FROM SIGNS TO SEQUENCES: COMPOSITIONAL SYNTHESIS FOR CUNEIFORM RECOGNITION ACROSS PERIODS AND COLLECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training data that pair cuneiform line images with sign sequences are scarce, even though isolated sign images and machine-readable texts are available separately. This scarcity keeps large collections of digitized tablets difficult to search and analyze automatically. We introduce a compositional supervision method that combines these independently available resources. Given a normalized sign sequence from a text corpus, we sample corresponding real sign crops from a tablet-disjoint visual dictionary (augmented for rare signs) and compose them into a synthetic line image. These image–sequence pairs pre-train three recognition architectures—a CRNN, an SVTRv2-inspired CTC model, and PaddleOCR-VL—before adaptation on a limited number of annotated tablets. In the principal evaluation, sign locations are supplied by ground-truth annotations and the input lines are reconstructed from their crops. With a fixed sign inventory, adaptation on 100 Ur III tablets yields sign error rates (SER) of 18–25% on in-period data. Removing compositional pre-training increases error by 24–52 percentage points across architectures and evaluation domains. Controlled experiments within the SVTRv2-inspired model further show that multi-sign sequence training provides substantially greater benefits than isolated-sign training, whereas preserving attested sign order and co-occurrence produces smaller, domain-dependent gains. Adaptation primarily improves recognition within the target period, while cross-period and cross-collection transfer remains difficult. An exploratory modular pipeline with predicted sign locations and reading order reaches 38.05% whole-side SER on reused in-period tablets, versus 68.92% for its strongest original geometric baseline. These results establish a practical way to repurpose existing Assyriological image and text collections as sequence-level supervision, while clarifying the remaining challenges for scalable cuneiform image-to-text systems.

1 INTRODUCTION

Cuneiform tablets are being digitized at scale in repositories such as the Cuneiform Digital Library Initiative (CDLI) (CDLI contributors, 2026) and the Electronic Babylonian Library (eBL) (Che et al., 2026), and automated reading could accelerate scholarly access. Annotated sign crops and machine-readable texts exist, but real line images aligned to their sign sequences are scarce. We ask: *how well can these resources support sequence recognition, and what does limited annotation of target tablets add?* Our principal task is to recognize sign sequences from lines reconstructed by stitching ground-truth sign crops; the output is a sequence of sign names, not the full scholarly transliteration, which adds phonetic and editorial choices. Complete tablet-side images with predicted sign locations are evaluated separately (§6.3).

Constructing compositional supervision. We build the missing supervision from real isolated sign crops and text-derived sign sequences (Figure 1). A text corpus supplies normalized sign sequences; a tablet-disjoint dictionary supplies real sign crops, plus augmented variants for rare signs. For each token we sample an exemplar, height-normalize it, and stitch the crops into a line image labeled with the sequence. The line arrangement is synthetic; its sign crops come from existing images or their augmentations. We pre-train sequence recognizers on these synthetic pairs (Stage-A) without paired

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

real-line images, then adapt them on annotated target tablets (Stage-B). Stage-B adaptation and the principal evaluation also use stitched lines, built from each real line's own ground-truth crops. Such composition has precedent in other scripts (Senthil et al., 2020); here we ask what it is worth for cuneiform across recognition paradigms, and what the supervision design and the target annotation each contribute.

Recognition from existing resources and limited annotation. We test two connectionist temporal classification (CTC) recognizers, CRNN and SVTRv2* (an SVTRv2-inspired variant), and an autoregressive vision-language model (VLM), PADDLEOCR-VL. On reconstructed lines with a closed vocabulary, adaptation on 100 Ur III tablets brings all three to 18–25% sign error rate (SER) in-period. Removing Stage-A raises SER by 24–52 pp across recognizers and domains (§6.1), including for the VLM with its strong pre-trained base. Target-tablet adaptation improves in-period SER by 12–15 pp, but off-domain gains are smaller and uneven, and cross-collection SER stays at 66–74% even after out-of-inventory signs are excluded.

What makes the supervision useful? To understand the gain, we compare what the same CTC architecture (SVTRv2*) learns from individual signs and from composed multi-sign lines. With 100 adaptation tablets, pre-training on shuffled multi-sign lines lowers SER by 23–46 pp compared with the same token groups rendered as single-sign images (Figure 2). The comparison uses test lines that share no exact match or 4-gram with the source text. Attested co-occurrence and order add a smaller, domain-dependent benefit, not clearly supported in-period. Given known sign boundaries, a strong per-sign classifier has lower off-domain SER than SVTRv2* (§6.1).

