RERANK-TTA: LEARNING TO RERANK AT TEST TIME FOR ROBUST TEXT RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text recognition faces deployment-time distribution shifts arising from changes in visual appearance, writing styles, image acquisition, and language. When target labels and source training data are unavailable, source-free test-time adaptation (TTA) offers a way to adapt deployed recognizers. However, common confidence-, entropy-, and consistency-based objectives do not explicitly optimize the relative ordering of competing transcriptions, while decoding-time search and LM reranking leave the recognizer unchanged. We propose **ReRank-TTA**, a source-free TTA framework that turns hypothesis preferences into relative supervision for autoregressive text recognition. Given unlabeled target images, ReRank-TTA generates candidate transcriptions, combines recognizer scores with a frozen language prior to select preferred and competing hypotheses, and updates an episodic recognizer with a pairwise sequence-ranking loss. This objective learns from relative preferences without treating any candidate as ground truth. Across IAM, RIMES, George Washington, and Bentham, the full pipeline consistently improves four TrOCR checkpoints over frozen greedy decoding, reducing CER/WER by an average of **3.74/5.57** absolute points, with mean relative improvements of **23.2%/17.9%**. Matched-decoding ablations demonstrate additional gains from adaptation beyond beam search and LM reranking. Experiments with PARSeq and PARSeq-tiny on SVTP and IC15 further demonstrate applicability to scene text recognition.

1 INTRODUCTION

Text recognition converts images of written text into machine-readable sequences. Despite substantial progress from convolutional, recurrent, transformer-based, and vision-language recognizers (Puigcerver, 2017; Michael et al., 2019; Coquenet et al., 2022; d'Arce et al., 2022; Li et al., 2025b; Kang et al., 2022; Barrere et al., 2022; 2024; Li et al., 2023; Bautista & Atienza, 2022), robust recognition remains difficult under deployment-time distribution shifts. In handwritten text recognition (HTR), a recognizer trained on one collection may encounter unseen writers, writing instruments, scanning artifacts, historical scripts, languages, and transcription conventions at test time (Garrido-Munoz & Calvo-Zaragoza, 2025). Scene text recognition introduces further variation in typography, perspective, and capture conditions. Although large-scale pretraining provides strong visual and linguistic representations (Li et al., 2023; Bautista & Atienza, 2022), performance can still degrade when target inputs differ from the training distribution. In many practical settings, labeled target data are unavailable, and the original training data cannot be accessed due to privacy, licensing, or storage constraints (Li et al., 2024). This motivates *source-free test-time adaptation* (TTA): adapting a deployed recognizer using unlabeled target inputs without source-domain replay or target annotations (Wang et al., 2020). Here, "source-free" means that no source training images, source labels, or target transcriptions are accessed during adaptation; an external frozen text-only language prior may still be used. Adaptation must therefore extract useful information from the target inputs themselves while preserving the recognition knowledge contained in the pretrained model, without requiring a new supervised training stage.

We study this problem primarily through handwriting recognition, with an additional evaluation on scene text. Recent HTR methods demonstrate the value of adapting to unlabeled test inputs or individual writers (Gu et al., 2025a;b). However, cross-collection deployment can involve simultaneous changes in writing style, language, and acquisition conditions (Garrido-Munoz & Calvo-Zaragoza, 2025). We therefore examine adaptation across multiple target collections and recognizer checkpoints

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

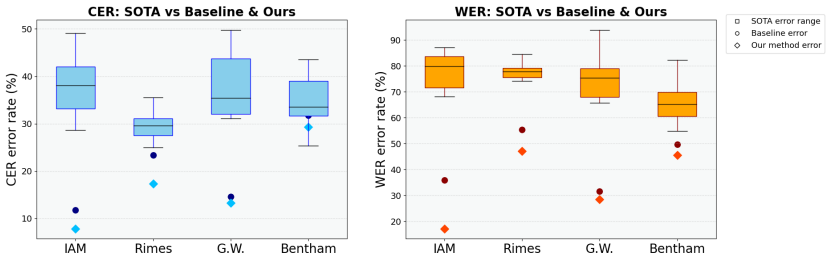

Figure 1: **CER/WER comparison across four handwriting benchmarks.** Markers compare frozen **TrOCR-stage1** with the full ReRank-TTA pipeline, which combines adaptation and enhanced decoding. Boxes summarize published results from specialized HTR systems and provide contextual references because training and evaluation protocols differ. The paired TrOCR results use the same pretrained checkpoint without source data or target labels at deployment.

rather than assuming that the shift is limited to writer identity. TrOCR provides a suitable backbone because its public checkpoints support evaluation before and after IAM handwriting fine-tuning (Li et al., 2023). Under our line-level IAM protocol, the tested generic OCR/VLM systems also remain behind a specialized TrOCR baseline (Appendix B). These observations motivate adapting existing recognizers, while Figure 1 illustrates the improvements obtained by the full ReRank-TTA pipeline across handwriting collections.

The central challenge concerns the sequence-level decision made by an autoregressive recognizer. Unlike closed-set classification, autoregressive text recognition constructs variable-length transcriptions through conditional predictions (Li et al., 2023; Michael et al., 2019). Under distribution shift, the highest-scoring transcript may be unreliable, while beam decoding can expose plausible alternatives that differ in ambiguous characters, word boundaries, insertions, deletions, or linguistic plausibility (Kang et al., 2021; Wick et al., 2022). These candidates reveal uncertainty over complete transcriptions. Common TTA objectives instead optimize prediction entropy, confidence, or visual consistency (Wang et al., 2020; Niu et al., 2022; Ebrahimi et al., 2022). Such signals can support adaptation, but they do not explicitly optimize the relative ordering of competing decoded hypotheses. In particular, increasing prediction confidence does not establish which alternative transcription should be preferred. This raises a question: can the hypotheses already produced during decoding provide a useful supervision structure for adapting the recognizer at test time?

Language priors offer a complementary perspective from the decoding side. Language models and sequence-level rescoring have long supported OCR and HTR through constrained decoding, candidate fusion, shallow fusion, and post-recognition reranking (Scheidl et al., 2018; Kang et al., 2021; Tomeh et al., 2013; Wick et al., 2022; Tarride et al., 2024). These approaches combine visual evidence with linguistic plausibility to distinguish among alternative transcriptions. Candidate hypotheses may differ in lexical validity, skipped or repeated words, and the interpretation of visually ambiguous characters (Kang et al., 2021; Wick et al., 2022). When used solely for decoding, however, the language prior guides hypothesis exploration or selection without updating the recognizer. We investigate whether this candidate-level information can also support adaptation. Specifically, language-prior-augmented scores can establish relative preferences among image-conditioned hypotheses, making the candidate set a source of supervision for recognizer updates. The language prior thus contributes to how the recognizer scores sequences, beyond its conventional role in selecting final outputs. At the sequence level, a preference can indicate how two plausible outputs should be ordered even when neither is known to be correct. This provides a concrete target for adaptation that is aligned with the recognizer's final transcription decision.

