

# CCS: A Continuous Spatial-Semantic Concordance Score for Robust Evaluation of Object Detection Models

Quoc Thai Mai

23133071@student.hcmute.edu.vn

Ho Chi Minh City University of Technology and Engineering

---

## Research Article

**Keywords:** Object detection evaluation, continuous metric, Gaussian overlap, semantic similarity, Wu-Palmer similarity, YOLO, tongue diagnosis

**Posted Date:** July 31st, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10531409/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

**Version of Record:** A version of this preprint was published at Signal, Image and Video Processing on September 23rd, 2026. See the published version at <https://doi.org/10.1007/s11760-026-05721-8>.

# CCS: A Continuous Spatial-Semantic Concordance Score for Robust Evaluation of Object Detection Models

Quoc Thai Mai<sup>1\*</sup>

<sup>1\*</sup>Faculty of IT, HCMC University of Technology and Engineering, No. 1 Vo Van Ngan Street, Thu Duc City, Ho Chi Minh City, 700000, Vietnam.

Corresponding author(s). E-mail(s): [23133071@student.hcmute.edu.vn](mailto:23133071@student.hcmute.edu.vn);

## Abstract

Standard object detection evaluation relies on hard-threshold metrics (mAP, F1) at a fixed IoU threshold. Their binary nature produces unstable rankings, penalizes near-correct detections as harshly as complete misses, and ignores semantic label relationships. We propose CCS, a Continuous Spatial-Semantic Concordance Score with two complementary terms: (i)  $C_{sp}$  models bounding boxes as 2D Gaussians and computes closed-form normalized cosine similarity, smoothly penalizing spatial deviation without a hard threshold; (ii)  $C_{sem}$  uses taxonomy-driven Wu-Palmer similarity to award partial credit when a predicted label is semantically close to ground truth (e.g., “red-spotted tongue” vs. “red tongue”). We benchmark CCS against IoU, GIoU, DIoU, NWD, and COCO mAP@[0.5:0.95] across six YOLO variants (v8/v10/v11) on an eight-class tongue-symptom dataset with severe per-class difficulty imbalance. F1@0.5 ranges 0.47–0.50 and degrades sharply at stricter thresholds; CCS stays stable at 0.62–0.65.  $C_{sem}$  (0.67–0.70) exceeds  $C_{sp}$  (0.57–0.60), showing semantic proximity rescues clinically meaningful near-misses. Five model pairs reverse ranking across IoU thresholds, which CCS avoids by construction; Spearman  $\rho = \mathbf{0.94}$  with F1@0.5 confirms CCS preserves overall ranking. CCS ranking is stable for  $\alpha \in [\mathbf{0.5}, \mathbf{0.7}]$ ; entropy weighting ( $\alpha = \mathbf{0.64}$ ) reproduces the default ranking. These results position CCS as a threshold-independent candidate for detection evaluation in class-imbalanced, semantically structured label spaces.

**Keywords:** Object detection evaluation; continuous metric; Gaussian overlap; semantic similarity; Wu-Palmer similarity; YOLO; tongue diagnosis

# 1 Introduction

Tongue inspection is a core non-invasive diagnostic technique in Traditional Chinese Medicine (TCM), yet it relies on subjective visual interpretation: even trained practitioners show low inter- and intra-rater agreement on standard features such as coating, cracks, and color [1]. This variability has motivated deep learning-based object detection for automated tongue-symptom localization, aided by recent large benchmarks such as TCM-Tongue [2].

Standard practice, inherited from COCO [3], declares a predicted box a match only if IoU with ground truth exceeds a fixed threshold (commonly 0.5, or averaged over 0.5:0.95). This binary decision, underlying Precision, Recall, F1, and mAP, has two damaging consequences in clinical settings: (i) it collapses continuous localization quality into a yes/no decision, so a box just over the threshold is “correct” and one just under is “wrong”; (ii) it treats every incorrect label as equally wrong, discarding clinical proximity between, e.g., a uniformly red tongue and a red-spotted one.

Several works have addressed instability (i). GIoU [4] and DIoU/CIoU [5] give differentiable signal even for non-overlapping boxes, primarily as training losses. For small objects—common in tongue symptoms—IoU-family metrics are extremely sensitive to centroid shifts, motivating Normalized Wasserstein Distance (NWD): boxes are modeled as 2D Gaussians and compared via Wasserstein distance [6]. Our  $C_{sp}$  builds on the same representation but uses closed-form cosine similarity between Gaussian densities, decomposing into scale-penalty and distance-decay terms. Crucially, none of GIoU, DIoU, or NWD address consequence (ii): all remain purely spatial and blind to label semantics.

The complementary idea—that not all label confusions are equally wrong—has a longer history in classification, where hierarchy-aware metrics use graph distance over a taxonomy to award partial credit for semantically “sensible” mistakes [7], building on Wu-Palmer similarity [8]. Adapting this to object detection remains an open challenge, as conventional protocols treat all misclassifications as uniform binary failures.

Consequence (ii) is not specific to TCM. Any detection task with classes along a clinically continuous spectrum faces the same gap—e.g., dermatological lesion subtypes under a diagnostic hierarchy [9] or radiological findings structured by shared ontology [10]. TCM tongue-symptom detection is a natural testbed: it combines severe per-class imbalance with an explicit label hierarchy and a large benchmark [2], but the gap extends beyond this domain.

CCS closes both gaps within a single score, evaluated here on TCM tongue-symptom detection as an initial testbed.  $C_{sp}$  replaces the hard IoU threshold with closed-form Gaussian cosine similarity;  $C_{sem}$  replaces binary label correctness with taxonomy-driven Wu-Palmer similarity, so a red-spotted tongue predicted where a red tongue was annotated receives partial credit rather than being scored as a plain miss. We combine them linearly,  $CCS = \alpha \cdot C_{sp} + \beta \cdot C_{sem}$ , and validate the default  $\alpha = \beta = 0.5$  against data-driven entropy weighting [11], which assigns weights proportional to each term’s information content.

We benchmark CCS against IoU, GIoU, DIoU, NWD, and COCO mAP@[0.5:0.95] on six detectors spanning three YOLO generations—YOLOv8 [12], YOLOv10 [13], and YOLO11 [14]—trained on the eight-class tongue-symptom subset of TCM-Tongue [2].

Hard-threshold F1 is both compressed (0.47–0.50 at IoU 0.5) and unstable—five model pairs invert ranking depending on the IoU cutoff—while CCS is stable (0.62–0.65) and consistent with F1@0.5 (Spearman  $\rho = 0.94$ ,  $n = 6$ ). The gap between  $C_{\text{sem}}$  (0.67–0.70) and  $C_{\text{sp}}$  (0.57–0.60) suggests a non-trivial share of what hard metrics record as error is near-miss localization rescued by semantic proximity. Entropy weighting recovers  $\alpha = 0.64$ , close to and reproducing the default  $\alpha = 0.5$  ranking. These results are evidence CCS is a viable candidate, not a finished general-purpose solution: all detectors are YOLO variants on a single dataset; extension to other architectures, domains, and clinical validation is left to future work (Section 5).

Section 2 reviews related work. Section 3 formulates CCS. Section 4 presents setup and results. Section 5 discusses implications and limitations. Section 6 concludes.

## 2 Related works

### 2.1 Spatial overlap metrics

IoU-thresholded matching (PASCAL VOC, COCO [3]) remains the de facto standard for deciding true positives. Continuous relaxations—GIoU [4], DIoU/CIoU [5], NWD [6]—were introduced primarily as training losses, not evaluation protocol replacements. Even detectors trained with smooth losses are still scored by thresholding IoU into a binary decision.  $C_{\text{sp}}$  targets this mismatch, reusing the Gaussian-box representation [6] but defined for evaluation rather than as a loss.

### 2.2 Joint quality metrics

Panoptic Quality (PQ) [15] combines localization and recognition by factorizing into Segmentation Quality (mean IoU over matched pairs) and Recognition Quality (F-score-like). Conceptually the closest prior two-term score, PQ still requires  $\text{IoU} > 0.5$  to decide which pairs are matched, inheriting the same discontinuity. CCS avoids this by construction:  $C_{\text{sp}}$  requires no matching threshold—every label is scored via Gaussian overlap (or  $s_k = 0$  if no counterpart exists) rather than a binary match/no-match gate.

