Ground Before Correcting: Evidence-Guided Visual Chinese Grammatical Error Correction at Multi-LLL 2026

Xiaochuan Xie **Keran Huang**
Chongqing University of Posts and Telecommunications
Chongqing, China

Abstract

We describe the GroundCorrect system of team xc-皮波 for the **Sentence-Level Track of VisCGEC-v2.0** at Multi-LLL 2026. The system combines frozen Qwen3-VL-4B-Instruct global recognition, rank-8 LoRA verification of full-line and local-crop evidence, and YOLO11s character detection with deterministic edit grounding. According to the retained team record, the official V3 competition submission ranked **second** with score **0.5114**. This result belongs exclusively to the official submission, not to the local reconstruction or the post-submission V3+ extension. Our primary local adapter comparison holds inputs, prompt, decoding, and grounding fixed: the development proxy rises from 0.1254 to 0.2234. Other configuration changes are reported as system development. V3+ is an exploratory case study that changes one development example, with no internal-selection or grounded-edit improvement. Local results are conditional configuration analyses: selection images also train the detector, and development data overlap training at the essay-group level.

1 Introduction

Team xc-皮波 participated in the **Sentence-Level Track of VisCGEC-v2.0** at the Multi-LLL 2026 Shared Task with GroundCorrect. VisCGEC takes learner handwriting and requires corrected Chinese text with spatially grounded edits (Wang et al., 2025). Unlike text-only Chinese grammatical error correction (CGEC), it couples recognition, correction, and localization: even a correct sentence can have incorrect edit coordinates.

Three difficulties motivate our design. Global resizing can obscure similar characters in long, narrow images. A vision–language model (VLM) can correct text yet generate unreliable coordinates. Further correction can introduce unsupported rewrites or inconsistent edit lists.

Our pipeline (Figure 1) separates *global understanding*, *local visual verification*, and *coordinate assignment*. A frozen pass proposes source and target sentences; a format-matched LoRA inspects retrieved regions with the full line. A separate detector grounds text differences, and guarded finalization handles output risks. An optional post-submission rank check is analyzed separately.

Competition submission and later analyses. The retained team record identifies V3 as the August 29 official submission: second place, score 0.5114. The September 11 local reconstruction is not assumed byte-identical: original weights and the complete August run directory were not preserved (Appendix D). V3+ is the September 12 exploratory extension without an official hidden-test result. The rank and score belong only to the official submission.

Our contributions are threefold:

- An evidence-guided correction pipeline that combines a frozen global pass with format-matched local verification.
- A deterministic grounding mechanism that separates linguistic edits from coordinate generation through character detection and monotonic alignment.
- A controlled local adapter comparison with fixed inputs, prompt, and grounding, accompanied by explicit data-use and evaluation disclosures. System-development records and the exploratory V3+ case study are reported separately.

2 Related Work

Chinese GEC. NLPCC 2018 established a common Chinese GEC setting (Zhao et al., 2018); FCGEC supplies fine-grained operations (Xu et al.,

Proceedings of the 1st Workshop on Multi-modality and Multi-lingual Language Learning, pages 33–46
November 9, 2026 ©2026 Association for Computational Linguistics

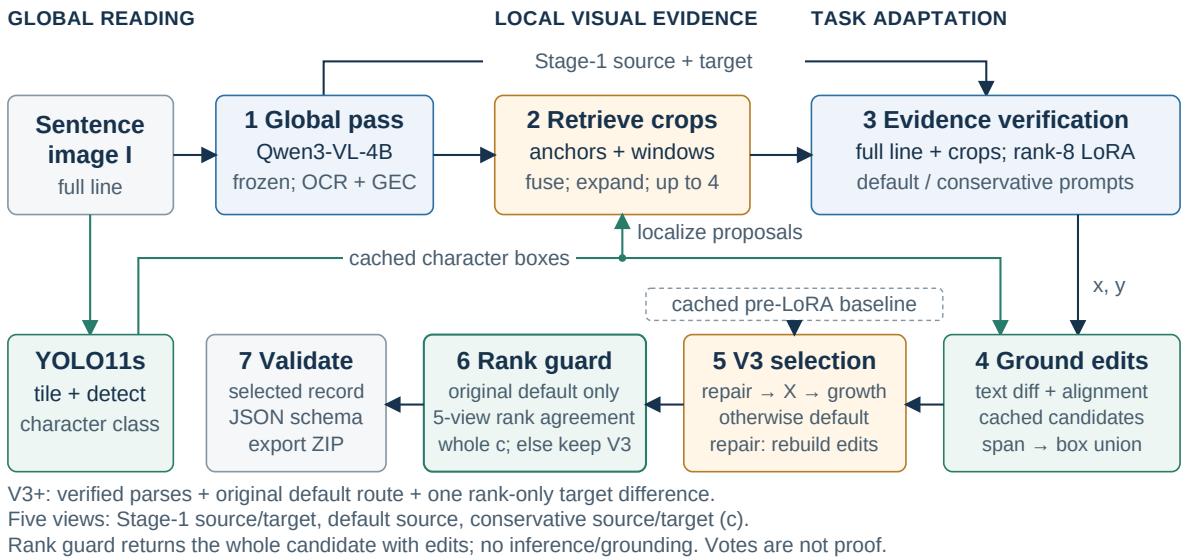


Figure 1: GroundCorrect: frozen global recognition, shared Stage-2 LoRA verification, YOLO11s grounding, and V3 candidate finalization. The rank-check block is optional post-submission V3+ work, not part of the official V3 submission. Candidates undergo schema validation.

2022), and MuCGEC emphasizes multi-reference, character-level evaluation (Zhang et al., 2022). These resources contextualize VisCGEC's image-grounded setting; we do not add their data to training.

Local visual evidence. TrOCR combines pre-trained image and text Transformers for recognition (Li et al., 2023). LLaVA-UHD handles high-resolution, variable-aspect-ratio images through slices (Guo et al., 2025). Our crops instead use task hypotheses and coverage windows outside the VLM. We adapt Qwen3-VL (Bai et al., 2025) with LoRA (Hu et al., 2022) and use a separate YOLO11 character detector (Jocher and Qiu, 2024) for grounding.

