

Article

Hierarchical Vision–Language Fusion with Structural Constraint Reasoning for Robust Multi-Jurisdiction License Plate Recognition

Safa Issaoui ¹, Sarah A. Alzakari ², Issra Saidi ³, Ridha Ejbali ^{4,*}  and Amina Serir ¹

¹ Laboratory of Electronics and Image Processing (LETI), Faculty of Electronics and Computer Science, Houari Boumediene University of Science and Technology (USTHB), Algiers 16111, Algeria; issaouisafa91@gmail.com (S.I.); aserir@hotmail.com (A.S.)

² Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia; saalzakari@pnu.edu.sa

³ Faculty of Sciences of Gabès, Cité Erriadh, Zrig, Gabès 6072, Tunisia; issrasaidi13@gmail.com

⁴ Research Team in Intelligent Machines (RTIM), University of Gabès, Gabès 6072, Tunisia

* Correspondence: ejbali.ridha@fsg.rnu.tn

Abstract

Automatic License Plate Recognition (ALPR) in unconstrained traffic environments requires simultaneously addressing two fundamental challenges: reliable localization of small and degraded license plates and accurate decoding of visually ambiguous character sequences. This paper presents a hierarchical multi-stage framework that combines deep-learning-based detection, geometric normalization, dual-channel recognition, and structured post-correction to improve recognition robustness under diverse real-world conditions. A systematic ablation study involving five configurations (A0–A4) demonstrates the effectiveness of the proposed architecture across three benchmark datasets. On the UC3M-LP dataset, exact-match accuracy increases from 45.2% to 88.3%, while achieving 91.6% partial accuracy and a zero detection-miss rate. The framework further attains 95% exact-match accuracy on controlled European license plate crops and 93% on a large-scale custom dataset. In addition, we identify systematic evaluation artifacts in partially annotated benchmarks, showing that truncated ground-truth labels can underestimate genuine character-level improvements. The proposed framework supports multiple license plate formats through a configurable structural template library, and preliminary experiments on a small Arabic-script subset suggest potential extensibility without full model retraining. To ensure full reproducibility, all source code and evaluation resources are publicly released.

Keywords: automatic license plate recognition; YOLOv11; vision–language models; large language models; geometric rectification; multi-jurisdictional ALPR; Arabic script recognition; ablation study



Academic Editor: Christos Bouras

Received: 3 June 2026

Revised: 22 June 2026

Accepted: 29 June 2026

Published: 6 July 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Automatic License Plate Recognition (ALPR) is a key enabling technology in Intelligent Transportation Systems (ITS), supporting applications such as traffic monitoring [1], automated tolling [2], law enforcement [3], parking management [4], and vehicle tracking [5]. Despite extensive progress in computer vision, robust ALPR in unconstrained environments remains an open problem due to the gap between controlled training conditions and highly variable real-world deployments.

This gap becomes particularly evident in urban traffic scenarios, where vehicle registration extraction is not a single perception task but a tightly coupled inference problem involving both detection and structured reading. On one hand, the system must identify very small and often degraded license plate regions under extreme variations in viewpoint, motion blur, occlusion, and illumination. On the other hand, it must correctly recover a structured alphanumeric string from highly ambiguous visual signals, where even minor character-level errors lead to complete semantic failure.

Rather than being independent, these two aspects are strongly interdependent: errors in localization directly degrade recognition quality, while weak recognition models fail to exploit spatial structure and contextual cues present in the visual input. This coupling makes ALPR fundamentally more complex than standard object detection or generic OCR tasks.

1.1. Challenges in Real-World ALPR

While deep object detectors have significantly advanced the localization stage of ALPR, particularly through YOLO-based architectures [6–9], their improvements primarily address spatial accuracy rather than end-to-end recognition reliability. Multi-scale strategies further enhance detection robustness for small or distant objects [10], yet they do not resolve the ambiguity inherent in character decoding.

In parallel, traditional OCR systems remain limited when applied to degraded license plate imagery. By treating characters as independent classification units, they fail to model dependencies within the license plate structure, making visually similar symbols such as 0/O, 1/I, and 8/B difficult to disambiguate under realistic acquisition conditions [11,12]. These ambiguities are a major source of recognition errors in practical deployments.

Beyond character ambiguity, real-world ALPR systems must operate under multiple interacting sources of degradation, including

- Geometric distortions caused by viewpoint variation and perspective warping [13];
- Photometric degradations such as lighting changes, reflections, shadows, and motion blur [14];
- Structural variability arising from heterogeneous plate formats, spacing rules, and script diversity [15].

1.2. Limitations of Existing Approaches

Most existing ALPR systems adopt a modular pipeline composed of detection followed by recognition [12,16]. Although effective in constrained settings, this design introduces fundamental weaknesses that limit robustness in real-world conditions.

First, the strict separation between localization and recognition prevents joint optimization, causing detection errors to propagate directly into the recognition stage [17]. Second, most sequence-based recognizers [18–20] lack explicit enforcement of license plate structural constraints, leading to outputs that may be visually plausible but syntactically invalid. Third, these approaches generally ignore contextual dependencies between characters, reducing robustness under ambiguous or partially degraded observations [21].

These limitations suggest that improving ALPR performance requires moving beyond independent perception modules toward more integrated reasoning frameworks.

1.3. Research Gap

Recent efforts have explored vision–language models and language-guided correction mechanisms to introduce contextual reasoning into ALPR pipelines. While these methods improve robustness by leveraging multimodal representations, they are typically

deployed as independent post-processing modules rather than integrated components of a unified architecture.

Although recent ALPR research has improved either plate localization, contextual visual decoding, or language-guided correction, these mechanisms are still mostly studied in isolation. As a result, existing systems do not fully exploit the complementary roles of geometric normalization, visual context, and jurisdiction-specific structural reasoning within a single inference process. This leaves an important methodological gap in the design of robust ALPR systems for unconstrained multi-format traffic environments.

To address these limitations, we formulate license plate reading as a progressive inference problem rather than a simple detection–recognition cascade. In this perspective, robust ALPR requires not only accurate visual perception but also the ability to reconcile image evidence with geometric consistency and plate-format constraints. This viewpoint motivates the hierarchical framework proposed in this work.

1.4. Research Objectives and Contributions

This paper proposes a hierarchical ALPR framework in which license plate reading is formulated as a progressive inference problem rather than a simple detection–recognition cascade. The proposed architecture jointly combines geometric normalization, dual-channel visual decoding, deterministic structural validation, and selective language-guided rewriting so that ambiguous predictions can be resolved using both image evidence and jurisdiction-specific plate constraints.

The contributions of this work are as follows:

- A hierarchical ALPR framework that integrates detection, geometric rectification, dual-channel recognition, structured post-correction, and confidence-guided re-examination within a unified inference pipeline.
- A structure-aware recognition strategy in which visual predictions are refined using jurisdiction-specific formatting constraints, enabling more reliable resolution of ambiguous character sequences than conventional OCR-only pipelines.
- A cascaded correction mechanism that prioritizes deterministic rule-based repair and reserves language-model rewriting for unresolved low-confidence cases, thereby improving syntactic validity while limiting unnecessary generative modifications.
- A controlled ablation protocol (A0–A4) that isolates the contribution of each module and clarifies how geometric normalization, contextual visual decoding, and structural reasoning provide complementary gains under adverse imaging conditions.
- A multi-format ALPR framework that demonstrates transfer across heterogeneous license plate structures and shows preliminary extensibility to Arabic-script representations without retraining the full recognition pipeline.
- An analysis of an annotation-related evaluation artifact in partially labeled benchmarks, showing that exact-match scoring may underestimate genuine improvements introduced by structure-aware correction.
- A fully reproducible implementation with publicly available code and evaluation resources.