### 2.3 Semantic and hierarchy-aware evaluation

Independently, classification literature has argued that not all misclassifications should be penalized equally. Wu-Palmer similarity [8] quantifies concept proximity via taxonomy distance; hierarchy-aware metrics score mistakes by graph distance rather than treating every wrong label identically [7]. This has not been extended to object detection, where a predicted label is either an exact match or a full miss—even when labels are adjacent taxonomy nodes (e.g., “red tongue” vs. “red-spotted tongue”).  $C_{\text{sem}}$  imports this idea, applying Wu-Palmer similarity over a provisional eight-class taxonomy.

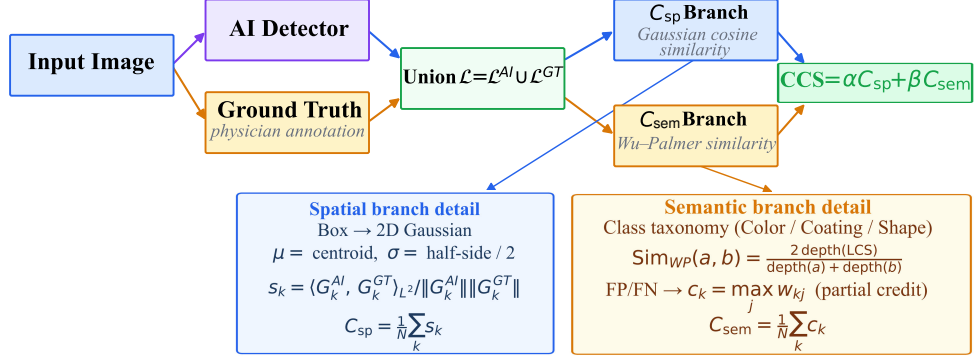


Fig. 1: CCS Framework: Continuous Spatial–Semantic Concordance Score

## 2.4 Detection evaluation in TCM and related medical imaging

Prior studies applied object detection to TCM tongue features: Chang et al. [17] used YOLOv4-tiny for fissures, tooth marks, coatings (AP 47.67–71.25%); Jiang et al. [18] used Faster R-CNN over seven shape/texture categories; the TCM-Tongue benchmark [2] evaluated nine detectors under COCO protocol. All use threshold-based, label-exact evaluation: a red-spotted tongue detected as red tongue is scored identically to a complete miss.

This gap extends beyond TCM. Dermatology organizes lesion labels under expert hierarchies [9]; radiology uses ontologies like RadLex [10]. Yet object detection in these domains still uses threshold-based, label-exact protocols. No prior work combines continuous spatial and semantic-proximity terms in a single detection score. CCS is proposed to close this gap, evaluated here on TCM as an initial testbed.

## 3 Proposed method

### 3.1 CCS: A Weighted Spatial–Semantic Combination

For each image, let  $\mathcal{L}^{AI}$  and  $\mathcal{L}^{GT}$  denote the predicted and annotated label sets,  $\mathcal{L} = \mathcal{L}^{AI} \cup \mathcal{L}^{GT}$  with  $N = |\mathcal{L}|$ . CCS is a convex combination of two per-image terms in  $[0, 1]$ :

$$\text{CCS} = \alpha \cdot C_{sp} + \beta \cdot C_{sem}, \quad \alpha + \beta = 1, \quad (1)$$

where  $C_{sp}$  (Section 3.2) scores spatial region overlap, and  $C_{sem}$  (Section 3.3) scores label proximity in a clinical taxonomy, independent of box overlap. Both terms use the union  $\mathcal{L}$ , so false positives and false negatives contribute to both rather than being discarded. We report  $C_{sp}$  and  $C_{sem}$  separately alongside CCS to distinguish localization errors from label confusions.

### 3.2 $C_{\text{sp}}$ : Spatial Concordance via Gaussian Distance Decay

For a box  $B_k = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ , we represent it as an axis-aligned 2D Gaussian centered at the centroid with  $\sigma$  set from half-width/height:

$$\mu_x = \frac{x_{\min} + x_{\max}}{2}, \quad \mu_y = \frac{y_{\min} + y_{\max}}{2}, \quad \sigma_x = \frac{x_{\max} - x_{\min}}{4}, \quad \sigma_y = \frac{y_{\max} - y_{\min}}{4}, \quad (2)$$

so  $[\mu \pm 2\sigma]$  spans the box extent.  $\sigma$  is computed independently for AI and GT boxes:

$$G_k^S(x, y) = \frac{1}{2\pi\sigma_x^S\sigma_y^S} \exp\left(-\frac{(x - \mu_x^S)^2}{2(\sigma_x^S)^2} - \frac{(y - \mu_y^S)^2}{2(\sigma_y^S)^2}\right), \quad S \in \{AI, GT\}. \quad (3)$$

The pairwise spatial concordance  $s_k$  is the normalized inner product (cosine similarity) between  $G_k^{AI}$  and  $G_k^{GT}$ :

$$s_k = \frac{\iint G_k^{AI} G_k^{GT} dx dy}{\sqrt{\iint (G_k^{AI})^2 dx dy \cdot \iint (G_k^{GT})^2 dx dy}}. \quad (4)$$

Assuming  $x$ - $y$  independence, Eq. (4) admits a closed form factorizing into scale penalty and distance decay:

$$s_k = \underbrace{\sqrt{\frac{2\sigma_x^{AI}\sigma_x^{GT}}{(\sigma_x^{AI})^2 + (\sigma_x^{GT})^2}} \sqrt{\frac{2\sigma_y^{AI}\sigma_y^{GT}}{(\sigma_y^{AI})^2 + (\sigma_y^{GT})^2}}}_{\text{scale penalty}} \cdot \underbrace{\exp\left(-\frac{\Delta x^2}{2((\sigma_x^{AI})^2 + (\sigma_x^{GT})^2)} - \frac{\Delta y^2}{2((\sigma_y^{AI})^2 + (\sigma_y^{GT})^2)}\right)}_{\text{Gaussian distance decay}}, \quad (5)$$

with  $\Delta x = \mu_x^{AI} - \mu_x^{GT}$ ,  $\Delta y = \mu_y^{AI} - \mu_y^{GT}$ . The scale penalty equals 1 when  $\sigma^{AI} = \sigma^{GT}$  and decays as scales diverge; the distance-decay falls off smoothly with centroid distance normalized by pooled variance. For equal-sized boxes, the scale penalty collapses to 1, giving  $s_k = \exp(-\Delta x^2/(4\sigma_x^2) - \Delta y^2/(4\sigma_y^2))$ .

#### **Relation to NWD.**

Both  $C_{\text{sp}}$  and NWD [6] model boxes as 2D Gaussians to avoid IoU discontinuities for small objects. NWD uses 2-Wasserstein distance mapped to  $[0, 1]$  via  $\exp(-\sqrt{\text{NWD}}/C)$  with a tunable calibration constant  $C$ .  $s_k$  is derived directly as cosine similarity between density functions, giving a closed form in  $\mu^{AI}, \mu^{GT}, \sigma^{AI}, \sigma^{GT}$  with no free constant, and factorizing into scale/distance terms.  $s_k$  is *not* the Bhattacharyya coefficient (which has  $4(\sigma^{AI})^2 + 4(\sigma^{GT})^2$  in the denominator rather than  $2(\cdot)$ ); we refer to it as ‘‘Gaussian cosine similarity’’ throughout. We also note  $C_{\text{sp}}$  sets  $\sigma = \text{width}/4$  (so

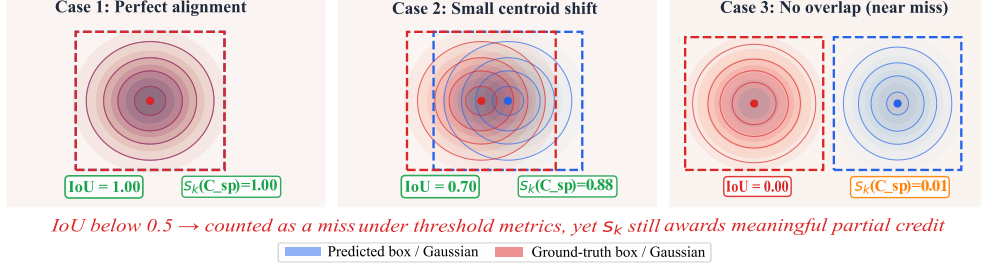


**Fig. 2:** Gaussian Cosine Similarity ( $s_k$ ) vs. IoU: soft spatial credit without hard threshold.

$[\mu \pm 2\sigma]$  spans the box) while NWD uses  $\sigma = \text{width}/2$ ; each is internally consistent, and the two are reported as independently defined similarity functions (Section 4.7).