Conservative correction and grounding. Align-and-Predict decoding adjusts the precision–recall trade-off in GEC (Sun and Wang, 2022), and Alirector uses alignment to address overcorrection in Chinese GEC (Yang and Quan, 2024). The exploratory V3+ case study selects complete candidates after generation. Slicing-aided detection (Akyon et al., 2022) offers a related motivation for preserving small visual details, although our sentence tiling is independently implemented. TrOCR, LLaVA-UHD, APD, Alirector, and slicing-aided detection are related work, not used components or evaluated baselines.

3 Method

3.1 Task and system overview

Given a sentence image $I \in \mathbb{R}^{H\times W\times 3}$, the system returns corrected text $\hat{y}$ and grounded edits $\hat{E} = \{(o_k, t_k, b_k)\}_{k=1}^{K}$, where o_k, t_k, and b_k are an operation, replacement text, and pixel-space box. Internally, a candidate is $q = (x, y, E)$: retaining the recognized source x makes its connection to the edit list explicit. The output schema allows Insert, Delete, Substitute, and WordOrder; the released sentence annotations contain only the first three types.

Figure 1 summarizes three functional modules. Evidence-guided correction produces a source–target pair from the full line and local crops. Deterministic grounding maps its text differences to character boxes. V3 finalization selects or repairs candidates before schema validation; the optional V3+ rank check is a later extension. Only the Stage-2 adapter and the character detector are trained; the base VLM and its visual encoder remain frozen.

3.2 Evidence-guided correction

Global hypothesis generation. The first pass uses frozen Qwen3-VL-4B-Instruct. A Chinese prompt asks the model to recognize the handwriting and make only high-confidence grammatical changes while preserving meaning. Its JSON output contains the recognized source, provisional target, and provisional edits. Writing G_θ for greedy

generation and p_0 for the global prompt, we retain the text fields

$$(x^{(0)}, y^{(0)}) = \pi_{x,y} G_\theta(I, p_0). \tag{1}$$

The projection $\pi_{x,y}$ parses the JSON fields, rather than taking the VLM's provisional boxes as final edits. This full-line view supplies global syntax and identifies regions for closer inspection.

Contextual region retrieval. We form candidate anchors from two test-time signals: localized differences between the provisional source and target (weight 1.0), and coordinates directly proposed by the VLM (weight 0.55). Proposals merge when their IoU is at least 0.15 or horizontal centers differ by at most 2.5% of image width. Deterministic overlapping coverage windows supplement a short candidate list. We retain at most four regions and expand each to the full image height and at least 22% of its width, preserving neighboring text (Figure 2, left). Reference targets and gold edit boxes are not used for test-time retrieval.

Format-matched verification. Stage 2 receives the full image, up to four crops, a manifest of their global coordinates, and the Stage-1 source and target. The prompt treats retrieved regions as evidence, not guaranteed errors: local views resolve ambiguous characters, while the full line supplies grammatical context.

Let $z = (I, C, x^{(0)}, y^{(0)}, p_2)$ denote this input, including crop set C and the default or conservative prompt p_2. For each adapted linear map, LoRA (Hu et al., 2022) parameterizes

$$W' = W + \frac{\alpha}{r} BA, \qquad r = 8, \quad \alpha = 16, \tag{2}$$

with $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$. The base W is frozen. This adds $r(d_{\text{in}} + d_{\text{out}})$ parameters instead of updating $d_{\text{in}} d_{\text{out}}$ entries per map. We adapt attention and feed-forward projections, with dropout 0.05; the visual encoder remains frozen.

Training uses the default prompt and the inference input format. The assistant answer is text-only JSON containing `source_text` and `target_text`. Let $a_{1:T}$ be the unmasked assistant suffix, including the chat-template terminator. With trainable adapter parameters ϕ, the token-average objective is

$$\mathcal{L}(\phi) = -\frac{1}{T} \sum_{t=1}^{T} \log p_{\theta,\phi}(a_t \mid z, a_{<t}). \tag{3}$$

This follows the autoregressive factorization of the supervised answer. Prompt labels are masked to -100, so image and prompt tokens condition generation but receive no direct prediction loss. Reference text occurs only in the assistant answer; there is no coordinate loss or joint backpropagation through retrieval and grounding. Stage 2 is accepted only if parsing succeeds without an error flag and both strings are nonempty. Otherwise, the system uses the cached Stage-1 hypothesis and rebuilds its edits. A conservative prompt provides an alternative candidate using the same adapter, not a second trained model.

3.3 Deterministic edit grounding

Linguistic correctness and spatial correctness need not coincide. We therefore discard the VLM's provisional edit semantics and coordinates when constructing grounded candidates. Instead, character-level sequence alignment between source x and target $\hat{y}$ determines INSERT, DELETE, and SUBSTITUTE spans.

A single-class YOLO11s detector supplies m boxes sorted by horizontal center. For $n, m > 0$, let source-character width w_i be 0.45 for whitespace, 0.60 for ASCII alphanumerics, and 1 otherwise. Expected character centers and normalized box centers are

$$u_i = \frac{\sum_{k<i} w_k + w_i/2}{\sum_{k=1}^{n} w_k}, \qquad v_j = \frac{c_j - c_{\min}}{c_{\max} - c_{\min}}, \tag{4}$$

where c_j is a box's horizontal center and $v_j = 0.5$ if all centers coincide. This compares relative positions without requiring a detector class for every character identity.

Let $D_{i,j}$ be the minimum cost of consuming i characters and j boxes. A monotonic alignment ends in exactly one of three transitions: match both, skip a character, or skip a box. Thus,

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} + 3(u_i - v_j)^2, \\ D_{i-1,j} + 0.10, \\ D_{i,j-1} + 0.10. \end{cases} \tag{5}$$

Initialize $D_{0,0} = 0$, $D_{i,0} = 0.10i$, and $D_{0,j} = 0.10j$. DP uses $O(nm)$ time and space; backtracking takes $O(n + m)$. Without detections, characters map to null. Optimality concerns the geometric cost, not recognition (Appendix A.5).