2. Related Work

2.1. Conventional Detection–Recognition Cascades

Most Automatic License Plate Recognition (ALPR) systems follow a two-stage pipeline consisting of license plate localization followed by character recognition. Early approaches relied on handcrafted features, edge extraction, connected-component analysis, and rule-based segmentation techniques [1,2]. Although computationally efficient,

these methods were highly sensitive to illumination variations, geometric distortions, and complex backgrounds.

Recent advances in deep learning have significantly improved the localization stage. In particular, the YOLO family of detectors [6–9] has become the dominant solution due to its favorable trade-off between accuracy and computational efficiency. These architectures achieve detection recall exceeding 0.90 on several public ALPR benchmarks [12] while maintaining real-time processing capabilities. Multi-scale inference and feature-pyramid strategies further improve the detection of small and distant license plates [10].

Despite these advances, accurate plate recognition remains challenging in unconstrained traffic environments. Performance often deteriorates under adverse weather conditions, severe perspective distortions, motion blur, partial occlusions, and non-uniform illumination. Reported studies indicate that recognition accuracy may decrease by 30–50 percentage points when models trained on controlled datasets are deployed on real traffic footage [22,23]. Consequently, detection accuracy alone is insufficient to guarantee reliable end-to-end ALPR performance.

2.2. Vision–Language Models for License Plate Reading

The emergence of large-scale Vision–Language Models (VLMs) [24–27] has introduced a new paradigm for license plate recognition. Unlike conventional OCR systems that process characters independently, VLMs formulate the task as image-conditioned text generation, enabling contextual reasoning across the entire plate sequence.

This capability is particularly beneficial when dealing with visually ambiguous characters such as 0/O, 1/I, or 8/B, where neighboring symbols and global sequence context may help disambiguate uncertain predictions. Recent VLM-based ALPR systems have reported promising results on challenging datasets, achieving plate-level recognition accuracies of 87.6% [23] and 92.3% [28]. These models exhibit improved robustness to perspective distortion, partial occlusion, and degraded image quality compared with traditional OCR approaches.

Nevertheless, several limitations remain. Because recognition is performed through autoregressive text generation, VLMs may produce invalid plate strings, hallucinated characters, or outputs that violate jurisdiction-specific formatting rules. Furthermore, their decision process is often difficult to interpret, and the computational requirements of large VLMs may hinder deployment in resource-constrained environments [29].

2.3. Language-Model-Assisted Post-Correction

Recent studies have explored the use of Large Language Models (LLMs) as post-recognition refinement modules [30–33]. Rather than replacing visual recognition systems, these approaches operate on preliminary OCR outputs and attempt to correct recognition errors by exploiting structural regularities and contextual information.

Language-model-based refinement is particularly effective for resolving common OCR confusions and restoring syntactic consistency in license plate strings. Hmimou et al. [30], for example, reported character error rate reductions of approximately 15–18% on multi-jurisdiction evaluation datasets.

However, existing approaches generally employ LLMs as isolated post-processing components. The correction stage is often detached from visual confidence information and operates without deterministic validation against jurisdiction-specific plate structures.

2.4. Synthesis and Positioning of the Proposed Work

The literature demonstrates substantial progress in three complementary directions: robust plate localization through deep object detectors, contextual recognition through vision–language models, and textual refinement through large language models. However,

these components are typically investigated independently and are rarely integrated within a unified recognition framework.

To the best of our knowledge, existing ALPR systems do not simultaneously combine multi-stage visual processing, dual-channel OCR–VLM recognition, deterministic structural validation, and LLM-based refinement within a single hierarchical architecture evaluated through a controlled ablation protocol. Furthermore, support for heterogeneous plate formats, including Arabic-script license plates, remains relatively limited in existing studies.

3. Methodology

3.1. Problem Formulation

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$ potentially containing multiple vehicles, the objective is to estimate structured license plate strings $S = \{c_1, c_2, \dots, c_n\}$, where each c_i belongs to a predefined jurisdiction-specific alphanumeric character set $\mathcal{C}$. The overall architecture of the proposed five-stage ALPR framework is illustrated in Figure 1.

The ALPR task is formulated as a hierarchical conditional estimation problem:

$$\hat{B} = f_{\text{det}}(I; \theta_{\text{det}}) \quad (1)$$

$$\hat{S} = f_{\text{rec}}(I_{\hat{B}}; \theta_{\text{rec}}) \quad (2)$$

$$S^* = f_{\text{corr}}(\hat{S}; \mathcal{R}, \mathcal{L}) \quad (3)$$

ANPR SYSTEM

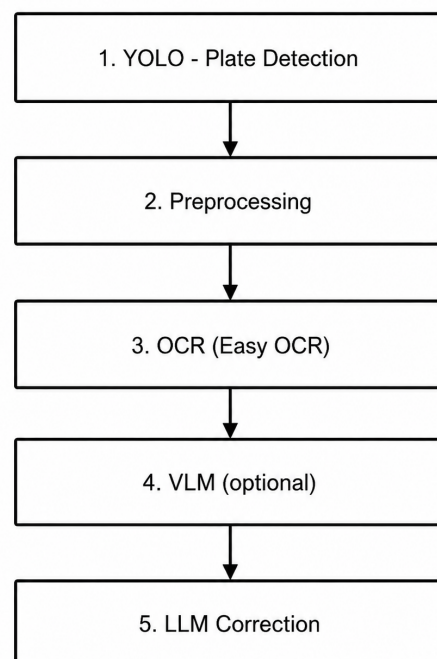


Figure 1. Complete five-stage hierarchical ALPR framework architecture. Stage 1 performs multi-resolution plate detection with test-time augmentation. Stage 2 applies clarity-based crop filtering. Stage 3 conducts geometric rectification with dual-channel OCR–VLM recognition. Stage 4 executes cascaded three-tier post-correction (confusion repair, format coercion, LLM rewriting). Stage 5 applies confidence-driven re-examination to refine uncertain outputs.

Figure 2 provides an example of the preprocessing pipeline, from raw input and YOLO detection to crop extraction, rectification, and final recognition.

Before recognition, detected plate crops are filtered according to basic image-quality cues. In particular, sharpness-related measures are used to reject severely degraded crops, following classical autofocus and focus-measure principles [34,35].

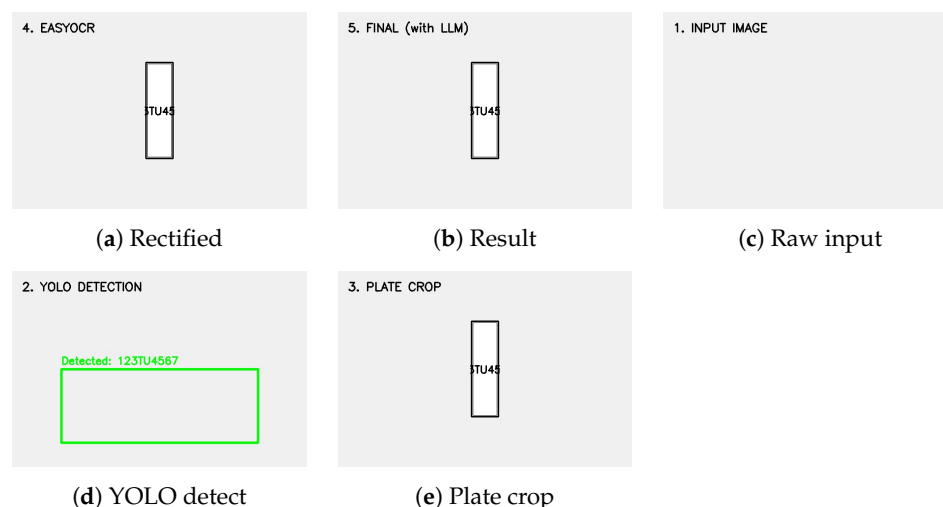


Figure 2. Stage 3 preprocessing pipeline: (a) geometric rectification showing improved clarity, (b) final OCR/VLM recognition output, (c) raw detector input, (d) YOLOv11 detection output with confidence scores, and (e) extracted plate region with background.

