



ARTICLE

MIRA-Pill: Multimodal Imprint Reading and Confidence-Gated Re-Ranking for Fine-Grained Pill Identification

Linh Nguyen Thi My^{1,2,*}, Nhat Trinh Ngoc Minh², Tham Vo¹ and Vinh Truong Hoang^{3,*}

¹Faculty of Information Technology, Nguyen Tat Thanh University, 300A Nguyen Tat Thanh Street, Xom Chieu Ward, Ho Chi Minh City, Vietnam

²Faculty of Information Technology, Van Lang School of Technology, Van Lang University, 69/68 Dang Thuy Tram Street, Binh Loi Trung Ward, Ho Chi Minh City, Vietnam

³Faculty of Information Technology, Ho Chi Minh City Open University, 35–37 Ho Hao Hon Street, Cau Ong Lanh Ward, Ho Chi Minh City, Vietnam

*Corresponding Authors: Linh Nguyen Thi My. Email: 2311561010@nttu.edu.vn; Vinh Truong Hoang. Email: vinh.th@ou.edu.vn

Received: 21 June 2026; Accepted: 19 August 2026; Published: 22 October 2026

ABSTRACT: Automated pill identification is a safety-critical task whose hardest cases involve visually near-identical medications that differ only in their printed or debossed imprint code. The widely used ePillID benchmark explicitly identified the reliable reading of such imprints as the most important open problem, but left it unsolved because the optical character recognition available at the time was not dependable on small, low-contrast pill surfaces. We revisit this problem with modern vision–language models (VLMs) and propose MIRA-Pill (Multimodal Imprint Reading and Confidence-Gated Re-Ranking), a framework that couples (i) a visual metric-learning branch (ResNet50 with Compact Bilinear Pooling) with (ii) a dedicated imprint-reading branch built on a compact VLM (Qwen2-VL-2B) fine-tuned with Low-Rank Adaptation, and fuses the two through a confidence-gated, log-linear product-of-experts re-ranking rule. The fusion is derived from a conditional-independence assumption and is accompanied by two provable properties: a no-harm guarantee when the imprint cannot be read, and a precise rescue condition that explains which errors the imprint corrects. Fine-tuning lifts imprint token recall on unseen pill types by 19.5 percentage points (pp) on hard consumer images. On ePillID (four cross-validation rounds over a fixed held-out test fold, both-sides protocol), MIRA-Pill raises Top-1 accuracy from 90.34% to 94.33% (+3.99 pp), a gain that is statistically significant at every test we apply (McNemar $p = 8.4 \times 10^{-21}$; bootstrap 95% confidence interval (CI) of $\Delta\text{Top-1} = [3.19, 4.83]$ pp), whereas a prior single-benchmark improvement of +0.29 pp was not significant. The method generalises across datasets: +7.29 pp on CURE with an OCR-consensus imprint database, and no aggregate degradation on the imprint-free OGYEIV2 dataset, where only 5 of 480 queries are hurt. The size of the gain correlates with imprint availability, confirming that the imprint signal, rather than incidental tuning, drives the improvement, while the confidence gate keeps the observed harm rate low when no imprint is present. For comparison, a zero-shot CLIP baseline reaches only 10.54% Top-1 under the same protocol, so the gain stems from the specialised reader and fusion rather than from generic vision–language capability. These results indicate that imprint reading, long considered unreliable, is now a dependable source of identification evidence on retrospective benchmarks; because the framework has not yet been evaluated in a real-world clinical or pharmacy setting, we position it as a promising candidate component, rather than a validated tool, for future automated pharmacy-verification and medication-safety systems.

KEYWORDS: Pill identification; fine-grained recognition; vision–language model; imprint reading; multimodal fusion; metric learning; ePillID

1 Introduction

Medication errors remain a persistent threat to patient safety, and the visual verification of a dispensed tablet or capsule against its intended identity is one of the last lines of defence against them [1,2]. Reports attribute several thousand deaths every year directly to medication errors, and the World Health Organization dedicated World Patient Safety Day 2022 to the theme “Medication Without Harm” [2]; a system that can distinguish visually confusable medications before they reach the patient therefore has direct clinical and economic value. Automated pill identification, the recognition of a pill from an ordinary photograph, has therefore grown into an active research area at the intersection of computer vision, optical character recognition (OCR), and clinical informatics [2]. The defining difficulty of the task is that it is fine-grained: thousands of distinct medications share a small vocabulary of shapes and colours, so that many pills are nearly indistinguishable by gross appearance and differ only in the alphanumeric imprint stamped or printed on their surface [3].

The ePillID benchmark [3] crystallised this challenge into a realistic, low-shot, fine-grained retrieval problem: a large gallery of clean reference images (two sides per pill) must be matched against a much smaller set of unconstrained consumer photographs, with many classes represented by only a single reference appearance. State-of-the-art systems on ePillID are bilinear metric-learning networks with strong appearance embeddings, yet the benchmark authors were explicit that the dominant remaining failure mode is the confusion of pills identical in shape and colour but with different imprints, and named the integration of a reliable imprint reader as the most important direction for future work, unpursued because the OCR of 2020 was not dependable on the small, curved, low-contrast surfaces of real pills.

We argue that this open problem is now solvable: modern vision–language models (VLMs) read text from difficult natural images more reliably than the OCR pipelines available when ePillID was released, which makes imprint reading worth revisiting. Building on a sustained line of pill-recognition research spanning a comprehensive survey [2], multi-stream class-incremental recognition [4], and a real-time multimodal edge recogniser [5], we propose MIRA-Pill, a framework that treats the imprint as a separate signal and fuses it with a conventional visual branch.

MIRA-Pill couples a visual branch (V), a ResNet50 backbone with Compact Bilinear Pooling [6,7] trained with metric learning and reused from the strongest publicly available ePillID configuration, with an imprint branch (I), a compact VLM (Qwen2-VL-2B [8]) fine-tuned with Low-Rank Adaptation (LoRA) [9] to read the imprint string from both sides of a pill. A confidence-gated, log-linear product-of-experts re-ranking rule combines the two: under a conditional-independence assumption it is the correct Bayesian fusion of the two evidence sources, and it degrades gracefully to the visual baseline whenever the reader returns nothing. The imprint string is a domain-invariant attribute: the code “TV 7296” is the same on a studio reference image and on a blurry phone photograph, which is exactly why it repairs the appearance-domain gap that degrades the visual branch.

The contributions of this paper are as follows. (1) We reframe ePillID’s stated open problem as a VLM imprint-reading task and show, through a feasibility study, that an off-the-shelf compact VLM already recovers 63.0%–65.3% of imprint tokens, establishing a clear signal to exploit. (2) We fine-tune the imprint reader with LoRA and demonstrate that it transfers to entirely unseen pill types, improving imprint token recall on hard consumer images by 19.5 pp. (3) We formalise the fusion as a log-linear product-of-experts model and prove a no-harm property and a rescue condition that are directly confirmed by our experiments. (4) We report a complete evaluation on ePillID, where fusion raises Top-1 accuracy by 3.99 percentage points with statistical significance under every test we apply, together with a failure-case study, open-set and zero-shot capabilities, full ablations, and VLM baselines, and we demonstrate cross-dataset generalisation (a gain of 7.29 percentage points on CURE) and no aggregate degradation on the imprint-free OGYEIV2 dataset.

[Section 2](#) reviews related work; [Section 3](#) presents the framework and its theoretical properties; [Section 4](#) describes the datasets, protocol, and implementation; [Section 5](#) reports the results; and [Section 6](#) concludes.

2 Related Work

We organise prior work into three threads (pill image recognition, imprint and text reading, and vision–language models with multimodal fusion), following the taxonomy of the recent comprehensive survey [2].

2.1 Pill Image Recognition

Early systems described the three canonical attributes (shape, colour, and imprint) with hand-crafted features and were brittle under unconstrained imaging [2]. Deep learning replaced these descriptors with features learned end-to-end: an early journal study already demonstrated fine-grained pill identification from consumer photographs with a deep convolutional network [10], and backbones such as ResNet [7] now underpin dedicated benchmarks and inspection systems, including ePillID [3], deep pill-detection models [11], and journal comparisons of real-time detectors for pill identification [12]. Because medications are fine-grained, multi-stream and multimodal variants followed: class-incremental multi-stream fusion [4], metric-based recognition combining textual and visual cues [13], and a real-time edge recogniser with adaptive lightweight attention [5]. Recent systems push towards deployable pipelines: code-free deep learning [14], mobile identification [15], and real-time inspection [16]. Recent journal studies go further and combine detection, attribute classification, and OCR-based imprint analysis inside a single pipeline [17,18], which confirms that the imprint is increasingly treated as a first-class identification cue rather than an optional attribute. The low-shot regime was targeted by Few-Shot Pill Recognition [19] and by ePillID [3], which we adopt as our base setting and baseline.

2.2 Imprint and Text Reading on Pills

The imprint is the most discriminative attribute for visually similar pills but also the hardest to read: small, often debossed rather than printed, low in contrast, and curved across the pill surface. Classical binarise-and-segment pipelines collapsed under unconstrained imaging [2], which is why the ePillID authors [3] explicitly deferred imprint OCR, identifying it as the key unsolved component of fine-grained pill identification. Our work takes up this thread, replacing fragile character segmentation with a generative vision–language reader and, unlike systems that concatenate recognised text as just another feature, fusing the imprint as a calibrated probabilistic factor behind an explicit confidence gate.