A non-insertion edit receives the bounding union of available boxes in its source span. An insertion uses its source index, capped at the final

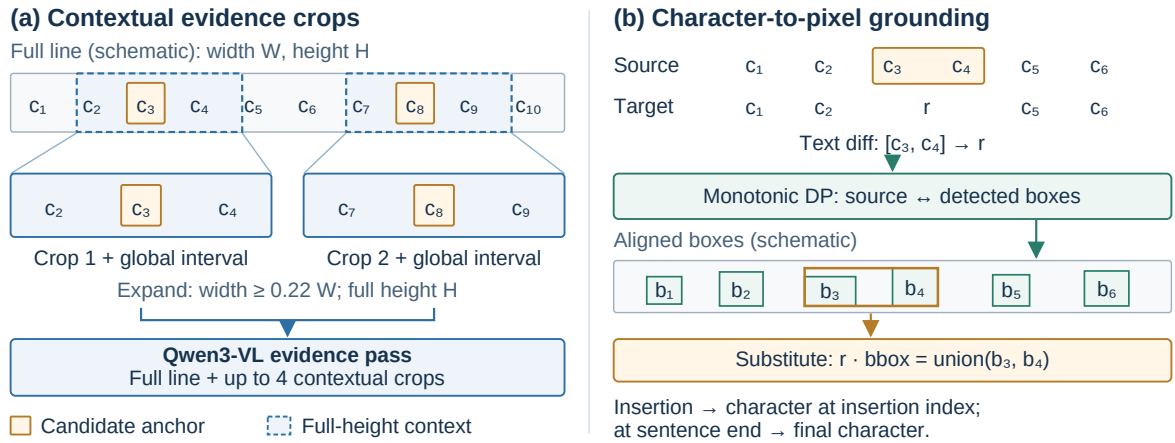


Figure 2: Evidence retrieval and edit grounding (schematic). Left: narrow anchors expand into full-height contextual crops. Right: a text difference identifies a source span, and monotonic character–box alignment maps it to a box union. Symbols illustrate the mechanism, not benchmark examples.

character. If no required aligned box exists, the implementation uses a single uniform-width source cell over the full image height. Partially matched spans use their available boxes without filling every unmatched character. Text differences use deterministic sequence matching, not an optimization of the downstream correction score. Figure 2 illustrates this separation of text differencing from coordinate assignment.

3.4 Priority-based V3 finalization

The finalizer applies rules in a fixed, first-match order. It first checks two narrow consistency repairs: restoring a Stage-1 comma at a preserved *if*-clause boundary, and completing an asymmetric comparison from the activity stated on its other side. A successful repair returns immediately with edits rebuilt from the default source and cached detector boxes.

Otherwise, it checks for an uppercase `X` newly introduced relative to the Stage-1 target, followed by target–source growth of at least three characters. A flagged output selects the conservative candidate if that candidate removes the specific trigger; otherwise it selects the retained baseline. Unflagged outputs retain the default candidate. These selection branches preserve their existing grounded edit lists. Length growth is a risk signal, not proof of hallucination.

We call the default matched-LoRA configuration V1, the risk guard V2, and the finalizer with both repairs V3. The optional V3+ post-check leaves earlier repair and fallback routes unchanged.

3.5 Exploratory post-submission rank check

V3+ was developed after submission to examine a known same-length rank substitution that the length and placeholder checks missed. It compares rank expressions across existing predictions and, under narrow parsing, agreement, and single-change conditions, selects the complete conservative candidate only on V3's unchanged default route. This keeps each selected sentence together with its own source and grounded edits. The observed evidence is one changed development example, with unchanged internal-selection and grounded-edit scores; it supports a targeted case study rather than a general factual-preservation claim. Section 4.5 reports its scope, and Appendix B contains the rule-level definitions and equations.

4 Experiments

4.1 Experimental setup

Data and preparation. We use the released Vis-CGEC sentence-level train, dev, and test splits (Wang et al., 2025), summarized in Table 1. Train/dev provide source transcriptions, target text, edit types, and boxes; the shared-task test input provides identifiers and images, not labels. Official train/dev character annotations are converted to a single detection class. Tiling at a 12:1 aspect ratio with 20% horizontal overlap produces 4,431 training tiles with 63,545 box instances and 573 dev tiles with 8,269 instances. No public test targets or hidden labels are used for training or scoring.

(a) Released annotations

Split	Images	Edits	Ins.	Del.	Sub.
Train	1,960	3,533	1,079	761	1,693
Dev	245	435	134	89	212
Test	246	–	–	–	–

(b) Component and selection uses

Subset	Images	LoRA	Detector	Other use
LoRA fit	1,746	Fit	Fit	Training
Selection	214	–	Fit	Select/report
Dev	245	–	Validate	Select/report
Test	246	–	Infer	Official eval.

Table 1: Released data and component-specific uses. The two train subsets sum to 1,960; both fit the detector. Selection excludes LoRA fitting only; dev validates the detector. Selection/dev also guide configurations and local reporting. Test labels are unavailable. Neither local split is an independent end-to-end generalization test.

Training and inference. The local reconstruction uses one NVIDIA GeForce RTX 5090 (32GB). LoRA trains 16,515,072 parameters (0.371%) for one epoch on 1,746 examples from 499 essay groups, reserving 214 examples from 55 groups. It uses learning rate 5×10^{-5}, batch 1, accumulation 8, cosine decay, bfloat16, and seed 20260829. Implementation uses PyTorch (Paszke et al., 2019), Transformers (Wolf et al., 2020), PEFT, and Ultralytics.

YOLO11s is reconstructed from public weights for 30 epochs using 1,280-pixel tiles, batch 16, and seed 3407. Tiled FP16 inference uses 160×1280 inputs and global NMS IoU 0.50. VLM decoding is greedy, with 12,544–200,704 pixels per view and token caps 384/256 for Stages 1/2. Appendix A records complete settings, software versions, and cache dependencies.

Evaluation measures. Because the private evaluator is unavailable, we implement an equal-weight local proxy for text correction and grounded edits:

$$S_{\text{proxy}} = \frac{1}{N} \sum_{i=1}^{N} \frac{F^{\text{text}}_{0.5,i} + F^{\text{edit}}_{0.5,i}}{2}, \qquad (6)$$

where $F_{0.5} = 1.25PR/(0.25P + R)$. Both component scores are computed per sentence and then macro-averaged; the composite averages their per-sentence mean, rather than pooling edit counts over the corpus.

Text scoring. Reference and predicted edits are extracted from the same released *gold source transcription*, `source_ground_truth`, to the reference and predicted targets respectively. Character-level `SequenceMatcher` with `autojunk=False` yields insertion, deletion, and substitution spans. Matches require identical type, half-open source span, and replacement string; multiset intersection limits repeated matches to reference multiplicity. We use this local differencer, not ChERRANT (Zhang et al., 2022) or ERRANT (Bryant et al., 2017).