If class  $k$  appears in only one side (false positive or false negative), there is no counterpart box, so  $s_k = 0$ . Since  $s_k \in (0, 1]$  when both boxes exist ( $s_k = 1$  for identical Gaussians), this is the only discontinuity in  $C_{\text{sp}}$ —a missing class is qualitatively different from a mislocalized box.

Aggregation:

$$C_{\text{sp}} = \frac{1}{N} \sum_{k \in \mathcal{L}} s_k, \quad \mathcal{L} = \mathcal{L}^{AI} \cup \mathcal{L}^{GT}. \quad (6)$$

For multiple instances of one class, we keep only the best-matching (AI, GT) pair;  $C_{\text{sp}}$  is thus class-level, not instance-level (Section 3.4).

### 3.3 $C_{\text{sem}}$ : Semantic Concordance via Taxonomy and Wu–Palmer Similarity

To distinguish “close” from “unrelated” label confusions, we organize the eight classes into a shallow hierarchy with three root branches—**Color**, **Coating**, **Shape** (Fig 3). *Color* groups **hongshe** (uniformly red) and **hongdianshe** (red-spotted), differing only in spatial distribution. *Coating* splits into *nature* (**botaishe**) and *color* (**baitaishe**, **huangtaishe**). *Shape* groups **pangdashe**, **liewenshe**, and **chihenshe**. This taxonomy is *provisional*—not validated by a TCM expert—and its grouping is a first-order determinant of  $C_{\text{sem}}$  (Section 3.4), not a neutral constant.

For labels  $a, b$ , let  $\text{LCS}(a, b)$  denote their Lowest Common Subsumer. Wu–Palmer similarity [8] is

$$\text{Sim}_{WP}(a, b) = \frac{2 \cdot \text{depth}(\text{LCS}(a, b))}{\text{depth}(a) + \text{depth}(b)}, \quad \text{Sim}_{WP}(a, a) := 1. \quad (7)$$

Equivalently, with  $d_{r,\ell} = \text{depth}(\text{LCS}(a, b))$ ,  $d_{a,\ell} = \text{depth}(a) - d_{r,\ell}$ ,  $d_{b,\ell} = \text{depth}(b) - d_{r,\ell}$ :

$$\text{Sim}_{WP}(a, b) = \frac{2 d_{r,\ell}}{2 d_{r,\ell} + d_{a,\ell} + d_{b,\ell}}. \quad (8)$$

**Table 1:** Definitions, feature scopes, and Traditional Chinese Medicine (TCM) clinical significance of the eight tongue-symptom classes [2].

| Label              | English Description   | Scope  | TCM Clinical Significance                                                                                                                                                     |
|--------------------|-----------------------|--------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>hongshe</b>     | Red tongue            | Global | Erythema indicates the progression of heat syndrome (excess fire/qi). Color intensity correlates with the severity of pathogenic heat; e.g., deep red indicates extreme heat. |
| <b>hongdianshe</b> | Red-spotted tongue    | Local  | Hyperplasia of fungiform papillae presenting as small dot-like lesions. Often associated with toxic heat accumulation or blood stasis.                                        |
| <b>botaishe</b>    | Peeled tongue coating | Local  | Partial or complete loss of the tongue coating (geographic tongue), typically appearing pale white. Signifies Spleen-Stomach Qi deficiency and exhaustion of Yin fluid.       |
| <b>baitaishe</b>   | White tongue coating  | Global | Physiologically, the coating is thin and white. Pathologically, a thick white coating indicates obstruction due to cold-dampness or food stagnation.                          |
| <b>huangtaishe</b> | Yellow tongue coating | Global | The transition from a white to a yellow coating indicates increasing internal heat, such as Stomach heat or active inflammatory processes.                                    |
| <b>pangdashe</b>   | Enlarged / Fat tongue | Global | An enlarged or protruding tongue body. Indicates accumulation of cold-dampness due to Spleen or Kidney Yang deficiency, leading to internal fluid retention.                  |
| <b>liewenshe</b>   | Fissured tongue       | Local  | Cracks or grooves of various shapes on the tongue surface. May signify the depletion of essence and blood or chronic Yin deficiency.                                          |
| <b>chihenshe</b>   | Teeth-marked tongue   | Local  | Indentations on the tongue margins caused by pressure from the teeth. A hallmark sign of Spleen Qi or Yang deficiency, often accompanied by internal dampness accumulation.   |

When  $\text{LCS}(a, b)$  is the root ( $d_{r,\ell} = 0$ ),  $\text{Sim}_{WP} = 0$  regardless of clinical relatedness—hence the branch assignment (Fig 3) is a consequential modeling decision [7].

Applying Eq. (7) to all class pairs yields  $W_{\text{sem}} \in [0, 1]^{8 \times 8}$  (Table 3). Non-zero off-diagonal entries occur only within a branch or between **botaishe** and coating-color classes (sharing the *Coating* ancestor); cross-branch pairs receive similarity 0.

For  $k \in \mathcal{L}$ , let  $w_{kj} = \text{Sim}_{WP}(k, j)$ . Per-class semantic credit is:

$$c_k = \begin{cases} 1, & k \in \mathcal{L}^{AI} \cap \mathcal{L}^{GT} \\ \max_{j \in \mathcal{L}^{GT}} w_{kj}, & k \in \mathcal{L}^{AI} \setminus \mathcal{L}^{GT} \\ \max_{j \in \mathcal{L}^{AI}} w_{kj}, & k \in \mathcal{L}^{GT} \setminus \mathcal{L}^{AI} \end{cases} \quad (9)$$

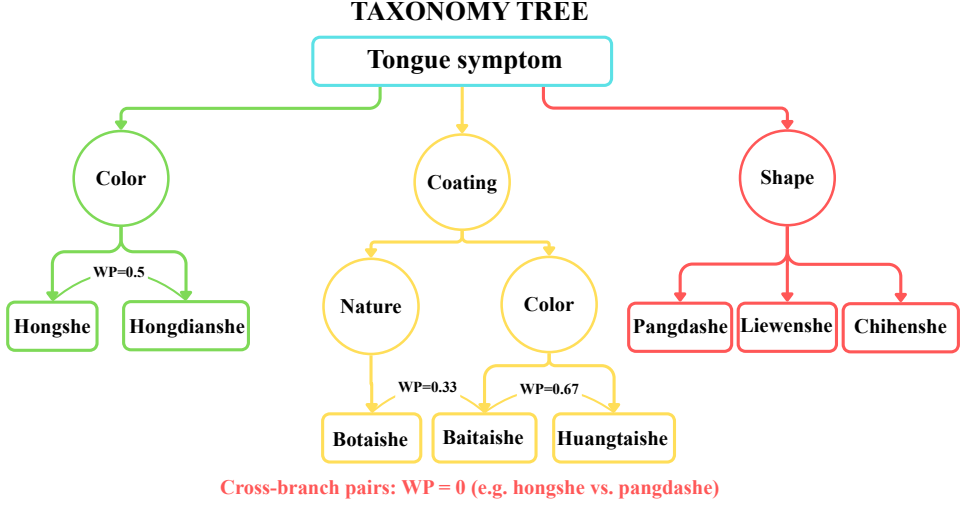


**Fig. 3:** Provisional taxonomy of tongue symptoms used to compute semantic concordance through Wu–Palmer similarity.

aggregated as  $C_{\text{sem}} = \frac{1}{N} \sum_{k \in \mathcal{L}} c_k$ . Unlike conventional label accuracy, a false positive/negative is not automatically 0: it receives credit equal to its highest similarity against the other side, so a predicted `hongdianshe` where the physician annotated `hongshe` scores  $c_k = 0.5$ . This mirrors  $C_{\text{sp}}$ ’s symmetric treatment of missing counterparts.