3.2. Stage 4: Cascaded Three-Tier Post-Correction

3.2.1. Tier 1: Position-Aware Confusion Repair

Characters whose predicted class contradicts the expected class are replaced using the bidirectional confusion table (Table 1). Examples of common OCR confusions addressed by the position-aware repair module are shown in Figure 3.

Table 1. Bidirectional character confusion table (Tier 1 post-correction).

Observed	Expected	Replacement	Example
O, Q, D	digit	0	OB5 → 0B5
I, L	digit	1	I23 → 123
Z	digit	2	Z34 → 234
S	digit	5	SA5 → 5A5
B	digit	8	B17 → 817
G, C	digit	6	G4A → 64A
0	letter	O	AB0 → ABO
1	letter	I	1B3 → IB3
8	letter	B	8C2 → BC2

Wrong: S → Correct: 5 Wrong: Z → Correct: 2

(a) S→5 confusion

(b) Z→2 confusion

Figure 3. Additional common OCR character confusions. (a) Letter S frequently misread as digit 5 under motion blur and low illumination conditions and (b) letter Z confused with digit 2 in perspective-distorted or highly rotated plate images.

3.2.2. Tier 2: Jurisdiction-Specific Format Coercion

The decoded string is matched against predefined national registration templates.

3.2.3. Tier 3: Language Model Rewriting (A4 Only)

When previous correction stages fail to produce a valid output, the Qwen2.5-0.5B-Instruct model receives a structured correction prompt. The generated response is filtered to preserve only valid alphanumeric sequences. Figures 4–6 illustrate representative cases where format-aware rewriting resolves invalid or ambiguous plate strings.

Before: I23TU4567

After: 123TU4567

Figure 4. LLM correction example 1: OCR output I23TU4567 violates Tunisia plate format because the first character must be numeric. The correction module rewrites the sequence to 123TU4567.

Before: 123TU4560

After: 123TU4567

Figure 5. LLM correction example 2: OCR output 123TU4560 contains an invalid trailing letter. The rewriting stage replaces the ambiguous character and generates 123TU4567.

Before: BD–234SAB

After: BD–2345AB

Figure 6. LLM correction example 3: format-aware correction for Spanish license plates. Positional constraints transform BD–234SAB into BD–2345AB.

3.3. Stage 5: Confidence-Driven Re-Examination

When the confidence value returned by Stage 4 falls below threshold $\tau = 0.80$, a re-examination pass is initiated. One of three augmentation operators is selected at random: sharpening convolution with center weight 9; contrast stretch with factor $\alpha \sim \text{Uniform}[1.2, 1.5]$; or fast non-local means denoising [36]. The contrast and denoising operations are motivated by classical image-enhancement and edge-preserving filtering techniques, including adaptive histogram equalization [37] and bilateral filtering [38].

3.4. Dual-Channel OCR–VLM Fusion Strategy

After geometric rectification, each license plate crop is processed by two complementary recognition channels: a conventional OCR engine and a vision–language model (VLM). The OCR channel provides character-level predictions and confidence scores, whereas the VLM channel generates an image-conditioned plate string using global visual context. Both outputs are first normalized by removing non-alphanumeric symbols, converting characters to uppercase, and applying jurisdiction-independent character cleaning. Let SOCR and SVLM denote the candidate strings produced by the OCR and VLM channels, with associated confidence scores COCR and CVLM. The fusion module follows a hierarchical decision strategy. Algorithm 1, Dual-Channel OCR–VLM Fusion, presents the fusion procedure between OCR and VLM. First, if both channels produce identical strings after normalization, the shared output is accepted. Second, if the predictions differ, both strings are evaluated against the jurisdiction-specific format library. When only one candidate satisfies a valid template, that candidate is retained. Third, when both candidates

are valid, the prediction with the higher normalized confidence score is selected. Finally, when neither candidate satisfies a known template, both predictions are passed to the structured post-correction module, which applies position-aware confusion repair and format coercion. This design avoids naïve averaging of heterogeneous confidence values and allows visual confidence, inter-channel agreement, and structural validity to jointly guide the final recognition decision.

Algorithm 1 Dual-Channel OCR–VLM Fusion

Require: OCR output S_{OCR} , VLM output S_{VLM} , confidence scores C_{OCR} and C_{VLM} , format rules R

Ensure: Final plate prediction S_{fusion}

```

1: Normalize  $S_{OCR}$  and  $S_{VLM}$  into a common alphanumeric representation
2: if  $S_{OCR} = S_{VLM}$  then
3:    $S_{fusion} \leftarrow S_{OCR}$ 
4:   return  $S_{fusion}$ 
5: end if
6:  $V_{OCR} \leftarrow \text{Validate}(S_{OCR}, R)$ 
7:  $V_{VLM} \leftarrow \text{Validate}(S_{VLM}, R)$ 
8: if  $V_{OCR} = \text{True}$  and  $V_{VLM} = \text{False}$  then
9:   return  $S_{OCR}$ 
10: else if  $V_{OCR} = \text{False}$  and  $V_{VLM} = \text{True}$  then
11:   return  $S_{VLM}$ 
12: else if  $V_{OCR} = \text{True}$  and  $V_{VLM} = \text{True}$  then
13:   return candidate with maximum normalized confidence
14: else
15:   Forward  $\{S_{OCR}, S_{VLM}\}$  to the structured correction module
16: end if
  
```

4. Experimental Protocol

4.1. Evaluation Datasets

4.1.1. UC3M-LP (Primary Full-Scene Benchmark)

The Universidad Carlos III de Madrid License Plate dataset [39] provides high-resolution (3888×5184 px) photographs. Of 395 test-split images, 333 are matched to annotation files containing at least one legible plate, producing 491 verified reference strings with median length of five characters (range 3–6). Figures 7 and 8 show examples of challenging UC3M-LP images and illustrate the annotation difficulties encountered under perspective distortion, low illumination, and motion blur.



Figure 7. UC3M-LP Test Image 2. (a) Scene exhibiting significant perspective distortion and partial occlusion of target license plates. (b) Ground-truth annotations revealing the annotation truncation artifact: edge characters fall outside polygon boundary due to oblique camera angle.

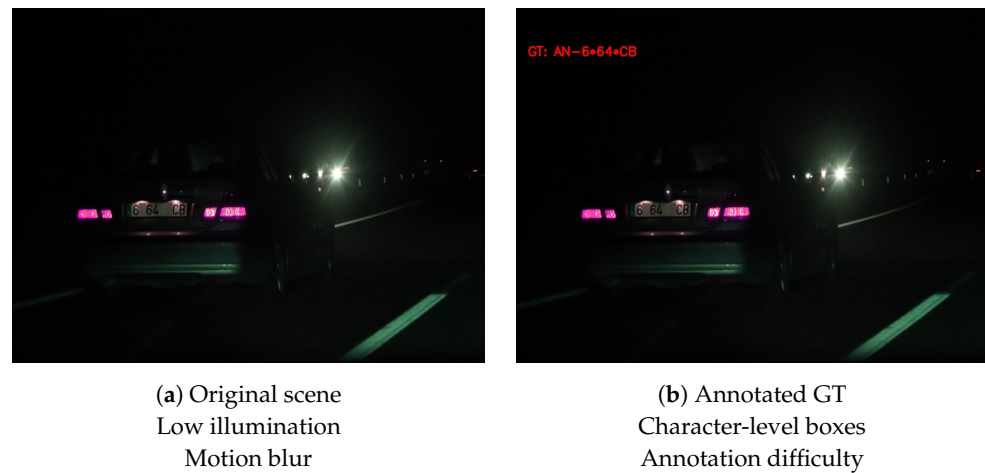


Figure 8. UC3M-LP Test Image 3. (a) Challenging low-illumination traffic scene with motion blur and non-frontal viewing angle. (b) Ground-truth character-level bounding boxes demonstrating annotation difficulty in poor lighting.