Grounded matching. Predictions are visited in stored order and greedily matched to the highest-IoU unused reference with identical type and replacement text, requiring IoU ≥ 0.5. Equal-IoU ties select the first eligible reference in annotation order. Each side can match once; this is a greedy, order-dependent assignment. Duplicate predictions remain in the precision denominator, including any copy without an eligible unused reference.

Empty sets and strings. For either component, two empty edit sets receive precision, recall, and $F_{0.5} = 1$; if only one is empty, all three are zero. Punctuation and whitespace participate literally in alignment and replacement matching. Text fields are not trimmed or normalized for Unicode, case, traditional/simplified characters, or punctuation. Other edge conventions are in Appendix A.4. Private evaluator conventions are unknown, so these remain local proxies.

Text evaluation uses gold source, whereas inference grounding uses each candidate's *predicted source*. Recognition can therefore hide a necessary source edit despite a correct target (Section 4.7).

Split and operating-point scope. Table 1 separates LoRA fitting, detector fitting/validation, and selection/reporting uses. Essay groups are identifier prefixes before the first underscore. Only 12 of the 245 dev sentences belong to essay groups absent from the full released training split; the other 233 belong to groups present there. The 214-example internal selection set excludes its 55 groups from LoRA training, but all its images occur in detector training, and it also guided model and rule selection. Development images serve detector validation and local configuration analysis. Neither local split is an untouched, fully group-disjoint end-to-end evaluation.

All matched-LoRA crop manifests use confidence-0.20 detections. Final grounding

Configuration	Text	Ground	Composite
Dev: 245 examples, grounding 0.50			
Frozen Stage 2	.2158	.1125	.1641
Matched LoRA (V1)	.2767	.1699	.2233
Guarded V2 / V3	.2787	.1699	.2243
V3+ (post-submission)	.2805	.1699	.2252
Selection holdout: 214, grounding 0.20			
Stage 1	.2135	.0794	.1464
Matched LoRA (V1)	.2654	.1416	.2035
Guarded V2 / V3	.2723	.1437	.2080
V3+ (post-submission)	.2723	.1437	.2080

Table 2: Local system-development record, not isolated component ablations or official scores. Dev's frozen-to-LoRA step changes both adapter and output prompt; dev/selection grounding confidences are 0.50/0.20 and guard fallbacks are frozen Stage 2/Stage 1. The selection set isolates LoRA training only. V3+ is exploratory post-submission work.

in the cached progression uses 0.50 on dev but 0.20 on holdout. The holdout guard falls back to Stage 1, whereas dev/test use the retained frozen Stage-2 baseline. The fresh adapter control instead fixes final grounding at 0.50 on both splits and applies no guard. We keep these protocols separate throughout.

4.2 Official competition outcome

According to the retained team record, the August 29 V3 submission ranked **second** with official score **0.5114**. Official component scores were unavailable. Hidden test labels and a private evaluator separate this outcome from our labeled dev/selection analyses. The reconstruction is not assumed byte-identical, and V3+ has no official score (Appendix D).

Our primary evidence for the adapter is the fixed-configuration comparison in Table 3, discussed first below. Table 2 is retained as a system-development record. Its rows can change prompts, grounding operating points, or fallback behavior and must not be read as isolated component ablations.

4.3 Primary comparison: controlled adapter effect

Table 3 is our primary evidence for the adapter's local effect. To separate it from the output-prompt change, we freshly generate both arms with one base model and the same preprocessed tensors per example, toggling only the LoRA adapter. Both use the default text-only prompt, identical cached Stage-1 outputs and crops, greedy decoding

Setting	Text	Ground	Composite
Dev adapter OFF	.2047	.0460	.1254
Dev adapter ON	.2770	.1699	.2234
Holdout adapter OFF	.2034	.0473	.1253
Holdout adapter ON	.2661	.1585	.2123

Table 3: Fresh paired adapter control. Both arms use the same text-only prompt and final grounding confidence 0.50, with no guard. The dev set contains 245 examples; the 214-example holdout isolates only LoRA training and previously served for selection. Its confidence differs from the original 0.20 holdout rows in Table 2.

capped at 256 tokens, and confidence-0.50 grounding, without V2/V3 post-processing. We disable TF32 and record the runtime, input, and weight hashes. This post-submission control does not reuse either arm's old predictions.

On dev, the adapter raises the composite by 0.0981 (Table 3); 59 examples improve, 16 worsen, and 170 tie. Both arms parse all 245 responses without fallback. A 10,000-replicate paired bootstrap over 188 essay groups gives a 95% percentile gain interval of [0.0621, 0.1337]. This estimates the adapter's effect under the text-only contract, not a gain over the best frozen prompt. On the 214-example holdout, the corresponding gain is 0.0870, with 51 improvements, 20 regressions, and 143 ties. All responses parse without fallback. The same bootstrap over its 55 groups gives [0.0456, 0.1327]. These results retain the split limitations described above. This comparison holds inputs, prompt, decoding, and grounding fixed; Table 2 instead records system development with configuration changes. The fresh dev ON output differs from the old cache on four targets and scores 0.2234 rather than 0.2233; we therefore keep the runs separate. Both intervals are conditional on a single trained adapter and fixed predictions, not training-seed uncertainty or an official-score estimate.

4.4 System-development record

Table 2 documents cached local configurations and their combined changes. On dev, the frozen Stage-2 baseline scores 0.1641, matched LoRA 0.2233, and V3 0.2243; on the internal selection set, matched LoRA scores 0.2035 and V3 0.2080. V3's two extra repair rules never trigger on either labeled split, so V2 and V3 tie there. These observations do not establish a measured benefit from the additional repairs.

The dev frozen-to-LoRA configuration differ-

ence is 0.0592 absolute (36.0% relative). A paired bootstrap over 188 essay groups, with 10,000 replicates, gives a 95% percentile difference interval of [0.0243, 0.0940], positive in 99.96% of replicates. This comparison changes both the adapter and the output prompt, from full JSON to text-only JSON. Its interval describes the combined configuration change, not an isolated adapter effect or training-seed uncertainty.