**Worked example.**

In Fig. 4,  $\mathcal{L}^{AI} = \{\text{hongshe}, \text{baitaishe}\}$ ,  $\mathcal{L}^{GT} = \{\text{hongshe}, \text{huangtaishe}, \text{chihenshe}\}$  ( $N = 4$ ). `hongshe` is an exact match ( $c = 1$ ). `baitaishe` (false positive) is credited  $\max(w_{\text{baitaishe}, \text{hongshe}}, w_{\text{baitaishe}, \text{huangtaishe}}, w_{\text{baitaishe}, \text{chihenshe}}) = \max(0, 0.667, 0) = 0.667$ . `huangtaishe` (false negative) is credited  $\max(w_{\text{huangtaishe}, \text{hongshe}}, w_{\text{huangtaishe}, \text{baitaishe}}) = 0.667$ . `chihenshe` (false negative) has no relation, credited 0. Hence  $C_{\text{sem}} = (1 + 0.667 + 0.667 + 0)/4 = 0.583$ , versus 0.25 under exact-match scoring.

As a reference, mean off-diagonal Wu–Palmer similarity across  $W_{\text{sem}}$  is  $\bar{w} = 0.119$ , well below the observed  $C_{\text{sem}}$  range of 0.67–0.70 (Table 4); the taxonomy does not trivially inflate credit. This is corroborated by Section 4.9: pooled across all detectors, only 39.5% of hard-class errors have any semantic overlap (Table 10).

### 3.4 Design Choices and Assumptions

- **Default weights.** We use  $\alpha = \beta = 0.5$  (Eq. (1)), reflecting no preference between spatial and semantic concordance. This value is *not* tuned on the data; Section 4 reports a sensitivity sweep and entropy weighting [11] as an independent check.

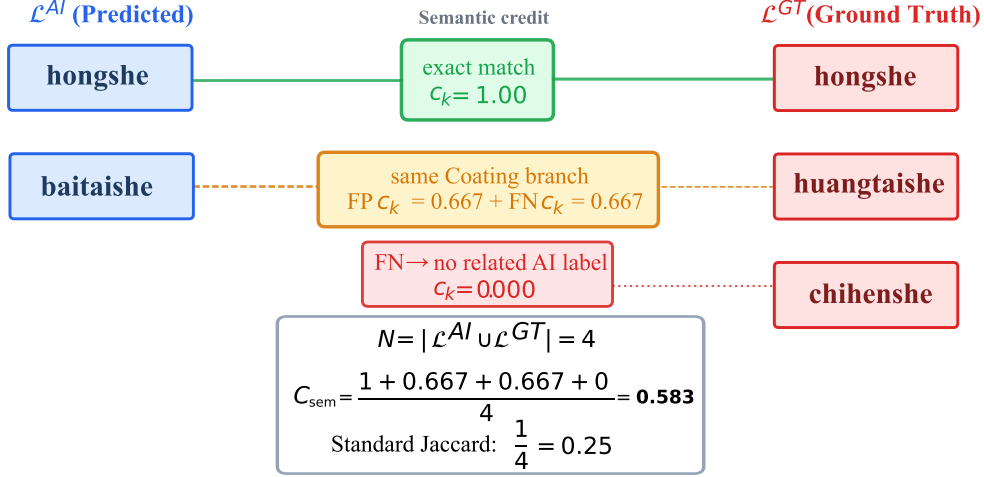


**Fig. 4:**  $C_{\text{sem}}$  Worked Example: Partial Credit for Semantically Close Label Errors.

- **Class-level, not instance-level, scoring.** Both terms score at most one region per class per image:  $C_{\text{sp}}$  keeps only the best-matching (AI, GT) pair. This matches the 8-class annotation convention (each symptom marked as at most one region per image), but CCS is not a drop-in replacement for instance-level metrics. Extending to instance level would require Hungarian matching before Eq. (5). Table 2 quantifies the uneven distribution: three hardest classes (**hongdianshe**, **chihenshe**, **botaishe**) have highest instance density, while whole-tongue classes (**hongshe**, **pangdashe**, **huangtaishe**) annotate  $\approx 1$  instance/image. **hongdianshe** is extreme: mean 6.33 (up to 63) instances per image, 84.2% of its GT regions exceed the single-instance convention, making CCS largely insensitive to how many spots the detector actually finds. This is the most significant open limitation.

**Table 2:** Ground-truth instance density per class and the resulting share of GT regions unscored under the class-level convention.

| Class       | Mean instances/image | Max/image | % GT unscored |
|-------------|----------------------|-----------|---------------|
| hongdianshe | 6.33                 | 63        | 84.2          |
| chihenshe   | 1.55                 | 7         | 35.5          |
| botaishe    | 1.48                 | 5         | 32.3          |
| liewenshe   | 1.18                 | 8         | 15.5          |
| baitaishe   | 1.01                 | 3         | 0.8           |
| hongshe     | 1.00                 | 1         | 0.0           |
| pangdashe   | 1.00                 | 1         | 0.0           |
| huangtaishe | 1.00                 | 1         | 0.0           |

- **Provisional taxonomy.** The grouping (Fig 3) was constructed to make  $C_{\text{sem}}$  computable, not validated by TCM expert consensus. We treat it as a revisable modeling assumption (Section 5.1) and expose the full similarity matrix (Table 3) so alternative taxonomies can be substituted.

## 4 Experiments and Results

### 4.1 Experimental Setup

We benchmark six detectors spanning three YOLO generations—YOLOv8n/m [12], YOLOv10n/m [13], YOLO11n/m [14]—on the eight-class tongue-symptom subset (Section 3) at  $1024 \times 1024$  resolution. All models are fine-tuned from pretrained weights for max 100 epochs with cosine LR schedule (initial 0.01, final fraction 0.01), 3.0 warmup epochs, weight decay 0.0005, and early-stopping patience 20 based on validation loss. Standard Ultralytics augmentations apply except HSV jitters (restricted to 0.015, 0.3, 0.2 respectively) to preserve diagnostic color features.

Detections are filtered at confidence 0.25 with NMS IoU 0.5 for all Precision/Recall/F1/CCS results; a near-zero threshold (0.001) is used only for the PR curve needed by COCO mAP@[.5:.95] (ten IoU thresholds 0.5 to 0.95, step 0.05). We emphasize these are front-end filtering steps required by all protocols—not the AI-to-GT matching decision that CCS replaces with continuous  $s_k$  (Eq. (5)). CCS, like F1, is reported at a single operating point (0.25) rather than swept; sensitivity to this choice is left to future work. All detectors share the same held-out split ( $n \approx 522$ –534 images per comparison;  $n = 3300$  pooled class-level samples for entropy weighting in Section 4.6).

For each detector we report: (i) IoU/GIoU/DIoU-matched Precision, Recall, F1, macro-F1 at  $\{0.3, 0.5, 0.7\}$  under greedy assignment; (ii) COCO mAP@[.5:.95] under each matching metric; (iii) CCS,  $C_{\text{sp}}$ ,  $C_{\text{sem}}$  at  $\alpha = \beta = 0.5$  (threshold-free by construction); (iv) NWD [6] as spatial baseline; (v) paired bootstrap CIs, Wilcoxon tests, and Spearman correlation across  $\binom{6}{2} = 15$  comparisons, Bonferroni-corrected at  $\alpha_{\text{corr}} = 0.05/15 \approx 0.0033$ .

### 4.2 Overall Ranking: CCS and mAP Disagree

Table 4 reports CCS, its components, mAP@[.5:.95], and F1@0.5 sorted by CCS. YOLOv8m has the highest CCS (0.6475) but fifth-highest mAP (0.3205); YOLOv10n has the highest mAP (0.3438) but lowest CCS (0.6191). Spearman  $\rho(\text{CCS}, \text{mAP}) = -0.37$ —a weak negative association—contrasting with  $\rho(\text{CCS}, \text{F1@0.5}) = 0.94$  (Section 4.8). Fig. 5 visualizes this: the model mAP selects as best (YOLOv10n) is weakest by CCS, because mAP weights strict IoU thresholds (up to 0.95), rewarding pixel-precise localization on well-handled classes while remaining blind to missed or semantically confused classes.



