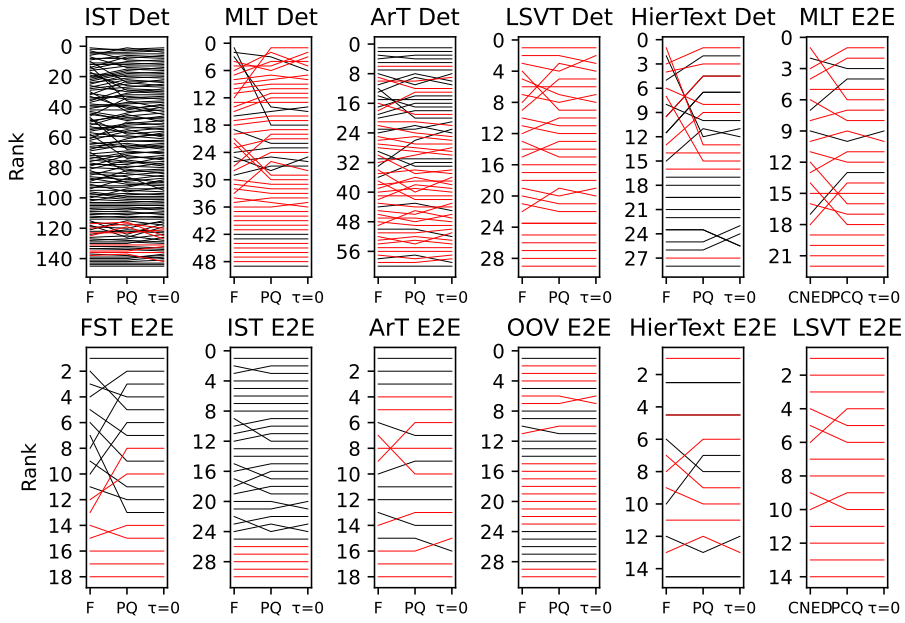

Fig. S12. Rank changes from updated protocol F-score to PQ metric or CNED to PCQ (at right, for MLT and LSVT) with additional change to $\tau = 0$. Red is competition entries; black are post-competition.