electronics

MDPI

Article

From Scene to Glyph: Hierarchical Retrieval for Multi-Modal In-Context Scene Text Recognition

Zhengwei Hu, Yanchun Jiang and Dajian Zhong *

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
* Correspondence: djzhong@shmtu.edu.cn

Abstract

Scene text recognition (STR) has benefited from vision–language modeling and multi-modal in-context learning, which enables models to exploit annotated demonstrations during inference without parameter updating. However, existing in-context STR methods mainly select demonstrations according to global visual similarity. Such image-level representations jointly encode textual content, font appearance, background, degradation, and geometric deformation, making it difficult to distinguish appearance compatibility from character-level relevance. As a result, globally similar samples may fail to provide effective evidence for ambiguous characters. To address this limitation, we propose Scene-to-Glyph, a hierarchical coarse-to-fine retrieval framework for multi-modal in-context scene text recognition. At the coarse-grained stage, an appearance-oriented scene representation is extracted from visual features to characterize shared conditions such as font style, degradation, deformation, and background appearance while reducing the influence of textual content. Scene-level similarity is then used to retrieve a compact set of appearance-compatible contextual candidates. At the fine-grained stage, uncertain character positions are identified from the recognition output, and decoder–vision correspondence is exploited to derive localized glyph representations. These representations are matched with character-level features from the candidate demonstrations, enabling the model to select demonstrations that are both scene-compatible and relevant to ambiguous characters. Experiments on conventional and challenging STR benchmarks validate the effectiveness of Scene-to-Glyph and the complementary roles of scene-level and glyph-level retrieval.

Keywords: scene text recognition;multi-modal in-context learning; hierarchical retrieval; scene-level representation; glyph-level matching

Check for updates

Academic Editor: Farhan Riaz

Received: 10 September 2026
Revised: 19 October 2026
Accepted: 21 October 2026
Published: 27 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Scene text recognition (STR) has achieved substantial progress, but robust recognition in complex visual conditions remains challenging. Text instances in natural scenes often exhibit large variations in font style, orientation, illumination, geometric deformation, image quality, and background appearance. These variations can significantly affect character representations and increase recognition ambiguity. Existing methods have addressed these challenges from different perspectives. ABINet [1] introduces bidirectional language modeling to refine ambiguous visual predictions. SVTR [2] models character dependencies with a purely visual architecture, while PARSeq [3] improves sequence modeling through permutation-based autoregressive prediction. More recent studies have explored multi-granularity prediction, character-level representation learning, compact visual representations, and vision-language pre-training [4–6]. Despite these advances, real-world

STR remains difficult because conventional benchmarks do not fully represent the visual diversity encountered in practical scenarios [7].

Multi-modal in-context learning provides a new paradigm for adapting STR models without updating their parameters. Unlike conventional adaptation methods that require additional training or fine-tuning, in-context learning allows a model to exploit a small number of annotated image-text demonstrations during inference. Zhao et al. introduced E^{2}STR to incorporate this capability into scene text recognition [8]. E^{2}STR first learns basic recognition ability and is subsequently trained with context-rich sequences. During inference, demonstrations are retrieved from an in-context pool and combined with the test sample. This design enables the recognizer to exploit contextual information without scenario-specific parameter optimization. Recent studies on multi-modal in-context learning have also shown that demonstration selection can strongly influence prediction quality [9,10]. Therefore, effective demonstration retrieval becomes an important component of multi-modal in-context STR.

Existing demonstration retrieval in E^{2}STR [8] is limited by its reliance on a single global image-level representation. E^{2}STR averages the visual token sequence to obtain an image embedding and retrieves demonstrations according to cosine similarity. This strategy is simple, computationally efficient, and already provides useful in-context evidence, as confirmed by our controlled ablation study (Section 4.4). The limitation of this strategy is therefore not that global similarity is harmful, but that it entangles textual identity and visual appearance into a single image-level criterion. In particular, the resulting representation may encode font style, background characteristics, degradation, geometric deformation, and word structure at the same time. Consequently, a high image-level similarity does not necessarily indicate that a demonstration is useful for the current recognition difficulty. This issue is particularly important because STR errors are often concentrated at a small number of ambiguous character positions. A globally similar image may provide little evidence for these characters, whereas another image with lower global similarity may contain a highly relevant local character pattern. Therefore, global image similarity alone cannot fully characterize contextual relevance. Several demonstration-selection methods have been proposed for multi-modal in-context learning, but they operate at granularity levels that do not resolve this issue. Learned retrievers select complete examples through model-derived supervision [11], Partial2Global formulates visual demonstration selection as a global ranking problem over complete examples [12], and TACO configures task-aware multi-modal in-context sequences at the example level [13]. EASE adopts a coarse-to-fine scheme whose local stage matches patch-level features [14]. None of these methods decomposes relevance into scene-level appearance compatibility and character-level evidence anchored at decoder-identified uncertain positions. Because STR errors concentrate at a few ambiguous characters, example-level or patch-level relevance cannot guarantee the character-level evidence that the recognizer actually needs. Figure 1 illustrates this discrepancy and motivates a hierarchical retrieval strategy that considers contextual relevance at different visual granularities.

In this work, we propose Scene-to-Glyph, a hierarchical coarse-to-fine retrieval framework for multi-modal in-context scene text recognition. The framework replaces single-stage global retrieval with two successive matching stages: scene-level coarse retrieval selects appearance-compatible candidates from the complete in-context pool using an appearance-oriented representation, and glyph-level fine retrieval matches decoder-derived local glyph representations of uncertain character positions against character-level features of the candidates. The proposed method is evaluated on both conventional and challenging STR benchmarks, together with ablation experiments to examine the effectiveness of the pro-

posed hierarchical retrieval strategy. The main contributions of this work are summarized as follows:

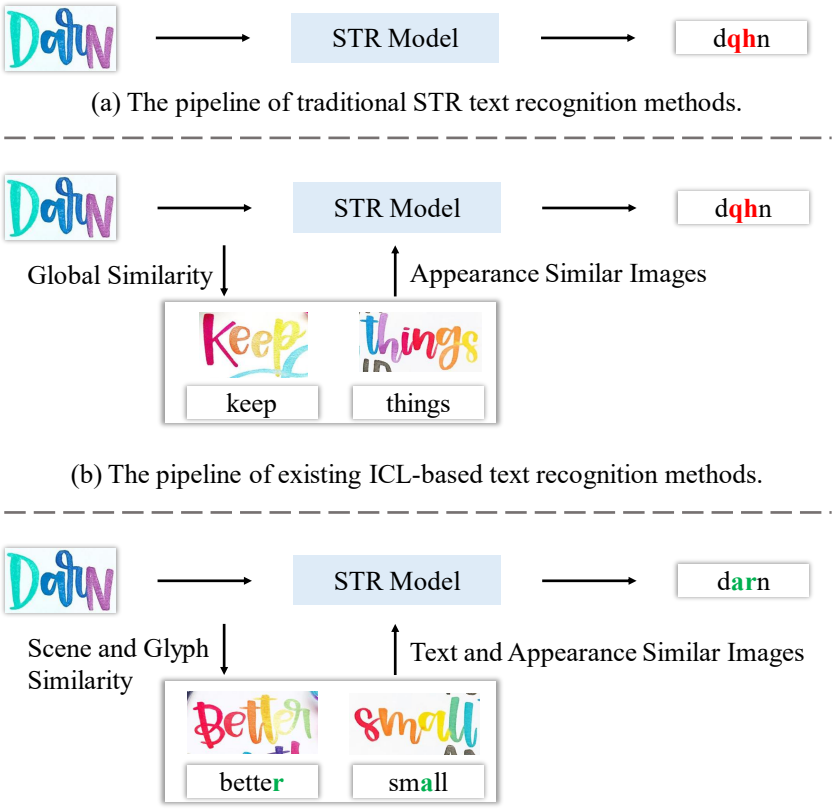


Figure 1. Motivation of the proposed Scene-to-Glyph retrieval strategy. Conventional global retrieval selects demonstrations according to image-level visual similarity, which may not provide relevant local evidence for hard characters. Scene-to-Glyph first retrieves appearance-compatible candidates and then performs glyph-level matching to identify demonstrations with relevant character evidence.

- We propose Scene-to-Glyph, a multi-modal in-context STR framework that retrieves demonstrations hierarchically: appearance-compatible candidates are selected first, and then candidates are re-ranked by their local evidence for uncertain character positions.
- We design an uncertainty-guided glyph-level matching procedure in which the glyph representation is obtained from decoder–vision attention at the two least confident positions.
- We introduce a relation-aware context-rich training strategy that constructs appearance-consistent, character-related, and hybrid contextual sequences.
- We conduct extensive experiments on conventional and challenging scene text recognition benchmarks. Experimental results demonstrate the effectiveness of our proposed method across regular, irregular, occluded, and highly stylized text.