4.5 Exploratory post-submission V3+ case study

The September 12 extension examines a known development error after competition submission. The rank-only and exploratory rank-plus-numbers variants route all 245 dev and 214 internal-selection records before reference labels are loaded for scoring; only predictions enter the selector, not IDs or correctness scores. This prediction-only execution does not remove the experiment's known-error-driven selection bias.

V3+ changes only `1587_15`, improving one dev sample with no regressions and 244 ties. All 214 internal-selection records remain identical. Its dev composite gain is 0.000907 absolute; grounded-edit scores are unchanged. The numeric extension produces exactly the same outputs, so it provides no measured benefit over the rank-only variant.

These observations support a targeted case study, not a general factual-preservation or end-to-end generalization claim. The unchanged grounded-edit score is important: correcting the rank in this case does not recover the missing necessary edit. Section 4.7 examines the case; rule definitions and replay checks are in Appendix B. V3+ has no official competition score.

4.6 Component analysis

Evidence reinspection. Figure 5(a) compares frozen Stage 1 with contextual crops plus a second VLM pass, holding predicted YOLO11n boxes fixed in a test-matched reconstruction. Text $F_{0.5}$ rises from 0.2076 to 0.2158, grounded-edit $F_{0.5}$ from 0.0944 to 0.0982, and the composite from 0.1510 to 0.1570 (0.0060 absolute; 4.0% relative). This isolates the combined reinspection step, not crops alone or the submitted detector choice.

Corpus OCR character error rate decreases from 0.0531 to 0.0524. Expanded crops cover the centers of 87.59% of gold edits, although only 13.79% of raw anchors reach IoU 0.5. This distinction supports using recall-oriented crops for evidence and a separate detector for precise final grounding.

Risk guard and additional probes. Figure 5(b) and Table 2 show that the V2 guard adds 0.0045 on holdout and 0.0010 on dev. V3's two extra repairs affect two test inputs, but no labeled examples; hidden test labels are unavailable, so we make no score-based claim for those changes. Appendix C separately records early auxiliary-context experiments, a discarded direct-image LoRA, threshold sweeps, and post-hoc detector-capacity probes. They are not additional controlled ablations of the final system.

4.7 Qualitative analysis

Figure 3 illustrates cached predictions at grounding confidence 0.50. The first two cases are unchanged from V3 to V3+; the third shows the patch's only observed intervention. They are selected examples, not an unbiased accuracy estimate or isolated adapter effect.

Text recovery versus localization. In `543_6`, V3 recovers the correct phrase about resting at home, but also normalizes the erroneous character in its predicted source. The source–target difference therefore emits no edit: the text is correct, but the gold correction is not localized. In `1969_7`, V3 instead inserts the missing character in "reason" and matches both the reference text and insertion location.

Exploratory rank correction without complete grounding. In `1587_15`, old V3 changes "runner-up" to "champion." Neither length nor placeholder checks detect the substitution. The five prediction views agree on "runner-up," so V3+ selects the complete conservative record and removes the erroneous rank-edit box. Text $F_{0.5}$ rises from 0.5556 to 1, but edit $F_{0.5}$ remains zero: both candidates already normalize a different erroneous source character in "championship," hiding that necessary correction from differencing. The composite therefore rises from 0.2778 to 0.5, not to 1. This one case illustrates the distinction between restoring a rank value, faithfully recognizing the source, and grounding all necessary edits. It does not establish general factual-preservation behavior.

5 Conclusion

Team xc-皮 波's GroundCorrect separates recognition, verification, and grounding for the

Sentence-Level Track of VisCGEC-v2.0. The retained record reports second place and 0.5114 for official V3. Fixed-configuration controls provide the adapter evidence; other runs document system development. Exploratory V3+ changes one dev example, without selection-set or grounded-edit improvement. Recognition can hide necessary edits despite correct targets. Broader generalization requires repeated training and fully separated end-to-end evaluation.

Limitations

Recognition and retrieval errors remain possible; the differencer does not explicitly recover WordOrder. The local scorer is not the private evaluator. Dev overlaps training essays; holdout excludes only LoRA training and informed selection. Controls use one checkpoint, not repeated training seeds. V3's two repairs never fire on labeled splits. The known-error-driven fact patch uses correlated views and restricted expressions: one improvement does not establish unseen-data safety. Missing original weights and run records prevent claims of byte-exact reproduction of the scored archive.

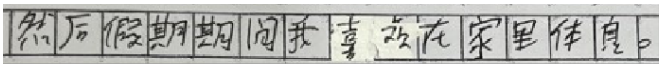
Ethical Considerations

We use released learner handwriting and train/dev annotations without personally identifying metadata. The archived run used no manual test corrections. Later reconstruction audits inspected some public test images, without test targets or hidden labels. This system should assist teachers and learners, not make high-stakes judgments without review.

Declaration on Generative AI

AI tools assisted writing, figures, and code. The authors verified results against saved artifacts and remain responsible for all claims, citations, and implementation.

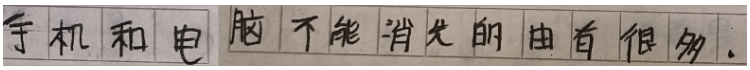
(a) OCR recovery: improved text, incomplete localization — 543_6



		F0.5: text / edit
Reference	然后假期期间我喜欢在家里休息。	
Frozen S2	然后假期期间我改为在家里休息。	0.56 / 0.00
V3+	然后假期期间我喜欢在家里休息。	1.00 / 0.00

V3+ = V3: text matches the reference, but source normalization hides the gold edit.

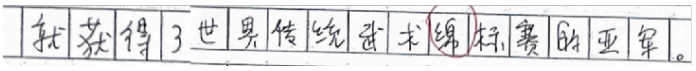
(b) Missing-character insertion: correct text and localization — 1969_7



		F0.5: text / edit
Reference	手机和电脑不能消失的理由有很多。	
Frozen S2	手机和电脑不能消失的有很多。	0.00 / 0.00
V3+	手机和电脑不能消失的理由有很多。	1.00 / 1.00

V3+ = V3: the missing character is inserted; the baseline deletes its neighbor.

(c) Rank protection: correct text, remaining localization failure — 1587_15



		F0.5: text / edit
Reference	就获得了世界传统武术锦标赛的亚军。	
Old V3	就获得了世界传统武术锦标赛的冠军。	0.56 / 0.00
V3+	就获得了世界传统武术锦标赛的亚军。	1.00 / 0.00

V3+ restores runner-up via the whole conservative record; the gold edit remains unlocalized.