Contributions. (i) *Compositional supervision for cuneiform*: we turn separately available glyph and text resources into paired sequence supervision through sign-name mapping, dictionary construction, rare-sign augmentation, and line rendering (§4). (ii) *Effectiveness across recognition paradigms*: in matched ablations, compositional pre-training lowers SER by 24–52 pp for CRNN, SVTRv2*, and PADDLEOCR-VL across periods and collections, given limited adaptation on target tablets (§6.1). (iii) *Evidence on supervision and annotation value*: within-SVTRv2* controls show a large effect of the multi-sign training format (23–46 pp) and a smaller contribution from attested text structure; across all three recognizers, adaptation on target tablets helps in-period recognition far more than off-domain recognition (§6.1, §6.2). (iv) *Exploratory whole-side recognition*: a modular pipeline with predicted sign locations and reading order reaches 38.05% whole-side SER on reused in-period tablets (§6.3).

2 RELATED WORK

Cuneiform visual recognition. Prior work on cuneiform image recognition (Bogacz & Mara, 2022) spans sign-level tasks: sign and wedge detection on photographs and 3D renderings (Dencker et al., 2020; Stötzner et al., 2023a;b; Cobanoglu et al., 2024), isolated-sign classification (Verwimp et al., 2025; Mikulinsky et al., 2025), and segmentation-guided per-sign classification (Poudel et al., 2026). A second line builds reading-order output from such predictions: detect-then-classify pipelines such as DeepScribe (Williams et al., 2025) report a whole-tablet transliteration character error rate, and the Electronic Babylonian Library pipeline (Che et al., 2026) scales sign detection to 87k fragments with heuristic line grouping. It reports normalized n-gram set-overlap of 0.23 on 26,790 sufficiently annotated fragments, excluding training fragments; this set-overlap measure is not directly comparable to sign-token edit distance. Line-level transliteration was also attempted from hand-drawn copies with sparse data (Bogacz et al., 2017). We instead study compositional line-level supervision for sequence recognizers, evaluated with sign-token edit distance (SER).

Data resources. MaiCuBeDa's sign and line annotations on 3D renderings (Homburg et al., 2025; Ahlborn & Homburg, 2025) supply real lines for adaptation and evaluation. The eBL and Old Babylonian sign corpora (Lewenstein et al., 2026; Smidt et al., 2026) supply isolated exemplars for composition; CuneiML (Chen et al., 2023) is a related multimodal tablet collection.

Concurrent tablet-level work. A concurrent submission predicts ordered sign sequences directly from Ur III tablet-side images with a single vision-language model trained progressively on signs,

synthetic lines, synthetic tablet sides, and real tablet sides (Anonymous, 2026). By contrast, our principal experiments isolate the effect of compositional line supervision under known sign locations across three recognizers, periods, and collections. Our whole-side experiment combines separately trained detection, classification, and sequence-verification components with geometric layout rules and language-model scoring. The studies share some source resources but test different training strategies and inference systems. Comparing their SERs requires aligning evaluation images, sign inventories, reference construction, and scoring conventions.

Synthetic and compositional supervision. Synthetic data bootstraps low-resource text recognition (Pippi et al., 2025; Baena et al., 2024); CuReD (Gordin et al., 2024) applies it to typeset cuneiform transliterations. For tablet images, Rusakov et al. (2019) generate signs with adversarial networks; Brandenbusch et al. (2021) generate class-controlled signs using surrounding visual context. ProtoSnap (Mikulinsky et al., 2025) uses prototype-to-crop skeleton alignments for structure-conditioned synthesis. StackMix (Shonenkov et al., 2021) recombines fragments weakly segmented from labeled handwriting lines using external text; Souibgui et al. (2022) build lines from generated and augmented symbols. DTLR (Baena et al., 2024) pre-trains on font-rendered and CASIA-based synthetic lines, then adapts to real line annotations. Closest to our setting, Senthil et al. (2020) compose synthetic Hindi word images from an independent handwritten-character library and a lexicon to pre-train a CNN–RNN/CTC recognizer for few-instance adaptation—the same resource-combination logic. What remains unclear for cuneiform is how much such composition contributes after adaptation to real tablets, and whether line format or attested sign order produces the benefit (§6.1, §6.2). For scene text, Baek et al. (2021) showed that pooled real labeled data trains competitive recognizers; we measure what target-tablet adaptation adds after compositional pre-training.