2. Related Work

2.1. Scene Text Recognition

Scene text recognition has evolved from sequence-based visual recognition toward stronger visual, linguistic, and cross-modal modeling. Early end-to-end methods formulated text recognition as image-to-sequence prediction. CRNN [15] combined convolutional features with recurrent sequence modeling and CTC decoding. Rectification-based methods were later developed to handle perspective and curved text. RARE [16] introduced a spatial transformation module before recognition, while ASTER [17] employed a flexible

thin-plate-spline rectification network. MORAN [18] further introduced multi-object rectification for irregular text. Attention-based methods also improved alignment between visual regions and output characters. SAR [19] used two-dimensional attention for irregular text, while Baek et al. [20] established a unified evaluation framework for major STR components. Semantic information was subsequently incorporated into the recognition process. SRN [21] modeled global semantic dependencies, whereas RobustScanner [22] combined contextual and positional cues for robust decoding. ABINet [1] introduced an autonomous bidirectional language model for iterative refinement. VisionLAN [23] instead integrated linguistic information directly into visual representations. These studies progressively reduced the dependence on purely local visual cues.

Recent STR research has focused on stronger representations, scalable architectures, and more challenging real-world distributions. Baek et al. [24] investigated training with real data and reduced annotation requirements. SVTR [2] demonstrated that a purely visual Transformer can effectively model inter-character dependencies. PARSeq [3] unified autoregressive and non-autoregressive inference through permutation language modeling. MATRN [25] further enabled interactive enhancement between visual and semantic representations. MGP-STR [4] exploited predictions at multiple semantic granularities. Jiang et al. [7] proposed that conventional benchmarks cannot fully represent the difficulty of real-world STR. CLIP4STR [26] later transferred pre-trained vision-language knowledge to text recognition. Large-scale studies also examined scaling behavior in STR [27]. Recent models have further improved decoder scaling, linguistic guidance, and CTC-based sequence modeling [28–30]. Diffusion-based synthesis has also been explored to generate realistic character-aware training images [31]. Although these advances have substantially improved recognition performance, they primarily focus on representation learning and recognition modeling, while the selection of informative demonstrations for in-context inference remains underexplored.

2.2. Multi-Granularity Representation Learning for Scene Text Recognition

Multi-granularity representation learning aims to preserve complementary information at global, regional, and character levels. Global features provide compact descriptions of complete text instances, but they may suppress local character differences. Character-level modeling addresses this limitation by explicitly preserving localized structures. CCD [5] learns character-to-character correspondences under flexible image transformations. MGP-STR [4] exploits several prediction granularities to improve recognition robustness. OTE [6] follows a different direction and compresses the complete text image into one global token. Its results show that global representations can retain rich text semantics when they are properly learned. Attention-based recognition also provides implicit regional representations. Focusing Attention reduces attention drift by introducing explicit supervision for character alignment [32]. These studies indicate that global and local representations encode different aspects of a text image.

Disentangled representation learning further separates textual identity from visual appearance. This distinction is important because the same character sequence can appear under very different fonts, backgrounds, colors, and geometric transformations. DARLING [33] explicitly disentangles text content and style for recognition, removal, and editing. MATRN [25] also demonstrates that visual and semantic representations provide complementary information. Recent character-aware synthesis methods preserve glyph structure while controlling contextual appearance [31]. These findings support the use of multiple representation levels in STR. However, existing multi-granularity methods mainly improve feature learning or decoding. They do not use global appearance and local character evidence as successive criteria for demonstration retrieval. This distinction moti-

vates the representation design adopted in Scene-to-Glyph. The character-level localization capability of CCD and the global-token formulation of OTE provide two complementary precedents for this separation.

2.3. Multi-Modal In-Context Learning

In-context learning enables a model to adapt its predictions through demonstrations without updating model parameters. Recent studies showed that few-shot predictions can be sensitive to demonstration content [34]. This paradigm was later extended to vision-language tasks. Frozen language models demonstrated multi-modal few-shot learning through interleaved visual and textual inputs [35]. Flamingo [36] further established a scalable architecture for multi-modal few-shot adaptation. More recent models have strengthened multi-modal in-context capabilities through larger generative architectures. Emu2 [37] showed that generative multi-modal models can acquire task-agnostic in-context behavior. Link-Context Learning [38] explored explicit relations between support examples and queries. Visual In-Context Prompting also extended contextual prompting to open-set visual tasks [39]. These studies establish demonstrations as an important inference-time information source.

Multi-modal in-context learning has recently been extended to STR and has also received increased attention in large multi-modal models. E^2STR [8] trains a regular-sized recognizer with context-rich image-text sequences. During inference, it retrieves demonstrations from a context pool and performs recognition without target-domain fine-tuning. Its experiments confirm that a small number of annotated examples can improve recognition in unseen domains. Recent multi-modal studies have investigated how models use visual demonstrations. SymDPO [40] strengthens the dependence on demonstration images through preference optimization. MimIC [41] models the hidden-state shifts induced by multi-modal demonstrations. CoMM [42] develops coherent interleaved image-text data to strengthen multi-modal contextual learning. ImageGen-CoT [43] extends contextual reasoning to text-to-image generation. Recent analyses have also identified limitations in cross-modal task mapping and transfer [44]. These studies indicate that multi-modal ICL depends on both demonstration content and demonstration organization. In STR, this dependence makes demonstration retrieval a central problem.

2.4. Demonstration Retrieval for In-Context Learning

Demonstration retrieval determines which examples are exposed to an in-context model and can strongly affect its prediction behavior. Early methods commonly relied on random sampling or semantic similarity. Learned retrieval was later introduced to identify examples that better support a query. Rubin et al. [11] trained a dense retriever using model-derived supervision. Active selection has also been formulated as a sequential decision problem [45]. Diversity provides another important selection criterion. Levy et al. [46] showed that diverse demonstrations can improve compositional generalization. More recent studies have reconsidered the relationship between a query, a model, and its demonstrations. Peng et al. [47] showed that effective selection can depend on both data and model characteristics. Partial2Global [12] formulates visual in-context demonstration selection as a global ranking problem. Reinforcement learning has also been used to jointly optimize demonstration relevance and diversity [10]. Multi-modal selection studies further show that visual and textual evidence can contribute differently to demonstration usefulness [9]. These results consistently indicate that nearest-neighbor similarity is not a complete measure of contextual utility.

Recent work has moved demonstration selection toward finer and more structured relevance modeling, but STR-specific hierarchical retrieval remains underexplored. Princi-

pled analyses of multi-modal demonstrations show that different modalities and example configurations affect contextual inference [48]. Learning-based selection methods also reveal that useful similarity can arise at several input and output levels [49]. TACO [13] further studies task-aware configuration of multi-modal in-context sequences. These methods mainly operate at the complete-example level. They do not explicitly decompose a text image into global appearance and local glyph evidence. Coarse-to-fine retrieval has been effective in other visual problems. EASE [14] first narrows the search space with global prototypes and then performs local feature retrieval. However, its local matching is defined at the patch or pixel level. It is not linked to character ambiguity in text recognition.

To position Scene-to-Glyph against these methods, Table 1 compares representative demonstration-selection approaches in terms of retrieval granularity, uncertainty awareness, the use of a learned retriever, and the inference cost described in the original studies. E2STR performs single-stage global image-level retrieval without a learned retriever; Partial2Global and TACO operate at the complete-example level; and EASE performs coarse-to-fine matching whose local stage is defined at the patch level. In contrast, Scene-to-Glyph combines scene-level appearance retrieval with uncertainty-guided glyph-level matching, which is specifically tailored to the character-level ambiguity of STR. To the best of our knowledge, coarse-to-fine demonstration retrieval has not been previously explored in the context of OCR or scene text recognition; the closest paradigm in this field is rectification-then-recognition pipelines such as RARE and ASTER [16,17], which are processing pipelines rather than retrieval mechanisms.

Table 1. Comparison of demonstration-selection approaches. "Not stated" means the description above does not specify the property.

Method	Retrieval Granularity	Uncertainty-Aware	Learned Retriever	Inference Cost (as Described)
E^2STR [8]	Image level (average-pooled visual tokens)	No	No (cosine similarity)	Efficient; single global search
Rubin et al. [11]	Example level	No	Yes (dense retriever)	Not stated
Partial2Global [12]	Example level (global ranking)	No	Not stated	Not stated
Wang et al. [10]	Example level	No	Yes (reinforcement learning)	Not stated
Chen et al. [9]	Example level	Not stated	Not stated	Not stated
TACO [13]	Sequence level (task mapping)	No	Not stated	Not stated
EASE [14]	Patch or pixel level, after prototype search	No	Not stated	Not stated
S2G (ours)	Scene level, then glyph level	Yes (two lowest confidence positions)	Yes (scene projector)	Extra initial decoding and glyph matching

It is also useful to distinguish three related but different lines of research. Representation learning for STR (e.g., CCD, MGP-STR, OTE, and DARLING [4–6,33]) improves the features used for recognition but does not address which demonstrations should be retrieved. Demonstration retrieval (e.g., [11,12,14]) determines which candidates are selected from a pool. In-context selection and organization (e.g., TACO, CoMM, and SymDPO [13,40,42]) studies how a model consumes organized demonstrations. Scene-to-Glyph belongs to the second line but is deliberately coupled to the third: its glyph-level stage is driven by the decoding uncertainty of the recognizer, so retrieval is directly conditioned on how the in-context model uses the demonstrations.