2.3 Vision–Language Models and Multimodal Fusion

Large vision–language models align images and text in a shared space and can read and reason about text in images. Contrastive image–text pre-training (CLIP) [20] popularised zero-shot transfer, while generative VLMs such as Qwen2-VL [8] read text from images at native resolution. General-purpose VLMs are nonetheless weak at fine-grained recognition out of the box, as our CLIP zero-shot baseline confirms ([Section 5](#)), motivating task-specific adaptation. The generative line of VLMs is evolving rapidly: instruction-tuned models such as LLaVA popularised visual instruction following [21], and successor families have continued to improve document and scene-text understanding; our framework is deliberately reader-agnostic, so newer readers can be substituted without changing the fusion rule. Journal surveys of vision–language modelling for vision tasks [22] and systematic reviews of visual–language foundation models in medical imaging [23,24] document both this rapid progress and the domain-adaptation gap that task-specific fine-tuning is required to close. We adapt the reader with LoRA [9], freezing the vision tower and training only low-rank projections;

for fusion we adopt a log-linear product-of-experts formulation, which under conditional independence is the exact posterior combination of the two likelihoods and yields the no-harm and rescue properties of Section 3. The visual branch follows metric-learning practice, combining a softmax cross-entropy head with an additive angular-margin objective (ArcFace) [25] on a compact bilinear embedding [6].

2.4 Summary and Positioning

The three threads above map directly onto the three components of our methodology: the pill-recognition thread supplies the visual metric-learning branch (Section 3.2), the imprint-reading thread motivates the VLM reader that replaces fragile character segmentation (Section 3.3), and the multimodal-fusion thread provides the probabilistic basis of the confidence-gated re-ranking rule (Section 3.4). Table 1 summarises all pill-identification systems discussed in this section along the technical dimensions that matter for this task. Two scoping rules keep the comparison fair. First, all listed works address the same application, image-based pill identification; the single detection-oriented entry [11] is included because pill detection is the front end of identification pipelines (we use such a step on CURE), and this role is stated in its row; the journal comparison of real-time detectors [12] benchmarks that same front end and is therefore not listed as a separate system. Second, generic backbones and adaptation techniques discussed above (ResNet [7], CBP [6], ArcFace [25], LoRA [9], Qwen2-VL [8], LLaVA [21]) are building blocks rather than complete pill-identification systems and are therefore not listed as separate rows; CLIP appears because we evaluate its zero-shot transfer on ePillID under the identical protocol (Section 5), and the survey [2] is the taxonomy source rather than a method. No prior system reads the imprint as an explicit string with a modern VLM and fuses it as a gated probabilistic factor; this combination is the specific gap that MIRA-Pill fills.

Table 1: Method-dimension matrix of representative pill-identification approaches.

Method	Visual Encoder	Imprint Use	Fusion	Main Limitation
Fine-grained CNN identification [10]	Deep CNN classifier	None	—	Gross appearance only; no imprint string reading
ePillID baseline [3]	ResNet50 + CBP metric learning	None	—	Imprint confusions remain (stated open problem)
Few-shot pill recognition [19]	Multi-stream CNN, few-shot metric learning	Imprint as visual stream, no string reading	Feature level	No explicit text matching or gating
Metric-based recognition [13]	CNN metric learning	Textual and visual cues	Feature level	No confidence gating; no probabilistic fusion
CG-IMIF [4]	Multi-stream incremental CNN	None	Feature level	Classification only; no imprint reading
Edge multimodal ALA [5]	Lightweight CNN streams	Text cue among streams	Adaptive attention	No string-level imprint matching

(Continued)

Table 1 (continued)

Method	Visual Encoder	Imprint Use	Fusion	Main Limitation
Deep pill detection [11]	Detection CNN (front end of identification)	None	—	Localises pills only; near-identical classes not separated
Code-free deep learning [14]	AutoML-built CNN classifier	None	—	Gross appearance only; no imprint reading
Mobile identification [15]	Lightweight mobile CNN	None	—	No explicit imprint string matching
Real-time inspection [16]	YOLOv5 detection pipeline	OCR-style text spotting (DTS + character RNN)	Pipeline level	No confidence-gated, database-level fusion of the read string
CLIP zero-shot [20] (evaluated on ePillID)	ViT contrastive	Implicit	Shared space	Weak at fine-grained classes
MIRA-Pill (this work)	ResNet50 + CBP	Explicit string read by fine-tuned VLM	Confidence-gated log-linear	Hard gate; single visual backbone

3 Proposed Method: MIRA-Pill

This section develops the framework in four steps. [Section 3.1](#) formalises the problem; [Sections 3.2](#) and [3.3](#) specify the visual and imprint branches; [Section 3.4](#) derives the confidence-gated fusion rule; and [Section 3.5](#) proves its two theoretical properties. The design is deliberately modular, so each block can be upgraded independently.

3.1 Overview and Problem Formulation

Let a gallery of reference images cover C drug classes, and let each query be a both-sides observation $q = (x_f, x_b)$ comprising the front and back images of one pill. The goal is a scoring function $S(q, c)$ that ranks the correct class c^* first. Each class c has a canonical imprint string T_c (from Pillbox metadata, or built by OCR consensus when metadata is unavailable). MIRA-Pill, shown in [Fig. 1](#), comprises a visual branch (V) that produces an appearance score $\hat{S}_{\text{vis}}(q, c)$, an imprint branch (I) that reads a string t_q and matches it against every T_c , and a confidence-gated fusion rule that re-ranks the top- K visual candidates.

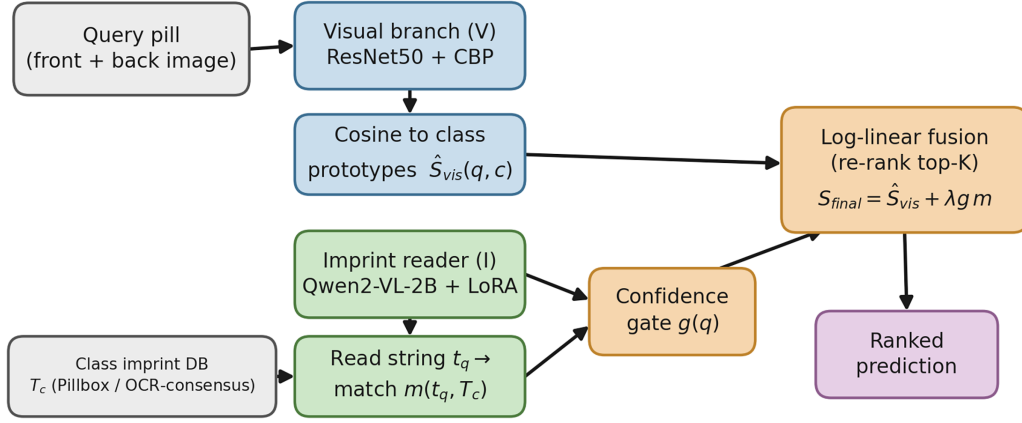


Figure 1: Overview of MIRA-Pill. The visual branch (blue) scores appearance similarity; the imprint branch (green) reads the imprint with a fine-tuned VLM and matches it to the class imprint database; a confidence gate (orange) suppresses the imprint term when nothing is read, and a log-linear rule re-ranks the top- K visual candidates.

3.2 Visual Branch

The visual encoder f_θ maps an image to $\mathbb{R}^d$ using a ResNet50 backbone [7] followed by Compact Bilinear Pooling (CBP) [6], which approximates second-order feature interactions with Count-Sketch projections and the fast Fourier transform (FFT):

$$\Psi(x) = \text{FFT}^{-1}(\text{FFT}(C_1 x) \odot \text{FFT}(C_2 x)), \quad \langle \Psi(x), \Psi(y) \rangle \approx \langle x, y \rangle^2. \quad (1)$$

The visual score is the cosine similarity between the side-averaged query embedding and the class prototype μ_c :

$$s_{\text{vis}}(q, c) = \cos(\bar{f}_q, \mu_c), \quad \bar{f}_q = \frac{1}{2}(f_\theta(x_f) + f_\theta(x_b)). \quad (2)$$

Training uses a multi-head metric-learning objective combining softmax cross-entropy, an additive angular-margin term (ArcFace) [25], triplet, and contrastive losses,

$$\mathcal{L}_{\text{vis}} = \lambda_{\text{ce}} \mathcal{L}_{\text{SCE}} + \lambda_{\text{arc}} \mathcal{L}_{\text{ArcFace}} + \lambda_{\text{tri}} \mathcal{L}_{\text{triplet}} + \lambda_{\text{con}} \mathcal{L}_{\text{contrastive}}, \quad (3)$$

where $\mathcal{L}_{\text{ArcFace}} = -\log[e^{s \cos(\theta_{c^*} + m)} / (e^{s \cos(\theta_{c^*} + m)} + \sum_{j \neq c^*} e^{s \cos \theta_j})]$. Fig. 2 illustrates the branch end to end, including the feature dimensions at each stage.

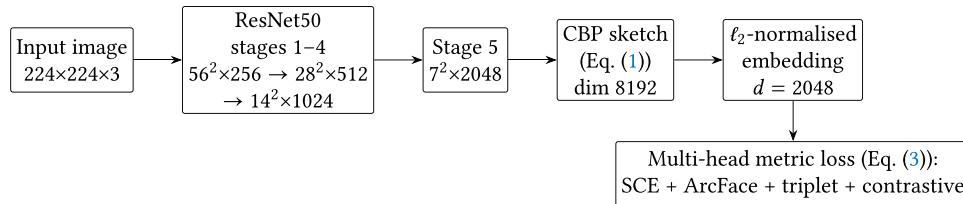


Figure 2: Architecture of the visual branch. A 224×224 input passes through the ResNet50 stages (256, 512, 1024, and 2048 channels), the final $7 \times 7 \times 2048$ feature map is aggregated by Compact Bilinear Pooling into a sketch vector, and an ℓ_2 -normalised embedding is trained with the multi-head metric-learning objective of Eq. (3). Class prototypes are the averaged reference embeddings and queries are scored by cosine similarity (Eq. (2)).