We propose **ReRank-TTA**, a source-free TTA framework that converts decoding-time hypothesis preferences into relative supervision for autoregressive text recognition. Its core mechanism, **Hypothesis Ranking Adaptation** (HRA), connects candidate ranking with episodic recognizer adaptation. For each unlabeled target batch, the pretrained recognizer generates candidate transcriptions. We combine recognizer scores with a frozen language prior to select a preferred hypothesis, and choose the highest-scoring text-different alternative under the recognizer's decoding score as its competitor. HRA updates a temporary recognizer with a pairwise sequence-ranking loss, encouraging a likelihood margin between the preferred and competing hypotheses. Both hypotheses originate from the same input image, so the supervision remains image-conditioned. After adaptation, the temporary

recognizer produces final predictions, optionally followed by LM reranking, and is discarded before the next batch.

HRA and enhanced decoding therefore operate through complementary mechanisms: beam search explores candidate sequences, LM reranking selects among them, and HRA modifies the recognizer's image-conditioned sequence likelihoods. This distinction also allows HRA to be evaluated with greedy final decoding, where no final LM reranking is applied, or together with stronger decoding strategies. Holding the final decoding configuration fixed measures the additional benefit of recognizer updates, while the complete pipeline measures their combined effect with search and selection. We evaluate ReRank-TTA on IAM, RIMES, GW, and Bentham using TrOCR-stage1, small, base, and large (Li et al., 2023). The full pipeline consistently improves over frozen greedy baselines. TTA comparisons on TrOCR-base and matched-decoding ablations on stage1 and base assess the contribution of adaptation separately from final search and reranking. We additionally study restricted parameter updates, language-model choice, and adaptation steps. Experiments with PARSeq and PARSeq-tiny (Bautista & Atienza, 2022) on SVTP and IC15 further extend the evaluation beyond handwriting. Our contributions are:

- We formulate source-free adaptation of autoregressive text recognizers as learning relative preferences among decoded transcription hypotheses, using candidate-level uncertainty as a supervision structure.
- We propose **ReRank-TTA**, whose **HRA** objective converts preferences constructed with a frozen language prior into pairwise sequence-ranking supervision for episodic recognizer updates.
- We demonstrate improvements across four handwriting benchmarks and two scene text benchmarks using multiple recognizer checkpoints, with matched-decoding ablations separating adaptation from beam search and LM reranking.

2 Related Work

Test-time adaptation for text recognition. Recent work has explored adapting text recognition and document models at deployment time. For text-line recognition, Tula et al. (2023) use source-free self-training with language-model feedback, while other studies investigate unsupervised writer adaptation (Kang et al., 2020) or few-shot supervised personalization (Bhunia et al., 2021). Related document adaptation methods employ masked visual–language modeling or meta-auxiliary self-supervised objectives (Ebrahimi et al., 2022; Gu et al., 2025b). These approaches demonstrate the value of deployment-time adaptation under different supervision and training assumptions. Our focus is on using competing decoded transcriptions as an explicit supervision structure. ReRank-TTA adapts an autoregressive text recognizer through relative hypothesis rankings constructed from unlabeled target images and a frozen language prior, with evaluations on both handwriting and scene text recognition.

Language models for sequence recognition. Language models have long been used in OCR and HTR to improve decoding through dictionary constraints, candidate fusion, shallow fusion, and n-best reranking (Scheidl et al., 2018; Kang et al., 2021; Tomeh et al., 2013; Wick et al., 2022; Li et al., 2025a; Tarride et al., 2024). These methods are effective because HTR errors are not purely visual: candidate transcriptions may differ in lexical validity, sequence plausibility, skipped or repeated words, and visually ambiguous characters (Scheidl et al., 2018; Kang et al., 2021; Wick et al., 2022). However, most LM-guided methods use the language model at decoding or rescoring time to select among fixed hypotheses, rather than using the LM signal to update the recognizer itself (Scheidl et al., 2018; Tomeh et al., 2013; Wick et al., 2022; Tarride et al., 2024). In contrast, ReRank-TTA uses language-prior-augmented sequence scores to construct relative rankings among candidate hypotheses and updates the recognizer itself, turning the language prior into weak source-free supervision rather than only a decoder-time reranker.

Ranking from candidate hypotheses. Ranking objectives learn from relative comparisons rather than absolute labels, as in pairwise learning-to-rank (Burges et al., 2005) and preference-based language generation (Christiano et al., 2017; Rafailov et al., 2023). Our setting differs from human-preference optimization: the comparisons are automatically constructed from the recognizer's own decoded hypotheses using a frozen language prior, without reward models, reference policies, or RL-style optimization. Prior OCR/HTR work shows that n-best candidates contain useful information for candidate fusion, reranking, and LM rescoring (Kang et al., 2021; Tomeh et al., 2013; Wick et al.,

2022), but these candidates are usually used only for inference-time selection. ReRank-TTA instead converts relative preferences among candidate hypotheses into source-free updates of the recognizer.

3 PRELIMINARIES

Problem setting. We study source-free test-time adaptation (TTA) for text recognition under deployment shifts in writing style, typography, image acquisition, and language (Sun et al., 2020; Wang et al., 2020; Garrido-Munoz & Calvo-Zaragoza, 2025). Let $X \in \mathbb{R}^{H\times W\times C}$ denote a cropped text image and $y = (y_1, \ldots, y_{|y|})$ a transcription sequence over the recognizer's vocabulary V. A pretrained autoregressive recognizer with parameters θ defines $p_\theta(y \mid X) = \prod_{t=1}^{|y|} p_\theta(y_t \mid y_{<t}, X)$, as in the autoregressive decoding of TrOCR and PARSeq (Li et al., 2023; Bautista & Atienza, 2022). At deployment, neither source training data nor target transcriptions are available, but a frozen text-only language prior may be used. Incoming text crops are processed in mini-batches $\mathcal{B} = \{X_i\}_{i=1}^{N}$. For each X_i, the recognizer produces candidate transcriptions $\mathcal{Y}_K(X_i;\theta) = \{y_i^{(1)}, \ldots, y_i^{(K)}\}$, where K denotes the adaptation candidate-set size. Our implementation generates these candidates by beam search, although the formulation only requires a finite set of decoded hypotheses. Our goal is to adapt a temporary learner θ' using the unlabeled batch and frozen language prior, then decode with θ' to improve recognition. We measure performance using CER/WER for handwriting and word accuracy for scene text.