**Table 4:** Overall ranking by CCS vs. mAP@[.5:.95] (IoU-matched). Highest value per column in **bold**.

| Model    | CCS           | $C_{sp}$      | $C_{sem}$     | mAP@[.5:.95]  | F1@0.5        |
|----------|---------------|---------------|---------------|---------------|---------------|
| YOLOv8m  | <b>0.6475</b> | 0.5940        | <b>0.7010</b> | 0.3205        | <b>0.4998</b> |
| YOLO11n  | 0.6453        | <b>0.5962</b> | 0.6945        | 0.3277        | 0.4952        |
| YOLO11m  | 0.6311        | 0.5800        | 0.6821        | 0.3209        | 0.4952        |
| YOLOv10m | 0.6308        | 0.5738        | 0.6877        | 0.3401        | 0.4845        |
| YOLOv8n  | 0.6232        | 0.5731        | 0.6734        | 0.3126        | 0.4764        |
| YOLOv10n | 0.6191        | 0.5719        | 0.6663        | <b>0.3438</b> | 0.4699        |

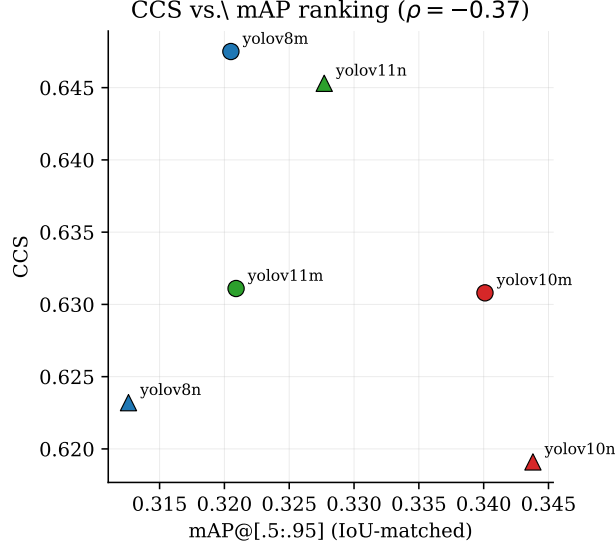


**Fig. 5:** CCS vs. mAP@[.5:.95] across six detectors. The model with the highest mAP@[.5:.95] (YOLOv10n) has the lowest CCS, illustrating that the two metrics rank models by different, only weakly related, notions of quality ( $\rho = -0.37$ ).

### 4.3 Threshold Sensitivity: Hard Metrics Degrade, CCS Does Not

Table 5 and Fig. 6(a) show IoU-matched F1 at  $\{0.3, 0.5, 0.7\}$ . F1 drops monotonically and substantially as threshold tightens: averaged over six detectors, F1 falls 15.0% relative from 0.3 to 0.7 (range 10.8–16.9%). CCS, computed once per image without any matching threshold, is identical across threshold columns; Fig. 6(b) shows all six CCS values in a narrow 0.6191–0.6475 band. Because  $s_k$  decays smoothly with centroid distance rather than being gated at a cutoff, CCS avoids the threshold-dependent instability motivating Section 4.4.

**Table 5:** IoU-matched F1 across matching thresholds vs. threshold-free CCS.  $\Delta$  is the relative drop in F1 from threshold 0.3 to 0.7.

| Model    | F1@0.3 | F1@0.5 | F1@0.7 | $\Delta(0.3 \rightarrow 0.7)$ | CCS    |
|----------|--------|--------|--------|-------------------------------|--------|
| YOLOv8n  | 0.5031 | 0.4764 | 0.4310 | -14.3%                        | 0.6232 |
| YOLOv8m  | 0.5279 | 0.4998 | 0.4395 | -16.8%                        | 0.6475 |
| YOLOv10n | 0.4889 | 0.4699 | 0.4362 | -10.8%                        | 0.6191 |
| YOLOv10m | 0.5082 | 0.4845 | 0.4321 | -15.0%                        | 0.6308 |
| YOLO11n  | 0.5251 | 0.4952 | 0.4390 | -16.4%                        | 0.6453 |
| YOLO11m  | 0.5247 | 0.4952 | 0.4363 | -16.9%                        | 0.6311 |

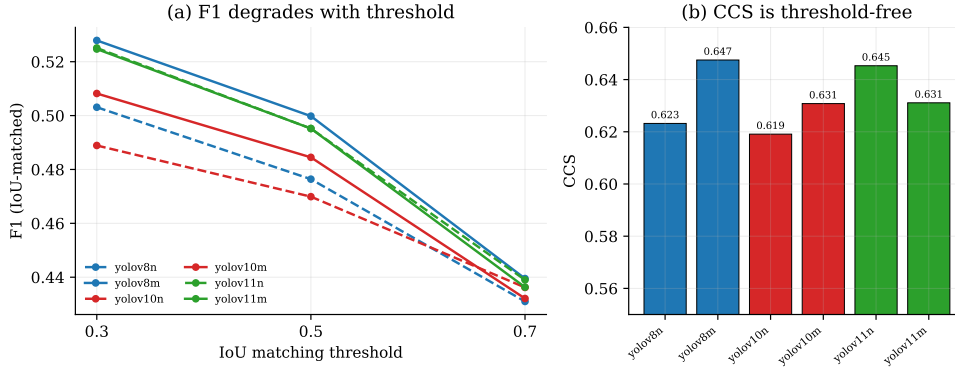


**Fig. 6:** (a) F1 (IoU-matched) declines steadily as the matching threshold tightens from 0.3 to 0.7 for all six detectors. (b) CCS, computed once per image without a matching threshold, is stable and does not depend on the choice of threshold in (a).

#### 4.4 Ranking Instability Across Thresholds: Flipped Pairs

For each detector pair, we compare the CCS winner against the winner under IoU/GIoU/DIoU-matched F1 at each threshold. Of 15 pairs, 5 (33%) contain at least one flip (Table 6). Four of five flips cluster at IoU = 0.7, consistent with IoU-family sensitivity to small centroid shifts [6]: YOLOv8n vs. YOLOv10n and YOLOv10n vs. YOLOv10m reverse under all three matching variants at 0.7; YOLOv10n vs. YOLO11m reverses under GIoU/DIoU at 0.7. The exception is YOLO11n vs. YOLO11m flipping at loose thresholds (0.3–0.5), showing instability is not confined to strict thresholds.

#### 4.5 Qualitative Analysis

We inspected images flagged by three criteria: (a) IoU < 0.45 but  $s_k > 0.6$ ; (b) false positive/negative with best semantic match > 0.4; (c) multiple boxes for the same

**Table 6:** Model pairs whose F1-based ranking disagrees with the CCS ranking at one or more IoU/GIoU/DIoU thresholds.

| Model A  | Model B  | $CCS_A$ | $CCS_B$ | Disagreements               |
|----------|----------|---------|---------|-----------------------------|
| YOLOv8n  | YOLOv10n | 0.6232  | 0.6191  | IoU/GIoU/DIoU@0.7           |
| YOLOv8m  | YOLO11n  | 0.6475  | 0.6453  | GIoU@0.7                    |
| YOLOv10n | YOLOv10m | 0.6191  | 0.6308  | IoU/GIoU/DIoU@0.7           |
| YOLOv10n | YOLO11m  | 0.6191  | 0.6311  | GIoU@0.7, DIoU@0.7          |
| YOLO11n  | YOLO11m  | 0.6453  | 0.6311  | DIoU@0.3, IoU@0.5, GIoU@0.5 |

class. Across six detectors, 44 image-model instances were flagged (18 each, with overlap).

*Localization near-misses rescued by  $C_{sp}$ .*

On `A(16).jpg`, YOLOv8m detects `chihenshe` and `baitaishe` plus false positive `huangtaishe` (Fig 7). The `chihenshe` box has  $IoU = 0.395$  (a miss at 0.5) yet  $s_k = 0.718$ , because centroids are close relative to pooled variance. The false-positive `huangtaishe` is credited  $c_k = 0.667$  against annotated `baitaishe` (coating-color siblings, Table 3). Combined:  $C_{sp} = 0.571$ ,  $C_{sem} = 0.889$ ,  $CCS = 0.730$ —well above a hard  $IoU@0.5$ -plus-exact-label score. The pattern recurs on `A(52).jpg` and `A(16).jpg` under YOLOv10n, confirming the effect is not model-specific.

