MIST: Multilingual Incidental Dataset for Scene Text Detection

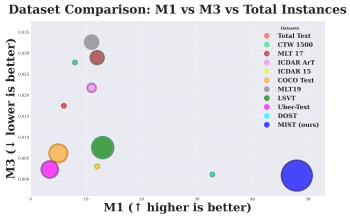

Figure 1. Clusters of scene text detection datasets based on M_1 and M_3. Bubble size represents the total number of text instances in the dataset.

Dataset	M_1 ↑	M_2 ↓	M_3 ↓	#images	#text instances
Total-Text	6	0.01093	0.01747	1555	9330
CTW1500	8	0.01796	0.02778	1500	12000
MLT17	12	0.01017	0.02893	18000	216000
ICDAR-ArT	11	0.01230	0.02174	10166	111826
ICDAR15	12	0.00204	0.00300	1500	18000
COCO-Text	5	0.00344	0.00613	63686	318430
MLT19	11	0.01125	0.03261	20000	220000
LSVT	13	0.01270	0.00750	30000	390000
Uber-Text	3.46	0.00277	0.00230	**82572**	285699
DOST	32.77	0.00090	0.00110	338	11076
MIST (ours)	**48**	**0.00067**	**0.00084**	12000	**576000**

Table 1. **Updated statistical comparison of scene text datasets.** Best values are **bolded**; second-best are underlined.

Model	P	R	F	F^{α}	F^{β}
DP-DETR	69.61	57.04	**62.70**	59.15	52.80
TBPN	70.87	47.75	57.06	52.09	44.68
MixNet	73.48	45.59	56.27	51.44	44.06
DB++	72.84	39.73	51.42	44.71	38.32

Table 2. **Updated benchmarking on MIST.** Precision (P), Recall (R), F-measure (F), and stratified evaluation scores

We thank the reviewers and the Area Chair (AC) for their insightful feedback and for acknowledging our contributions. We have revised the draft to improve its clarity per the suggestions. We address the reviewers' (RKHDz: R1, R4LEn: R2, RPacg: R3, RbCjV: R4) concerns below.

[R1,R2,R3,R4,AC]Multilingual nature, limited language diversity and geographic bias We acknowledge that MIST is not globally representative in language coverage and contains relatively few instances of arbitrarily shaped text. As the data were collected exclusively in India, the language diversity is inherently limited. The strength of MIST lies in its incidental nature, large-scale, real-world scenes with rich macro-level complexities.

[R1,R2,R3,R4,AC] Comparison with more datasets: We thank the reviewers for suggesting to compare with other datastes such as MLT19, DOST, Uber-Text and LSVT. Upon comparison, our stance of MIST being more incidental still holds true, as shown in Fig 1 and Table 1. We have revised the results in the main paper (Table 1, Fig 6, Fig 7).

[R1,R2,R3,R4,AC]Narrow Baseline Evaluation We appreciate the reviewers' feedback about the narrow baseline evaluation. Hence, we have added MixNet and DBNet++ as baselines. Even with the new baselines, MIST stands to be challenging. We were unable to include CPN(AAAI'24) and KACM(CVPR'24), scene text detection models from 2024, due to the lack of open-source code. We have contacted the authors to request code access; if it becomes available, we will report the benchmark results.

[R4, R1] Related Works and Clarity of Presentation We have rewritten Sec 2.1 to explain MIST's improvements. We have also clarified the experimental setup and visualizations (Fig. 1; Secs. 3.5, 4.1, 5.2), and provide training configurations in the supplementary material.

[R2] Overfitting risk We have provided the per-bin F-Measure in Fig. 9b in the main paper, showing the skewed performance in regions of higher M3.

[R4] Geographic bias By "generality", we mean the *generalizability of model trained on* MIST. Sec 5.2 shows the model transfers well to out-of-domain benchmarks, which while not eliminating the bias, mitigates concerns about overfitting to a single region.

[R1,R4] Dataset Sampling and Split We collected 150,h of video across 10 regions (15,h/region; 15 sequences/region) and sampled 12k frames (1,200/region). For each region (M=15), we assigned equal per-sequence quotas (b=80) and used stratified uniform sampling: partition each sequence into B temporal bins and select one frame per bin. The B frames were then shuffled and split into train/val/test sets. Full details are provided in the supplementary material.

[R2] Underrepresentation of Arbitrary Shaped Text in MIST The underrepresentation of curved text in MIST is by design, as its central purpose is to isolate robustness to incidental scene factors. Including arbitrary shapes would confound subsequent analysis, making it difficult to attribute model weaknesses. We position MIST as complementary to arbitrary shape focused benchmarks. Incidental robustness remains a primary bottleneck for current detectors (Sec 3.5 in the main paper), hence, we advocate for prioritizing improvements on benchmarks like MIST until performance plateaus, following which a composite benchmark can stress both complexities.

[R4] Incidental datasets contain more text instances *Incidental* acquisition records the broader scene rather than a targeted crop. If a focused crop contains n text instances, an incidental view spanning $k \geq 1$ comparable subareas contains at least n (typically $> n$) instances; thus incidental scene images exhibit higher text–instance counts.

[R1] Benefits over MLT17 and MLT19 Fig. 1 and Table 1 shows MIST as a more incidental dataset than MLT17/MLT19. In the main paper, Fig 8b and Fig 10b show that models trained on MLT17 attain high F-measure, whereas the same architectures trained on MIST saturate around $F \approx 0.60$, even at the highest M_3 (i.e., the most focused scenes). This indicates that even with dataset-specific training, achieving top performance on MIST remains difficult, underscoring its value as a challenging benchmark.

[R2] Error Analysis Sec 5.1 in the revised paper presents an extensive error analysis on texts suffering low resolution, occlusion, and poor illumination.

[AC] VLMs for Detection Evaluation of ChatGPT-5 and Gemini 2.5 Pro on text detection revealed poor performance. The models exhibited a significantly low recall rate of 15%, and the predicted bounding boxes were not scaled to the original image resolution to account for input compression. Visual results are provided in supplementary