Motivation. Common TTA objectives optimize confidence- or sample-level surrogates. For example, Tent minimizes prediction entropy (Wang et al., 2020), while EATA selects reliable, non-redundant samples and regularizes important parameters to reduce forgetting (Niu et al., 2022). These objectives do not explicitly optimize the relative ordering of alternative transcription hypotheses. Language-guided decoding and n-best reranking provide complementary information by combining recognition scores with linguistic plausibility (Scheidl et al., 2018; Kang et al., 2021; Tomeh et al., 2013; Wick et al., 2022), but leave the recognizer unchanged when used solely for inference. This motivates **Hypothesis Ranking Adaptation** (HRA): we use language-prior-augmented scores to select a preferred transcription, pair it with a high-scoring competing alternative, and update the temporary learner with a pairwise sequence-ranking loss. The resulting supervision operates on complete image-conditioned hypotheses, aligning adaptation with the sequence-level decisions of autoregressive text recognition.

4 METHOD

Overview. We propose **ReRank-TTA**, a source-free test-time adaptation framework for autoregressive text recognition. At the core of ReRank-TTA is **Hypothesis Ranking Adaptation** (HRA), which constructs a preferred transcription hypothesis and a text-different competing hypothesis from the recognizer's own decoded candidates, and updates a temporary test-time learner with a sequence-level pairwise ranking loss. As summarized in Algorithm 1, HRA first builds valid hypothesis pairs using language-prior-augmented sequence scores, then adapts the recognizer episodically before final decoding. We describe hypothesis ranking construction and the HRA loss in Sec. 4.1, followed by episodic adaptation and final shallow-fusion reranking in Sec. 4.2.

4.1 HYPOTHESIS RANKING ADAPTATION

HRA is a test-time pairwise ranking objective over candidate transcription hypotheses. Let X denote a text image and let $p_\theta(y|X)$ be the sequence probability assigned by the autoregressive recognizer. At test time, no ground-truth transcription is available. ReRank-TTA first generates a finite set of candidate hypotheses from the pretrained recognizer before the inner-loop update: $\mathcal{Y}_K(X;\theta) = \{y^{(1)}, y^{(2)}, \ldots, y^{(K)}\}$, where K is the number of candidate hypotheses used for adaptation. In our implementation, these hypotheses are generated by beam search, but the formulation only requires a decoded candidate set. Hypothesis pairs are constructed once per test batch using the pretrained recognizer θ, and the resulting pairs are reused for U adaptation steps of the temporary learner θ'.

For each candidate $y \in \mathcal{Y}_K(X;\theta)$, we compute a language-prior-augmented ranking score:

$$S(y|X) = d_\theta(y|X) + \beta_{\mathrm{LM}} s_{\mathrm{LM}}(y). \quad s_{\mathrm{LM}}(y) = \frac{1}{|y|}\sum_{t=1}^{|y|} \log p_{\mathrm{LM}}(y_t \mid y_{<t}). \tag{1}$$

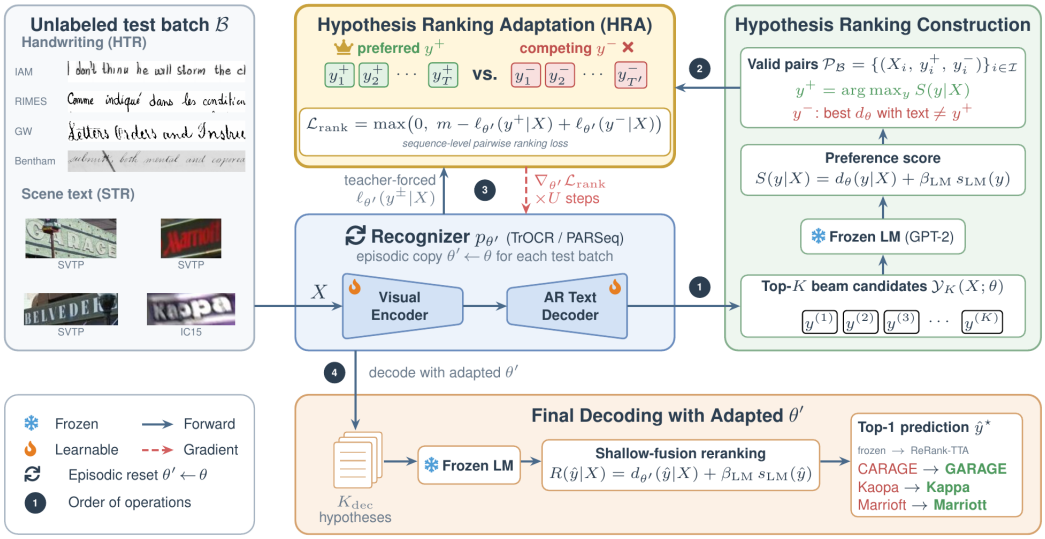

Figure 2: **Overview of ReRank-TTA. Autoregressive text recognizer:** a temporary copy of the pretrained recognizer serves as the episodic learner θ' for each unlabeled test batch. **Hypothesis ranking construction:** the pretrained recognizer generates top-K candidate hypotheses. The fused recognizer–LM score selects the preferred hypothesis y^+, while the recognizer's beam score selects the highest-scoring text-different candidate as y^-. The resulting valid pairs form the adaptation mini-batch. **HRA update:** Hypothesis Ranking Adaptation updates θ' with a sequence-level pairwise ranking loss that increases the relative likelihood of y^+ over y^-. **Final decoding:** the adapted recognizer generates final K_{dec} hypotheses, which are reranked with the frozen LM under shallow fusion to select the top-1 prediction.

Here $d_\theta(y|X)$ denotes the beam-decoding score returned by the recognizer, $s_{\text{LM}}(y)$ is the length-normalized log-likelihood under the frozen LM, and β_{LM} controls the strength of the language prior. The frozen LM is pretrained independently and is used only for scoring; no source-domain images, source labels, or target transcriptions are used during adaptation.

For brevity, we write $\mathcal{Y}_K$ for $\mathcal{Y}_K(X;\theta)$ when the image and recognizer are clear from context. The preferred hypothesis is selected as

$$y^+ = \arg\max_{y\in\mathcal{Y}_K} S(y|X). \tag{2}$$

The competing hypothesis y^- is selected as a hard text-different alternative: among candidates whose decoded text differs from y^+, we choose the one with the highest beam-decoding score $d_\theta(y|X)$.