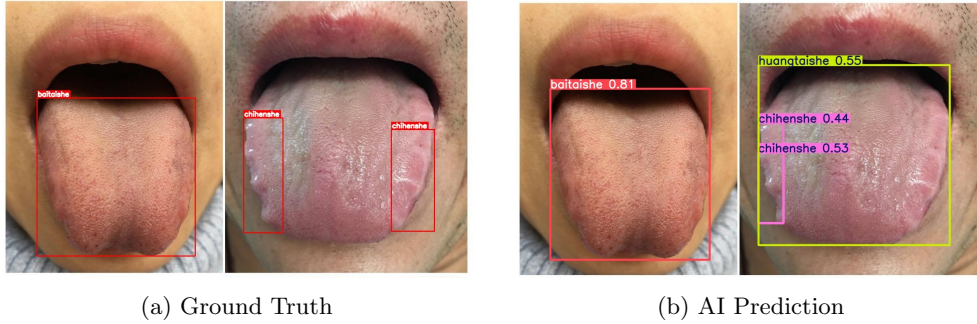


**Fig. 7:** Comparison between the ground truth (GT) and the AI model predictions on image `A(16).jpg` (`A(16).jpg` has 2 component images).

*Semantic near-misses rescued by  $C_{sem}$ .*

On `A(17).jpg`, YOLOv10n predicts `hongshe` and `baitaishe`, sharing no label with ground truth (`liewenshe`, `huangtaishe`) (Fig 8).  $C_{sp} = 0$  due to complete label mismatch. Yet under the taxonomy (Fig 3), each predicted label carries partial credit: `hongshe` and `liewenshe` share the Shape branch, `baitaishe` and `huangtaishe` share Coating-color, so  $c_{sem} = 0.333$  and  $CCS = 0.167$  rather than 0. The same holds

for YOLOv11n/m (0.333), rising to 0.444 for YOLOv8n. A related case, A(14).jpg (YOLOv10n), has one exact match (**huangtaishhe**) and one near-miss (**hongshe** predicted for **hongdianshe**, same branch):  $c_{\text{sem}} = 0.667$ , higher than a naive label-overlap score (0.333), illustrating CCS rewards clinically adjacent misclassifications more than unrelated ones.

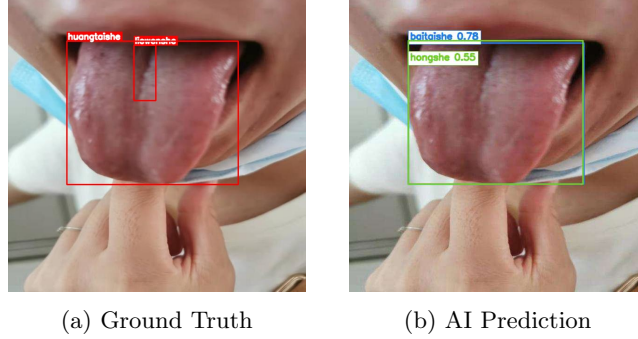


**Fig. 8:** Comparison between the ground truth (GT) and the AI model predictions on image A(17).jpg.

*Class-level scoring limitation (Section 3.4).*

On A(12).jpg, the physician annotated 3 **chihenshe** and 25 **hongdianshe** instances; YOLOv8n predicted one box each for **hongshe**, **chihenshe**, **baitaishhe** (missing **hongdianshe** and **liewenshe**) (Fig 9). Because CCS scores at most one region per class, the predicted **chihenshe** box matches its best GT instance ( $s_k = 0.645$ ) with no penalty for the two undetected instances. This blind spot is most severe for patchy classes (**hongdianshe**, **chihenshe**) whose annotations naturally produce many small regions per image: 84.2% of **hongdianshe** GT regions exceed the single-instance convention and receive no credit or penalty regardless of detector performance.

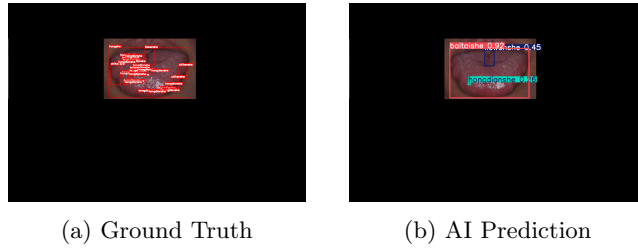


**Fig. 9:** Comparison between the ground truth (GT) and the AI model predictions on image A(12).jpg.

**Table 7:** Default vs. data-driven  $(\alpha, \beta)$  weighting schemes.

| Scheme                                    | $\alpha$ | $\beta$ |
|-------------------------------------------|----------|---------|
| Default (equal weight)                    | 0.500    | 0.500   |
| Standard-deviation-proportional weighting | 0.530    | 0.470   |
| Entropy weighting [11]                    | 0.639    | 0.361   |

#### 4.6 Weight Sensitivity: $\alpha, \beta$ Ablation and Data-Driven Validation

Across all detectors,  $C_{\text{sem}}$  exceeds  $C_{\text{sp}}$  by 0.08–0.13 (Table 4), so CCS decreases monotonically as  $\alpha$  increases (Fig. 10). Sweeping  $\alpha \in [0, 1]$  (step 0.05) and re-ranking, the ranking changes at only two transition points ( $\alpha = 0.5$ : YOLOv10m/YOLO11m swap 3–4;  $\alpha = 0.75$ : YOLOv8m/YOLO11n swap 1–2) and is stable throughout  $\alpha \in [0.5, 0.7]$ . Table 7 compares the default against two data-driven schemes over  $n = 3300$  pooled class-level samples: entropy weighting [11] (more weight to the term with more variation) yields  $\alpha = 0.639, \beta = 0.361$ ; standard-deviation-proportional weighting yields  $\alpha = 0.530, \beta = 0.470$ . Both fall in or adjacent to the stable range  $[0.5, 0.7]$ , and the entropy-weighted ranking exactly reproduces the  $\alpha = 0.5$  ranking—evidence that  $\alpha = \beta = 0.5$  is defensible without tuning.

Concretely, for criterion  $j$  over the  $n = 3300$  samples:

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^n p_{ij} \ln p_{ij}, \quad p_{ij} = \frac{x_{ij}}{\sum_{i=1}^n x_{ij}}, \quad (10)$$

with weights  $w_j = (1 - e_j) / \sum_j (1 - e_j)$ ; standard-deviation weighting sets  $w_j \propto \sigma_j$ .

#### 4.7 Relation to the Normalized Wasserstein Distance

As discussed,  $C_{\text{sp}}$  and NWD [6] both model boxes as 2D Gaussians. NWD requires a calibration constant  $C$  (set to mean  $\sqrt{wh}$  over all GT annotations, yielding  $C = 182.60$  px) inside  $\exp(-\sqrt{\text{NWD}}/C)$ , while  $s_k$  is closed-form and calibration-free. Table 8 shows they nonetheless agree closely: mean  $C_{\text{sp}}$  and mean NWD correlate at Pearson  $r \in [0.980, 0.986]$ , and NWD-matched F1 shows the same threshold-degradation pattern. This validates that  $C_{\text{sp}}$  captures the same spatial-quality signal as an established baseline, while its primary contribution is the calibration-free closed form that integrates cleanly with  $C_{\text{sem}}$  in a single combined score—the combination, not  $C_{\text{sp}}$  alone, is where CCS’s contribution lies.