VLMs and language priors. VLMs support vision-language tasks (Liu et al., 2023; Chen et al., 2024), while OCR benchmarks report difficulties across Unicode scripts and ancient writing (Kargaran et al., 2026; Chen et al., 2025). Language priors can affect OCR outputs, with model-specific visual-grounding failures (Karamolegkou et al., 2026; Liang et al., 2026). Fine-tuned VLMs and lightweight OCR models are competitive in their settings (Chung & Choi, 2026; Cui et al., 2026).

Evaluation contamination. Our overlap audit connects to work on cross-language recognition, distribution shift, and evaluation contamination (Crosilla et al., 2025; Recht et al., 2019; Levchenko, 2025). Synthetic lines rendered from a text corpus can expose evaluation sign sequences even when tablets and images are disjoint, so tablet-disjoint splits do not guarantee unseen lines (§5.3).

3 Task and Available Resources

3.1 Recognition Task and Input Conditions

Line-level cuneiform sign-sequence recognition maps a reconstructed tablet line image to Assyriological sign names (e.g., `LUGAL`, `ŠUM`$_2$, `KI`; CDLI/ePSD convention, CDLI contributors, 2026). The inventory is fixed at 749 labels: 747 canonical sign types, `X` (illegible), and `lacuna` (reconstructed); "out of vocabulary" (OOV) means outside this inventory. Stage-A supervision draws on two independent resources combined in §4.1—a tablet-disjoint dictionary of real isolated-sign crops and Ur III sign sequences from CDLI texts (CDLI contributors, 2026)—while annotated MaiCuBeDa tablets (Homburg et al., 2025; Ahlborn & Homburg, 2025) supply Stage-B adaptation and evaluation (coverage in §5.1).

Oracle-localized input. The controlled experiments supply ground-truth (GT) sign locations. We crop sign boxes derived from GT polygons and order them by horizontal box center on MaiCuBeDa Hilprecht tablets (in-period and cross-period sets), or by annotation character index on MaiCuBeDa-HT (cross-collection set). The crops are height-normalized and stitched into a reconstructed line; OOV signs are excluded from both image and label. These sign-localized, stitched inputs reflect the current annotation scale and isolate recognition from localization. A natural-line diagnostic is reported in Appendix I.

Automatic whole-side input. The exploratory system (§4.3) instead reads tablet-side images with predicted boxes and reading order, without annotation-based filtering (§5.4, §6.3).