Sequence-level ranking loss. Given y^+ and y^-, HRA updates the temporary learner using a pairwise margin objective:

$$\mathcal{L}_{\text{rank}} = \max\left(0, m - \ell_{\theta'}(y^+|X) + \ell_{\theta'}(y^-|X)\right), \quad \ell_{\theta'}(y|X) = \frac{1}{|y|}\sum_{t=1}^{|y|} \log p_{\theta'}(y_t \mid y_{<t}, X), \tag{3}$$

where $m > 0$ is the margin, and $\ell_{\theta'}(y|X)$ is the length-normalized teacher-forced log-likelihood assigned by the temporary learner to a complete candidate sequence. This objective is inspired by RankSVM-style pairwise ranking losses (Joachims, 2002). When the margin is violated, the gradient increases the recognizer's image-conditioned likelihood of y^+ relative to y^-, directly changing the sequence ordering used at inference. Both hypotheses originate from the recognizer's candidate set for the same image, while the frozen LM contributes only to selecting the preferred candidate and receives no updates.

HRA therefore uses complete-sequence comparisons as an adaptation signal rather than only for selection. In contrast to sequence-level rescoring (Wick et al., 2022), which chooses among fixed candidates, the ranking objective updates the recognizer so that its own decoding scores better reflect the constructed preference.

Hypothesis-pair validity. Valid hypothesis pairs are constructed only when the candidate set contains at least one text-different alternative to y^+. If no such alternative exists, the sample is skipped. The resulting valid pairs form a batch-level set

$$\mathcal{P}_{\mathcal{B}} = \left\{ (X_i, y_i^+, y_i^-) \,:\, i \in \mathcal{I} \right\}, \tag{4}$$

where $\mathcal{I}$ indexes samples with valid hypothesis pairs.

4.2 Episodic Adaptation and Final Decoding

At deployment, ReRank-TTA follows an episodic source-free TTA protocol. For each incoming mini-batch $\mathcal{B} = \{X_i\}_{i=1}^{N}$, we clone the pretrained recognizer parameters θ to obtain a temporary learner θ'. Hypothesis pairs are first constructed once using the pretrained recognizer, forming the batch-level valid pair set $\mathcal{P}_{\mathcal{B}}$. We then perform U inner-loop HRA updates using only the unlabeled images and valid hypothesis pairs in $\mathcal{B}$:

$$\theta' \leftarrow \theta' - \eta \nabla_{\theta'} \mathcal{L}_{\text{rank}}(\mathcal{B}; \theta'), \tag{5}$$

where η is the adaptation learning rate and U is the number of test-time adaptation steps. The batch objective averages valid pairwise losses:

$$\mathcal{L}_{\text{rank}}(\mathcal{B}; \theta') = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathcal{L}_{\text{rank}}\big(X_i, y_i^+, y_i^-; \theta'\big), \tag{6}$$

Algorithm 1: Episodic ReRank-TTA with Hypothesis Ranking Adaptation

Require: Pretrained recognizer p_θ, frozen LM p_{LM}, test batch $\mathcal{B}$, adaptation candidate-set size K, final decoding size K_{dec}, inner-loop steps U, learning rate η

for each test batch $\mathcal{B} = \{X_i\}_{i=1}^{N}$ **do**
  Initialize temporary learner $\theta' \leftarrow \theta$
  Initialize valid hypothesis-pair set $\mathcal{P}_{\mathcal{B}} \leftarrow \emptyset$
  for each image $X_i \in \mathcal{B}$ **do**
    Generate adaptation hypotheses $\mathcal{Y}_K(X_i; \theta)$
    Compute ranking score $S(y|X_i)$ for $y \in \mathcal{Y}_K(X_i; \theta)$ using Eq. (1)
    Select preferred hypothesis y_i^+ and competing hypothesis y_i^-
    Add (X_i, y_i^+, y_i^-) to $\mathcal{P}_{\mathcal{B}}$ if valid
  end for
  for $u = 1, \ldots, U$ **do**
    Compute batch loss $\mathcal{L}_{\text{rank}}(\mathcal{B}; \theta')$ on $\mathcal{P}_{\mathcal{B}}$
    $\theta' \leftarrow \theta' - \eta \nabla_{\theta'} \mathcal{L}_{\text{rank}}(\mathcal{B}; \theta')$
  end for
  for each image $X_i \in \mathcal{B}$ **do**
    Decode final hypotheses $\{\hat{y}_i^{(b)}\}_{b=1}^{K_{\text{dec}}}$ using adapted learner θ'
    Compute reranking score $R(\hat{y}_i^{(b)}|X_i)$ for $b = 1, \ldots, K_{\text{dec}}$
    Select $\hat{y}_i^\star = \arg\max_{1 \le b \le K_{\text{dec}}} R(\hat{y}_i^{(b)}|X_i)$
  end for
  Discard θ'
end for

where $\mathcal{I}$ indexes samples with valid hypothesis pairs. If no valid pair is found, the update is skipped for that batch.

Final decoding and reranking. After adaptation, each text image is decoded with the adapted learner θ'. Given final candidates $\{\hat{y}^{(1)}, \ldots, \hat{y}^{(K_{\text{dec}})}\}$, our main setting applies shallow-fusion LM reranking for final selection:

$$R(\hat{y}^{(b)}|X) = d_{\theta'}(\hat{y}^{(b)}|X) + \beta_{\text{LM}} s_{\text{LM}}(\hat{y}^{(b)}), \quad \hat{y}^\star = \arg \max_{1 \le b \le K_{\text{dec}}} R(\hat{y}^{(b)}|X). \tag{7}$$

Here s_{LM} is the same length-normalized LM score used in Eq. (1), and the same LM weight β_{LM} is used for both hypothesis ranking construction and final reranking. The frozen LM is used for scoring only and is never updated. After decoding, θ' is discarded and the next batch starts again from the pretrained θ, preventing cross-batch error accumulation.

5 Experiments

5.1 Datasets and Setup

Benchmarks and evaluation protocols. We evaluate on four handwriting corpora and two scene text benchmarks. The handwriting datasets include **GW**, comprising 18th-century English letters (Rath et al., 2004); **IAM**, containing modern English handwriting from 657 writers (Marti & Bunke, 2002); **RIMES**, comprising French mail and serving as a cross-language target (Kermorvant & Louradour, 2010); and **Bentham**, containing 19th-century English manuscripts (Causer & Wallace, 2012). We use the Aachen split for IAM and the official splits for GW, RIMES, and Bentham. To evaluate applicability beyond handwriting, we additionally test PARSeq and PARSeq-tiny (Bautista & Atienza, 2022) on **SVTP** (Quy Phan et al., 2013), featuring perspective-distorted scene text, and **IC15**, containing incidental scene text (Karatzas et al., 2015). We use a learning rate of 1×10^{-4} and a mini-batch size of 4 for all experiments. We report CER and WER for handwriting recognition and word accuracy (%) for scene text recognition.