3.3 Imprint Branch

The imprint reader is a compact VLM, Qwen2-VL-2B [8], fine-tuned with LoRA [9] by conditional maximum likelihood over the imprint character sequence:

$$t_q = \arg \max_t P_\psi(t \mid x_f, x_b), \quad \mathcal{L}_{\text{imp}} = - \sum_k \log P_\psi(y_k \mid y_{<k}, x_f, x_b). \quad (4)$$

Because ePillID images are single-sided and many sides are blank, each example is supervised with the complete imprint of the pill given both sides, which yields clean labels and matches the both-sides evaluation; only reference images are used for fine-tuning, with consumer images held out. The match between a read string t_q and a class imprint T_c is an asymmetric token-overlap function (let $\text{tok}(\cdot)$ denote the token set and $\subseteq$ denote substring containment after whitespace removal):

$$m(t_q, T_c) = 0.85 \cdot \frac{|\text{tok}(t_q) \cap \text{tok}(T_c)|}{|\text{tok}(t_q)|} + 0.15 \cdot \mathbf{1}[t_q \subseteq T_c] \in [0, 1]. \quad (5)$$

Fig. 3 shows how the two side images enter the VLM: both sides are supplied in one conversation turn, the vision tower is kept frozen, and only the LoRA projections are trained.

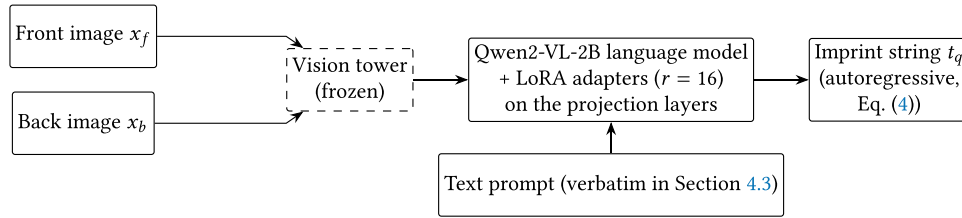


Figure 3: Architecture of the imprint-reading branch. The two side images of a query are encoded by the frozen vision tower of Qwen2-VL-2B; the language model, adapted with rank-16 LoRA on its projection layers (18.5M trainable parameters, 0.83% of 2.2B), decodes the imprint string autoregressively under the prompt reproduced in Section 4.3. Both sides are supplied in a single conversation turn so that the model reads the complete imprint of the pill.

3.4 Confidence-Gated Log-Linear Fusion

Assuming conditional independence between visual appearance and the imprint string given the class, which is reasonable because they are different physical aspects of the pill, the posterior factorises as

$$P(c \mid x, t_q) \propto P_{\text{vis}}(c \mid x) \cdot P_{\text{imp}}(t_q \mid c)^\lambda. \quad (6)$$

Taking logarithms gives an additive product-of-experts form, $\log P(c \mid x, t_q) = \log P_{\text{vis}}(c \mid x) + \lambda \log P_{\text{imp}}(t_q \mid c) + \text{const}$, where the additive constant does not depend on the class c . MIRA-Pill is a practical instantiation of this expression in which the min-max normalised cosine $\hat{S}_{\text{vis}}$ stands in for $\log P_{\text{vis}}$, the match score $m(t_q, T_c)$ stands in for $\log P_{\text{imp}}$, and a confidence gate $g(q)$ handles the unreadable-imprint case (an uninformative, flat likelihood that must not perturb the ranking):

$$S_{\text{final}}(q, c) = \hat{S}_{\text{vis}}(q, c) + \lambda g(q) m(t_q, T_c), \quad g(q) = \mathbf{1}[t_q \neq \emptyset]. \quad (7)$$

Fig. 4 traces the complete path from the visual score, the read string, and the class imprint database to the final score. The hard gate $g \in \{0, 1\}$ generalises to a soft gate $g(q) = \sigma(\text{mean log-probability of the generated string})$, explored in Section 5.9. The imprint string is domain invariant, so $P_{\text{imp}}(t_q \mid c)$ is stable under the

appearance-domain shift that degrades $P_{\text{vis}}(c | x)$; fusion therefore supplies evidence exactly where the visual branch is weakest.

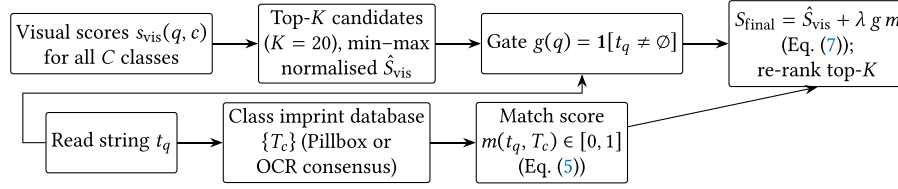


Figure 4: The fusion and re-ranking module. The visual branch supplies normalised scores for the top- K candidates; the read string is matched against every class imprint in the database; the gate suppresses the imprint term when the read is empty; and the log-linear rule of Eq. (7) produces the final ranking.

3.5 Theoretical Properties

Two properties follow directly from Eq. (7) and are verified empirically in Section 5.

Proposition 1: If $g(q) = 0$ (the imprint cannot be read), then $S_{\text{final}}(q, \cdot) = \hat{S}_{\text{vis}}(q, \cdot)$, so the ranking is identical to the visual baseline.

Proof: Substitute $g(q) = 0$ into Eq. (7). $\square$

The gate blocks the imprint term whenever the read is empty, so a pill whose imprint cannot be read is never harmed; this guarantee is exercised on the imprint-free OGYEIV2 data, whereas on ePillID, where every read is non-empty, the low observed hurt rate comes from the match term rarely outranking the correct class (Section 5).

Proposition 2: Consider a query whose correct class is c^* but which the visual branch ranks below an incorrect class $\hat{c}$, with visual margin deficit $\delta = \hat{S}_{\text{vis}}(q, \hat{c}) - \hat{S}_{\text{vis}}(q, c^*) \geq 0$. Let the imprint match margin be $\Delta m = m(t_q, T_{c^*}) - m(t_q, T_{\hat{c}})$. After fusion, c^* outranks $\hat{c}$ if and only if

$$\lambda g(q) \Delta m > \delta. \quad (8)$$

Proof: c^* outranks $\hat{c}$ iff $S_{\text{final}}(q, c^*) > S_{\text{final}}(q, \hat{c})$, i.e., $\hat{S}_{\text{vis}}(q, c^*) + \lambda g m(t_q, T_{c^*}) > \hat{S}_{\text{vis}}(q, \hat{c}) + \lambda g m(t_q, T_{\hat{c}})$, which rearranges to $\lambda g \Delta m > \delta$. $\square$

Proposition 2 characterises the errors the imprint corrects: visually near-identical pills (small δ) whose imprints differ (large Δm), exactly the failure mode highlighted by the ePillID benchmark. The geometry of this condition is shown in Fig. 5: a wrong visual ranking is rescued if and only if its imprint margin exceeds its visual margin deficit, and the cases observed in our experiments (e.g., TV 7296, TEVA 7238) lie squarely in the rescued region. At inference the reader runs once per query and the re-ranking touches only the top- K visual candidates, adding $O(K)$ cost ($K = 20$) with no change to gallery construction; when the imprint is unreadable the system reduces to the visual baseline (Proposition 1), which is a desirable property for eventual deployment.

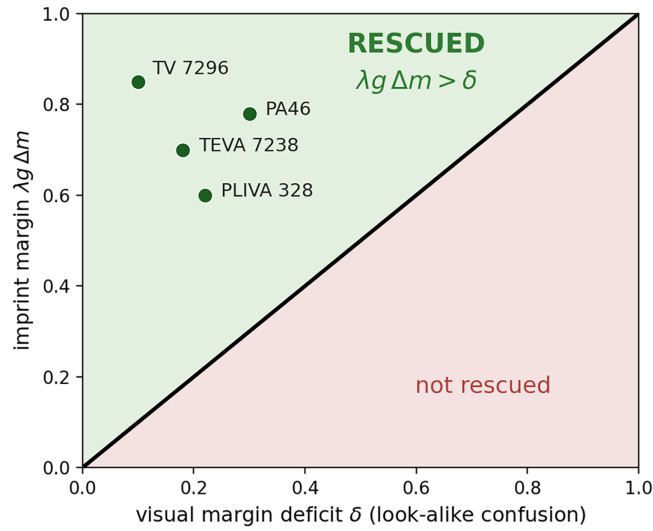


Figure 5: Geometry of the rescue condition (Proposition 2). The decision line $\lambda g \Delta m = \delta$ separates rescued queries (green: imprint margin exceeds visual deficit) from unchanged ones (red). Observed rescued cases lie in the green region: visually near-identical (small δ) yet imprint-separable (large Δm). By Proposition 1, queries with an unreadable imprint ($g = 0$) never move.

4 Experiments

4.1 Datasets

We evaluate on three datasets that span the full range of imprint availability. ePillID [3] is the primary benchmark: 9804 reference and 3728 consumer images over 4902 appearance classes. Ground-truth imprints are not in the public release and online drug APIs return empty records for the discontinued 2016-era National Drug Code (NDC) identifiers; we therefore recover imprint labels from the public Pillbox master data, joining on product code to cover 99.2% of pill types (9730 reference and 3605 consumer images with imprint ground truth). CURE [19] (196 classes, 8971 images) has cluttered backgrounds and free lighting; it has no imprint labels, so we build an imprint database by OCR consensus (the fine-tuned reader applied to each class’s clean reference images), and we add a pill-detection and cropping step ($\sim 97\%$ detection) because the images range from 640^2 up to 2976^2 pixels and the pill occupies only a small fraction of each capture. OGYEIV2 [13] (112 Hungarian-market classes, single-pill images) consists mostly of plain or scored tablets with no imprint, and serves as a no-harm stress test for Proposition 1 (Fig. 6).