3. Materials and Methods

3.1. Framework Overview

The proposed Scene-to-Glyph (S2G) framework improves demonstration retrieval for multi-modal in-context scene text recognition. As illustrated in Figure 2, the overall framework contains a context-rich training stage and a hierarchical inference stage. During the training stage, context-rich sequences are constructed according to appearance-level and character-level relations. Appearance-consistent samples preserve similar visual conditions across different textual contents, while character-related samples introduce local correspondence across different words. These samples are integrated into hybrid context-rich sequences for autoregressive in-context learning. The recognizer is optimized with autoregressive decoding. It enables the model to exploit both visual inputs and contextual demonstrations. During the inference stage, an annotated context pool is provided for each query. S2G retrieves demonstrations through two successive stages. Scene-level retrieval first selects candidates with similar appearance conditions. These conditions include font style, degradation, and geometric deformation. Glyph-level retrieval then evaluates local visual evidence at uncertain character positions. The final demonstrations are combined with the query for context-conditioned recognition.

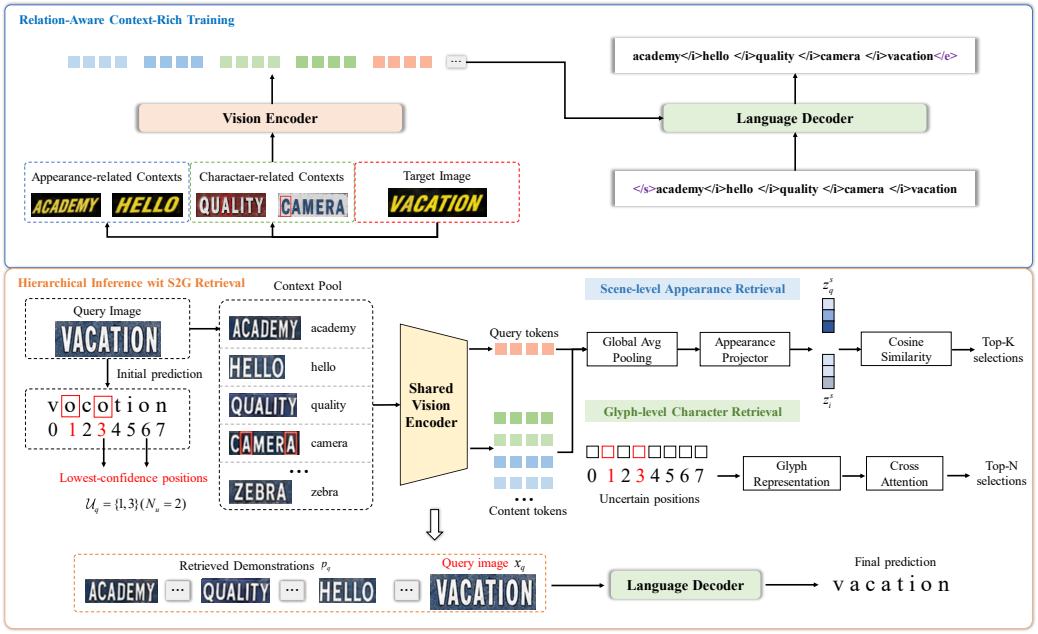


Figure 2. Overall framework of the proposed Scene-to-Glyph method. During training, appearance-consistent and character-related samples are integrated into context-rich sequences. During inference, scene-level retrieval first selects appearance-compatible candidates from the context pool. Glyph-level retrieval further evaluates local visual relevance at uncertain character positions. The selected demonstrations are finally combined with the query for in-context recognition.

Regarding component responsibilities, the overall training process of Scene-to-Glyph consists of two sequential stages, as summarized in Algorithm 1. The vision encoder and the autoregressive decoder follow the E2STR recognizer architecture [8], initialized with an MAE-pretrained ViT and an OPT-125M language decoder, respectively. In the first stage, the recognizer is trained on Union14M-L using the vanilla scene text recognition objective; the newly introduced appearance projector $P_s(\cdot)$ is then trained with the triplet objective while the vision encoder is frozen. In the second stage, the recognizer is further optimized through relation-aware in-context training with context-rich sequences. The

glyph representations introduce no additional parameters, as they are derived on the fly from decoder–vision cross-attention, and all parameters are frozen during inference.

Algorithm 1 Scene-to-Glyph: Relation-aware training and hierarchical retrieval.

Require: Training set $\mathcal{D}$, annotated context pool $\mathcal{C} = \{(x_i, y_i)\}_{i=1}^{M}$, query image x_q, hyperparameters K, N, N_u
Ensure: Final transcription $\hat{y}_q$
Stage 1: Vanilla recognition training and appearance projector training
Train the recognizer on $\mathcal{D}$ with the standard autoregressive objective (Equation (1))
Train the appearance projector $P_s(\cdot)$ with controlled triplets (Equation (7)), with the vision encoder frozen
Stage 2: Relation-aware in-context training
Construct context-rich sequences with the Hybrid strategy, combining appearance-consistent and character-related samples
Train the recognizer with the in-context objective (Equation (2))
Feature-bank construction (once per benchmark)
Compute scene embeddings $z_i^s = P_s(\bar{f}_i)$ for all context samples and store the scene bank $\mathcal{Z}^s$
Extract glyph representations $g_{i,k}$ of all character positions via teacher-forced cross-attention
Hierarchical retrieval and in-context decoding (per query)
Perform the initial demonstration-free recognition of x_q
Select the N_u lowest-confidence positions $\mathcal{U}_q$ (Equation (10))
Compute the query glyphs $g_{q,t}$, $t \in \mathcal{U}_q$ (Equation (11))
Rank the pool by scene similarity S_i^s and keep the top-K candidates (Equation (9))
Rank the candidates by glyph similarity S_i^g and keep the top-N demonstrations $\mathcal{D}_q$ (Equation (12))
return the final transcription $\hat{y}_q = D_{\mathrm{ICL}}(\mathcal{S}_q)$

3.2. Relation-Aware Context-Rich Training

The recognition model is optimized across the two training stages summarized in Algorithm 1. The first stage establishes the basic mapping from a text image to its transcription; the newly introduced appearance projector is additionally trained within this stage, with the vision encoder frozen. For an image-text pair (x, y), the model is trained with the standard autoregressive objective

$$\mathcal{L}_{\mathrm{STR}} = -\sum_{l=1}^{T} \log p(y_l \mid y_{<l}, x), \tag{1}$$

where T denotes the length of the target transcription. This stage provides the fundamental recognition capability. The second stage aims to equip the recognizer with the ability to exploit relations among different image-text samples. Instead of the random sampling strategy, we explicitly introduce three strategies to construct context sequences: the Appearance-Consistent Strategy, the Character-Related Strategy, and the Hybrid Strategy.

The Appearance-Consistent Strategy. As shown in Figure 3a, the target image 'VACATION' is paired with context samples such as 'ACADEMY', 'HELLO', and 'QUALITY'. Although these samples contain different textual contents, they share similar visual conditions with the target. The shared conditions can include font style, text color, background appearance, blur, and geometric deformation. In practice, these factors are controlled by a font-generation pipeline based on DG-Font [50]: the target image serves as the style reference, and each context sample is synthesized with a different transcription in the same visual style through deformable glyph generation. This strategy therefore constructs contextual relations from appearance consistency rather than lexical identity. It encourages the recognizer to associate different text instances that exhibit similar visual characteristics.

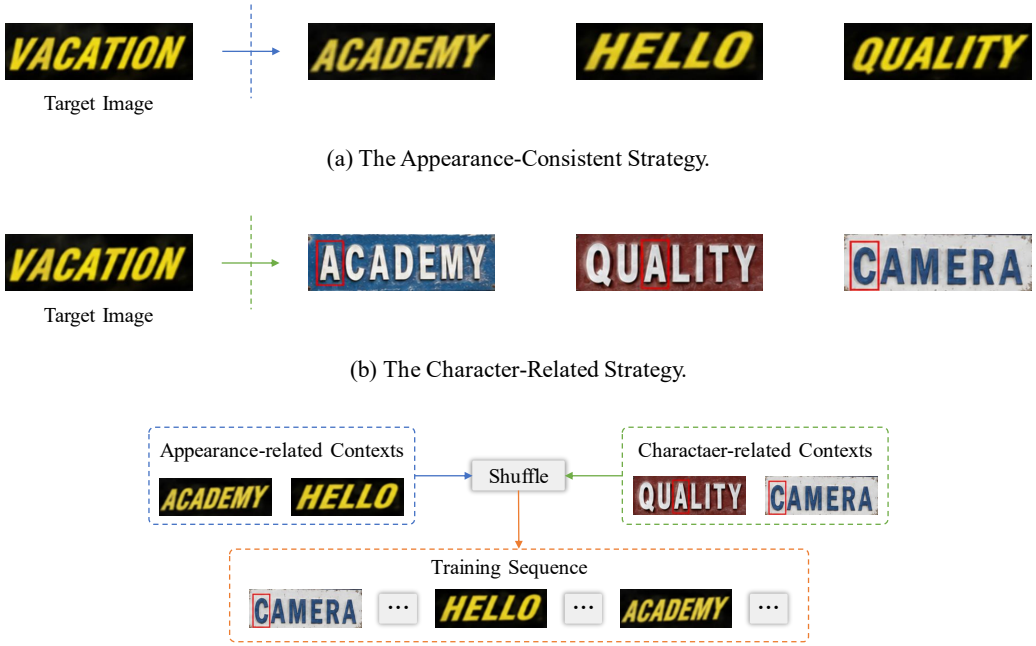


Figure 3. Illustration of the relation-aware context construction strategies. (**a**) The Appearance-Consistent Strategy constructs context samples with different textual contents but similar visual conditions. (**b**) The Character-Related Strategy selects different words that share character-level evidence with the target. Shared characters are highlighted by red boxes. (**c**) The Hybrid Strategy combines appearance-related and character-related samples and shuffles them to form the context-rich training sequence.