#### 4.8 Statistical Significance

Spearman  $\rho$  between mean CCS and mean F1@0.5 is 0.94; exact two-sided  $p = 0.0167$  (full enumeration of  $6! = 720$  permutations, as asymptotic  $t$ -approximation is inappropriate at  $n = 6$ ). Correlation with mAP@[.5:.95] is  $\rho = -0.3714$ ,  $p =$

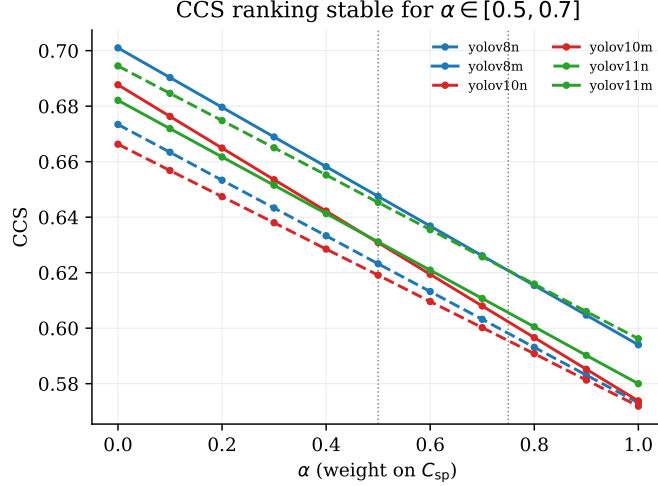


**Fig. 10:** CCS as a function of  $\alpha$  for all six detectors. Dotted vertical lines mark the two ranking-transition points ( $\alpha = 0.5, 0.75$ ); the ranking is otherwise constant across  $\alpha \in [0, 1]$ .

**Table 8:** Correlation between  $C_{sp}$  and the Normalized Wasserstein Distance [6] ( $C = 182.60$ ).

| Model    | Mean $C_{sp}$ | Mean NWD | Pearson $r$ | F1 <sub>NWD</sub> @0.3 | F1 <sub>NWD</sub> @0.7 |
|----------|---------------|----------|-------------|------------------------|------------------------|
| YOLOv8n  | 0.5731        | 0.5554   | 0.9837      | 0.5319                 | 0.4995                 |
| YOLOv8m  | 0.5940        | 0.5782   | 0.9809      | 0.5711                 | 0.5286                 |
| YOLOv10n | 0.5719        | 0.5567   | 0.9825      | 0.5225                 | 0.4925                 |
| YOLOv10m | 0.5738        | 0.5596   | 0.9795      | 0.5410                 | 0.5096                 |
| YOLO11n  | 0.5962        | 0.5798   | 0.9837      | 0.5557                 | 0.5229                 |
| YOLO11m  | 0.5800        | 0.5654   | 0.9863      | 0.5616                 | 0.5274                 |

0.4972, confirming the divergence is not a small-sample artifact, though the CCS-F1 agreement should still be read as suggestive given  $n = 6$ .

For each of the 15 pairwise comparisons we run a paired bootstrap (95% percentile CI) and Wilcoxon signed-rank test. Table 9 summarizes significance before and after Bonferroni correction ( $\alpha_{\text{corr}} = 0.0033$ ). At uncorrected 0.05, 4/15 CCS-bootstrap and 8/15 F1-bootstrap pairs are significant; after correction, no CCS pair remains significant versus 4/15 for F1@0.5. CCS differences are directionally consistent with F1 ( $\rho = 0.94$ ) but at  $n \approx 530$  images are not individually distinguishable from noise after strict correction.

Notably, Wilcoxon flags 5/15 CCS pairs at corrected threshold, including three (YOLOv8n vs. YOLOv8m, YOLOv8n vs. YOLO11n, YOLOv8m vs. YOLOv10n)

**Table 9:** Pairwise significance summary over 15 model comparisons ( $\alpha_{\text{corr}} = 0.05/15 \approx 0.0033$ ).

|                                         | CCS (bootstrap) | F1@0.5 (bootstrap) |
|-----------------------------------------|-----------------|--------------------|
| Significant, uncorrected ( $p < 0.05$ ) | 4/15            | 8/15               |
| Significant, Bonferroni-corrected       | 0/15            | 4/15               |

that bootstrap does not. Inspecting the per-image CCS difference distributions: skewness near zero (0.001, 0.059,  $-0.220$ ) rules out asymmetric outliers; only 3.5–4.7% of images have exactly zero difference, ruling out a large tied-at-zero mass. Instead, near-zero skewness with mild excess kurtosis (1.00–2.38) indicates a symmetric but moderately heavy-tailed distribution. Under this condition, the Wilcoxon signed-rank test’s asymptotic relative efficiency relative to a mean-based test never falls below 0.864 and grows without bound as tails grow heavier than Gaussian [19], so a rank-based test can detect a location shift that a percentile bootstrap CI (which absorbs the heavier-tail variance) fails to distinguish. We treat the bootstrap result as more conservative and recommend larger evaluation sets before treating CCS differences between closely-ranked detectors as decisive.

## 4.9 Class Imbalance

The eight-class subset is severely imbalanced in per-class difficulty (Fig. 11): averaged over all models, **baitaishe** and **huangtaishe** reach mean  $\text{AP}@[.5:.95]$  of 0.802 and 0.629, while **botaishe**, **hongdianshe**, **liewenshe**, **chihenshe** remain below 0.13 (lowest **chihenshe** at 0.061). This  $> 13\times$  gap means a macro-averaged mAP is dominated by well-localized classes. To test whether low AP reflects near-miss confusions that  $C_{\text{sem}}$  credits, we quantified for every false positive/negative among the four hardest classes whether its best semantic match (Table 3) fell on the same branch (near-miss,  $w_{ij} > 0$ ) or an unrelated branch (far-miss,  $w_{ij} = 0$ ), pooled over all detectors (Table 10). The picture is uneven: **botaishe** errors are overwhelmingly near-miss (94.5%), as its Coating-nature siblings (**huangtaishe**/**baitaishe**) absorb most confusions; **hongdianshe** and **chihenshe** also show substantial near-miss shares (41.5% and 36.3%). **liewenshe** is predominantly far-miss (76.7%): despite taxonomy siblings (**pangdashe**/**chihenshe**), its errors do not land there, indicating genuine detection failure rather than rescuable near-miss. The near-miss share is consistent across detectors (35.8–42.0%), suggesting a dataset property rather than a model artifact. Pooled: 39.5% of hard-class errors (589/1492) are near-miss confusions—non-trivial but concentrated unevenly across classes.

## 5 Discussion

Across six YOLO detectors, hard-threshold F1 is both compressed and unstable (drops up to 16.9% as IoU threshold tightens; ranking flips for 5/15 model pairs). CCS, computed without any matching threshold, avoids this instability while remaining consistent with F1@0.5 ( $\rho = 0.94$ ,  $p = 0.0167$ ). The  $C_{\text{sem}} - C_{\text{sp}}$  gap and the near-miss

**Table 10:** Share of detection errors (false positives/negatives) among the four hardest classes that are near-miss confusions (best semantic match on the same taxonomy branch) vs. far-miss (unrelated branch), pooled across all six detectors.

| Class              | Total errors | Near-miss | % Near-miss |
|--------------------|--------------|-----------|-------------|
| <b>botaishe</b>    | 145          | 137       | 94.5        |
| <b>hongdianshe</b> | 443          | 184       | 41.5        |
| <b>chihenshe</b>   | 441          | 160       | 36.3        |
| <b>liewenshe</b>   | 463          | 108       | 23.3        |
| Pooled (all four)  | 1492         | 589       | 39.5        |

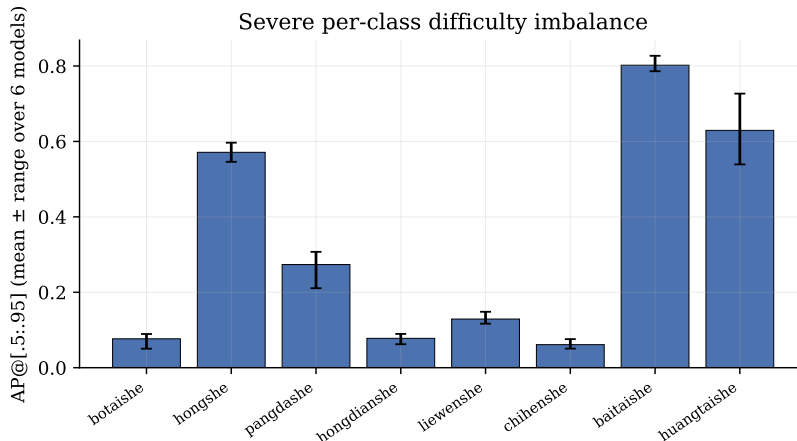


**Fig. 11:** Mean AP@[.5:.95] per class, averaged over six detectors (error bars show the min–max range across models). **botaishe**, **hongdianshe**, **liewenshe**, and **chihenshe** remain hard for every detector, independent of architecture or size.

analysis (39.5% of hard-class errors) indicate that a meaningful share of what mAP/F1 record as pure error reflects semantically adjacent confusions rather than clinically arbitrary mistakes—a distinction hard-threshold metrics cannot express. Practically, CCS is best used alongside existing metrics: mAP and F1 remain informative for pixel-precise localization, while CCS surfaces whether errors are clinically benign near-misses or genuine failures—critical when detector output feeds a diagnostic workflow rather than a leaderboard.