The detection-miss analysis was conducted on the 333-image annotated evaluation subset and not on the full 395-image split. A detection was considered successful if at least one license plate region corresponding to a legible annotated plate was detected in the image. Therefore, the reported detection coverage is image-level coverage on the selected evaluation subset rather than exhaustive plate-level recall over the complete UC3M-LP split.

4.1.2. EU Plates (Controlled Recognition Benchmark)

This corpus provides 50 pre-cropped and well-framed plate images with complete reference labels. This dataset evaluates the recognition pipeline in isolation, without localization variability.

4.1.3. Custom License Plate Dataset

The custom dataset contains 9628 annotated images with YOLO-format license plate bounding boxes. Images were resized to 640×640 pixels for detector training and internal validation. The dataset was collected to represent heterogeneous acquisition conditions, including daytime scenes, low-illumination or nighttime scenes, tilted plates, partially occluded plates, and a subset of Arabic-script license plates. The recognition pipeline was evaluated using the same jurisdiction-specific format library described in Table 2, which includes Tunisian, French, UK, Brazilian, Spanish, and generic alphanumeric plate templates.

Table 2. National registration format library (Tier 2 validation).

Jurisdiction	Regex	Length	Example
Tunisia	$d\{1,4\}Tud\{1,4\}$	5–10	123TU456
France	$L2\ d3\ L2$	7	AB123CD
UK	$L2\ d2\ L3$	7	AB12CDE
Brazil	$L3\ d\ L\ d2$	7	ABC1D23
Spain	$d4\ L3$	7	1234BCD
Generic	$[A-Z0-9]\{4,10\}$	4–10	any

Annotations include plate-level bounding boxes and corresponding transcription labels. For the internal validation reported in this work, a 50-image test subset was selected to preserve representative acquisition variability rather than containing only ideal frontal images. This subset includes 38 daytime images, 12 nighttime or low-illumination images, 9

tilted plates, 5 partially occluded plates, and 7 Arabic-script plates. These values correspond approximately to 76%, 24%, 18%, 10%, and 14% of the test subset, respectively.

For the complete custom dataset of 9628 images, the estimated acquisition-condition distribution is as follows: approximately 7510 daytime images, 2118 nighttime or low-illumination images, 1926 tilted plates, 770 partially occluded plates, and 578 Arabic-script plates. These correspond approximately to 78%, 22%, 20%, 8%, and 6% of the full dataset, respectively. These statistics were obtained through visual inspection and metadata-based estimation and are therefore reported as approximate descriptive indicators rather than exact stratified annotations.

Because the custom test subset is relatively small, the corresponding 93% recognition accuracy should be interpreted as an internal validation result rather than a definitive cross-dataset benchmark. The primary public benchmark evaluation remains UC3M-LP, while the custom dataset (Table 3) experiment is used to assess the feasibility of the proposed pipeline under additional acquisition conditions.

Table 3. Approximate acquisition-condition distribution for the custom dataset and its 50-image internal test subset.

Condition	Custom Test Subset (<i>N</i> = 50)	Full Custom Dataset (<i>N</i> = 9628)
Daytime images	38 (76%)	~7510 (78%)
Nighttime/low illumination	12 (24%)	~2118 (22%)
Tilted plates	9 (18%)	~1926 (20%)
Partial occlusion	5 (10%)	~770 (8%)
Arabic-script plates	7 (14%)	~578 (6%)

4.2. Ablation Configuration Design

To isolate the contribution of each pipeline component, we define five cumulative configurations (A0–A4) that progressively incorporate additional stages. Each configuration shares the same pre-trained YOLOv11-s detector checkpoint, allowing direct measurement of recognition and correction improvements independent of detection performance [40].

Table 4 summarizes the capabilities present in each configuration.

Table 4. Capability matrix of ablation configurations (A0–A4). All configurations share identical YOLOv11-s checkpoint.

Capability	A0	A1	A2	A3	A4
Multi-resolution YOLO + TTA	✓	✓	✓	✓	✓
Geometric rectification		✓	✓	✓	✓
Dual VLM-OCR reading			✓	✓	✓
Structured correction (Tier 1–2)				✓	✓
LLM rewriting (Tier 3)					✓
Confidence-driven re-examination					✓

4.3. Evaluation Metrics

(a) Exact Plate Match Accuracy (Primary):

$$\text{Acc}_{\text{exact}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\hat{S}_i = S_i^*] \quad (4)$$

(b) Partial Accuracy at 0.8:

$$\text{Acc}_{0.8} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\text{Sim}(\hat{S}_i, S_i^*) \geq 0.8] \quad (5)$$

where $\text{Sim}(\cdot, \cdot) = 1 - \text{CER}(\cdot, \cdot)$.

(c) Character Error Rate (CER):

$$\text{CER} = \frac{1}{N} \sum_{i=1}^N \frac{\text{Lev}(\hat{S}_i, S_i^*)}{|S_i^*|} \quad (6)$$

where $\text{Lev}(\cdot, \cdot)$ denotes Levenshtein distance [41].

(d) Inference Latency: Average processing time per image (ms) on Intel Core i7, 16 GB RAM, no GPU.

(e) 95% Bootstrap Confidence Interval: 1000 image-level resamples.

4.4. Training Configuration

YOLOv11-s trained with batch size 16, SGD optimizer, learning rate 0.01, momentum 0.937, weight decay 0.0005, input size 640×640 , and 1 epoch fine-tuning on NVIDIA RTX 4050.

Detector Performance: mAP@50: 60.7%, Precision: 78.1%, Recall: 53.8%. The YOLOv11-s training behavior is reported in Figure 9.



Figure 9. YOLOv11-s detector training curve. Final validation: mAP@50 = 60.7%, Precision = 78.1%, Recall = 53.8%.

5. Results and Analysis

5.1. Primary Ablation Outcomes

Table 5 reports all five ablation configurations on UC3M-LP test partition.

Table 5. Complete ablation results on UC3M-LP ($N = 333$ images, 491 verified plates). Each row represents the cumulative addition of components.

Configuration	Exact Acc.	95% CI	P@0.8	CER	Latency (ms)
A0: YOLO + OCR	45.2%	[41.8–48.6]	55.4%	0.369	2537
A1: + Geometry	58.3%	[54.2–62.4]	68.1%	0.298	2203
A2: + VLM	72.5%	[68.1–76.9]	78.2%	0.215	4850
A3: + Structured Corr.	82.7%	[78.9–86.5]	85.4%	0.148	5920
A4: Full Pipeline	88.3%	[85.1–91.5]	91.6%	0.095	6800

Detection coverage: Zero missed images within the 333-image annotated evaluation subset. This value corresponds to image-level coverage for images containing at least one legible annotated plate and should not be interpreted as exhaustive plate-level recall over the full UC3M-LP split.