The Character-Related Strategy. Figure 3b illustrates the second type of contextual relation. The target word 'VACATION' is associated with different words that contain shared characters. For example, 'ACADEMY' and 'QUALITY' contain the character 'A', while 'CAMERA' contains both 'A' and 'C'. The shared characters are highlighted in Figure 3b. The complete words remain different from the target word. This design introduces local character correspondence without relying on identical word-level patterns. No individual character crop is provided to the recognizer. Instead, the model learns character-level relations through interactions among complete text instances. In practice, context words that share at least one character with the target transcription are drawn uniformly at random from the candidate set so that the character distribution of the constructed contexts remains consistent with that of the training data.

The Hybrid Strategy. In practice, the above two types of contextual samples are combined within the same training sequence. As shown in Figure 3c, appearance-related contexts and character-related contexts are first collected separately. The selected samples are then shuffled before being organized into a context-rich sequence. This construction allows a single training sequence to contain both global appearance relations and local character relations. The recognizer is therefore exposed to complementary contextual evidence during autoregressive learning. These two contextual relations provide the basis for the hierarchical retrieval design used during inference. Scene-level retrieval evaluates appearance compatibility, while glyph-level retrieval evaluates local visual correspondence at uncertain character positions.

These relations are explicitly constructed rather than left to random sampling, as is commonly done in existing in-context training pipelines. The appearance-consistent samples are generated by the DG-Font-based pipeline described above: each context image

is synthesized with a different transcription so that its appearance configuration, including the font family, the color palette, the blur, and the geometric distortion, follows that of the target, with bounded perturbations that keep the visual conditions similar without making them identical. In practice, every training sequence is assembled with the Hybrid strategy so that the two relation types are combined within each individual sequence rather than sampled across sequences: the context slots are filled with both appearance-consistent and character-related samples, which are then shuffled before being organized into the context-rich sequence. Within the Character-Related strategy, the relation type is explicitly defined by shared characters, while the individual context words are drawn uniformly at random from the candidates, without frequency-based reweighting or overlap constraints. This keeps the character distribution of the constructed contexts consistent with the natural distribution of the training data. Since each sequence already contains both relation types, the balance between lexical overlap and appearance similarity is regulated by the Hybrid construction itself rather than by per-sample filtering.

Given the image sequence X and text sequence Y, the in-context training objective is

$$\mathcal{L}_{\mathrm{ICL}} = -\sum_{l=1}^{T_3} \log p\left(Y_l \mid Y_{<l}, X_{\leq l}\right), \tag{2}$$

where $X_{\leq l}$ denotes the visual information available before the l-th target token. The context construction process establishes appearance-level and character-level relations during training. It does not perform retrieval over the inference context pool. During inference, these relations are identified dynamically through the proposed Scene-to-Glyph retrieval process.

The relation-aware construction also establishes consistency between training and inference. The Hybrid strategy exposes the decoder to exactly the two relation types used by the retrieval stages at test time. Appearance-consistent contexts correspond to what scene-level retrieval returns, and character-related contexts correspond to what glyph-level retrieval returns. The inference pool is constructed with the same composition rules. The retrieved demonstrations are also placed into the same context-rich sequence format as during training. Retrieval therefore changes only the samples in the context slots. It does not change the sequence structure or the conditional distribution seen by the decoder. Hierarchical retrieval thus samples from the same relation-aware subspace that the second training stage optimizes. This explains why relation-aware training prepares the decoder better for retrieved demonstrations than random sampling. The two retrieval stages always combine both criteria in the contexts they return. A training sequence with only one relation type would therefore present a one-sided context distribution. The Hybrid construction is accordingly not one option among equals. It is the sequence-level organization whose context distribution matches the one produced by two-stage retrieval.

3.3. Scene-to-Glyph Retrieval for Inference

The proposed S2G strategy treats demonstration relevance as a hierarchical property rather than a single image-level similarity. A useful demonstration should first share compatible visual conditions with the query. It should also contain local visual evidence related to the recognition difficulty of the query. These requirements correspond to scene-level appearance compatibility and glyph-level character relevance, respectively. The two criteria are evaluated sequentially instead of being combined through a weighted score. This design preserves their different functions within the retrieval process. Scene retrieval reduces the search space, while glyph retrieval performs targeted fine-grained selection within that space.

3.3.1. Scene-Level Appearance Retrieval

Given a query image x_q, let $\mathcal{C} = \{(x_i, y_i)\}_{i=1}^{M}$ denote the annotated context pool, where M is the number of context samples, x_i is the i-th text image, and y_i is its transcription. The scene-level stage retrieves context images with visual conditions similar to those of the query. In this work, 'scene'refers to the appearance configuration of a cropped text instance. Relevant factors include font and stroke characteristics, foreground color, background appearance, degradation, blur, and geometric deformation. These factors determine how textual content is visually presented. Therefore, two different words can be scene-compatible when they exhibit similar appearance conditions.

Visual Feature Extraction. Given a text image x_i, the shared vision encoder $E(\cdot)$ converts it into a sequence of visual tokens:

$$F_i = E(x_i) = [f_{i,1}, f_{i,2}, \ldots, f_{i,n}] \in \mathbb{R}^{n \times d}, \tag{3}$$

where n denotes the number of visual tokens and d is the feature dimension. Each token encodes visual information from the input image at a different spatial or semantic location. The same vision encoder is used for both the query and the context samples. This shared encoding space provides a common visual basis for subsequent retrieval. The visual token sequence contains both textual and appearance-related information. To obtain a compact representation for scene-level matching, the visual tokens are aggregated through global average pooling:

$$\bar{f}_i = \frac{1}{n} \sum_{j=1}^{n} f_{i,j}. \tag{4}$$

The pooled feature summarizes the overall visual characteristics of the text image. However, it may still contain substantial information about textual identity. Directly using this representation for retrieval can therefore favor samples with similar word structures rather than similar visual conditions.

Appearance Projection. To emphasize appearance information, we introduce a lightweight appearance projector $P_s(\cdot)$:

$$z_i^s = P_s(\bar{f}_i), \qquad z_i^s \in \mathbb{R}^{d_s}, \tag{5}$$

where d_s denotes the dimension of the scene representation. The projector consists of two linear transformations with a nonlinear activation between them. To reduce the influence of textual identity, the appearance space is learned using controlled triplets. Each triplet contains an anchor, a positive sample, and a negative sample. The anchor and positive contain different textual contents but share similar appearance conditions. In contrast, the anchor and negative contain the same textual content under different appearance conditions. For example, an anchor sample displaying 'HELLO' and a positive sample displaying 'VACATION' are generated under the same font, color, and degradation conditions. In contrast, a negative sample displaying 'HELLO' is generated under a substantially different appearance configuration. This construction encourages the retrieval representation to focus on appearance consistency rather than lexical identity.

Formally, let c denote textual content and ω denote an appearance configuration. For each training triplet, we construct

$$x_a = (c_a, \omega_1), \qquad x_p = (c_p, \omega_1), \qquad x_n = (c_a, \omega_2), \tag{6}$$

where x_a, x_p, and x_n denote the anchor, positive, and negative samples, respectively. c_a and c_p represent different textual contents, while ω_1 and ω_2 represent different appearance configurations. The constraints $c_a \neq c_p$ and $\omega_1 \neq \omega_2$ are imposed during triplet construction. The corresponding appearance embeddings are denoted by z_a^s, z_p^s, and z_n^s, while the appearance projector is optimized using a margin-based triplet objective:

$$\mathcal{L}_s = \max\left[0, m + D_{\cos}(z_a^s, z_p^s) - D_{\cos}(z_a^s, z_n^s)\right], \tag{7}$$

where m is the margin and $D_{\cos}(\cdot,\cdot)$ denotes the cosine distance. The projector is trained within the first stage, following the vanilla recognition training. Besides, the vision encoder is frozen during projector training.