Table 1: **Contextual comparison with specialized HTR systems.** The upper block reports published cross-dataset OOD results from Garrido-Munoz & Calvo-Zaragoza (2025); the lower block reports our reproduced TrOCR checkpoints and ReRank-TTA results under a unified evaluation protocol. For IAM-finetuned TrOCR-small/base/large, IAM is in-domain and RIMES/GW/Bentham are OOD; TrOCR-stage1 is pretrained-only before IAM fine-tuning. Because protocols differ across blocks, the upper block is used as contextual reference rather than a strict head-to-head comparison. Within each TrOCR checkpoint pair, the better result is shown in **bold**.

Model	IAM		Rimes		GW		Bentham	
	CER↓	WER↓	CER↓	WER↓	CER↓	WER↓	CER↓	WER↓
CRNN (Puigcerver, 2017)	34.9	68.2	25.0	74.2	31.1	68.8	25.3	54.9
VAN (Coquenet et al., 2022)	28.6	76.8	21.3	66.9	32.0	74.7	26.6	62.1
C-SAN (d'Arce et al., 2022)	31.5	83.6	29.8	80.7	49.8	93.9	38.9	82.2
HTR-VT (Li et al., 2025b)	33.7	83.7	28.3	78.7	38.6	83.3	33.3	70.8
Kang (Kang et al., 2022)	42.1	87.1	32.0	78.3	44.0	77.6	39.4	68.2
Michael (Michael et al., 2019)	49.1	82.9	35.5	84.5	43.6	76.1	43.5	69.6
LT (Barrere et al., 2022)	42.0	72.0	30.8	77.4	32.3	65.7	33.8	60.6
VLT (Barrere et al., 2024)	41.3	70.3	29.4	76.1	32.1	65.8	33.3	60.5
TrOCR-stage1 (Li et al., 2023)	11.78	35.86	23.32	55.30	14.61	31.66	31.77	49.63
ReRank-TTA	**7.81**	**17.09**	**17.36**	**47.14**	**13.31**	**28.53**	**29.35**	**45.59**
TrOCR-small (Li et al., 2023)	7.13	15.49	28.99	67.64	22.35	36.22	20.55	31.69
ReRank-TTA	**5.20**	**11.30**	**23.76**	**62.13**	**17.52**	**30.41**	**16.97**	**27.79**
TrOCR-base (Li et al., 2023)	5.10	11.95	27.54	65.25	16.86	29.07	17.55	28.02
ReRank-TTA	**3.49**	**8.61**	**20.90**	**58.36**	**13.19**	**25.76**	**12.63**	**23.01**
TrOCR-large (Li et al., 2023)	3.71	9.58	23.94	59.82	15.18	26.21	11.76	19.99
ReRank-TTA	**2.63**	**7.05**	**18.11**	**52.85**	**11.53**	**21.65**	**8.47**	**17.06**

Backbone recognizer. While many VLM/LLM backbones exist, few are tailored to handwriting. We therefore adopt *TrOCR* (Li et al., 2023), a handwriting-specific Transformer encoder–decoder with a ViT visual encoder and an autoregressive text decoder. As shown in Table 7, recent generic OCR/VLM systems such as Donut, GOT-OCR2, and Qwen3-VL-4B still yield higher CER/WER on IAM than a specialized TrOCR recognizer (Kim et al., 2022; Wei et al., 2024; Yang et al., 2025). This suggests that current general-purpose OCR/VLMs do not yet remove the need for specialized HTR backbones; we therefore study source-free test-time adaptation on top of TrOCR rather than replacing the recognizer with a generic OCR/VLM. In our *source-free* setting, we evaluate four public TrOCR checkpoints: **TrOCR-stage1**, the stage-1 pretrained checkpoint before downstream IAM handwriting fine-tuning, and **TrOCR-small/base/large**, which are IAM-finetuned handwriting recognizers. This setup lets us test ReRank-TTA both before and after IAM handwriting fine-tuning. Notably, ReRank-TTA further improves TrOCR-large to **2.63** CER and **7.05** WER on IAM, a highly competitive result among line-level HTR systems. No target-domain labels or source training data are used at deployment.

Language model teacher. We use a *frozen* autoregressive language model to score candidate hypotheses for HRA and shallow-fusion reranking. Unless otherwise specified, we use **GPT-2 Large**(Radford et al., 2019) and compute length-normalized left-to-right log-likelihoods as the language-prior score. The LM is never updated; gradients are applied only to the HTR recognizer. We also evaluate smaller language models, including DistilGPT-2 and GPT-2 French-small, to test sensitivity to the language prior (Table 5). Additional implementation details, reproducibility information, and limitations are provided in Appendix A.

5.2 Comparison with Specialized HTR Systems

Table 1 evaluates **ReRank-TTA** on four public TrOCR checkpoints: **stage1**, **small**, **base**, and **large**. ReRank-TTA improves every checkpoint across IAM, RIMES, GW, and Bentham. Averaged over the four datasets, it reduces CER/WER by **3.41/8.52** points for TrOCR-stage1, **3.89/4.85** for TrOCR-small, **4.21/4.64** for TrOCR-base, and **3.46/4.25** for TrOCR-large. Across all checkpoints and datasets, the average reduction is **3.74** CER and **5.57** WER, corresponding to **23.2%/17.9%** relative improvements. These results demonstrate the effectiveness of the full ReRank-TTA pipeline across

Table 2: **Comparison of test-time adaptation strategies on TrOCR-base.** CER (%, lower is better) across four HTR benchmarks. The final column reports the percentage of adapted parameters; LN/bias HRA updates only LayerNorm affine parameters and trainable biases.

TTA Methods	IAM CER↓	RIMES CER↓	GW CER↓	Bentham CER↓	Adapt. Params % updated
TrOCR-base (Li et al., 2023)	5.10	27.54	16.86	17.55	N/A
TENT-base (Wang et al., 2020)	4.77	23.92	16.28	15.71	3.4%
EATA-base (Niu et al., 2022)	4.76	26.26	15.61	14.99	3.4%
LN/bias HRA only	4.50	23.61	14.75	14.79	9.91%
HRA only	**4.36**	**23.48**	**14.65**	**14.74**	100%

Table 3: **Cross-domain results on scene text recognition.** Word accuracy (%, ↑) on SVTP and IC15. Gains over the frozen recognizer are shown in parentheses.

Method	SVTP	IC15
PARSeq	95.04	89.74
ReRank-TTA	**97.21** (+2.17)	**90.95** (+1.21)
PARSeq-tiny	92.40	88.11
ReRank-TTA	**95.35** (+2.95)	**89.50** (+1.39)

Table 4: **Ablation of ReRank-TTA components across four datasets.** We disentangle the effects of hypothesis ranking adaptation (HRA), beam-search decoding, and shallow-fusion LM reranking. Lower CER/WER indicates better performance.