Figure 3: Selected dev cases with unretouched images and cached outputs. Ruled lines and marks are original; amber marks baseline differences, green/red matching/nonmatching outputs. Scores are text/edit $F_{0.5}$. Only the third case changes under exploratory V3+.

References

Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. 2022. Slicing aided hyper inference and fine-tuning for small object detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 966–970. IEEE.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2025. LLaVA-UHD: An LMM perceiving any aspect ratio and high-resolution images. In *Computer Vision – ECCV 2024*, volume 15141 of *Lecture Notes in Computer Science*, pages 390–406, Cham. Springer Nature Switzerland.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Glenn Jocher and Jing Qiu. 2024. Ultralytics YOLO11.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13094–13102.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xin Sun and Houfeng Wang. 2022. Adjusting the precision-recall trade-off with align-and-predict decoding for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Dublin, Ireland. Association for Computational Linguistics.

Xiaoman Wang, Dan Yuan, Xin Liu, Yike Zhao, Xiaoxiao Zhang, Xizhi Chen, and Yunshi Lan. 2025. VisCGEC: Benchmarking the visual Chinese grammatical error correction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5054–5068, Albuquerque, New Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. FCGEC: Fine-grained corpus for Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1900–1918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haihui Yang and Xiaojun Quan. 2024. Alirector: Alignment-enhanced Chinese grammatical error corrector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2531–2546, Bangkok, Thailand. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339. PMLR.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States. Association for Computational Linguistics.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *Natural Language Processing and Chinese Computing*, volume 11109 of *Lecture Notes in Computer Science*, pages 439–445, Cham. Springer.

A Reproducibility Configuration

This appendix specifies the preserved local reconstruction and its evaluation inputs. The accompanying supplement contains complete UTF-8 prompt templates, an anonymized machine-readable configuration, split identifiers, artifact checksums, command examples, and a result index. These materials document the available reconstruction; they do not establish byte identity with the scored archive. They are supplied with the paper, rather than described as an already-public repository.

A.1 Verified training configuration

Table 4 combines the final training report, adapter configuration, and the saved training arguments at step 219. This distinction matters: the executed run saved every 75 steps, whereas the current command-line default is 100. Stage-1 training-data preparation used an inference batch of eight; the LoRA collator uses one example per device.

Parameter	Preserved value
Base / visual encoder	Frozen Qwen3-VL-4B
LoRA rank / alpha	8 / 16
LoRA dropout / bias	0.05 / none
Target projections	q, k, v, o; gate, up, down
Trainable parameters	16,515,072 (0.371%)
Device batch / accumulation	1 / 8
Epochs / optimizer steps	1 / 219
Learning rate	5×10^{-5}
Schedule / warmup ratio	Cosine / 0.05
Optimizer / weight decay	Fused AdamW / 0
Adam betas / epsilon	$(0.9, 0.999)$ / 10^{-8}
Gradient clipping norm	1.0
Precision	bfloat16; TF32 enabled
Checkpointing	Enabled; non-reentrant
Training / data seed	20260829 / 20260829
Save / logging interval	75 / 5 steps

Table 4: Verified local LoRA settings. The effective batch size is eight on one device. Projection names denote the corresponding `_proj` modules.

The assistant target is a compact JSON object containing only `source_text` and `target_text`. The collator verifies that the tokenized prompt is an exact prefix of the complete conversation and sets all prompt-token labels to -100. Consequently, image and user-prompt tokens receive no direct language-model loss. The supervised suffix includes the assistant answer and its chat-template terminator. The base parameters and visual encoder remain frozen.

The local environment used Python 3.11, PyTorch 2.8.0 with CUDA 12.8, Transformers 4.57.3, PEFT 0.18.1, Accelerate 1.10.1, Pillow 11.3.0, and Ultralytics 8.4.118. The full base-model revision is `ebb281ec70b05090aa6165b016eac8ec08e71b17`. The adapter tensor SHA-256 is `20ceecc9db7e62fd26a6003842d87b5a5586b9022e55ebc70d0abd32707ff578`. Hashes of the reconstructed detector, fallback predictions, and final output are supplied in the configuration.

A.2 Prompt and inference contract

Both passes use the same Chinese system instruction. Stage 1 receives the full image followed by its user instruction. Stage 2 receives the full image, up to four crops, and then a compact context object containing crop coordinates, retrieval scores and sources, and the Stage-1 source and target. The retained `pegasus_target_text` field is an empty string in this path. No reference target is placed in this context. The default Stage-2 prompt is used during LoRA training; the conservative prompt is an inference-only alternative using the same adapter.

The reproduced inference settings use greedy decoding, bfloat16, scaled dot-product attention, and per-image pixel bounds of 12,544–200,704. The preserved scripts specify 384 generated tokens for Stage 1 and 256 for text-only Stage 2. These token limits are made explicit in the supplied commands, because the saved inference reports did not serialize the complete original command line. Invalid or empty Stage-2 responses fall back to the cached Stage-1 hypothesis before deterministic grounding.

Candidate retrieval uses cached YOLO11s detections at confidence 0.20. The released-dev main comparison re-grounds fixed predictions at confidence 0.50; the recorded internal holdout instead uses the confidence-0.20 detections stored in its training manifest. Final test grounding uses a separate confidence-0.50 cache. These operating points should not be silently unified when comparing the preserved scores.

A.3 Split scope and cached dependencies

Essay groups are the identifier prefix before the first underscore. The training script sorts the 554 groups, shuffles them using seed 20260829, and reserves 55 groups, yielding 1,746 LoRA-training and 214 holdout examples. The supplement lists the reserved groups and hashes the ordered exam-

ple IDs independently of machine-specific paths. This isolation applies to LoRA training only: the detector was trained using annotations from the complete released training split, including these holdout images. The same holdout also informed guard selection; it is not an untouched end-to-end test set.

V3 selects among the default LoRA output, the conservative output, and a retained baseline submission at test time. For the labeled guard comparison, the holdout fallback is cached Stage 1, whereas dev uses frozen Stage 2. The supplement includes a CPU-only replay and per-example routing records for both operating points. The test baseline's complete prediction records, including boxes, are an input dependency. Only the linguistic-repair branch recomputes edits with the cached detector boxes. Exact output reproduction therefore requires the cached Stage-1 outputs, crop manifests, detector boxes, and baseline JSON, not merely the adapter weights. The result index separates available artifact evidence from command defaults and notes dependencies that are not bundled. Fresh inference and retraining can test the procedure but are not promised to reproduce all stored bytes.