This construction makes textual identity a nuisance factor for the scene representation rather than a useful signal. Because the anchor and the positive share the appearance configuration but carry different transcriptions, any content-dependent direction in the embedding increases their distance and therefore increases the triplet loss, so content variation within an appearance class is actively suppressed. Because the anchor and the negative share the same transcription under a different appearance configuration, identical content is explicitly trained as a dissimilarity relation and cannot serve as a similarity cue. The projector is optimized solely on this objective with a frozen vision encoder and receives no content supervision, since no transcription or character label enters the projector objective; the scene space therefore retains only the directions that the appearance-consistency objective selects. At the system level, character-level evidence is delegated to the glyph stage, and the controlled ablation in Section 4.4 shows that the two stages provide complementary gains: if the scene embedding still encoded word content, the subsequent glyph-level matching would be largely redundant.

Scene-Level Coarse Retrieval. After training, the appearance embeddings of all context samples are pre-computed and stored in a scene feature bank:

$$\mathcal{Z}^s = \{z_1^s, z_2^s, \ldots, z_M^s\}, \tag{8}$$

For a query image x_q, its scene representation z_q^s is obtained using the same feature extraction and projection pipeline. The similarity between z_q^s and the i-th context representation z_i^s is measured by cosine similarity:

$$S_i^s = \frac{(z_q^s)^{\mathrm{T}} z_i^s}{\|z_q^s\|_2 \|z_i^s\|_2}. \tag{9}$$

A larger S_i^s indicates stronger appearance compatibility between the query and the corresponding context sample. All context samples are ranked according to S_i^s and the top-K samples are retained as the final scene-compatible candidate set. The scene stage therefore reduces the retrieval space from M context samples to K appearance-compatible candidates. This stage evaluates only global appearance compatibility. The retained candidates are subsequently passed to the glyph-level stage for fine-grained matching.

3.3.2. Glyph-Level Character Retrieval

The glyph-level character retrieval stage further evaluates the local relevance of the K candidates retained by scene-level retrieval. In this work, the 'glyph' is defined as the local visual representation associated with a character position, rather than the semantic identity of the character itself. The same character may exhibit different local patterns under changes in font, blur, or geometric deformation, while different characters may become visually similar under severe degradation. Therefore, glyph-level retrieval measures the visual correspondence between local character representations, which enables the retrieval process to identify context samples that provide relevant local evidence for uncertain character predictions.

Uncertainty-Guided Glyph Representation. Given a query image x_q, an initial prediction is first performed without contextual demonstrations. At each decoding time step t, the decoder produces a probability distribution over the character vocabulary. The maximum class probability is used as the confidence of the current prediction. Let N_u denote

the number of uncertain character positions considered for glyph-level retrieval. The N_u positions with the lowest confidence are selected as

$$\mathcal{U}_q = \text{Uncertain}\Big(\max_{v\in\mathcal{V}} p_{q,t}(v), N_u\Big), \tag{10}$$

where $\mathcal{V}$ denotes the character vocabulary and $\mathcal{U}_q$ contains the selected uncertain positions. In this work, N_u is set to 2.

The choice of $N_u = 2$ follows from the error structure of STR. Word-level recognition failures are typically caused by local factors such as attention drift at a few character positions [51]. A misrecognized word therefore usually differs from its ground truth in only one or two characters. Two positions are sufficient to cover the dominant failure modes. A single position would be brittle because the lowest-confidence position may reflect the decoding order rather than the true recognition difficulty. Larger values increase the retrieval cost because each additional position requires one more nearest-neighbor search over the characters of every candidate. More importantly, the additional positions are usually decoded correctly. Their glyph representations do not describe the recognition difficulty and dilute the averaged relevance score with non-ambiguous evidence. We also prefer a fixed number of positions over a fixed confidence threshold. A threshold selects a variable number of positions. Easy queries may have no position below the threshold, and glyph-level retrieval then has no evidence to use. Hard queries may have many positions below it, which changes both the retrieval cost and the meaning of the averaged score. The threshold also depends on the confidence scale of the recognizer and must be re-tuned for every model. Selecting the two lowest-confidence positions avoids these problems. It requires no threshold, keeps the retrieval budget constant, and keeps the meaning of the score fixed across queries. The same principle holds for the number of demonstrations. As reported in Section 4.4, two demonstrations already obtain most of the available gain.

The local visual representations associated with the selected uncertain positions are derived from decoder-vision cross-attention. Let $\alpha^q_{t,j}$ denote the normalized attention weight assigned to the j-th visual token at decoding position t. For each $t \in \mathcal{U}_q$, the corresponding query glyph representation is obtained by attention-weighted aggregation:

$$g_{q,t} = \sum_{j=1}^{n} \alpha^q_{t,j} f_{q,j}, \qquad t \in \mathcal{U}_q. \tag{11}$$

The resulting representations form the query glyph set $\mathcal{G}_q$. Each element captures the local visual evidence associated with an uncertain character position. Since the attention response may also involve neighboring strokes or adjacent regions, these features are treated as decoder-conditioned local character representations. The same procedure is applied to the annotated context samples. Their transcriptions enable teacher-forced decoding to obtain the attention response at each character position. The corresponding glyph representations are generated using the same attention-weighted aggregation as in Equation (11). These features are pre-computed and stored before inference. During retrieval, only the glyph representations of the K scene-compatible candidates are accessed for subsequent fine-grained matching.

Glyph-Level Fine Retrieval. For each scene-compatible candidate, the glyph representations associated with the uncertain query positions are compared with all character representations of the candidate. For each query glyph, the most similar local representation in the candidate is first identified. The candidate relevance is then obtained by averaging these best matches over the N_u uncertain positions:

$$S^g_i = \frac{1}{N_u} \sum_{t\in\mathcal{U}_q} \max_{1\le k\le T_i} \cos\big(g_{q,t}, g_{i,k}\big), \tag{12}$$

where T_i denotes the number of characters in the i-th context sample. This formulation evaluates whether a candidate provides relevant local evidence for the selected uncertain positions. A larger S_i^g indicates stronger local visual correspondence between the uncertain query regions and the candidate. The K scene-compatible candidates are ranked according to S_i^g, and the top-N samples are retained as the final demonstrations. The two retrieval stages therefore serve complementary roles: scene-level retrieval constrains the candidate space according to global appearance compatibility, whereas glyph-level retrieval further distinguishes candidates according to local character relevance. The selected demonstrations remain complete image-text pairs and are subsequently combined with the query for context-conditioned recognition.

3.4. In-Context Recognition with Retrieved Demonstrations

The final output of S2G is a set of complete image-text demonstrations rather than isolated glyph representations. After glyph-level fine retrieval, the K scene-compatible candidates are ranked according to S_i^g, and the top-N samples are selected as

$$\mathcal{D}_q = \{(x_{p_1}, y_{p_1}), (x_{p_2}, y_{p_2}), \ldots, (x_{p_N}, y_{p_N})\}, \tag{13}$$

where N denotes the number of demonstrations used for final recognition. The scene-level and glyph-level representations are used only to evaluate demonstration relevance. They are not directly inserted into the recognition sequence. Each selected demonstration therefore retains its complete text image and corresponding transcription, which keeps the inference format consistent with the context-rich sequences used during training.

The selected demonstrations are then combined with the query image to construct the context-rich input:

$$\mathcal{S}_q = [(x_{p_1}, y_{p_1}), \ldots, (x_{p_N}, y_{p_N}), x_q]. \tag{14}$$

The trained in-context decoder produces the final transcription as

$$\hat{y}_q = D_{\mathrm{ICL}}\left(\mathcal{S}_q\right), \tag{15}$$

where $D_{\mathrm{ICL}}(\cdot)$ denotes the in-context decoding process and $\hat{y}_q$ is the final recognition result.

The inference context pool is constructed independently for each benchmark. Each pool contains annotated image-text pairs, randomly sampled from the corresponding training data. For WordArt and OST, we use the official training splits. For the six conventional benchmarks, we sample from Union14M-L. All pool samples have dataset-provided transcriptions, and no test images are included. We do not apply additional duplicate filtering. Global retrieval, scene-level retrieval, and the full Scene-to-Glyph use the same pool, recognizer, and inference settings in the ablation study. This ensures a fair comparison of the retrieval strategies.

4. Results

4.1. Datasets and Evaluation Metric

Datasets. We adopt Union14M-L [7] as the training dataset. It is a large-scale real-world scene text corpus comprising approximately four million labeled text images collected from diverse visual scenarios. The dataset covers substantial variations in text appearance, background complexity, imaging conditions, and geometric layouts, providing a broad training distribution for scene text recognition. Union14M-L is used to establish the fundamental recognition capability of the model and for subsequent relation-aware in-context training.

We evaluate the proposed method on nine scene text recognition benchmarks. Six conventional benchmarks are first employed to evaluate general recognition performance. IIIT5K [52] contains 3000 test images collected from natural scenes and born-digital images.