4.2 Data Splits and Evaluation Protocol

Table 2 traces the data from the raw release to the final test queries and makes every evaluation denominator explicit. The 4902 appearance classes comprise 3942 reference-only classes and 960 classes that also carry consumer images, and the official release partitions the consumer images of those 960 classes into five disjoint folds of 192 classes each (745–746 images per fold). This fold structure is not our design choice but the official protocol released with the benchmark [3], adopted unchanged so that our numbers are directly comparable with published results. Every class contributes its two reference images to the gallery and to training, because the benchmark is a low-shot retrieval task in which a consumer photograph is matched against a reference gallery that necessarily covers all classes; what is held out is therefore not a class but the consumer images of a class. Fold 4 is held out as the test set in every round, the four cross-validation rounds rotate which of folds 0–3 provides the validation set, and the remaining consumer images (2237 or 2238,

covering 576 classes) join the 9804 reference images for training. The four rounds consequently share one test set (hence the identical test columns of Table 2), so the reported spread across rounds reflects sensitivity to the training/validation partition rather than to the test sample. Images without Pillbox imprint metadata are retained in the evaluation and treated as carrying no database imprint.

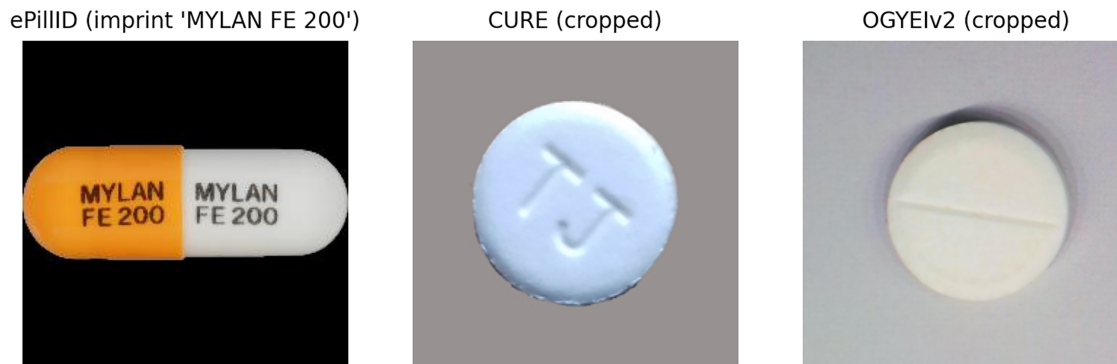


Figure 6: Sample pills spanning the imprint spectrum. Left: ePillID carries clear debossed alphanumeric imprints with Pillbox ground truth (e.g., “MYLAN FE 200”). Centre: CURE pills bear imprints (e.g., “TJ”) but have no imprint metadata, so the database is built by OCR consensus. Right: OGYElv2 pills are predominantly plain or score-lined with no alphanumeric imprint, serving as a no-harm stress test.

Table 2: Data flow from the public ePillID release to the final test queries, per round. The test fold is fixed across the four rounds, so the test columns are identical by construction.

Stage	Round 1	Round 2	Round 3	Round 4
Validation fold (rotating)	0	1	2	3
Appearance classes in the release	4902	4902	4902	4902
Reference images (all classes; gallery and training)	9804	9804	9804	9804
Consumer images used for training (classes)	2237 (576)	2237 (576)	2237 (576)	2238 (576)
Validation consumer images (classes)	746 (192)	746 (192)	746 (192)	745 (192)
Test consumer images, fixed fold 4 (classes)	745 (192)	745 (192)	745 (192)	745 (192)
classes with both a front and a back image	145	145	145	145
their consumer images	630	630	630	630
images with no counterpart side (not pairable)	115	115	115	115
Both-sides queries (all front–back combinations)	626	626	626	626
Excluded for a blank side or an unreadable imprint	0	0	0	0
Final test queries (classes)	626 (145)	626 (145)	626 (145)	626 (145)

The query set is built by the ePillID evaluator, which enumerates, within each pill type, every pair of consumer images that carry different side labels. A pill type can therefore contribute a query only if its consumer images include at least one front and one back: of the 192 classes in the test fold, 145 satisfy this and account for 630 of the 745 consumer images, while the remaining 115 images have no counterpart side and cannot be paired. Enumerating all front–back combinations of those 630 images yields the 626 both-sides queries, which therefore span 145 classes; a class photographed several times on one face contributes more than one pairing, which is why 626 queries draw on 630 images rather than 2×626 . Every query thus observes both faces of the pill, and none is scored single-sided. No query is excluded on the grounds of a

blank side, an unreadable imprint, or a failed read: such cases simply activate the gate $g(q) = 0$ of Eq. (7) and fall back to the visual ranking, so difficult imprint cases remain in the denominator and cannot inflate the reported benefit of the imprint branch. Pooling the four rounds gives 2504 query evaluations, namely the same 626 queries scored once under each round's model; because these evaluations are not independent, the clustered bootstrap of Section 5.3 resamples whole pill classes and fold-class cells rather than individual queries.

A single imprint reader serves all four rounds, fine-tuned exclusively on reference images under the 3593/1197 split of Table 3 (75%/25% by pill type, drawn from the 4790 pill types whose cleaned Pillbox imprint is non-empty): no consumer image enters reader fine-tuning at any point, and the reader's exposure to a test-fold class is limited to the same studio reference images that the gallery itself draws on. This 75%/25% split is not the cross-validation partition, because it serves a different purpose: the four-fold structure operates at the level of the official consumer folds described above, whereas the 3593/1197 split partitions the reader's fine-tuning pool at the pill-type level into an adaptation set and a set of 1197 pill types withheld entirely from adaptation, on which we measure how well the reader generalises to pill types it has never seen (Tables 4 and 5). Nothing is selected on these 1197 types: the reader's hyper-parameters are fixed a priori (Section 4.3) and the evaluated checkpoint is the fixed end-of-training checkpoint, so the split provides a held-out measurement set rather than a model-selection set. Scoring follows the ePillID both-sides protocol: the strings read from the two sides of a pill are merged before scoring against the ground-truth imprint, which is the correct setting because each image shows only one side and many sides are blank. We report Top-1 and Top-5 accuracy and mean average precision (MAP); on ePillID these are given as mean $\pm$ standard deviation over the four cross-validation rounds. For CURE and OGYEIV2 the test classes are unseen during training (a low-shot setting), with the gallery formed from reference images.

Table 3: Configuration of the LoRA-fine-tuned imprint reader.

Component	Detail
Fine-tuning/held-out split	3593/1197 pill types (75%/25%, split by type; no leakage). Held-out types measure generalisation only; distinct from the cross-validation folds (Section 4.2)
Per example Method	Two side images + cleaned imprint label (e.g., "A117", "ADOXA 150MG") LoRA ($r = 16$) on the LLM projections; vision tower frozen
Trainable params	18.5 M (0.83% of 2.2 B); fits an 8 GB GPU (fp16 + gradient checkpointing)
Backbone	Qwen2-VL-2B [8]

Table 4: Progression of LoRA fine-tuning of the imprint reader.

Checkpoint	Training Loss	Held-Out Token Recall (%)	Held-Out Complete-Imprint (%)
Start (epoch 0)	0.67	56.3	42.9
End of epoch 2 (final)	0.29	75.8	64.8

Table 5: Imprint reader before and after LoRA fine-tuning, on the 1197 unseen held-out pill types (both-sides).

Metric (Unseen Held-Out Types)	Baseline	Fine-Tuned	Gain
Reference token recall	64.2	84.7	+20.5
Consumer token recall	56.3	75.8	+19.5
Consumer complete-imprint	42.9	64.8	+21.9

Algorithm 1 summarises which data are used at each stage of one cross-validation round.

Algorithm 1: Per-fold protocol of MIRA-Pill

Input: fold k with training types $\mathcal{C}_k^{tr}$ and test types $\mathcal{C}_k^{te}$ (official ePillID split).

1. Train the visual branch on the images of $\mathcal{C}_k^{tr}$ (100 epochs).
 2. Apply the shared imprint reader, LoRA-fine-tuned once on reference images only (Table 3).
 3. Build the class-imprint database $\{T_c\}$ from Pillbox metadata (all classes; metadata, not images).
 4. Fix $\lambda = 0.2$ and $K = 20$ a priori (no per-fold tuning; Section 4.3).
 5. Score the both-sides consumer queries of $\mathcal{C}_k^{te}$ with Eq. (7) and report Top-1, Top-5, and MAP.
-

4.3 Implementation Details

The visual branch is the ResNet50 + CBP configuration from the original ePillID code base, trained for 100 epochs per fold; we deliberately omit the gradient-reversal and coordinate-attention modules of our earlier work so that a clean baseline isolates the backbone-independent fusion contribution. The imprint reader is Qwen2-VL-2B [8] fine-tuned with LoRA for two epochs (~ 2 h 37 m on a single RTX 4060; AdamW, learning rate 10^{-4} , effective batch size 8, loss converging smoothly from 0.67 to 0.29); Table 3 gives the full configuration and Table 4 the training progression. The fusion weight, $\lambda = 0.2$ on ePillID and $\lambda = 0.3$ on the cross-dataset benchmarks, is the only hyper-parameter introduced by the fusion: we fix it a priori from the product-of-experts interpretation (the imprint term should refine, not dominate, the normalised visual score) rather than tuning it per test fold, and likewise fix the re-ranking window at $K = 20$; the sensitivity of the results to both choices is analysed in Section 5.