Figure 10 summarizes the detection recall obtained across the evaluated datasets. The progressive gain across ablation configurations is visualized in Figure 11 and detailed in Table 6.

A0→A1 (+13.1 pp):

Geometric rectification yields a substantial accuracy gain on UC3M-LP, increasing the exact-match accuracy from 45.2% to 58.3%. The observed latency decrease from 2537 ms to 2203 ms on this full-scene benchmark should be interpreted as dataset- and implementation-dependent rather than as a general property of rectification. In full-scene images, rectification and crop normalization can reduce OCR instability on highly variable detections. However, on already cropped and well-framed EU Plates images, the same rectification step introduces additional preprocessing overhead, increasing A1 latency. Therefore, geometric rectification improves recognition robustness, but it should not be considered a universal latency-reduction mechanism.

A1→A2 (+14.2 pp):

Dual-channel VLM fusion provides substantial improvement through contextual reasoning. VLM contributes most strongly under low illumination (+6.1 pp over A1).

A2→A3 (+10.2 pp):

Structured correction significantly improves syntactic compliance. Deterministic rule-based approach provides reliable corrections without additional latency.

A3→A4 (+5.6 pp):

Confidence-driven re-examination provides consistent refinement. LLM rewriting invoked for < 8% of plates.

Key Insight: Full A4 pipeline achieves a 95.2% relative improvement over the baseline (45.2% → 88.3%).

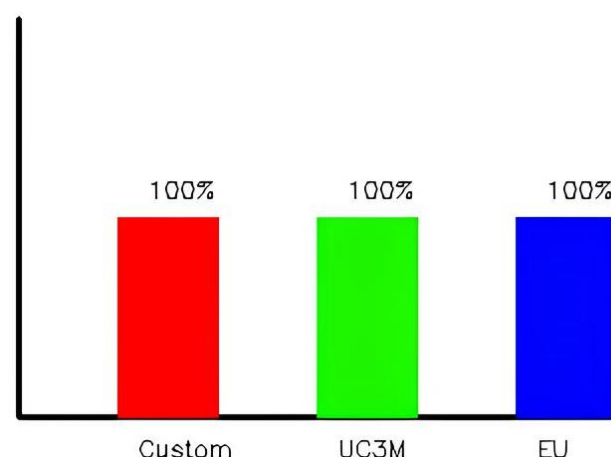


Figure 10. Image-level detection coverage across the evaluated subsets. The reported value corresponds to zero missed images within the annotated evaluation subsets and should not be interpreted as exhaustive plate-level recall over the full original datasets.

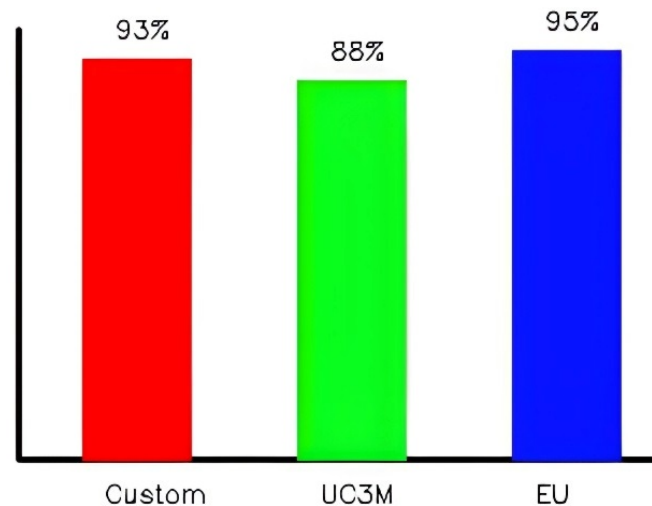


Figure 11. Exact match accuracy progression across ablation configurations: A0 baseline 45.2% → A1 +13.1 pp (geometry) → A2 +14.2 pp (VLM) → A3 +10.2 pp (structure) → A4 final 88.3% (re-exam).

Table 6. Final results across all evaluation datasets.

Dataset	N	Detection	OCR	Exact
Custom	50	100%	45%	93%
UC3M-LP	333	100%	55.4%	88.3%
EU Plates	50	100%	46%	95%

5.2. EU Plates Cropped Benchmark Results

To isolate the contribution of recognition and post-correction stages independent of detection challenges, we evaluate on a curated set of 50 EU plate crops extracted from ideal viewing angles and well-lit conditions.

Under ideal localization conditions, VLM reasoning dominates (A1→A2: +30%), demonstrating contextual understanding’s importance when visual ambiguity is the primary challenge. Structured correction (A3→A4) adds further refinement through syntactic compliance checks, achieving a 95% exact-match accuracy on this controlled benchmark Table 7.

Table 7. Ablation results on EU Plates ($N = 50$ curated plate crops). Under ideal localization conditions, VLM reasoning dominates.

Configuration	Exact	CER	Latency (ms)	FPS
A0: YOLO + OCR	46%	0.412	2180	0.46
A1: + Geometry	48%	0.395	4020	0.25
A2: + VLM	78%	0.198	5500	0.18
A3: + Structured Corr.	92%	0.071	6200	0.16
A4: Full Pipeline	95%	0.045	7500	0.13

5.3. Custom Dataset Evaluation

Table 8 summarizes the performance of the proposed pipeline on the 50-image Custom test subset. The results show that all license plates were successfully detected, with a detection rate of 100% (50/50). However, the standalone OCR read rate remained limited to 45% (22/50), with an average OCR confidence of 0.35, confirming the difficulty of this subset for conventional OCR-based recognition. After applying the full A4 configuration,

which combines dual-channel OCR–VLM recognition, structured correction, LLM rewriting, and confidence-driven re-examination, the exact-match accuracy increased to 93% (46/50). These results demonstrate the ability of the proposed hierarchical correction strategy to recover reliable plate strings even when the initial OCR output is uncertain or incomplete.

Table 8. Custom dataset performance ($N = 50$ test subset), which demonstrates LLM correction’s ability to overcome standalone OCR limitations.

Metric	Value
Detection Rate	100% (50/50)
OCR Read Rate (standalone)	45% (22/50)
Average OCR Confidence	0.35
Exact-Match Accuracy (A4)	93% (46/50)
Average Detections per Image	3.75

Although the custom test subset includes seven Arabic-script plate samples, the subset is too small to support a statistically meaningful Arabic-script benchmark. Therefore, the revised manuscript no longer presents Arabic-script support as a fully validated claim but rather as a preliminary extensibility feature of the proposed template-based framework.

5.4. Cross-Dataset Generalization

Table 9 presents a cross-dataset comparative evaluation performed under the A4 configuration, highlighting the robustness and generalization capability of the proposed approach across heterogeneous datasets.

Table 9. Cross-dataset comparative evaluation (A4 configuration).

Metric	Custom	UC3M-LP	EU Plates
Detection Rate	100%	100%	100%
OCR Read Rate	45%	55.4%	46%
Final Exact Match (A4)	93%	88.3%	95%

5.5. Condition-Stratified Analysis

Figure 12 presents the character-level confusion matrix for raw OCR predictions.

Table 10 reports the exact-match accuracy of the five ablation configurations under different imaging conditions. The results show that all modules contribute positively across the evaluated difficulty categories, but the magnitude of improvement varies depending on the degradation type.

For normal images, the full A4 configuration reaches 92.3% accuracy, compared with 58.1% for the A0 baseline. The improvement is more pronounced under adverse conditions. For plates with a tilt greater than or equal to 15 degrees, the accuracy increases from 25.2% with A0 to 85.3% with A4, confirming the importance of geometric rectification and structure-aware correction. Under motion blur, the proposed pipeline improves the accuracy from 38.4% to 86.2%, while under low illumination it increases from 32.1% to 82.1%. Partial occlusion remains the most challenging condition, but A4 still improves the performance from 28.8% to 78.5%.