A.4 Local evaluation edge conventions

Text edits are compared as exact tuples of operation, half-open source span, and replacement text. Text precision and recall use multiset intersection, so repeated exact predictions match only up to gold multiplicity. Grounded matching is greedy in prediction order; equal-IoU ties retain the first unused eligible reference in annotation order. Coordinate IoU uses differences without an inclusive-endpoint $+1$; a zero-union pair has IoU zero. An empty dataset returns zero macro aggregates. These implementation conventions do not claim identity with the private evaluator.

JSON parsing trims its outer generated representation and Markdown wrappers, not parsed string values. Scoring retains literal punctuation, whitespace, Unicode forms, case, and traditional/simplified characters. The source field for text scoring is the released gold transcription, whereas candidate grounding uses predicted source; no gold source is introduced into test-time grounding.

A.5 Alignment optimality and complexity

For Equation 5, removing the last match/skip transition leaves an optimal consumed-prefix alignment: a cheaper prefix would otherwise lower the full path cost. This gives the dynamic program's optimal substructure. The algorithm uses $O(nm)$ time and space, with $O(n + m)$ backtracking. Its optimum concerns the defined positional cost, not the correctness of source recognition or detector boxes.

B V3 Finalization and Exploratory V3+ Details

Figure 4 shows V3's first-match routing followed by V3+'s default-only fact check. Candidates d and c use default and conservative prompts with the same LoRA; b is the retained pre-LoRA baseline.

B.1 Priority consistency repairs

V3 first checks two narrow repairs:

1. Restore a Stage-1 comma at a preserved *if*-clause boundary.
2. Complete an asymmetric comparison's missing activity from the opposite predicate, limited to five activities: calling, spending time together, communicating, chatting, and meeting.

A match returns the repaired default target, rebuilding edits from the default source and cached detector boxes. Only this branch recomputes edits; all other branches retain their grounded records.

B.2 Placeholder and length routing

Otherwise, the V2 guard first checks for uppercase `X` in the default target but absent from the Stage-1 target. When triggered, it returns c if the conservative target has no `X`, and b otherwise.

Only without this trigger does it check default length growth of at least three characters. Growth is each candidate's target length minus its own source length. This branch returns c if its growth is less than three, and b otherwise. Growth signals risk, not proven hallucination. Without any trigger, return d. Clearing a branch's specific trigger suffices; returning a conservative candidate after the placeholder check does not require passing the later length check. Both risk branches select complete cached records rather than merging targets with another candidate's edits.

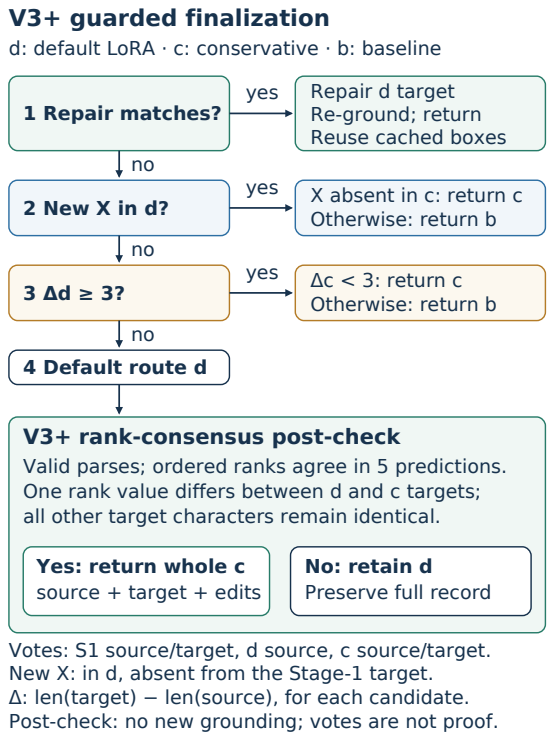


Figure 4: V3+ selection. Original V3 repairs and fallbacks are preserved; only the original default route reaches the new rank check. A fact-triggered switch copies the entire conservative candidate without recomputing edits.

B.3 Observed routing and scope

The unpatched V3 local test audit counts 232 default, two repaired, five conservative, and seven baseline records. The original holdout proxy improves from 0.2035 to 0.2080, using confidence-0.20 grounding and Stage-1 fallback; dev/test use frozen Stage-2 fallback. This holdout isolates LoRA, not detector training, and served model/guard selection. Neither V3 repair fires on dev/holdout, so V2 and V3 tie there. The two test interventions have no measured hidden-test score improvement.

B.4 Prediction-only fact post-check

The post-check is exploratory work from September 12, after competition submission. The rule-level definitions below are retained for reproducibility, not as evidence of a general factual-preservation capability.

The current V3+ configuration enables rank protection only. It canonicalizes the three named podium ranks and supported positive ordinal-place numerals. Specified compound suffixes such as those denoting competitions or teams are excluded. Mixed Arabic/Chinese or malformed ordinal expressions reject the entire decision rather than contributing partial consensus.

The Stage-1 parse-success field must exist and indicate no error. Both Stage-2 candidates must have the expected generation provenance, directly parseable raw JSON, and source/target strings matching the preserved cache. Missing fields, additional target rewrites, fact-count changes, or conflicting signatures preserve the V3 record. The exploratory numeric extension covers only Arabic quantities with enumerated units, treats the literal sign/unit as part of the type, and rejects comma-separated number fragments.

The dev audit changes one of 237 original default routes to the conservative candidate: final provenance counts are 236 default, five conservative, and four baseline records. Holdout preserves 203 default, eight conservative, and three baseline records. Neither variant changes any earlier repair or fallback branch. Both variants' full outputs, sample scores, and routing audits are included in the supplement. These counts do not describe a new official test submission.

B.5 Formal post-check definition

Let q_3 be V3's selected complete record and ρ_3 its recorded route. Let $F(s)$ contain the ordered recognized fact types and normalized values in string s; spans and literal surfaces are stored separately. For default and conservative records $q_d = (x_d, y_d, E_d)$ and $q_c = (x_c, y_c, E_c)$, the support views are

$$\mathcal{V} = (x^{(0)}, y^{(0)}, x_d, x_c, y_c). \tag{7}$$

Consensus $\mathcal{A}$ requires $F(\mathcal{V}_1)$ to be nonempty and equal to $F(v)$ for every $v \in \mathcal{V}$. These are correlated predictions, not independent votes or a calibrated correctness probability.