ICDAR2013 (IC13) [53] contains 1015 cropped word images and mainly consists of regular horizontal text. Street View Text (SVT) [54] contains 647 word images collected from Google Street View, including samples affected by blur, noise, and low resolution. ICDAR2015 (IC15) [55] contains 2077 test images captured under incidental acquisition conditions and exhibits substantial variations in orientation, perspective, and image quality. Street View Text Perspective (SVTP) [56] consists of 645 cropped word images characterized by strong perspective distortion. CUTE80 [57] contains 288 cropped word images and primarily evaluates curved text recognition.

To further evaluate recognition under challenging visual conditions, we employ the Occlusion Scene Text (OST) benchmark [23] and WordArt [58]. OST contains weakly and heavily occluded subsets, denoted as WOST and HOST, respectively, with 2416 images in each subset. These samples contain partially missing character-level visual cues and are therefore suitable for evaluating recognition robustness under occlusion. WordArt contains 6316 artistic text images, including 4805 training images and 1511 test images. Its test set exhibits highly customized fonts, decorative effects, character connections or overlaps, and substantial background interference. In our experiments, WordArt is used only for evaluation without benchmark-specific fine-tuning. Regarding data splits, the recognizer is trained only on the Union14M-L training split. WordArt and OST are used for evaluation, with their official training splits used only to build the in-context pools and their test splits reserved for evaluation. For the six conventional benchmarks, the pools are sampled from Union14M-L, and only the test splits are used for evaluation.

Evaluation Metric. Following the standard evaluation protocol for scene text recognition, word accuracy is adopted as the primary evaluation metric. A prediction is considered correct only when the complete predicted word exactly matches its ground-truth transcription. The word accuracy is calculated as

$$\mathrm{Acc} = \frac{N_{\mathrm{correct}}}{N_{\mathrm{total}}} \times 100\%, \tag{16}$$

where N_{correct} denotes the number of correctly recognized text instances and N_{total} is the total number of evaluated samples. Higher accuracy indicates better recognition performance. All reported word accuracies follow the same normalization protocol. Predictions are compared with ground truth using case-insensitive exact string matching. No lexicon is used in any experiment. Punctuation and other non-alphanumeric characters are retained in both predictions and ground truth. Images with non-alphanumeric ground-truth transcriptions are also retained. The same protocol is applied to all methods and tables.

4.2. Implementation Details

All experiments are implemented in PyTorch (version 2.0.1) and conducted on two NVIDIA GeForce RTX 4090 GPUs, each with 24 GB of memory. The recognizer consists of a vision encoder and a language decoder. The vision encoder is a 12-layer ViT with a model dimension of 768, initialized with MAE-pretrained weights. The language decoder is a 12-layer autoregressive Transformer based on OPT-125M. Input images are resized to 128×32. The output vocabulary contains 94 characters plus an end-of-sequence token, and the maximum transcription length is 100 characters. We train the model in two stages. In the first stage, the recognizer is trained for 10 epochs using the vanilla scene text recognition objective. In the second stage, relation-aware in-context training is performed for 5 epochs. AdamW is used in both stages, with a weight decay of 0.01 and a cosine learning rate scheduler without warm-up. The initial learning rates are 1×10^{-4} and 5×10^{-6}, respectively. The effective batch size is 64 in the first stage and 8 in the second stage.

The appearance projector $P_s(\cdot)$ consists of two linear layers with a GELU activation, and its output dimension is set to $d_s = 256$. It is trained after the initial recognition training

within the first stage. During projector training, the vision encoder is frozen. We use a margin-based triplet loss with $m = 0.2$ and approximately 100k triplets sampled from the DG-Font-based appearance generation pipeline. The projector is trained for 5 epochs using AdamW, with a learning rate of 1×10^{-4} and a weight decay of 0.01. Each positive sample shares the anchor's appearance configuration but has a different transcription. Each negative sample retains the anchor's transcription but uses a substantially different appearance configuration.

During inference, the context pool size is set to $M = 300$. We retrieve $K = 10$ scene-level candidates and select $N_u = 2$ uncertain character positions. Scene and glyph representations of context samples are pre-computed to accelerate retrieval. All reported accuracies are averaged over three independent runs.

4.3. Comparison with Existing Methods

We compare the proposed method with representative scene text recognition approaches on both conventional and challenging benchmarks. As shown in Table 2, our method achieves accuracies of 96.8%, 91.3%, and 96.2% on IIIT5K, SVT, and IC13, respectively, demonstrating the effectiveness of our proposed method on regular-text datasets. On the irregular-text datasets, our method obtains 85.6% on IC15 and 87.8% on SVTP, with a particularly notable gain of 2.4 percentage points on SVTP. This suggests that informative in-context demonstrations are especially beneficial for text affected by perspective distortion and irregular layouts. On CUTE80, our method achieves 97.4%, which is slightly lower than the 97.7% reported by Buoy et al., but remains highly competitive. Overall, the results on the conventional benchmarks demonstrate that the proposed method achieves strong recognition performance across diverse text layouts while maintaining performance on curved text.

Table 2. Comparison with representative scene text recognition methods. '*' represents that the method is reproduced.

Methods	Regular Text			Irregular Text		
	IIIT5K	**SVT**	**IC13**	**IC15**	**SVTP**	**CUTE80**
Yue et al. [22]	95.3	88.1	94.8	77.1	79.5	90.3
Wang et al. [59]	90.5	82.2	–	–	–	83.3
Long et al. [60]	93.7	88.9	92.4	76.6	78.8	86.8
Wang et al. [61]	94.3	89.2	93.9	74.5	80.0	84.4
Luo et al. [62]	–	–	–	76.1	79.2	84.4
Aberdam et al. [63]	82.9	87.9	–	–	–	–
Baek et al. [24]	92.1	88.9	93.1	74.7	79.5	78.2
Buoy et al. [64]	96.1	90.6	95.4	84.9	85.4	**97.7**
Zhao et al. * [8]	96.6	90.9	96.0	85.9	87.2	97.1
Ours	**96.8**	**91.3**	**96.2**	**85.6**	**87.8**	97.4

Table 3 further evaluates the proposed method on HOST, WOST, and WordArt, which involve severe occlusion or substantial appearance variations. Our method achieves 57.8% on HOST, outperforming SATRN by 1.1 percentage points, while obtaining 75.4% on WOST, which is 0.4 percentage points below ABINet. On WordArt, the proposed method reaches 77.0%, exceeding SATRN by 1.3 percentage points. These results indicate that the proposed Scene-to-Glyph retrieval strategy is particularly effective for visually ambiguous text. Scene-level retrieval identifies demonstrations with compatible overall appearance, while glyph-level retrieval further emphasizes local character correspondence. Competitive results on three benchmarks demonstrate the effectiveness of hierarchical demonstration retrieval

under challenging recognition conditions. We note that most of the baseline numbers in Tables 2 and 3 are taken from the E2STR. Within that source, they share the same training data and the same evaluation protocol.

Table 3. Comparison with representative methods on challenging scene text recognition benchmarks. '*' represents that the method is reproduced.

Method	HOST	WOST	WordArt
ASTER [17]	43.3	64.7	65.6
NRTR [65]	50.8	71.5	64.1
SAR [19]	46.4	70.3	72.4
SATRN [66]	56.7	75.6	75.7
ABINet [1]	56.5	**75.8**	75.3
Zhao et al. * [8]	56.9	74.6	76.3
Ours	**57.8**	75.4	**77.0**

4.4. Ablation Study

To evaluate the contribution of the proposed hierarchical retrieval strategy, we conduct ablation experiments on WordArt, which contains substantial variations in font style, character appearance, decorative effects, and background interference. As shown in Table 4, the Base model without retrieved demonstrations achieves a word accuracy of 75.8%. Introducing conventional global visual retrieval improves the accuracy to 76.3%, confirming the benefit of incorporating relevant in-context demonstrations. Replacing global retrieval with the proposed scene-level retrieval further increases the accuracy to 76.5%, indicating that appearance-oriented matching provides more suitable contextual examples than direct global feature similarity. When glyph-level retrieval is additionally introduced, the accuracy reaches 77.0%. These progressive improvements indicate that scene-level appearance compatibility and glyph-level local correspondence provide complementary criteria for demonstration selection, supporting the effectiveness of the proposed Scene-to-Glyph retrieval strategy.

The global-retrieval variant in Table 4 is a direct re-implementation of the E2STR retrieval strategy within our own codebase. It selects demonstrations by image-level similarity over average-pooled visual tokens. It uses the same recognizer, training data, context pool, and inference protocol as the proposed method, so the gap to the published E2STR numbers stems mainly from differences in implementation details rather than from the retrieval rule itself. We therefore report the reproduced numbers for the controlled comparison.

Table 4. Ablation study of the proposed Scene-to-Glyph retrieval strategy on WordArt.

Method	Word Accuracy (%)
Base	75.8
Base + Global Retrieval	76.3
Base + Scene Retrieval	76.5
Base + Scene + Glyph Retrieval	**77.0**

We further examine the influence of the number of in-context demonstrations on WordArt by varying N while keeping the remaining settings unchanged. As shown in Table 5, the accuracy increases from 75.8% without demonstrations to 76.6% with one demonstration and further reaches 77.0% when $N = 2$. Increasing the number of demonstrations to four yields only a marginal improvement to 77.1%, indicating that most of the performance gain can be obtained with a small number of relevant contextual examples. Considering

the negligible difference between $N = 2$ and $N = 4$ and the additional computational cost introduced by longer context sequences, we set $N = 2$ as the default configuration in the remaining experiments.