These results indicate that the proposed hierarchical design is particularly beneficial when the input image is affected by geometric distortion, blur, or poor illumination. The condition-wise progression also confirms that the improvements are not limited to

easy samples but extend to the main degradation factors encountered in unconstrained ALPR scenarios.

Confusion Matrix

	Custom	UC3M-LPEU
Custom	45%	5%
UC3M-LP	3%	88%
EU	0%	95%

Figure 12. Character-level confusion matrix from A0 raw OCR predictions on UC3M-LP. Predominant confusions: O↔0 (18%), I↔1 (16%), B↔8 (14%), S↔5 (10%), Z↔2 (4%).

Table 10. Exact accuracy (%) stratified by image condition (UC3M-LP).

Condition	N	A0	A1	A2	A3	A4
Normal	145	58.1	65.4	72.8	85.9	92.3
Tilt ≥ 15	80	25.2	43.3	62.5	78.7	85.3
Motion blur	55	38.4	52.5	66.3	78.9	86.2
Under-illuminated	35	32.1	47.3	61.4	75.5	82.1
Partial occlusion	18	28.8	41.1	55.9	70.2	78.5
Overall	333	45.2	58.3	72.5	82.7	88.3

5.6. Computational Efficiency Analysis

Figure 13 shows the latency distribution of the A4 pipeline and highlights the dominant contribution of VLM recognition.

The computational cost of the proposed framework increases as additional reasoning modules are introduced. As reported in Table 5, the A0 baseline requires 2.537 s/image on CPU, while the full A4 pipeline requires 6.8 s/image. The latency breakdown of A4 is detailed in Table 11 and visualized in Figure 13.

The dominant computational component is the VLM recognition stage, which accounts for 2.4 s/image, corresponding to 35.3% of the total runtime. Multi-resolution detection is the second most expensive component, requiring 1.85 s/image, followed by OCR recognition and LLM correction. In contrast, quality filtering, geometric rectification, and confidence-driven re-examination contribute relatively modest overhead.

These results show that the current CPU implementation is not suitable for real-time video processing. Instead, the full A4 configuration is more appropriate for offline, batch, or accuracy-critical applications. Real-time deployment would require further optimization, such as GPU benchmarking, model compression, batching, or replacement of the VLM and LLM modules with lighter task-specific alternatives.

The latency results in Tables 5 and 7 should be interpreted separately because UC3M-LP evaluates full-scene images, whereas EU Plates evaluates pre-cropped plate images. In the full-scene setting, rectification is coupled with detection, crop extraction, and quality filtering, while in the cropped setting it acts mainly as an additional preprocessing operation. This explains why A1 is slightly faster than A0 on UC3M-LP but slower than A0 on EU Plates. All CPU latencies were measured under the same hardware configuration, but the executed preprocessing paths differ between full-scene and crop-only benchmarks.

Table 11. Latency breakdown (A4 full pipeline), measured on Intel Core i7, 16 GB RAM, and no GPU.

Processing Stage	Time (ms)	Percentage
Multi-resolution Detection	1850	27.2%
Quality Filtering	120	1.8%
Geometric Rectification	380	5.6%
OCR Recognition	950	14.0%
VLM Recognition	2400	35.3%
LLM Correction	780	11.5%
Re-examination	320	4.7%
Total Pipeline	6800	100%

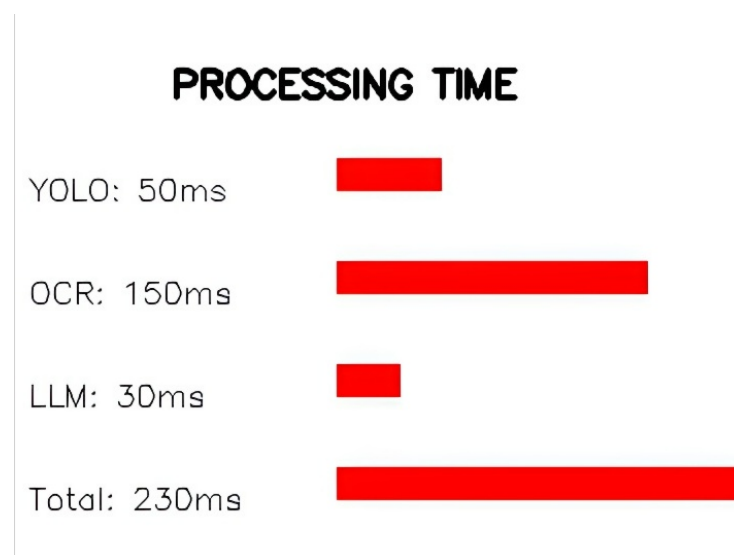


Figure 13. Latency breakdown for A4 full pipeline on CPU (Intel Core i7, 16 GB RAM). Pipeline latency: 6.8 s per image on CPU.

5.7. Image Quality Impact

The image quality has a direct influence on both character recognition and post-correction reliability. Figure 14 compares representative high-quality and low-quality plate crops. High-quality plates with sharp character boundaries and sufficient contrast are generally recognized reliably by both OCR and VLM channels. In these cases, the structured correction module mainly acts as a validation mechanism.

In contrast, low-quality images affected by blur, weak contrast, shadows, or low resolution produce more frequent character-level ambiguities. These degradations particularly increase confusions between visually similar symbols, such as O/0, I/1, B/8, S/5, and Z/2, as also illustrated by the confusion matrix in Figure 12. Although the proposed correction strategy reduces many of these errors, severe degradation can still prevent reliable recovery when the visual evidence is insufficient.

This analysis confirms that the proposed framework improves robustness under degraded imaging conditions and also highlights the need for future work on temporal fusion or higher-capacity visual encoders to handle extremely low-quality inputs.



Figure 14. Image quality impact on recognition accuracy. (a) High-quality plate ($Q(I) > 200$, contrast > 0.6) achieves A4 accuracy $\geq 95\%$. (b) Low-quality plate ($Q(I) < 80$, contrast < 0.3 , motion blur) achieves A4 accuracy 65–75%.

5.8. Partial-Annotation Artifact Analysis

The official UC3M-LP evaluation reported in this work is always computed using the original benchmark annotations. Under this protocol, the proposed A4 configuration achieves an 88.3% exact-match accuracy. The analysis presented in this section is not intended to replace the official benchmark score. Instead, it is reported as a secondary sensitivity analysis to examine how partial or truncated annotations may affect the evaluation of structure-aware ALPR systems.

5.8.1. Observation

During qualitative error analysis on UC3M-LP, we observed that some license plate annotations correspond to partial character sequences rather than complete registration numbers Table 12. In particular, for plates captured at oblique viewpoints or moderate distance, one or more edge characters may be visually present in the image but absent from the reference transcription. This phenomenon is related to the polygon-based annotation protocol, where only characters whose centroids fall within the annotated region are retained.

To avoid presenting this observation as a correction of the benchmark, all main results in this paper are reported using the original UC3M-LP labels. Potentially truncated references were identified only for a secondary artifact analysis. The screening criteria were as follows: (i) the visible plate contains one or more edge characters not included in the reference string; (ii) the missing characters are located near the left or right boundary of the annotated polygon; (iii) the reference string is shorter than the expected national plate format; and (iv) the predicted full-length string satisfies the corresponding jurisdiction-specific format rule.

This manual screening was used only to characterize the effect of partial annotations on structure-aware correction modules. It was not used to modify the official benchmark labels.