Backbone	HRA	Beam	LM	IAM		RIMES		GW		Bentham	
				CER	WER	CER	WER	CER	WER	CER	WER
TrOCR-stage1	—	—	—	11.78	35.86	23.32	55.30	14.61	31.66	31.77	49.63
	—	✓	—	8.76	19.90	18.62	50.54	13.73	30.14	30.12	46.67
	✓	—	—	9.14	20.27	19.41	52.99	14.44	31.66	30.98	49.26
	—	✓	✓	8.07	17.66	17.89	48.23	**13.05**	28.73	29.71	45.91
	✓	✓	—	8.36	18.84	18.04	49.64	13.73	29.96	29.70	46.30
	✓	✓	✓	**7.81**	**17.09**	**17.36**	**47.14**	13.31	**28.53**	**29.35**	**45.59**
TrOCR-base	—	—	—	5.10	11.95	27.54	65.25	16.86	29.07	17.55	28.02
	—	✓	—	4.32	10.69	23.97	65.26	15.98	27.75	13.73	25.36
	✓	—	—	4.36	10.73	23.48	64.03	14.65	28.44	14.74	26.71
	—	✓	✓	3.70	8.96	21.55	59.66	15.13	25.94	12.86	23.40
	✓	✓	—	3.99	9.93	21.46	60.45	13.93	26.65	13.41	24.63
	✓	✓	✓	**3.49**	**8.61**	**20.90**	**58.36**	**13.19**	**25.76**	**12.63**	**23.01**

pretrained and IAM-finetuned checkpoints. The contribution of HRA under matched decoding is examined in Table 4.

Positioning against established HTR systems. The upper block of Table 1 reports representative specialized HTR systems reproduced from Garrido-Munoz & Calvo-Zaragoza (2025). Since training data, model capacity, and decoding protocols differ across these systems, we treat them as contextual reference points rather than strict SOTA comparisons. Notably, ReRank-TTA with TrOCR-large reaches **2.63/7.05** CER/WER on IAM and improves OOD performance on RIMES, GW, and Bentham to **18.11/52.85**, **11.53/21.65**, and **8.47/17.06**, respectively.

TTA strategy comparison. Table 2 compares HRA with standard TTA baselines on TrOCR-base under greedy decoding. TENT and EATA provide moderate gains, but their improvements are smaller and less consistent. Full-parameter HRA achieves the strongest overall results on TrOCR-base, suggesting that hypothesis-level ranking is a more effective TTA signal than confidence- or entropy-based updates in this setting.

Parameter-efficient update variant. Table 2 also reports a parameter-efficient *LN/bias-update-only* variant, which updates only LayerNorm affine parameters and bias terms while using the same frozen GPT-2-Large teacher. This variant updates only **9.91%** of TrOCR-base parameters, yet remains close to full-parameter HRA under greedy decoding: the CER gaps are only **0.14**, **0.13**, **0.10**, and **0.05** points on IAM, RIMES, GW, and Bentham, respectively. These results show that the HRA signal can support conservative and deployment-friendly adaptation.

Table 5: **External LM variants**. †RIMES uses GPT-2 French-small; the other datasets use DistilGPT-2. GPT-2-large is our final choice.

LM Teacher	IAM		RIMES		GW		Bentham	
	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓
Baseline w/o LM	11.78	35.86	23.32	55.30	14.61	31.66	31.77	49.63
DistilGPT-2 / GPT-2 French-small†	8.26	18.12	17.58	47.44	13.98	29.25	29.77	45.99
GPT-2-large	**7.81**	**17.09**	**17.36**	**47.14**	**13.31**	**28.53**	**29.35**	**45.59**

Table 6: **Inner-step ablation on TrOCR-stage1.** We evaluate ReRank-TTA with different numbers of HRA updates. Lower CER/WER is better.

Inner Steps	IAM		RIMES		GW		Bentham	
	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓
1 step	7.94	17.43	17.51	47.55	13.33	28.59	29.64	45.81
3 steps	7.89	17.23	17.48	47.28	13.33	28.62	29.48	45.65
5 steps	**7.81**	**17.09**	**17.36**	**47.14**	**13.31**	**28.53**	**29.35**	**45.59**

Generalization to scene text recognition. Table 3 extends our evaluation beyond handwriting. ReRank-TTA improves PARSeq word accuracy from 95.04% to 97.21% on SVTP and from 89.74% to 90.95% on IC15. For PARSeq-tiny, the corresponding gains are 2.95 and 1.39 percentage points. These consistent improvements across both benchmarks and model sizes support the applicability of ReRank-TTA to scene text recognition.

5.3 Ablation Studies

Ablations: HRA, beam search, and LM reranking. Table 4 separates adaptation from final decoding. With either greedy or beam decoding, HRA improves all eight metrics for TrOCR-base and seven of eight for TrOCR-stage1, with the remaining metric unchanged. Adding HRA to beam search with LM reranking improves all eight metrics for TrOCR-base and seven of eight for TrOCR-stage1; for the remaining stage1 metric, GW CER increases from 13.05 to 13.31. These results support complementary roles for HRA and enhanced decoding while making the single stage1 degradation explicit.

Language-model choice for ReRank-TTA. Table 5 studies the external LM used for hypothesis ranking construction and final reranking. Besides GPT-2 Large, we evaluate compact teachers: GPT-2 French-small for RIMES and DistilGPT-2 for IAM, GW, and Bentham. These compact LMs already improve over the no-LM baseline on all four datasets, indicating that ReRank-TTA does not rely exclusively on a large teacher model. GPT-2 Large gives the strongest overall performance and is therefore used as the default teacher.

Effect of inner-loop update steps (U). Table 6 reports CER/WER for different numbers of inner-loop HRA updates using the **TrOCR-stage1** backbone. A few adaptation steps are sufficient to obtain most of the gains: performance generally improves from one to three or five steps, while improvements between consecutive step counts are small and dataset-dependent. Five steps gives the strongest overall results across all four datasets and is used as the default. This suggests that ReRank-TTA provides a stable test-time adaptation signal without requiring many update steps.

Discussion. The ablations support complementary roles for adaptation and decoding. Beam search explores candidate sequences, LM reranking selects among them, and HRA updates the recognizer's sequence likelihoods. The full method improves over decoding-only variants in most evaluated settings. The LM and inner-step experiments further show that improvements are retained with compact language priors and a small number of adaptation steps. *See Appendix Fig. 3 for qualitative examples; Appendix B provides additional experiments on generic OCR/VLM baselines and cross-domain robustness.*

6 Conclusion

We introduced **ReRank-TTA**, a source-free test-time adaptation framework for autoregressive text recognition. Its HRA objective converts preferences among candidate transcriptions into pairwise supervision for episodic recognizer updates. Experiments across four handwriting benchmarks and multiple TrOCR checkpoints show consistent improvements, while matched-decoding ablations isolate the additional contribution of adaptation. Results with PARSeq and PARSeq-tiny on SVTP and IC15 further support applicability beyond handwriting.