The exact prompt supplied to the imprint reader for every query is reproduced verbatim below:

“This is a close-up of a single pill/tablet/capsule. Almost all pills carry imprinted text: letters, numbers or a score line, often faint, debossed (pressed in) or low-contrast. Look very carefully, including faint or partially lit characters, and transcribe EVERY character you can see. Output ONLY the imprint characters, uppercase, separated by spaces, with no explanation. Make your best attempt even if faint; only output NONE if the surface is truly blank.”

The imprint labels from the Pillbox master data were cleaned with the following rules before fine-tuning and matching: (i) convert to uppercase; (ii) treat the side separators ‘;’, ‘|’, and ‘/’ as spaces so both faces of a pill form one token set; (iii) keep only the characters A–Z, 0–9, and space, replacing any other symbol with a space; (iv) collapse repeated whitespace; and (v) remove non-visual Pillbox placeholder tokens such as “LOGO” that denote a graphic rather than readable text. Empty fields yield an empty string, for which the confidence gate is inactive ($g(q) = 0$ in Eq. (7)) and the query falls back to the visual ranking. Token sets in Eq. (5) are obtained by splitting on whitespace after this normalisation.

Finally, the software environments: the two branches run under different ones, and the reader’s adapter is sensitive to the library version in a way that is easy to miss. The visual branch follows the environment of the ePillID release (Python 3.6.7, PyTorch 1.3.1, torchvision 0.4.2, CUDA toolkit 10.1). The imprint reader was fine-tuned and evaluated, and all experiments reported in this paper were run with PyTorch 2.5.1 (CUDA 12.1), Transformers 4.46.2, PEFT 0.13.2, and safetensors 0.7.0 on a single 8 GB GPU. One incompatibility deserves an explicit warning. It was exposed when an adapter produced by a preliminary fine-tuning run on a newer software stack (PEFT 0.19.1 with the correspondingly newer Transformers release) was later loaded in the older, frozen environment listed above; that frozen environment is the one in which every result reported in this paper was produced, the newer stack having been used only in exploratory runs before the versions were fixed for reproducibility. The mechanism is as follows: across Transformers releases the Qwen2-VL language tower changed its module path, newer releases inserting an additional language-model level into

the layer names, so LoRA adapter keys saved by a newer release do not match the names expected by an older one. When the names do not match, PEFT attaches no weights and reports no error, and the adapter silently behaves like the unadapted base model, reproducing the baseline rather than the fine-tuned rows of Table 5. Practitioners who observe baseline-level imprint accuracy from a fine-tuned adapter should therefore verify that the adapter keys match the module names used by the version of the Hugging Face Transformers library installed in their environment and, if they do not, rename the keys accordingly.

5 Results and Discussion

5.1 Feasibility and Fine-Tuning of the Imprint Reader

Before investing in fusion, we tested whether a compact VLM can read pill imprints at all. Off the shelf, Qwen2-VL-2B reaches 63.0% token recall on reference and 65.3% on consumer images under the both-sides protocol (Table 6). When an imprint is visible the reader is often exactly correct (e.g., “MYLAN FE 200”); most apparent “errors” are blank pill sides for which an empty read is in fact correct, which is precisely why the both-sides protocol is necessary. A separate image-preprocessing study (contrast-limited adaptive histogram equalisation, CLAHE, plus 2× upscaling) did not help, reducing reference token recall by 3.8 pp while adding only 1.3 pp on consumer images, indicating that the bottleneck is the model rather than image quality and motivating fine-tuning.

Table 6: Off-the-shelf imprint-reading feasibility (Qwen2-VL-2B, both-sides, 50 pills per domain).

Domain	Token Recall (%)	Complete-Imprint (%)
Reference	63.0	50.0
Consumer	65.3	54.0

To connect this feasibility result with the OCR literature discussed in Section 2, Table 7 compares the VLM against two widely used OCR engines on the same feasibility subset (50 pills per domain): the classical Tesseract engine [26] and the modern lightweight PP-OCrv3 system [27]. Both fail almost completely on this material: Tesseract 5.4 recovers only 6.5% of reference and 4.0% of consumer imprint tokens, and PP-OCrv3’s text detector finds no legible text at all. Debossed, low-contrast, curved imprints are far from the printed-document setting these engines target, which is why the generative VLM is the appropriate reader for this task.

Table 7: Imprint token recall of OCR baselines and the compact VLM on the feasibility subset of Table 6 (50 pills per domain, both-sides).

Reader	Reference (%)	Consumer (%)
Tesseract 5.4 [26]	6.5	4.0
PP-OCrv3 [27]	0.0	0.0
Qwen2-VL-2B (off-the-shelf)	63.0	65.3

LoRA fine-tuning produces large, transferable gains: on the 1197 held-out pill types the reader has never seen, consumer token recall rises from 56.3% to 75.8% (+19.5 pp) and complete-imprint accuracy from 42.9% to 64.8% (+21.9 pp) (Table 5). Because the reader is trained only on reference images yet its consumer-image reading improves by 19.5 pp on pill types it has never seen, the gain is genuine domain transfer rather than

memorisation. Table 4 reports the fine-tuning progression: the loss decreases smoothly over the two epochs with no sign of divergence or over-fitting on the held-out types.

5.2 Main Result on ePillID

Table 8 reports the main result over the four cross-validation rounds under the both-sides protocol. MIRA-Pill raises Top-1 accuracy from 90.34 ± 1.23 to 94.33 ± 1.06 (+3.99 pp) and MAP from 94.60 ± 0.86 to 96.89 ± 0.66 (+2.29 pp). The visual-only baseline (90.34%) is consistent with the published ResNet50 + CBP result ($\approx 91.4\%$), confirming a faithful reproduction; all four rounds improve by +2.9 to +5.0 pp (Fig. 7), far more than the inter-round standard deviation. A prior single-benchmark variant of this pipeline reported a non-significant +0.29 pp gain; MIRA-Pill improves the same baseline by roughly fourteen times as much. Table 9 places these numbers next to the published state of the art: our reproduction is within about one point of the published figure, and the imprint branch adds a gain roughly four times that residual gap.

Table 8: Main result on ePillID (four cross-validation rounds over a fixed held-out test fold, both-sides).

Model (Both-Sides)	Top-1	Top-5	MAP
Visual-only (ResNet50 + CBP)	90.34 ± 1.23	99.36 ± 0.47	94.60 ± 0.86
MIRA-Pill (visual + imprint)	94.33 ± 1.06	99.80 ± 0.15	96.89 ± 0.66
Improvement	+3.99	+0.44	+2.29

Note: Values in bold indicate the best performance achieved by the proposed method.

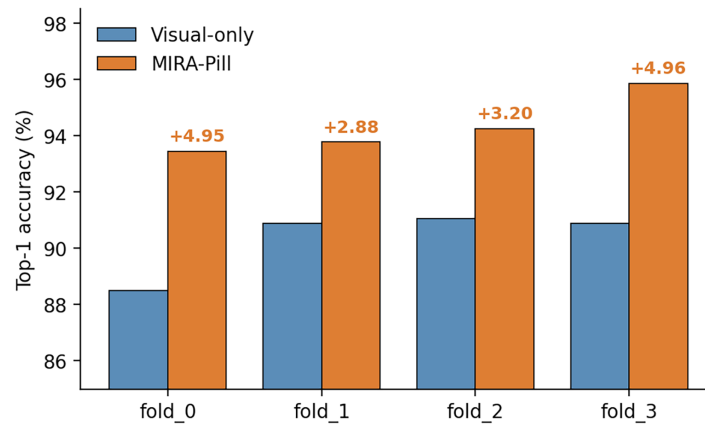


Figure 7: Per-round Top-1 accuracy on ePillID, labelled by the rotating validation fold. All four rounds improve consistently after fusion (+2.9 to +5.0 pp).

Table 9: Comparison with published results on ePillID (both-sides Top-1 accuracy, %).

Method	Top-1 (%)
CLIP zero-shot [20] (our evaluation, fold 3)	10.54
ResNet50 + CBP, as published [3]	≈ 91.4
ResNet50 + CBP, our reproduction (four rounds)	90.34 ± 1.23
MIRA-Pill (this work, four rounds)	94.33 ± 1.06

5.3 Statistical Significance

Because both models are evaluated on the same queries, we use paired tests on the pooled 2504 queries from the four rounds (Table 10). The improvement is significant at every level we test, with p -values from 6×10^{-3} to 1×10^{-21} ; the bootstrap 95% confidence intervals of $\Delta\text{Top-1}$, [3.19, 4.83] pp, and of ΔMAP , [1.84, 2.76] pp, both exclude zero, in direct contrast to the earlier non-significant +0.29 pp result.

Table 10: Statistical tests for MIRA-Pill vs. the visual-only baseline (pooled 2504 queries).

Test	Statistic	p -Value
McNemar (Top-1; $b = 106$, $c = 6$) [28]	$\chi^2 = 87.5$	8.4×10^{-21}
Paired t -test (per-query AP)	$t = 9.7$	1.1×10^{-21}
Wilcoxon (per-query AP)	—	1.0×10^{-19}
Bootstrap 95% CI, $\Delta\text{Top-1}$	[3.19, 4.83] pp	Excludes 0
Bootstrap 95% CI, ΔMAP	[1.84, 2.76] pp	Excludes 0
Fold-level paired t -test ($n = 4$)	Cohen's $d = 3.58$	0.0056

Following standard practice for confirmatory analyses, we designate paired Top-1 accuracy on ePillID as the primary endpoint; all remaining tests are secondary robustness checks. Because queries are nested within pill classes and folds, the pooled McNemar test may understate uncertainty; we therefore add a clustered bootstrap that resamples whole pill classes (and, as a variant, fold–class cells) with 10,000 replicates. We also repeat training with additional random seeds, since both branches involve stochastic optimisation, and we extend the paired analysis to every major claim. We de-emphasise the per-query paired t -test on average precision, which is bounded in $[0, 1]$ and non-normal; the Wilcoxon and bootstrap results are the appropriate ones. Table 11 reports the extended analysis.