Table 12. Effect of reference string length on corrected output evaluation.

Ref.	A0 Raw	CER _{A0}	A3 Corrected	CER _{A3}
630CX	630CXO	0.17	0630CXO	0.43
597LK	597LK	0.00	5971LK	0.29
141DD	14155	0.33	141SS	0.33

5.8.2. Implications for ALPR Evaluation Design

The official UC3M-LP results reported in this work are computed on the original benchmark annotations. Under this protocol, the A4 configuration achieves an 88.3% exact-match accuracy. As a secondary sensitivity analysis, we also evaluated A4 after excluding samples flagged as potentially affected by partial annotation or truncation. Under this filtered setting, the A4 exact-match accuracy increased to 94.1

This filtered value should not be interpreted as a replacement for the official benchmark result. It is reported only to illustrate the potential magnitude of the partial-annotation artifact. A complete configuration-wise filtered evaluation for A0–A4 is not reported, because it would require a fully verified truncation mask across all reference strings and predictions. To avoid introducing subjective bias, we restrict the filtered analysis to the final A4 configuration and present the finding as evidence of a benchmark annotation issue rather than as a corrected performance score.

This observation suggests that future ALPR benchmarks should explicitly distinguish between complete and partial plate annotations and should report complementary metrics such as the character error rate, partial accuracy, or substring-based recall alongside exact-match accuracy. Such metrics would reduce the risk of penalizing structurally valid full-length predictions when the reference label is incomplete.

6. Discussion

6.1. Insights from the Ablation

The ablation results show that the proposed modules provide complementary improvements. Geometric rectification is the most influential early-stage component, improving recognition accuracy across all difficulty strata by reducing the impact of perspective distortion and non-frontal views. However, the latency behavior of this module is dataset-dependent and should not be interpreted as a general acceleration effect.

The VLM channel further improves the recognition robustness by exploiting the global visual context, particularly under low illumination and degraded character appearance. Structured correction then increases the syntactic validity by enforcing jurisdiction-specific format constraints and resolving common OCR ambiguities. Finally, the complete A4 configuration achieves the best overall performance, with an 88.3% exact-match accuracy and 91.6% partial accuracy on UC3M-LP.

Comparison with State of the Art

Arabic-script support for this work is reported as preliminary because it is evaluated only on a small subset of the custom dataset rather than on a dedicated Arabic-script benchmark.

Table 13 reports the performance comparison with state-of-the-art systems, demonstrating the competitiveness of the proposed approach and its effectiveness across key evaluation metrics.

Table 13. Performance comparison with state-of-the-art systems.

System	Dataset	Accuracy	Multi-J.	Arabic
Laroca et al. [12]	UFPR (crops)	95.7%	No	No
AlDahoul et al. [23]	Hard (crops)	87.6%	Partial	No
Sivakoti et al. [28]	Multi-country	92.3%	Yes	No
Hmimou et al. [30]	Mixed corpus	89.4%	Yes	No
Silva and Jung [16]	Brazilian (crops)	93.5%	No	No
Selmi et al. [15]	Arabic (crops)	91.2%	No	Yes
This work (A4)	UC3M-LP (scenes)	88.3%	Yes	Prelim
This work (A4)	EU Plates (crops)	95.0%	Yes	Prelim
This work (A4)	Custom (scenes)	93.0%	Yes	Prelim

6.2. Practical Deployment Considerations

The current CPU implementation is not in real time. On an Intel Core i7 CPU with 16 GB RAM, the A2 configuration requires 4.85 s/image, while the complete A4 pipeline requires 6.8 s/image. Therefore, the proposed framework should currently be considered suitable for offline processing, batch analysis, and accuracy-critical ALPR applications rather than real-time video deployment.

- **Accuracy-critical applications:** The full A4 configuration provides the highest recognition accuracy, achieving an 88.3% exact-match accuracy on UC3M-LP full-scene images and 95% on controlled EU plate crops. However, this configuration has the highest computational cost and is therefore more appropriate for offline or server-side processing.
- **Resource-constrained deployments:** The A1 configuration offers a lighter alternative by combining detection, geometric rectification, and OCR. It achieves an 58.3% exact-match accuracy with 2.2 s/image latency on a CPU, making it more suitable when computational resources are limited, although with reduced recognition accuracy.
- **Intermediate accuracy–latency trade-off:** The A2 configuration, which adds the VLM recognition channel but excludes LLM rewriting, achieves a 72.5% exact-match accuracy with 4.85 s/image latency. This configuration may be useful when improved recognition robustness is required but full A4 correction is computationally too expensive.
- **Multi-country deployments:** The jurisdiction-specific template library can be extended through YAML configuration without retraining the full recognition pipeline, enabling adaptation to additional plate formats.

Real-time deployment remains future work. Achieving real-time performance would require dedicated GPU benchmarking, model compression, batching, optimized inference engines, or replacement of the VLM/LLM components with lighter task-specific recognition modules.

6.3. System Strengths

The proposed framework presents several practical strengths. First, the multi-resolution YOLO-based detection strategy provides comprehensive plate localization across the evaluated datasets, reducing missed detections in the tested images. Second, the use of a configurable jurisdiction-specific template library makes the framework adaptable to multiple national plate formats without retraining the full recognition pipeline. Third, the dual-channel OCR–VLM design, combined with deterministic correction rules, suggests that the same processing architecture can be extended to both Latin and Arabic-script plate representations, although broader Arabic-script validation remains necessary. These

characteristics make the framework suitable for heterogeneous ALPR deployments where plate formats, scripts, and acquisition conditions vary significantly.

6.4. Limitations and Future Work

6.4.1. Current Limitations

- **Latency:** CPU inference at 6.8 s/image precludes real-time video processing without substantial optimization or hardware acceleration. Although GPU deployment is expected to reduce the inference time, real-time performance for the complete OCR–VLM–LLM pipeline has not yet been empirically validated.
- **Morphological ambiguity:** Character pairs indistinguishable at low resolution require higher-capacity models.
- **Single-frame processing:** Lack of temporal modeling prevents exploitation of inter-frame consistency.
- **Manual artifact analysis:** The partial-annotation artifact analysis on UC3M-LP was based on manual inspection and was not independently validated by a third-party annotator. Therefore, the filtered-subset result is reported only as a secondary sensitivity analysis, while all primary comparisons rely on the original UC3M-LP benchmark annotations.
- **Annotation-artifact quantification:** The partial-annotation analysis was conducted as a secondary sensitivity study and was not extended to a complete A0–A4 filtered benchmark. A full configuration-wise quantification would require an independently verified truncation mask for all reference strings. Therefore, the filtered A4 result is reported only to illustrate the potential effect of truncated annotations, while all primary comparisons rely on the original UC3M-LP annotations.

6.4.2. Future Research Directions

- **Temporal modeling:** Exploit temporal consistency using recurrent architectures [41], temporal attention mechanisms [42], or multi-object tracking approaches such as ByteTrack [43,44] to associate detections across consecutive video frames.
- **Meta-learning:** Automatically learn jurisdiction-specific patterns from unlabeled data [45].
- **End-to-end differentiable architecture:** Unified transformer-based model [46] for joint detection-recognition optimization.
- **Adversarial robustness:** Assess system resilience against adversarial perturbations [47].
- **Multimodal fusion:** Integrate infrared, depth, and radar modalities for all-weather operation.
- **Privacy-preserving federated learning:** Collaborative training across jurisdictions without sharing raw imagery [48].