Define $M(s)$ as a literal text skeleton obtained by replacing each recognized span with its fact type. Condition $\mathcal{B}$ requires y_d, y_c to contain the same ordered fact types, exactly one different value, exactly one different fact surface, and $M(y_d) = M(y_c)$. Punctuation, non-fact wording, and all other fact surfaces must remain unchanged. With P denoting verified parsing and valid inputs, the post-check is

$$q_+ = \begin{cases} q_c, & \rho_3 = d \ \wedge \ P \ \wedge \mathcal{A} \wedge \mathcal{B}, \\ q_3, & \text{otherwise.} \end{cases} \tag{8}$$

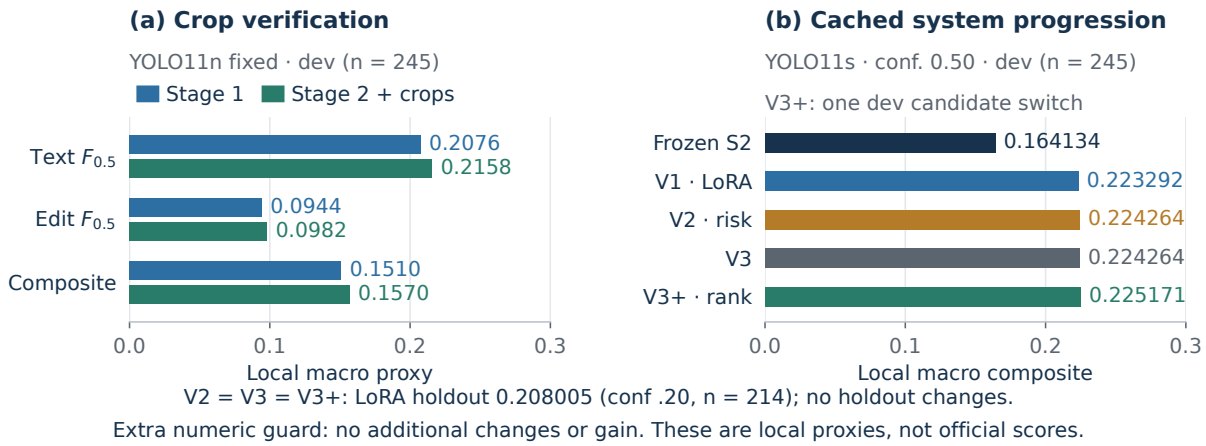


Figure 5: Two local dev comparisons with separate grounding setups. (a) Frozen Stage 1 versus evidence reinspection with fixed YOLO11n boxes. (b) Cached system progression through V3+ with YOLO11s confidence 0.50; the frozen-to-LoRA step changes the output prompt as well as the adapter. The adapter-only control is Table 3. The footer's selection score uses grounding 0.20.

Here $\rho_3 = d$ means the original, unrepaired default route, not merely a text match. Parsing must explicitly succeed at Stage 1; both Stage-2 outputs must be non-fallback generations whose parsed source/target fields equal the cache.

Equation 8 preserves every non-default V3 route and ensures $q_+ \in \{q_3, q_c\}$. Selecting y_c also selects x_c and E_c; no candidate's sentence is paired with another candidate's edits. These are structural guarantees, not guarantees of semantic accuracy: shared OCR errors and missing edits can survive consensus. Output schema checks cover IDs, operation types, nonempty text, integer coordinates, and image bounds.

The replay verifies all original V3 per-sample scores and aggregates with zero differences and checks whole-record identity. The retained experiment reports 31 passing synthetic boundary tests covering parsing, invalid facts, extra rewrites, and preservation of earlier routes. These implementation checks do not extend the empirical result beyond the single changed development example.

C Auxiliary Development Analyses

These development-only and post-submission diagnostics do not measure the scored archive.

C.1 Development-only textual context and oracle boxes

An early configuration used public OCR disagreement, a Pegasus target (Zhang et al., 2020) for proposals and Stage-2 context, public source transcription, and annotated character boxes. Neither textual input was available to the submitted test runner. Its proxy rose from 0.1561 at Stage 1 to 0.1637 after verification. These are auxiliary, not controlled or main development results. Gold targets and edit boxes are never used for test-time crop selection.

C.2 Early direct-image adaptation

A one-epoch direct-image Qwen3-VL LoRA trained 11.8M language-attention parameters (0.265% of 4.45B), freezing the visual encoder (Hu et al., 2022). On the same first 64 auxiliary-dev examples, its evidence-pipeline proxy was 0.1586 versus frozen 0.1876. Training–inference prompt mismatch is one possible explanation; the adapter was discarded. The later adapter differs in output prompt, trainable modules, and evaluation setup, preventing attribution solely to format matching.

C.3 Fixed-prediction confidence sweep

With fixed frozen-Qwen predictions and reconstructed YOLO11s boxes, thresholds 0.45, 0.48, 0.50, and 0.52 all score 0.1641. Five-fold essay-group selection chooses 0.50 throughout, with no concatenated held-out gain. Retaining 0.50 does not establish a threshold benefit for V3.

C.4 Detector-capacity probe

Post-submission YOLO11n/s retrainings used identical data and 30 epochs, yielding dev $\text{mAP}_{50:95}$ of 0.801/0.800. Frozen Qwen outputs scored 0.1570/0.1641 with n/s localization. This

compares post-hoc retrainings, not an n-to-s upgrade of the award system, which already used s.

D Retained-Artifact Integrity Checks

The September 12 V3+ experiment preserves all older archives and records hashes and predictions for 245 dev and 214 holdout examples. Its official hidden-test score is unknown; its local gain is not attributed to older runs.

The retained August 29 candidate has 246 unique records, 245 edits, 85 empty edit lists, nonempty targets, and valid in-image integer boxes. Its timestamps predate upload, but do not prove that its exact bytes were scored. The September 11 reconstructed V3 archive has 246 records and 252 edits with the same schema checks. Both have one root member, `submission_sentence.json`.

Hashes and the older frozen-pipeline audit are in the companion supplement; that audit is not evidence for V3+ accuracy.