Table 11: Extended statistical analysis (primary endpoint: paired $\Delta\text{Top-1}$ on ePillID; 10,000 bootstrap replicates).

Analysis	Result
Class-level clustered bootstrap, 95% CI of $\Delta\text{Top-1}$ (pp)	[1.88, 6.54]
Fold–class clustered bootstrap, 95% CI of $\Delta\text{Top-1}$ (pp)	[2.71, 5.40]
Clustered bootstrap, 95% CI of $\Delta\text{Top-5}$ (pp)	[0.08, 0.95]
Clustered bootstrap, 95% CI of ΔMAP (pp)	[1.09, 3.65]
Multi-seed (3 seeds, fold 3): Top-1, mean $\pm$ std	94.20 ± 1.77
CURE: paired bootstrap 95% CI of $\Delta\text{Top-1}$ (pp)	[3.33, 11.32]
OGYElv2: paired bootstrap 95% CI of $\Delta\text{Top-1}$ (pp)	[0.42, 6.04]
Fine-tuned vs. off-the-shelf reader (fold 3): McNemar p	0.34 ($\chi^2 = 0.9$; n.s.)

5.4 Failure-Case Analysis

Pooling the four rounds, fusion rescues 106 queries (4.2%) while hurting only 6 (0.2%), a net of +100 correct matches that accounts for the +3.99 pp gain (Table 12); the 0.2% hurt rate shows that an active imprint term rarely overrides a correct visual decision, while the no-harm guarantee of Proposition 1 concerns empty reads and is exercised on the imprint-free OGYElv2 data (Table 13). The rescued cases are exactly the near-identical-imprint confusions named by the ePillID benchmark (Table 14): the visual branch selects a near-twin and the imprint read flips the decision back, an instance of the rescue condition of Proposition 2 (small δ , large Δm); Fig. 8 shows representative examples at the image level. Table 15 breaks the same pooled

queries down by gate state: all six hurt cases arise from a misread or hallucinated imprint that matches a wrong class better than the true one (four are the same dosage near-twin, LAMICTAL XR 100 misread as “XR 50”, recurring across rounds; in one the true class carries no imprint metadata, so the spurious read has nothing correct to match), none from a correct read whose string is shared by several classes.

Table 12: Effect of fusion on individual queries (pooled four rounds, 2504 queries).

Case	Count	Rate
Rescued (visual wrong → MIRA correct)	106	4.2%
Hurt (visual correct → MIRA wrong)	6	0.2%
Net rescued	+100	—
Both correct/both wrong	2256/136	—

Table 13: No-harm test on the imprint-free OGYEIV2 dataset (holdout, $\lambda = 0.3$).

Model (OGYEIV2 Holdout)	Top-1	Top-5	MAP
Visual-only (ResNet50 + CBP)	11.25	27.29	21.06
MIRA-Pill (IDF + gating)	14.17	33.33	24.98
Δ	+2.92	+6.04	+3.92

Note: Values in bold indicate the best performance achieved by the proposed method.

Table 14: Qualitative rescued cases: visually near-identical pills disambiguated by the imprint.

Imprint Read (Query)	Visual Prediction (Wrong)	Corrected
TV 7296	TV 7295	TV 7296
TEVA 1084 7238	TEVA 7239	TEVA 7238
T PLIVA 328	PLIVA 467	PLIVA 328
PA46 L	PA83	PA46
15 A	IG 276	A 15

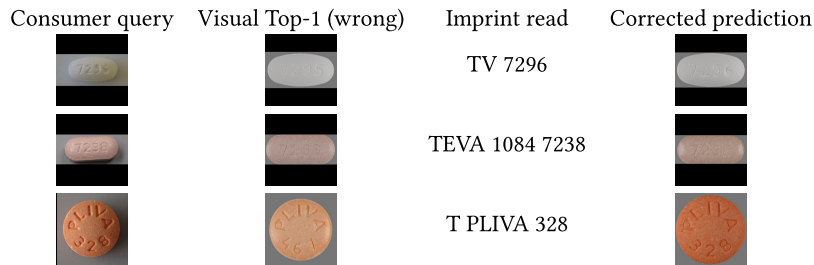


Figure 8: Qualitative examples of rescued queries. Each row shows the consumer query, the visually near-identical class chosen by the visual branch, the imprint string read by the fine-tuned VLM, and the corrected prediction after fusion.

Table 15: Confidence-gate activity on the pooled ePillID queries (2504 queries).

Gate state	Count	Share
Active ($g = 1$: non-empty read)	2504	100.0%
of which final Top-1 correct	2362	94.3%
of which rescued/hurt	106/6	4.2%/0.2%
Inactive ($g = 0$: empty read, visual fallback)	0	0.0%

5.5 Ablation Study

Table 16 ablates the two branches on fold 3. The imprint-only branch has low Top-1 (35.1%) because a single read is ambiguous among 4902 classes, and its MAP (84.5%) should be read as an optimistic bound, since the metric assigns rank one to every class whose database imprint matches the read equally well; the visual branch is strong at Top-1 but fails on near-identical pills; their fusion reaches 95.85%, reflecting the complementary modalities (consistent with Proposition 2). The weak stand-alone performance of the imprint branch and its large contribution in fusion are two sides of the same mechanism, and the reason is ambiguity rather than illegibility. Used alone, the read string must discriminate among all 4902 classes, and an imprint is simply not a unique key at that scale: on fold 3 the set of classes whose database imprint matches the read string as well as the true class does contains on average 565 classes, and exceeds ten classes for 238 of the 626 queries. A branch that cannot separate the true class from several hundred equally matching ones necessarily has low Top-1, which is what Table 16 reports. We stress that the stand-alone Top-1 of this branch is consequently sensitive to how such ties are broken, so it should be read as an order of magnitude rather than a precise operating point; the same caveat applies to its MAP, which counts every tied class as rank one and is therefore an optimistic bound.

Table 16: Branch ablation on ePillID fold 3 (both-sides, $\lambda = 0.2$).

Branch	Top-1	Top-5	MAP
Visual-only	90.89	99.36	94.85
Imprint-only	35.14	62.94	84.52
MIRA-Pill (fusion)	95.85	99.84	97.80

Note: Values in bold indicate the best performance achieved by the proposed method.

In fusion the branch faces a far easier problem, and the same measurement makes this precise: restricted to the $K = 20$ visually shortlisted candidates, the mean number of equally matching classes falls from 565 to 4.52. The shortlist removes almost all of the ambiguity, and the surviving confusable classes are near-twins that differ precisely in their imprint, so even a partial read produces a large match margin Δm exactly where the visual margin deficit δ is small, which is the rescue condition of Proposition 2. Because the imprint term is added to the visual score rather than replacing it, the visual ranking also breaks whatever ties remain inside the shortlist. The net effect on the pooled queries is 106 rescued against 6 hurt (Table 12). The two branches also fail on substantially different queries, the visual branch on imprint-only near-twins and the reader wherever the imprint is not discriminative, which is why combining them helps; on ePillID every query yields a non-empty read, so the confidence gate is always active here (Table 15) and its no-harm role is exercised on the imprint-free OGYElv2 data instead. In short, the imprint branch is not a competing classifier but a high-precision disambiguator whose value is realised once the visual branch has reduced the

candidate set. The fusion weight λ has a single-peaked, robust response around 0.1–0.3 (Fig. 9), and the result is insensitive to the re-ranking window ($K = 5/10/20/50$ all give 95.7%–95.9%).

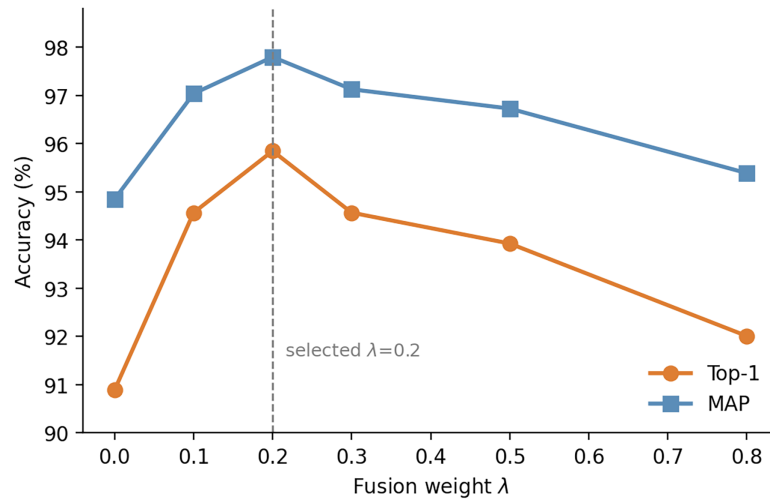


Figure 9: Sensitivity to the fusion weight λ on ePillID fold 3. The response is single-peaked and robust around $\lambda = 0.2$.

Table 17 separates the contribution of fusion from that of fine-tuning. Fusing the visual branch with the off-the-shelf reader already yields +4.32 pp, showing the fusion mechanism is robust even with an unadapted VLM; fine-tuning the reader adds a further +0.64 pp on this fold, a difference that does not reach significance on 626 queries (McNemar $p = 0.34$, Table 11). The principal, directly measured value of fine-tuning is the large improvement in reading quality itself (+19.5 pp consumer token recall, Table 5); its incremental effect on final Top-1 is small because the shortlisted near-twins are often disambiguated even by a partial read.

Table 17: Reader ablation on ePillID fold 3 ($\lambda = 0.2$): fusion vs. fine-tuning.

Configuration	Top-1	Top-5	MAP
Visual-only (no imprint)	90.89	99.36	94.85
Fusion + off-the-shelf reader	95.21 (+4.32)	99.84	97.48
Fusion + fine-tuned reader	95.85 (+4.96)	99.84	97.80

Note: Values in bold indicate the best performance achieved by the proposed method.