## 5.1 Limitations

**Class-level scoring.** Both terms score at most one pair per class per image (Section 3.4); Table 2 shows this is concentrated in patchy classes (**hongdianshe**:

84.2% of GT regions unscored). A further consequence is imperfect granularity match: our F1/mAP figures use greedy instance-level assignment (Section 4.1), so they penalize missed instances CCS does not. The threshold-instability and semantic-proximity findings are unaffected (they hold at class level for all metrics), but an instance-level extension via Hungarian matching before Eq. (5) is needed for dense multi-instance datasets.

**Provisional taxonomy.** The three-branch taxonomy (Fig. 3) was constructed for computability rather than derived from a formal TCM ontology. The full similarity matrix (Table 3) is exposed so alternatives can be substituted.

**Statistical power.** With  $n = 6$  detectors, after Bonferroni correction no CCS pairwise difference remains significant ( $n \approx 522$ –534 images). The bootstrap/Wilcoxon discrepancy is best resolved by a larger evaluation set.

**Single institution, architecture, and operating point.** All detectors are YOLO variants on one dataset; generalizability to other architectures is untested. CCS is reported at a single confidence threshold (0.25); sensitivity to this choice is unevaluated.

**No clinical validation.** Whether  $C_{\text{sem}}$ ’s partial-credit values match a practitioner’s perceived clinical severity has not been tested and would require a dedicated expert-rating study.

These limitations point to an extension agenda: instance-level matching, expert taxonomy validation, evaluation across architectures and medical imaging domains (Section 2.4), confidence-threshold sensitivity analysis, and larger- $n$  or prospective clinical validation.

## 6 Conclusion

We introduced CCS, a continuous spatial–semantic concordance score replacing hard IoU-matching with closed-form Gaussian cosine similarity and binary label correctness with taxonomy-driven Wu–Palmer partial credit. Across six YOLO detectors on eight-class TCM tongue-symptom data, CCS remained stable (0.62–0.65) while F1 degraded up to 16.9% and reversed rankings for a third of detector pairs as IoU threshold changed; CCS ranking agreed with F1@0.5 ( $\rho = 0.94$ ,  $p = 0.0167$ ), and entropy weighting independently reproduced the equal-weight default. Quantified evidence—39.5% of hard-class errors pooled, up to 94.5% for individual classes—shows a non-trivial share of what hard metrics record as pure error is semantically near-miss, though this share is uneven across classes.

We present these results as evidence CCS is a viable evaluation approach for detection tasks with class imbalance and semantically structured labels, not a finished replacement for mAP: it is validated on a single domain with class-level scoring, a provisional taxonomy, and untested operating-point dependence (Section 5). We hope this testbed motivates extension of the same spatial–semantic combination to instance-level matching and other semantically structured medical imaging domains.

**Acknowledgements.** The author thanks Thanh Thao Tran and Ngoc Ai Duyen Phan for their inspiration and discussions, and the Faculty of IT, HCMC University of Technology and Engineering for GPU computing resources.

## Declarations

### *Funding*

The author received no financial support for the research, authorship, and/or publication of this article.

### *Conflict of interest*

The author declares that there are no competing interests.

### *Ethics approval and consent to participate*

Not applicable.

### *Consent for publication*

Not applicable.

### *Data availability*

The data used in this study are publicly available and were obtained from the dataset published in [2].

### *Materials availability*

Not applicable.

### *Code availability*

The source code is publicly available at <https://github.com/tmai110605/TongueTriage.git>.

### *Author contribution*

The sole author conceived the study, designed the methodology, implemented the algorithms, conducted the experiments, analyzed the results, wrote the manuscript, and approved the final version of the manuscript.

## Appendix A Per-Detector Near-Miss Breakdown

Table A1 reports the near-miss share of hard-class errors (Section 4.9) per detector. The share is stable across architectures and model sizes (35.8–42.0%), supporting the pooled analysis in Table 10 as representative.

## References

- [1] Kim, M., Cobbin, D., Zaslowski, C.: Traditional Chinese medicine tongue inspection: An examination of the inter- and intrapractitioner reliability for specific tongue characteristics. *Journal of Alternative and Complementary Medicine* 14(5), 527–536 (2008)

**Table A1:** Share of hard-class detection errors that are near-miss confusions, per detector.

| Model    | Total errors | Near-miss | Far-miss | % Near-miss |
|----------|--------------|-----------|----------|-------------|
| YOLOv8n  | 268          | 103       | 165      | 38.4        |
| YOLOv8m  | 224          | 94        | 130      | 42.0        |
| YOLOv10n | 274          | 98        | 176      | 35.8        |
| YOLOv10m | 232          | 97        | 135      | 41.8        |
| YOLOv11n | 233          | 95        | 138      | 40.8        |
| YOLOv11m | 261          | 102       | 159      | 39.1        |

- [2] Jin, X., Gao, L., Tong, A., Chen, Z., Kong, J., Sun, N., Ma, H., Wang, Q., Bai, Y., Su, T.: TCM-Tongue: A standardized tongue image dataset with pathological annotations for AI-assisted TCM diagnosis. arXiv preprint arXiv:2507.18288 (2025)
- [3] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 740–755 (2014)
- [4] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666 (2019)
- [5] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12993–13000 (2020)
- [6] Wang, J., Xu, C., Yang, W., Yu, L.: A normalized gaussian wasserstein distance for tiny object detection. arXiv preprint arXiv:2110.13389 (2021)
- [7] Bertinetto, L., Mueller, R., Tertikas, K., Samangoeei, S., Lord, N.A.: Making better mistakes: Leveraging class hierarchies with deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12506–12515 (2020)
- [8] Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 133–138 (1994)
- [9] Benyahia, S., Meftah, B., Lézoray, O.: Hierarchical approach for the classification of multi-class skin lesions based on deep convolutional neural networks. In: Pattern Recognition and Artificial Intelligence, Cham, pp. 139–149 (2022)
- [10] Rubin, D.L.: Creating and curating a terminology for radiology: Ontology

- modeling and analysis. *Journal of Digital Imaging* **21**(4), 355–362 (2008)
- [11] Hwang, C.-L., Yoon, K.: Multiple Attribute Decision Making: Methods and Applications. *Lecture Notes in Economics and Mathematical Systems*, vol. 186. Springer, Berlin, Heidelberg (1981)
  - [12] Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. <https://github.com/ultralytics/ultralytics>
  - [13] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: YOLOv10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458* (2024)
  - [14] Jocher, G., Qiu, J.: Ultralytics YOLO11. <https://github.com/ultralytics/ultralytics>
  - [15] Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9404–9413 (2019)
  - [16] Stillwagon, K., VandeLoo, A.D., Forest, C.R.: Maximum matching accuracy: An instance segmentation evaluation metric utilizing globally optimal matching. *arXiv preprint arXiv:2606.10107* (2026)
  - [17] Chang, W.-H., Chen, C.-C., Wu, H.-K., Hsu, P.-C., Lo, L.-C., Chu, H.-T., Chang, H.-H.: Tongue feature dataset construction and real-time detection. *PLoS ONE* **19**(3), 0296070 (2024)
  - [18] Jiang, T., Lu, Z., Hu, X., Zeng, L.-z., Ma, X., Huang, J., Cui, J., Tu, L., Zhou, C., Yao, X., Xu, J.-t.: Deep learning multi-label tongue image analysis and its application in a population undergoing routine medical checkup. *Evidence-Based Complementary and Alternative Medicine* **2022**, 3384209 (2022)
  - [19] Hodges, J.L., Lehmann, E.L.: The efficiency of some nonparametric competitors of the t-test. *The Annals of Mathematical Statistics* **27**(2), 324–335 (1956)