Table 5. Effect of the number of in-context demonstrations on WordArt.

Number of Demonstrations	0	1	2	4
Word Accuracy (%)	75.8	76.6	77.0	**77.1**

To further examine whether the decoder cross-attention reliably localizes the intended characters, Figure 4 visualizes the cross-attention maps of the decoder at each decoding step for a stylized sample from WordArt. At each step, the attention mass concentrates on the region of the character currently being decoded and decays smoothly toward its immediate neighborhood, whereas unrelated characters and the background receive little activation. This behavior is consistent with the construction of the glyph representations. Specifically, the character query at each decoding step encodes the textual identity predicted by the decoder at the previous step, and the visual tokens belonging to the corresponding character exhibit the strongest correspondence with this query, thereby dominating the attention distribution. These qualitative observations support the claim that the glyph representations derived from attention-weighted visual tokens are indeed character-localized.



Figure 4. Qualitative visualization of the decoder cross-attention on a stylized sample from WordArt. From left to right: the input image, followed by the cross-attention maps of the decoding steps for q, u, e, e, and n.

We briefly characterize the computational overhead of the hierarchical pipeline and the nature of the observed gains. Beyond the vanilla recognizer, the pipeline adds three operations: a preliminary decoding pass, scene-level retrieval over the pre-computed feature bank, and glyph-level matching over a small candidate set. Because the scene and glyph representations of the context samples are pre-computed, both retrieval steps reduce to similarity search over fixed-size vectors, and the search space is small, as the candidate set contains only $K = 10$ scene-level candidates and $N_u = 2$ uncertain positions; the ablation on the number of demonstrations further shows that most of the gain is obtained with the number of 2. The gains themselves are not uniform across benchmarks: they concentrate on the challenging sets, where the text is occluded or heavily stylized, while the margins on the conventional regular-text benchmarks are smaller. This pattern is consistent with the design, since the glyph stage is triggered only by the two lowest-confidence positions of the preliminary decoding pass, so the additional evidence is directed, by construction, at the positions that the recognizer itself finds most ambiguous.

4.5. Discussion on Failure Modes

The hierarchical design has several potential failure modes. First, error propagation: if the initial demonstration-free prediction is confidently wrong, the uncertainty selection

may miss the truly ambiguous position, and the glyph stage then retrieves evidence for an incorrectly localized difficulty. Second, mislocalized cross-attention: attention drift on severely degraded or connected characters weakens the character-localization property of the glyph representations, so the derived local features may cover neighboring strokes rather than the intended character. Third, rare characters absent from the pool: when none of the scene-compatible candidates contains the ambiguous character, glyph-level matching degenerates toward appearance-dominated selection and provides little character-level evidence. Fourth, appearance-projector mismatch: the projector is trained on triplets from the controlled rendering pipeline, so a domain gap between synthetic appearance configurations and real test conditions may degrade scene-level retrieval. Possible mitigations include calibrating the confidence of the initial decoder, constraining the cross-attention with positional priors, enriching the pool with rare-character samples, and fine-tuning the appearance projector on target-domain appearance triplets.

5. Conclusions

This work presents Scene-to-Glyph, a hierarchical retrieval framework for multi-modal in-context scene text recognition. Instead of relying on a single global similarity measure, the proposed method decomposes demonstration selection into two successive stages. Scene-level appearance retrieval first obtains a compact set of candidates that share compatible visual conditions with the query, and glyph-level character retrieval then matches the decoder-derived glyph representations of uncertain positions against the character-level features of those candidates. The two stages address complementary aspects of contextual relevance, namely, global visual compatibility and local evidence for ambiguous characters. Relation-aware context-rich training further aligns the recognizer with this process by exposing it to context-rich sequences that carry both appearance-level and character-level relations so that the demonstration organization seen at inference matches the relations the model has been trained to exploit. Experiments on conventional and challenging STR benchmarks demonstrate the effectiveness of our proposed method. The gains are most pronounced under occlusion and highly stylized appearance, namely, on HOST and WordArt. On the regular benchmarks, the margins are smaller. These results indicate that decomposing relevance into scene-level and glyph-level criteria provides a more effective mechanism for selecting informative in-context demonstrations than a single global similarity measure. The main limitation of this study is that hierarchical retrieval operates over an annotated context pool, so the quality of the selected demonstrations depends on the coverage of that pool; when scene-compatible candidates contain little character-level evidence for the ambiguous positions, glyph-level matching degrades toward appearance-dominated selection.

Author Contributions: Conceptualization, Z.H., Y.J. and D.Z.; methodology, Z.H. and Y.J.; software, Z.H.; validation, Y.J.; investigation, Z.H. and Y.J.; resources, D.Z.; data curation, Y.J.; writing—original draft preparation, Z.H. and Y.J.; writing—review and editing, D.Z.; funding acquisition, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shanghai Sailing Program under grant number 24YF2716400.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; Zhang, Y. Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7098–7107. [CrossRef]
2. Du, Y.; Chen, Z.; Jia, C.; Yin, X.; Zheng, T.; Li, C.; Du, Y.; Jiang, Y.G. SVTR: Scene Text Recognition with a Single Visual Model. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; pp. 884–890. [CrossRef] [PubMed]
3. Bautista, D.; Atienza, R. Scene Text Recognition with Permuted Autoregressive Sequence Models. In *Proceedings of the Computer Vision—ECCV 2022*; Springer: Cham, Switzerland , 2022; pp. 178–196. [CrossRef] [PubMed]
4. Wang, P.; Da, C.; Yao, C. Multi-Granularity Prediction for Scene Text Recognition. In *Proceedings of the Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 339–355. [CrossRef]
5. Guan, T.; Shen, W.; Yang, X.; Feng, Q.; Jiang, Z.; Yang, X. Self-Supervised Character-to-Character Distillation for Text Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 19473–19484.
6. Xu, J.; Wang, Y.; Xie, H.; Zhang, Y. OTE: Exploring Accurate Scene Text Recognition Using One Token. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 28327–28336.
7. Jiang, Q.; Wang, J.; Peng, D.; Liu, C.; Jin, L. Revisiting Scene Text Recognition: A Data Perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 20543–20554. [CrossRef]
8. Zhao, Z.; Tang, J.; Lin, C.; Wu, B.; Huang, C.; Liu, H.; Tan, X.; Zhang, Z.; Xie, Y. Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 15567–15576. [CrossRef]
9. Chen, C.; Zhai, Y.; Zhao, Y.; Gao, J.; Ding, B.; Li, J. Provoking Multi-modal Few-Shot LVLM via Exploration-Exploitation In-Context Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2025; pp. 3826–3835.
10. Wang, X.; Wu, J.; Yuan, Y.; Cai, D.; Li, M.; Jia, W. Demonstration Selection for In-Context Learning via Reinforcement Learning. In Proceedings of the 42nd International Conference on Machine Learning. PMLR , Vancouver, BC, Canada, 13–19 July 2025; Volume 267, pp. 64937–64954.
11. Rubin, O.; Herzig, J.; Berant, J. Learning To Retrieve Prompts for In-Context Learning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 2655–2671. [CrossRef]
12. Xu, C.; Liu, C.; Wang, Y.; Yao, Y.; Fu, Y. Towards Global Optimal Visual In-Context Learning Prompt Selection. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; Volume 37.
13. Li, Y.; Yang, J.; Yun, T.; Feng, P.; Huang, J.; Tang, R. TACO: Enhancing Multimodal In-context Learning via Task Mapping-Guided Sequence Configuration. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025; pp. 736–763. [CrossRef]
14. Du, J.; Hao, F.; Yu, M.; Kong, D.; Wu, J.; Wang, B.; Xu, J.; Li, P. Shift the Lens: Environment-Aware Unsupervised Camouflaged Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 19271–19282. [CrossRef]
15. Shi, B.; Bai, X.; Yao, C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2298–2304. [CrossRef] [PubMed]
16. Shi, B.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Robust Scene Text Recognition with Automatic Rectification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4168–4176.
17. Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2035–2048. [CrossRef] [PubMed]
18. Luo, C.; Jin, L.; Sun, Z. MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition. *Pattern Recognit.* **2019**, *90*, 109–118. [CrossRef]
19. Li, H.; Wang, P.; Shen, C.; Zhang, G. Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8610–8617. [CrossRef]
20. Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S.J.; Lee, H. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4715–4723. [CrossRef]
21. Yu, D.; Li, X.; Zhang, C.; Liu, T.; Han, J.; Liu, J.; Ding, E. Towards Accurate Scene Text Recognition with Semantic Reasoning Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12113–12122.