REFERENCES

Killian Barrere, Yann Soullard, Aurélie Lemaitre, and Bertrand Coüasnon. A light transformer-based architecture for handwritten text recognition. In *International Workshop on Document Analysis Systems*, pp. 275–290. Springer, 2022.

Killian Barrere, Yann Soullard, Aurélie Lemaitre, and Bertrand Coüasnon. Training transformer architectures on few annotated data: an application to historical handwritten text recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 27(4):553–566, 2024.

Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pp. 178–196. Springer, 2022.

Ayan Kumar Bhunia, Shuvozit Ghose, Amandeep Kumar, Pinaki Nath Chowdhury, Aneeshan Sain, and Yi-Zhe Song. Metahtr: Towards writer-adaptive handwritten text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15830–15839, 2021.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Tim Causer and Valerie Wallace. Building a volunteer community: results and findings from transcribe bentham. *Digital Humanities Quarterly*, 6(2), 2012.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Denis Coquenet, Clément Chatelain, and Thierry Paquet. End-to-end handwritten paragraph text recognition using a vertical attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):508–524, 2022.

Rafael d'Arce, Terence Norton, Sion Hannuna, and Nello Cristianini. Self-attention networks for non-recurrent handwritten text recognition. In *International Conference on Frontiers in Handwriting Recognition*, pp. 389–403. Springer, 2022.

Sayna Ebrahimi, Sercan O Arik, and Tomas Pfister. Test-time adaptation for visual document understanding. *arXiv preprint arXiv:2206.07240*, 2022.

Carlos Garrido-Munoz and Jorge Calvo-Zaragoza. On the generalization of handwritten text recognition models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15275–15286, 2025.

Wenhao Gu, Li Gu, Ching Yee Suen, and Yang Wang. Metawriter: Personalized handwritten text recognition using meta-learned prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23494–23504, 2025a.

Wenhao Gu, Li Gu, Ziqiang Wang, Ching Yee Suen, and Yang Wang. Docttt: Test-time training for handwritten document recognition using meta-auxiliary learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1904–1913. IEEE, 2025b.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, 2002.

Lei Kang, Marçal Rusinol, Alicia Fornés, Pau Riba, and Mauricio Villegas. Unsupervised writer adaptation for synthetic-to-real handwritten word recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3502–3511, 2020.

Lei Kang, Pau Riba, Mauricio Villegas, Alicia Fornés, and Marçal Rusiñol. Candidate fusion: Integrating language modelling into a sequence-to-sequence handwritten word recognition architecture. *Pattern Recognition*, 112:107790, 2021.

Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. Pay attention to what you read: non-recurrent handwritten text-line recognition. *Pattern Recognition*, 129:108766, 2022.

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)*, pp. 1156–1160. IEEE, 2015.

Christopher Kermorvant and Jérôme Louradour. Handwritten mail classification experiments with the rimes database. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, pp. 241–246. IEEE, 2010.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Hongliang Li, Dezhi Peng, and Lianwen Jin. Ego-lm: An efficient, generic, and out-of-the-box language model for handwritten text recognition. *Pattern Recognition*, 159:111130, 2025a.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5743–5762, 2024.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 13094–13102, 2023.

Yuting Li, Dexiong Chen, Tinglong Tang, and Xi Shen. Htr-vt: Handwritten text recognition with vision transformer. *Pattern Recognition*, 158:110967, 2025b.

U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5(1):39–46, 2002.

Johannes Michael, Roger Labahn, Tobias Grüning, and Jochen Zöllner. Evaluating sequence-to-sequence models for handwritten text recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1286–1293. IEEE, 2019.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Joan Puigcerver. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pp. 67–72. IEEE, 2017.

Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE international conference on computer vision*, pp. 569–576, 2013.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Toni M Rath, R Manmatha, and Victor Lavrenko. A search engine for historical manuscript images. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 369–376, 2004.

Harald Scheidl, Stefan Fiel, and Robert Sablatnig. Word beam search: A connectionist temporal classification decoding algorithm. In *2018 16th International conference on frontiers in handwriting recognition (ICFHR)*, pp. 253–258. IEEE, 2018.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Solène Tarride, Yoann Schneider, Marie Generali-Lince, Mélodie Boillet, Bastien Abadie, and Christopher Kermorvant. Improving automatic text recognition with language models in the pylaia open-source library. In *International Conference on Document Analysis and Recognition*, pp. 387–404. Springer, 2024.

Nadi Tomeh, Nizar Habash, Ryan Roth, Noura Farra, Pradeep Dasigi, and Mona Diab. Reranking with linguistic and semantic features for arabic optical character recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 549–555, 2013.

Debapriya Tula, Sujoy Paul, Gagan Madan, Peter Garst, Reeve Ingle, and Gaurav Aggarwal. Is it an i or an l: Test-time adaptation of text line recognition models. *arXiv preprint arXiv:2308.15037*, 2023.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.

Christoph Wick, Jochen Zöllner, and Tobias Grüning. Rescoring sequence-to-sequence models for text line recognition with ctc-prefixes. In *International Workshop on Document Analysis Systems*, pp. 260–274. Springer, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Table 7: **Generic OCR/VLM baselines vs. specialized HTR on IAM.** We report CER/WER (lower is better) on the IAM test set. Generic OCR/VLM systems remain less effective than a specialized TrOCR backbone for line-level handwritten text recognition.

	TrOCR-base (Li et al., 2023)		**Donut** (Kim et al., 2022)		**GOT-OCR2** (Wei et al., 2024)		**Qwen3-VL-4B** (Yang et al., 2025)	
	CER↓	WER↓	CER↓	WER↓	CER↓	WER↓	CER↓	WER↓
IAM	**5.10**	**11.95**	28.00	44.00	60.76	80.24	7.68	31.73

APPENDIX

A REPRODUCIBILITY, LIMITATIONS, AND DISCUSSION

Implementation details. All experiments follow publicly documented preprocessing and official dataset splits: the Aachen split for IAM and the official splits for GW, RIMES, and Bentham. We use released public TrOCR checkpoints and a frozen autoregressive language model as the sequence scorer. Unless otherwise specified, ReRank-TTA uses a learning rate of 1×10^{-4}, mini-batch size 4, five inner-loop adaptation steps, a final beam size of 5, and shallow-fusion LM reranking for final selection. Candidate hypotheses for HRA are generated by beam decoding and scored with the same frozen LM used for final reranking. Experiments are run on NVIDIA H100 MIG GPU instances with 20GB GPU memory, 4 CPU cores, and 80GB system memory per job. CER/WER are computed with a unified evaluation script across all methods. No source-domain images, source labels, or target transcriptions are used during adaptation. We will release the source code upon acceptance. All datasets, checkpoints, and external models are public research assets and are cited in the paper. We do not redistribute modified versions of these datasets.