5.6 Open-Set and Zero-Shot Capabilities

MIRA-Pill also enables capabilities the visual baseline lacks. For open-set recognition (rejecting queries whose class is not in the gallery), visual confidence already separates known from unknown very well (area under the receiver operating characteristic curve, AUROC, of 99.32%; the false-positive rate at 95% true-positive rate is denoted FPR@95TPR), so the imprint adds only +0.24 pp of AUROC; however, it clearly improves closed-set accuracy on registered queries, from 98.18% to 99.69% (Table 18). For pills absent from the gallery but present in the imprint database (a new drug with a catalogue record but no collected images), the purely visual system is helpless (0%), whereas imprint recovery alone reaches 35.1% Top-1 and 84.5% MAP (Table 19), the latter again an optimistic bound because equally matching classes share rank one: recognition without any gallery image, a capability the benchmark never evaluated.

Table 18: Open-set recognition on ePillID fold 3 (five splits, 50% unknown classes).

System	AUROC	FPR@95TPR	Closed-set acc.
Visual-only	99.32 $\pm$ 0.52	1.55%	98.18%
MIRA-Pill	99.56 $\pm$ 0.35	2.20%	99.69%

Note: Values in bold indicate the best performance achieved by the proposed method.

Table 19: Zero-shot identification of pills absent from the gallery (imprint record only, no images).

System	Top-1	Top-5	MAP
Visual-only (unregistered class)	0%	—	—
Imprint recovery (DB only, 0 images)	35.1%	62.9%	84.5%

Note: Values in bold indicate the best performance achieved by the proposed method.

5.7 VLM Baselines

To position the method against the “just use a VLM” objection, Table 20 reports two VLM baselines under the same identification metrics as every other row. CLIP zero-shot reaches only 10.54% Top-1. The off-the-shelf imprint reader, used alone as an identifier by matching its read string against the imprint database, reaches 26.20% Top-1: it recovers about 65% of imprint tokens (Table 6), but a read string alone is ambiguous among 4902 classes, and its MAP carries the same optimistic tie bound discussed for the fine-tuned imprint-only branch in Table 16. Placed inside the fusion, the same off-the-shelf reader yields 95.2% Top-1 (Table 17), and fine-tuning raises this to 95.9%: the value lies in the specialised framework (metric learning, fine-tuned reader, log-linear fusion) rather than in any single model.

Table 20: VLM baselines and the proposed method on ePillID fold 3.

Method	Top-1	Top-5	MAP
CLIP zero-shot [20]	10.54	30.67	22.12
Qwen2-VL off-the-shelf (imprint only) [8]	26.20	48.08	87.37
Visual-only (ResNet50 + CBP)	90.89	99.36	94.85
MIRA-Pill (proposed)	95.85	99.84	97.80

Note: Values in bold indicate the best performance achieved by the proposed method.

5.8 Cross-Dataset Generalisation: CURE and OGYElv2

To show that MIRA-Pill is not specific to ePillID, we evaluate on two further datasets (Fig. 10). On CURE [19], where we build the imprint database by OCR consensus and evaluate on unseen test classes, fusion improves Top-1 by +7.29 pp and MAP by +6.58 pp (Table 21). Because CURE has no canonical imprints, the reader occasionally hallucinates common tokens on blank surfaces; an inverse-document-frequency (IDF) weighting of the match score suppresses non-discriminative tokens and rewards rare codes, increasing rescues and reducing harm (on ePillID, where the reads are accurate, IDF is neutral and plain matching is kept). The consensus database itself is built as follows: the fine-tuned reader is applied to every clean reference image of a class, the reads are tokenised, and a token is retained for the class entry if it appears in any of the reads (the union of all reference reads with tokens de-duplicated and order preserved); entries are then normalised by the same uppercasing and separator rules as the ground-truth imprints (Section 4.3), with no further manual cleaning. We acknowledge that the same reader contributes both to

database construction and to query reading; the two roles differ in input domain (clean, well-lit reference images vs. unconstrained consumer photographs) and in aggregation (a consensus over several images vs. a single read), and any residual shared-error bias would depress rather than inflate the measured gain, because a systematic misread shared by database and query still has to outrank the visually preferred wrong class. On OGYEIV2 [13], which has essentially no imprints, MIRA-Pill shows no aggregate degradation of the visual baseline (only 5 of 480 queries are hurt), which is consistent with Proposition 1 because the gate leaves the ranking unchanged whenever the read is empty, and it even improves slightly (+2.9 pp) because the reader emits class-consistent tokens that act as a weak appearance descriptor (Table 13).

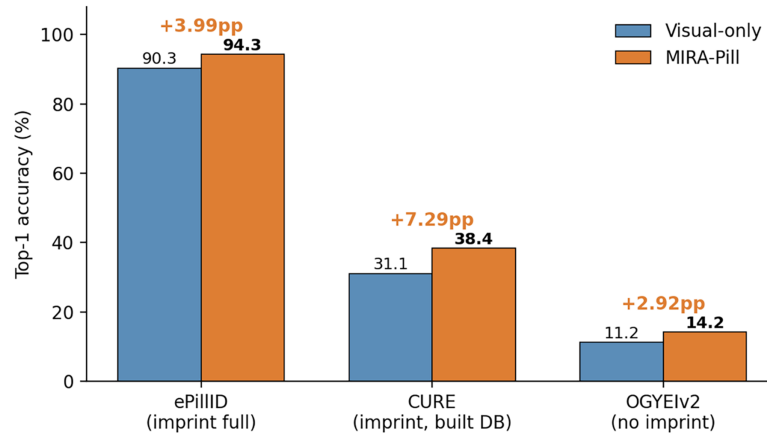


Figure 10: Top-1 accuracy across the three datasets. The gain grows with imprint availability: largest on CURE and ePillID, and smallest (but still no-harm) on the imprint-free OGYEIV2.

Table 21: Cross-dataset result on CURE (unseen test classes, IDF-weighted matching, $\lambda = 0.3$).

Model (Both-Sides)	Top-1	Top-5	MAP
Visual-only (ResNet50 + CBP)	31.09	56.80	43.49
Imprint-only	13.20	23.13	50.79
MIRA-Pill (fusion + IDF)	38.38	62.50	50.07
Improvement	+7.29	+5.70	+6.58

Note: Values in bold indicate the best performance achieved by the proposed method.

5.9 Additional Experiments: Modern Backbone and Soft Gate

Two further experiments complete the empirical picture, and both bear directly on the limitations discussed below.

Modern visual backbone. To confirm that the imprint branch adds value independently of the age of the visual encoder, we replace ResNet50 + CBP by a modern backbone and repeat the fold-3 experiment with the identical reader, database, and fusion rule (Table 22).

Table 22: Backbone replication on ePillID fold 3: the fusion gain persists on a modern visual encoder.

Configuration	Top-1 (%)	MAP (%)
ConvNeXt-T visual-only	89.14	94.27
ConvNeXt-T + imprint fusion	92.81	96.23
Gain from fusion	+3.67	+1.96

Soft confidence gate. We also replace the hard gate of Eq. (7) by the soft gate $g(q) = \sigma(\alpha \bar{\ell}_q + \beta)$, where $\bar{\ell}_q$ is the mean log-probability of the generated string and (α, β) are calibrated without labels from the distribution of the reader’s generation log-probabilities ($\beta = -\text{median}$, $\alpha = 4/\text{IQR}$), with no accuracy label involved. On fold 3 the soft gate yields 93.77% Top-1 vs. 95.85% for the hard gate under an identical reader pass (the hard baseline coincides with the fold-3 main result), and its main qualitative effect is to down-weight low-confidence reads emitted on blank surfaces, which the hard gate treats as full evidence. Under this preliminary distribution-based calibration the soft gate trades rescues (32 to 18) for fewer harmful flips (1 to 0) and therefore does not yet exceed the hard gate on Top-1; a label-calibrated soft gate is left to future work (Table 23).

Table 23: Hard vs. soft confidence gate on ePillID fold 3.

Gate	Top-1 (%)	Rescued	Hurt
Hard, $g = \mathbf{1}[t_q \neq \emptyset]$	95.85	32	1
Soft, $g = \sigma(\alpha \bar{\ell}_q + \beta)$	93.77	18	0

5.10 Discussion

The three datasets tell a consistent story: +3.99 pp on ePillID (full imprints), +7.29 pp on CURE (US imprints, self-built database), and a no-harm result (+2.9 pp) on imprint-free OGYEIV2. The magnitude of the improvement correlates with the availability of imprints, which supports, although a correlation alone cannot prove, the interpretation that the imprint signal rather than incidental tuning is the source of the gain, and the confidence gate keeps the observed harm rate low when imprints are absent. Table 24 tests this reading directly by stratifying the ePillID queries by imprint status: if the imprint signal drives the improvement, the gain should concentrate in the strata where an imprint exists and is actually read, and vanish where it does not; incidental tuning would instead spread uniformly.

Table 24: Gain stratified by imprint availability and readability (pooled ePillID queries).

Stratum	Queries	Visual Top-1 (%)	$\Delta\text{Top-1 (pp)}$
Imprint present, read non-empty	2408	90.49	+4.19
Imprint present, read empty	0	—	—
Imprint absent in metadata	96	86.46	−1.04

The absolute numbers on CURE and OGYEIV2 are lower than on ePillID because their test classes are unseen (a low-shot setting), the images single-sided, and the imprint databases self-built; these are robustness tests rather than leaderboard comparisons. Together with the theory of Section 3, the results show that

reading the imprint with a fine-tuned VLM and fusing it as a calibrated probabilistic factor addresses the ePillID open problem with a low observed harm rate across pills with and without imprints.