22. Yue, X.; Kuang, Z.; Lin, C.; Sun, H.; Zhang, W. RobustScanner: Dynamically Enhancing Positional Clues for Robust Text Recognition. In *Proceedings of the Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 135–151.
23. Wang, Y.; Xie, H.; Fang, S.; Wang, J.; Zhu, S.; Zhang, Y. From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 14194–14203.
24. Baek, J.; Matsui, Y.; Aizawa, K. What If We Only Use Real Datasets for Scene Text Recognition? Toward Scene Text Recognition with Fewer Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 3113–3122.
25. Na, B.; Kim, Y.; Park, S. Multi-modal Text Recognition Networks: Interactive Enhancements between Visual and Semantic Features. In *Proceedings of the Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 446–463. [CrossRef]
26. Zhao, S.; Quan, R.; Zhu, L.; Yang, Y. CLIP4STR: A Simple Baseline for Scene Text Recognition with Pre-trained Vision-Language Model. *IEEE Trans. Image Process.* **2024**, *33*, 6893–6904. [CrossRef] [PubMed]
27. Rang, M.; Bi, Z.; Liu, C.; Wang, Y.; Han, K. An Empirical Study of Scaling Law for Scene Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 15619–15629.
28. Maracani, A.; Ozkan, S.; Cho, S.; Kim, H.; Noh, E.; Min, J.; Min, C.J.; Park, D.; Ozay, M. Accurate Scene Text Recognition with Efficient Model Scaling and Cloze Self-Distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 14516–14526. [CrossRef]
29. Du, Y.; Chen, Z.; Xie, H.; Jia, C.; Jiang, Y.G. SVTRv2: CTC Beats Encoder-Decoder Models in Scene Text Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 20147–20156.
30. Zhang, Y.; Liu, C.; Wei, J.; Yang, X.; Zhou, Y.; Ma, C.; Ji, X. Linguistics-aware Masked Image Modeling for Self-supervised Scene Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 9318–9328.
31. Ye, X.; Du, Y.; Tao, Y.; Chen, Z. TextSSR: Diffusion-based Data Synthesis for Scene Text Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–25 October 2025; pp. 17464–17473.
32. Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; Zhou, S. Focusing Attention: Towards Accurate Text Recognition in Natural Images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5076–5084.
33. Zhang, B.; Xie, H.; Gao, Z.; Wang, Y. Choose What You Need: Disentangled Representation Learning for Scene Text Recognition, Removal and Editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 28358–28368.
34. Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; Singh, S. Calibrate Before Use: Improving Few-shot Performance of Language Models. In Proceedings of the 38th International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; Volume 139, pp. 12697–12706.
35. Tsimpoukelli, M.; Menick, J.L.; Cabi, S.; Eslami, S.M.A.; Vinyals, O.; Hill, F. Multimodal Few-Shot Learning with Frozen Language Models. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Volume 34.
36. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 28 November–9 December 2022; Volume 35.
37. Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; Wang, X. Generative Multimodal Models are In-Context Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 14398–14409.
38. Tai, Y.; Fan, W.; Zhang, Z.; Liu, Z. Link-Context Learning for Multimodal LLMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27176–27185.
39. Li, F.; Jiang, Q.; Zhang, H.; Ren, T.; Liu, S.; Zou, X.; Xu, H.; Li, H.; Yang, J.; Li, C.; et al. Visual In-Context Prompting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 12861–12871.
40. Jia, H.; Jiang, C.; Xu, H.; Ye, W.; Dong, M.; Yan, M.; Zhang, J.; Huang, F.; Zhang, S. SymDPO: Boosting In-Context Learning of Large Multimodal Models with Symbol Demonstration Direct Preference Optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 9361–9371.
41. Jiang, Y.; Fu, J.; Hao, C.; Hu, X.; Peng, Y.; Geng, X.; Yang, X. Mimic In-Context Learning for Multimodal Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 29825–29835. [CrossRef]
42. Chen, W.; Li, L.; Yang, Y.; Wen, B.; Yang, F.; Gao, T.; Wu, Y.; Chen, L. CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 8073–8082.

43. Liao, J.; Yang, Z.; Li, L.; Li, D.; Lin, K.; Cheng, Y.; Wang, L. ImageGen-CoT: Enhancing Text-to-Image In-context Learning with Chain-of-Thought Reasoning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–25 October 2025; pp. 17214–17223.
44. Wang, Y.; Li, S. Why Multimodal In-Context Learning Lags Behind? Unveiling the Inner Mechanisms and Bottlenecks. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, San Diego, CA, USA, 2–7 July 2026; pp. 13670–13685. [CrossRef]
45. Zhang, Y.; Feng, S.; Tan, C. Active Example Selection for In-Context Learning. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 9134–9148. [CrossRef]
46. Levy, I.; Bogin, B.; Berant, J. Diverse Demonstrations Improve In-context Compositional Generalization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 1401–1422. [CrossRef]
47. Peng, K.; Ding, L.; Yuan, Y.; Liu, X.; Zhang, M.; Ouyang, Y.; Tao, D. Revisiting Demonstration Selection Strategies in In-Context Learning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; pp. 9090–9101. [CrossRef]
48. Xu, N.; Wang, F.; Zhang, S.; Poon, H.; Chen, M. From Introspection to Best Practices: Principled Analysis of Demonstrations in Multimodal In-Context Learning. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Albuquerque, NM, USA, 29 April–4 May 2025; pp. 3299–3324. [CrossRef]
49. Liu, H.; Wang, W.; Sun, H.; Tian, C.X.; Kong, C.; Dong, X.; Li, H. Unraveling the Mechanics of Learning-Based Demonstration Selection for In-Context Learning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025; pp. 2623–2641. [CrossRef]
50. Xie, Y.; Chen, X.; Sun, L.; Lu, Y. DG-Font: Deformable Generative Networks for Unsupervised Font Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 5130–5140. [CrossRef]
51. Zhong, D.; Lyu, S.; Shivakumara, P.; Yin, B.; Wu, J.; Pal, U.; Lu, Y. SGBANet: Semantic GAN and Balanced Attention Network for Arbitrarily Oriented Scene Text Recognition. In *Proceedings of the Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 464–480. [CrossRef]
52. Mishra, A.; Alahari, K.; Jawahar, C.V. Scene Text Recognition Using Higher Order Language Priors. In Proceedings of the British Machine Vision Conference, Guildford, UK, 3–7 September 2012; pp. 127.1–127.11. [CrossRef]
53. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Gomez i Bigorda, L.; Robles Mestre, S.; Mas, J.; Fernandez Mota, D.; Almazan, J.; de las Heras, L.P. ICDAR 2013 Robust Reading Competition. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1484–1493. [CrossRef]
54. Wang, K.; Babenko, B.; Belongie, S. End-to-End Scene Text Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1457–1464. [CrossRef]
55. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.K.; Bagdanov, A.D.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.R.; Lu, S.; et al. ICDAR 2015 Competition on Robust Reading. In Proceedings of the 13th International Conference on Document Analysis and Recognition, Tunis, Tunisia, 23–26 August 2015; pp. 1156–1160. [CrossRef]
56. Phan, T.Q.; Shivakumara, P.; Tian, S.; Tan, C.L. Recognizing Text with Perspective Distortion in Natural Scenes. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 569–576. [CrossRef]
57. Risnumawan, A.; Shivakumara, P.; Chan, C.S.; Tan, C.L. A Robust Arbitrary Text Detection System for Natural Scene Images. *Expert Syst. Appl.* **2014**, *41*, 8027–8048. [CrossRef]
58. Xie, X.; Fu, L.; Zhang, Z.; Wang, Z.; Bai, X. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 303–321.
59. Wang, Q.; Huang, Y.; Jia, W.; He, X.; Blumenstein, M.; Lyu, S.; Lu, Y. FACLSTM: ConvLSTM with Focused Attention for Scene Text Recognition. *Sci. China Inf. Sci.* **2020**, *63*, 120103. [CrossRef]
60. Long, S.; Guan, Y.; Bian, K.; Yao, C. A New Perspective for Flexible Feature Gathering in Scene Text Recognition via Character Anchor Pooling. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 2458–2462.
61. Wang, T.; Zhu, Y.; Jin, L.; Luo, C.; Chen, X.; Wu, Y.; Wang, Q.; Cai, M. Decoupled Attention Network for Text Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12216–12224.
62. Luo, C.; Zhu, Y.; Jin, L.; Wang, Y. Learn to Augment: Joint Data Augmentation and Network Optimization for Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13746–13755.

63. Aberdam, A.; Litman, R.; Tsiper, S.; Anschel, O.; Slossberg, R.; Mazor, S.; Manmatha, R.; Perona, P. Sequence-to-Sequence Contrastive Learning for Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15302–15312.
64. Buoy, R.; Iwamura, M.; Srun, S.; Kise, K. Explainable Connectionist-Temporal-Classification-Based Scene Text Recognition. *J. Imaging* **2023**, *9*, 248. [CrossRef] [PubMed]
65. Sheng, F.; Chen, Z.; Xu, B. NRTR: A No-Recurrence Sequence-to-Sequence Model for Scene Text Recognition. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 781–786.
66. Lee, J.; Park, S.; Baek, J.; Oh, S.J.; Kim, S.; Lee, H. On Recognizing Texts of Arbitrary Shapes with 2D Self-Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 546–547.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.