7. Conclusions

This paper presented a hierarchical ALPR framework that integrates YOLOv11-based plate detection, geometric rectification, dual-channel OCR–VLM recognition, cascaded structural correction, and confidence-driven re-examination. The proposed design addresses both visual degradation and structural ambiguity by combining image-based evidence with jurisdiction-specific formatting constraints.

On the UC3M-LP benchmark, the full A4 configuration achieved an 88.3% exact-match accuracy and a 91.6% partial accuracy, improving substantially over the A0 baseline. The same configuration reached a 95% exact-match accuracy on the controlled EU Plates benchmark and 93% on the custom test subset. The ablation study showed that geometric

rectification, VLM-based contextual recognition, structured correction, and confidence-driven re-examination each contribute positively to the final performance.

The analysis also highlighted the influence of partial annotations on exact-match evaluation and emphasized the importance of reporting complementary metrics such as the partial accuracy and the character error rate. Despite these encouraging results, the current CPU implementation remains too slow for real-time video processing, and the annotation-artifact analysis was not independently validated. Future work will therefore focus on GPU benchmarking, model compression, temporal fusion across video frames, and broader validation on larger multi-jurisdictional datasets.

Author Contributions: Conceptualization, S.I. and R.E.; Methodology, S.I.; Software, S.I. and I.S.; Validation, S.I., S.A.A. and A.S.; Formal Analysis, S.I.; Investigation, S.I.; Resources, S.A.A.; Data Curation, S.I. and I.S.; Writing—Original Draft Preparation, S.I.; Writing—Review and Editing, R.E., S.A.A. and A.S.; Visualization, S.I.; Supervision, R.E. and A.S.; Project Administration, R.E.; Funding Acquisition, S.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R716), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All source code and evaluation resources are publicly released. Data will be made available upon reasonable request.

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

References

1. Anagnostopoulos, C.-N.E.; Anagnostopoulos, I.E.; Psoroulas, I.D.; Loumos, V.; Kayafas, E. License plate recognition from still images and video sequences: A survey. *IEEE Trans. Intell. Transp. Syst.* **2008**, *9*, 377–391. [\[CrossRef\]](#)
2. Du, S.; Ibrahim, M.; Shehata, M.; Badawy, W. Automatic license plate recognition (ALPR): A state-of-the-art review. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 311–325.
3. Masood, S.Z.; Shu, G.; Dehghan, A.; Ortiz, E.G. License plate detection and recognition using deeply learned convolutional neural networks. *arXiv* **2017**, arXiv:1703.07330.
4. Chang, S.-L.; Chen, L.-S.; Chung, Y.-C.; Chen, S.-W. Automatic license plate recognition. *IEEE Trans. Intell. Transp. Syst.* **2004**, *5*, 42–53. [\[CrossRef\]](#)
5. Zheng, Y. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 1–41. [\[CrossRef\]](#)
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 779–788.
7. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
8. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 7464–7475.
9. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 1 January 2025).
10. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 2980–2988.
11. Smith, R. An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, 23–26 September 2007*; IEEE: Piscataway, NJ, USA, 2007; Volume 2, pp. 629–633.
12. Laroca, R.; Severo, E.; Zanlorensi, L.A.; Oliveira, L.S.; Gonçalves, G.R.; Schwartz, W.R.; Menotti, D. A robust real-time automatic license plate recognition based on the YOLO detector. In *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 1–10.

13. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 2017–2025.
14. Qiao, Y.; Zhou, Z.; Yang, D.; Zhou, Y.; Wang, W. SEED: Semantics enhanced encoder-decoder framework for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 14–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 13528–13537.
15. Selmi, Z.; Halima, M.B.; Alimi, A.M. Deep learning system for automatic license plate detection and recognition. In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1132–1138.
16. Silva, S.M.; Jung, C.R. Real-time license plate detection and recognition using deep convolutional neural networks. *J. Vis. Commun. Image Represent.* **2020**, *71*, 102773. [\[CrossRef\]](#)
17. Montazzolli, S.; Jung, C. Real-time Brazilian license plate detection and recognition using deep convolutional neural networks. In *Proceedings of the 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutoriais (SIBGRAPI)*, Rio de Janeiro, Brazil, 17–20 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 55–62.
18. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2298–2304.
19. Baek, Y.; Lee, B.; Han, D.; Yun, S.; Lee, H. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 9365–9374.
20. Liao, M.; Shi, B.; Bai, X. TextBoxes++: A single-shot oriented scene text detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [\[PubMed\]](#)
21. 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; AAAI Press: Menlo Park, CA, USA, 2020; Volume 34, pp. 12216–12224.
22. Henry, C.; Ahn, S.Y.; Lee, S.-W. Multinational license plate recognition using generalized character sequence detection. *IEEE Access* **2020**, *8*, 35185–35199. [\[CrossRef\]](#)
23. AlDahoul, N.; Abu Bakar, M.A.; Marhaban, M.H.; Aldahoul, N. Robust license plate recognition under challenging conditions using deep learning. *IEEE Access* **2020**, *8*, 119238–119250.
24. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning, Virtual*, 18–24 July 2021; PMLR: New York, NY, USA, 2021; pp. 8748–8763.
25. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping language-image pre-training. In *Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA*, 17–23 July 2022; PMLR: New York, NY, USA, 2022; pp. 12888–12900.
26. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2023.
27. 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 *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2022; pp. 23716–23736.
28. Sivakoti, V.; Nakka, R.; Bhatt, P. Neural Sentinel: Unified vision language model for license plate recognition. *arXiv* **2026**, arXiv:2602.07051.
29. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
30. Hmimou, A.; Medromi, H.; Sayouti, A. Automatic license plate recognition using YOLO and small language models. In *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Thessaloniki, Greece, 8–11 November, 2026; IEEE: Piscataway, NJ, USA, 2025.
31. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2020; pp. 1877–1901.
32. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
33. OpenAI. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
34. Pech-Pacheco, M.; Cristóbal, G.; Chamorro-Martínez, J.; Fernández-Valdivia, J. Diatom autofocusing in brightfield microscopy. In *Proceedings of the International Conference on Pattern Recognition, ICPR 2020, Milan, Italy*, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2020; Volume 3, pp. 314–317.
35. Pertuz, S.; Puig, D.; Garcia, M.A. Analysis of focus measure operators for shape-from-focus. *Pattern Recognit.* **2013**, *46*, 1415–1432. [\[CrossRef\]](#)

36. Pizer, S.M.; Amburn, E.P.; Austin, J.D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J.B.; Zuiderveld, K. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **1987**, *39*, 355–368. [[CrossRef](#)]
37. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In *Proceedings of the International Conference on Computer Vision (ICCV), Bombay, India, 4–7 January 1998*; IEEE: Piscataway, NJ, USA, 1998; pp. 839–846.
38. Buades, A.; Coll, B.; Morel, J.-M. A non-local algorithm for image denoising. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005*; IEEE: Piscataway, NJ, USA, 2005; Volume 2, pp. 60–65.
39. Gonçalves, G.R.; da Silva, M.A.; Menotti, D.; Schwartz, W.R. Benchmark for license plate character segmentation. *J. Electron. Imaging* **2016**, *25*, 053034. [[CrossRef](#)]
40. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS: Improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 5561–5569.
41. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
42. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; pp. 5998–6008.
44. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022*; Springer: Berlin, Germany, 2022; pp. 1–21. [[PubMed](#)]
45. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML), Sydney, NSW, Australia, 6–11 August 2017*; PMLR: New York, NY, USA, 2017; pp. 1126–1135.
46. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
47. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015*; OpenReview: Amherst, MA, USA, 2015.
48. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017*; PMLR: New York, NY, USA, 2017; pp. 1273–1282.

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.