A Appendix

A.1 Comprehensive Text Reading Datasets (Worms) Composition

The following tables provide detailed statistics of the Worms training set. The dataset includes diverse subsets to cover various text reading tasks comprehensively.

Table 11: The statistics of the training set of our proposed comprehensive text reading datasets (Worms). Dataset marked with an * suggests that the annotations are inexact, potentially representing pseudo-labels generated through model-based pre-annotation. In the "Granularity" column, the presence of a blue star ⋆ signifies that the original dataset lacked annotations, which we have supplemented. Synth-arxiv with † means synthetic formula data generated by ourselves.

Dataset	Subset	Category	Writing Type	Granularity		Number
				word	line	
ICDAR2013 [28]	train	Natural scene full image	printed	✓	⋆	229
ICDAR2015 [27]	train			✓	⋆	1K
CTW1500 [44]	train			⋆	✓	1K
TotalText [12]	train			✓	⋆	1.3K
HierText [50]	trainval			✓	✓	10K
TextOCR [67]	train			✓	⋆	25K
Open Images V5 Text [32]	trainval			✓	✗	208K
Uber-Text [88]	trainval			✗	✓	83K
COCO-Text [70]	trainval			✓	✗	54K
Curved SynthText [46]	train			✓	✗	149K
MLT2017 [55]	train			✓	✗	10K
LAION-OCR [62]*	train			✓	✗	2.5M
PubLayNet [89]*	trainval	Document full image		✓	✓	352K
MJSynth [25]	train	Cropped text		none		8.9M
SynthText [20]	train					7.3M
SynthAdd[4]	train					1.2M
Union14M-L [26]	train					4.1M
OOV [18]	train					4.4M
LAION-OCR [62]*	train					11.2M
IAM [53]	trainval		handwritten			60K
CVL [31]	train					86K
RIMES [19]	trainval					52K
CROHME [77]	trainval	Cropped formula	none			11K
HME100K [86]	train					75K
LatexOCR[5]	trainval					165K
Synth-arxiv†	train					10.2M
Total	-	-	-	-		51.1M

[4] From https://mmocr.readthedocs.io/en/v0.6.3/datasets/recog.html#synthadd.
[5] Results are from: https://github.com/lukas-blecher/LaTeX-OCR