A direct numerical comparison with other recent pill-recognition systems is confounded by differing protocols (both-sides setting, fold construction, gallery definition, inclusion of consumer images), so reported accuracies are not directly commensurable. We therefore benchmark against a visual baseline reproduced under an identical protocol, and the contribution we claim is the improvement attributable to the imprint branch (+3.99 pp) rather than a leaderboard ranking; the imprint branch is backbone-agnostic and could be combined with any stronger visual model.

5.11 Limitations

Several limitations bound the present study and motivate future work. First, to isolate the contribution of the imprint branch we keep the original ResNet50 + CBP backbone; a stronger backbone would raise absolute accuracy, which the backbone-agnostic fusion readily accommodates. Second, the four main ePillID rounds use a single training seed; the three-seed replication on fold 3 (Table 11) supports stability, and the fold-level effect size is already very large ($d = 3.58$), but multi-seed runs across all rounds would tighten the variance estimates further. Third, the cross-dataset results on CURE and OGYEIV2 rely on a self-built OCR-consensus imprint database rather than curated ground truth, so dataset-specific reader fine-tuning would likely raise their absolute numbers. Fourth, the confidence gate is currently a hard binary; a soft gate calibrated on the mean generation log-probability is a natural extension, and would in particular down-weight low-confidence hallucinated reads on blank surfaces instead of treating them as full evidence. A preliminary experiment with soft log-linear thresholding on fold 3, together with a replication of the pipeline on a modern visual backbone, is reported in Section 5.9. Fifth, all evidence reported in this paper is retrospective and benchmark-based: the framework has not been evaluated in a live clinical or pharmacy workflow, and such an evaluation would involve questions beyond recognition accuracy (integration into dispensing practice, human oversight, latency, and regulatory validation); accordingly, statements about pharmacy verification and medication safety in this paper describe the motivating application and the framework's suitability as a candidate component, not demonstrated clinical utility. None of these affects the central finding, that reading the imprint with a fine-tuned VLM and fusing it as a calibrated probabilistic factor yields a statistically significant improvement with a low observed harm rate.

6 Conclusion

We revisited the imprint-reading problem that the ePillID benchmark identified as its most important unsolved component and proposed MIRA-Pill, which couples a visual metric-learning branch to a LoRA-fine-tuned vision-language imprint reader through a confidence-gated, log-linear product-of-experts re-ranking rule, supported by a no-harm guarantee and a precise rescue condition. Fine-tuning improves imprint token recall on unseen pill types by 19.5 pp, and on ePillID the framework raises Top-1 accuracy from 90.34% to 94.33% (+3.99 pp) with strong statistical significance, in contrast to a prior, non-significant +0.29 pp improvement. The method generalises (+7.29 pp on CURE) and shows no aggregate degradation on imprint-free pills (5 of 480 OGYEIV2 queries hurt), with the size of the gain tracking imprint availability. Future work will proceed along four concrete directions: (i) replacing the hard confidence gate by a soft gate calibrated on the mean generation log-probability, so that low-confidence reads are down-weighted continuously rather than discarded; (ii) pairing the backbone-agnostic fusion with modern visual encoders such as ConvNeXt or vision transformers to raise absolute accuracy (Section 5.9 verifies that the fusion gain persists on a modern backbone); (iii) dataset-specific reader fine-tuning and curated imprint

databases for CURE-like settings, together with multilingual imprints; and (iv) a prospective evaluation in a pharmacy-verification workflow with multi-seed and clustered statistical reporting.

Acknowledgement: The authors would like to thank Nguyen Tat Thanh University, Van Lang University, and Ho Chi Minh City Open University, Vietnam, for supporting the time and facilities used in this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Study conception and design, analysis and interpretation of results, and draft manuscript preparation: Linh Nguyen Thi My. System implementation, code development, and experiments: Nhat Trinh Ngoc Minh. Manuscript review: Tham Vo and Vinh Truong Hoang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The ePillID dataset is publicly available at <https://github.com/usuyama/ePillID-benchmark>. The CURE dataset is available at <https://github.com/suiyiling/Few-shot-pill-recognition>, and the OGYElv2 dataset at https://github.com/richardRadli/pill_detection. The cross-validation folds are those of the public ePillID release and are used unchanged, and the imprint-label cleaning rules and the exact reader prompt are reported in full in Section 4.3, so the imprint branch can be rebuilt from the public data alone. The trained models and the source code are not publicly available.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as an Editorial Board Member of this journal, Vinh Truong Hoang had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no conflicts of interest.

References

1. Yaniv Z, Faruque J, Howe S, Dunn K, Sharlip D, Bond A, et al. The national library of medicine pill image recognition challenge: an initial report. In: Proceedings of the 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR); 2016 Oct 18–20; Washington, DC, USA. p. 1–9.
2. Nguyen Thi My L, Le VT, Vo T, Truong Hoang V. A comprehensive review of pill image recognition. *Comput Mater Contin.* 2025;82(3):3693–740. doi:10.32604/cmc.2025.060793.
3. Usuyama N, Delgado NL, Hall AK, Lundin J. ePillID Dataset: a low-shot fine-grained benchmark for pill identification. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14–19; Seattle, WA, USA. p. 3971–7.
4. Nguyen TT, Pham HH, Nguyen PL, Nguyen TH, Do M. Multi-stream fusion for class incremental learning in pill image classification. In: Proceedings of the Computer Vision—ACCV 2022; 2023 Dec 4–8; Macao, China. p. 341–56.
5. Nguyen Thi My L, Vo T, Truong Hoang V. Real-time pill image recognition on edge devices with adaptive lightweight attention. *J Real-Time Image Process.* 2026;23(1):53. doi:10.1007/s11554-025-01831-7.
6. Gao Y, Beijbom O, Zhang N, Darrell T. Compact bilinear pooling. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 317–26.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
8. Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191. 2024.
9. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (ICLR 2022); 2022 Apr 25–29; Virtual; 2022.

10. Wong YF, Ng HT, Leung KY, Chan KY, Chan SY, Loy CC. Development of fine-grained pill identification algorithm using deep convolutional network. *J Biomed Inform.* 2017;74(1):130–6. doi:10.1016/j.jbi.2017.09.005.
11. Kwon HJ, Kim HG, Lee SH. Pill detection model for medicine inspection based on deep learning. *Chemosensors.* 2022;10(1):4. doi:10.3390/chemosensors10010004.
12. Tan L, Huangfu T, Wu L, Chen W. Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification. *BMC Med Inform Decis Mak.* 2021;21(1):324. doi:10.1186/s12911-021-01691-8.
13. Rádli R, Vörösházi Z, Czúni L. Metric-based pill recognition with the help of textual and visual cues. *IET Image Process.* 2024;18(14):4623–38. doi:10.1049/ipr2.13273.
14. Ashraf AR, Somogyi-Végh A, Merczel S, Gyimesi N, Fittler A. Leveraging code-free deep learning for pill recognition in clinical settings: a multicenter, real-world study of performance across multiple platforms. *Artif Intell Med.* 2024;150:102844.
15. Kim S, Chae M, Lee J, Lee H. Advanced pharmaceutical recognition system based on deep learning for mobile medication identification. *Appl Sci.* 2025;15(10):5644. doi:10.3390/app15105644.
16. Kavitha N, Madhumathy P. Real-time pill identification and classification using deep learning framework for medicine inspection systems. *Discov Electron.* 2025;2(1):80. doi:10.1007/s44291-025-00122-6.
17. Pujari JJ, Bikku T, Singamaneni KK, Priya Ranjani AC, Ide NAD, Thota S. A hybrid framework for pill identification using convolutional neural networks and optical character recognition. *Discov Artif Intell.* 2026;6(1):649. doi:10.1007/s44163-026-01405-x.
18. Sai Sachin A, Karthikeyan J. A lightweight hybrid deep learning framework for multi-pill detection, multi-attribute recognition, OCR-based imprint analysis, and metadata retrieval. *Front Artif Intell.* 2026;9:1890296. doi:10.3389/frai.2026.1890296.
19. Ling S, Pastor A, Li J, Che Z, Wang J, Kim J, et al. Few-shot pill recognition. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020 Jun 14–19; Seattle, WA, USA. p. 9786–95.
20. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*; 2021 Jul 18–24; Virtual. p. 8748–63.
21. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. *Adv Neural Inf Process Syst.* 2023;36:34892–916. doi:10.52202/075280-1516.
22. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: a survey. *IEEE Trans Pattern Anal Mach Intell.* 2024;46(8):5625–44. doi:10.1109/tpami.2024.3369699.
23. Sun Y, Wen X, Zhang Y, Jin L, Yang C, Zhang Q, et al. Visual-language foundation models in medical imaging: a systematic review and meta-analysis of diagnostic and analytical applications. *Comput Methods Programs Biomed.* 2025;268(9):108870. doi:10.1016/j.cmpb.2025.108870.
24. Ryu JS, Kang H, Chu Y, Yang S. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomed Eng Lett.* 2025;15(5):809–30. doi:10.1007/s13534-025-00484-6.
25. Deng J, Guo J, Yang J, Xue N, Kotsia I, Zafeiriou S. ArcFace: additive angular margin loss for deep face recognition. *IEEE Trans Pattern Anal Mach Intell.* 2022;44(10):5962–79. doi:10.1109/tpami.2021.3087709.
26. Smith R. An overview of the tesseract OCR engine. In: *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*; 2007 Sep 23–26; Curitiba, Brazil. p. 629–33.
27. Li C, Liu W, Guo R, Yin X, Jiang K, Du Y, et al. PP-OCrv3: more attempts for the improvement of ultra lightweight OCR system. *arXiv:2206.03001.* 2022.
28. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* 1947;12(2):153–7. doi:10.1007/bf02295996.