Limitations and discussion. ReRank-TTA is source-free with respect to recognition training data, but it relies on a frozen external language model as a scoring prior. The LM is never updated and does not access target labels, yet its coverage and biases can affect hypothesis selection, especially for low-resource languages, minority scripts, historical spelling variants, or domains poorly represented in the LM pretraining data. Another limitation is that HRA can only rank among hypotheses generated by the base recognizer; if the candidate set does not contain a plausible transcription, the ranking loss cannot recover it. Performance therefore still depends on the out-of-distribution robustness of the underlying recognizer. ReRank-TTA is not tied to a specific autoregressive recognizer, as long as the model can generate candidate hypotheses and provide sequence likelihoods for adaptation. Future work includes using stronger or more multilingual handwriting recognizers, domain-adapted language priors, and more efficient candidate generation to improve coverage and reduce test-time cost.

B ADDITIONAL EXPERIMENTS

Generic OCR/VLM baselines vs. specialized HTR. Table 7 compares representative generic OCR/VLM baselines with a specialized TrOCR recognizer on IAM line-level HTR. Despite recent progress in general-purpose OCR/VLM models, they remain less effective than a dedicated HTR backbone in this setting: TrOCR-base obtains $5.10/11.95$ CER/WER, whereas Donut and GOT-OCR2 degrade substantially, and Qwen3-VL-4B reaches $7.68/31.73$. This supports our choice of studying source-free adaptation on top of a specialized HTR recognizer rather than replacing it with a generic OCR/VLM.

Cross-domain robustness and source selection. Tables 8 and 9 summarize cross-domain HTR results reported by Garrido-Munoz & Calvo-Zaragoza (2025). Table 8 compares in-distribution (ID) and out-of-distribution (OOD) WER across several HTR architectures, where each OOD result uses the best-performing source domain for the target. OOD error rates remain substantially higher than ID results, highlighting the difficulty of deployment under domain shift. Several model–dataset pairs marked with * did not converge in the ID setting and are excluded from the reported averages.

Table 9 further shows that the best source domain varies across targets and architectures. This variability suggests that cross-domain transfer is not uniform and motivates source-free TTA methods that can adapt at deployment without assuming access to the original source data.

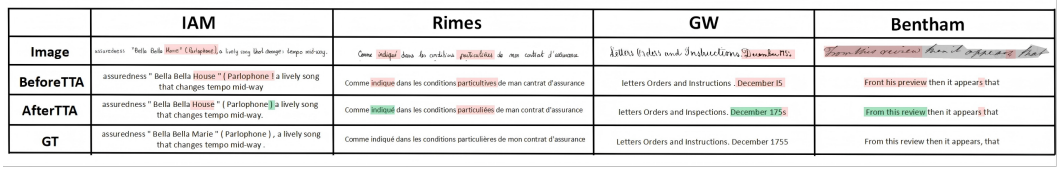

Figure 3: **Qualitative effects of ReRank-TTA across datasets.** Examples from **IAM**, **RIMES**, **GW**, and **Bentham**. Rows show the input line, transcript *before* and *after* adaptation, and ground truth (GT). Green highlights indicate characters corrected by ReRank-TTA; Red highlights mark remaining errors.

Table 8: **ID and OOD WER (%) for HTR models across datasets.** The OOD result uses the best-performing source domain for each target. Results marked with * indicate non-converged ID runs and are excluded from the average. All values are reproduced from Garrido-Munoz & Calvo-Zaragoza (2025).

Dataset	CRNN (Puigcerver, 2017)		VAN (Coquenet et al., 2022)		C-SAN (d'Arce et al., 2022)		HTR-VT (Li et al., 2025b)		Kang (Kang et al., 2022)		Michael (Michael et al., 2019)		LT (Barrere et al., 2022)		VLT (Barrere et al., 2024)	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
IAM	22.4	68.2	24.2	76.8	50.7	83.6	18.2	83.7	23.2	87.1	20.2	82.9	23.4	72.0	27.0	70.3
RIMES	11.5	74.2	18.2	66.9	45.1	80.7	24.5	78.7	14.8	78.3	20.0	84.5	13.2	77.4	13.3	76.1
GW	26.0	68.8	31.2	74.7	33.8	93.9	71.7*	83.3	104.3*	77.6	80.1*	76.1	195.4*	65.7	51.4*	65.8
Bentham	13.2	54.9	20.5	62.1	30.2	82.2	24.6	70.8	19.8	68.2	19.8	69.6	14.3	60.6	15.5	60.5

Qualitative examples. Fig. 3 illustrates qualitative improvements from ReRank-TTA on IAM, RIMES, GW, and Bentham. Rows show the input line, the transcript *before* and *after* adaptation, and the ground truth (GT). Across cases, ReRank-TTA corrects candidate-level errors such as word-internal characters, accents, historical numerals, punctuation, and visually confusable words. These examples are consistent with the role of HRA: the frozen LM helps identify more plausible hypotheses among the recognizer's candidates, while the ranking loss adapts the recognizer toward the preferred hypothesis without using target labels.

C Test-Time Adaptation Variants for TrOCR

TENT. TENT was originally designed for classification-style prediction, so we adapt it to TrOCR's encoder–decoder setting as a token-level entropy minimization baseline. For each unlabeled test batch, we decode candidate transcripts and apply a brief teacher-forced update using token-level confidence objectives. To keep the baseline stable, updates are restricted to LayerNorm affine parameters and decoder bias terms, and adaptation is performed episodically with no cross-batch accumulation.

EATA. We implement EATA as an entropy-based TrOCR adaptation baseline with an additional regularization term that keeps the model close to its pre-adaptation state. Following the same episodic protocol as the other TTA baselines, only a small subset of normalization and bias parameters is updated, and no source data or target labels are used. After adaptation, the model is decoded normally for evaluation.

Table 9: **Best source domain for each target dataset and architecture.** All values are reproduced from Garrido-Munoz & Calvo-Zaragoza (2025).

Dataset	CRNN (Puigcerver, 2017)	VAN (Coquenet et al., 2022)	C-SAN (d'Arce et al., 2022)	HTR-VT (Li et al., 2025b)	Kang (Kang et al., 2022)	Michael (Michael et al., 2019)	LT (Barrere et al., 2022)	VLT (Barrere et al., 2024)
IAM	Bentham	RIMES	RIMES	RIMES	RIMES	Bentham	Bentham	Bentham
RIMES	IAM	IAM	IAM	IAM	IAM	IAM	IAM	IAM
GW	IAM	Bentham	IAM	Bentham	IAM	IAM	IAM	IAM
Bentham	IAM	IAM	IAM	IAM	IAM	IAM	IAM	IAM