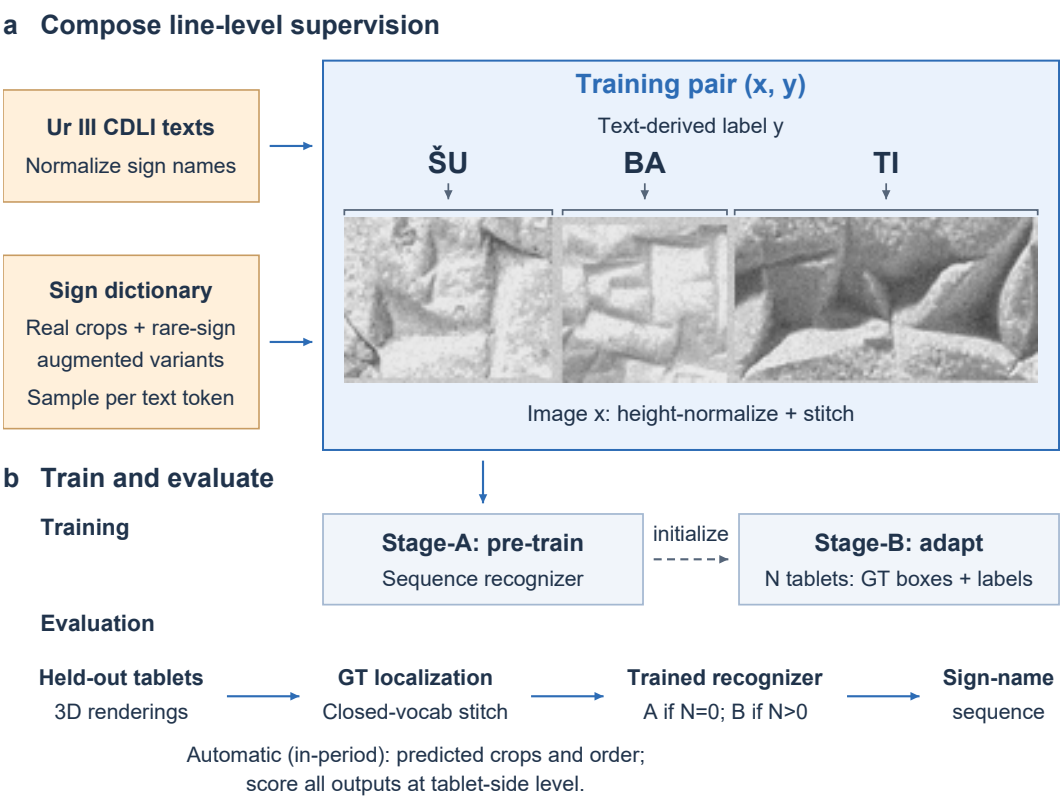


Figure 1: **Synthetic training pairs from real sign crops and existing texts.** **(a)** Texts supply sign sequences; a dictionary supplies a crop for each token. Height-normalizing and stitching the sampled crops creates a synthetic line image paired with the text-derived label. **(b)** Stage-A pre-trains the recognizer; Stage-B adapts on N annotated Ur III tablets (N=0 skips only Stage-B). Primary evaluation uses GT-localized, closed-vocabulary stitches of 3D renderings; the exploratory pipeline with predicted localization is in §4.3 and Figure 5 (Appendix D).

4 Compositional Supervision

4.1 Constructing Synthetic Training Pairs

We synthesize labeled line images for Stage-A pre-training as follows (Figure 1a). A sequence of canonical sign names is taken from an Ur III CDLI text and mapped to the label inventory (§3.1). For each sign, we sample one crop uniformly from its dictionary entry. The dictionary contains real glyph crops; under-represented sign types also include geometric and photometric augmentation variants within existing classes (Appendix B). We height-normalize each crop to 64 pixels and concatenate the crops left-to-right into a *stitched* line image. The normalized sign-name sequence from the text supplies its label; OOV signs are excluded from both image and label. The dictionary holds 199,243 images covering 744 of the 747 canonical sign types and contains no MaiCuBeDa-tablet images (provenance in Appendix B). The three empty entries (du, $\check{s}ar_2$, 600) occur in no Stage-A line or evaluation reference: normalization maps du and $\check{s}ar_2$ to DU and $\check{S}AR_2$, which have exemplars, and drops bare numerals.

Synthetic corpus and evaluation separation. After removing every composed line whose CDLI text belongs to a development, in-period, or cross-period tablet (Haft Tappeh texts are already absent), the Stage-A corpus contains 387,315 composed image–sequence pairs (deletion counts in Appendix A). They average 4.6 signs and retain all 285 corpus sign types. "Text-disjoint" means no shared CDLI *text*: repeated formulae can still reproduce an evaluation line verbatim (§5.3). All primary results use Stage-A models trained on this corpus. CTC models split it into 347,322 training and 39,993 validation pairs; PaddleOCR-VL uses 126,000 training records with low-rank adaptation (LoRA),

followed by the continuation documented in Appendix C. Texts of the Stage-B training-pool tablets are not removed.

4.2 Sequence Recognizers and Target-Tablet Adaptation

The primary experiments use three sequence recognizers: two CTC models—CRNN (Shi et al., 2017) and our SVTRv2*—and the autoregressive PaddleOCR-VL.

CTC recognizers. CRNN and SVTRv2* map a stitched line image to the sign sequence by CTC decoding (749 labels + blank); no sign-boundary labels are given to the decoder. The asterisk in SVTRv2* marks our SVTRv2-*inspired* variant, not the original: SVTR mixing blocks and SVTRv2 vertical feature rearrangement, but *no* semantic-guidance module (Du et al., 2022; 2025).

Vision-language model. PaddleOCR-VL (PaddlePaddle Team, 2026) is a 0.9 B multimodal model with a visual encoder and autoregressive decoder; Stage-A and Stage-B fine-tune LoRA adapters (Hu et al., 2022); configuration and decoding are in Appendix C. Its smaller Stage-A budget (§4.1) is a confound for cross-model comparison.

Stage-B fine-tunes each recognizer on N real MaiCuBeDa tablets. Inputs are GT-localized sign crops from each annotated line, stitched from left to right by bounding-box center. CTC updates all parameters; PaddleOCR-VL continues LoRA from the Stage-A adapter. Details are in Appendix C.1.

4.3 Recognition with Predicted Localization

For an exploratory whole-side evaluation, we assemble separately trained components. Given a tablet-side image and its surface type, a YOLO26n detector (Jocher et al., 2026) proposes sign boxes, a dictionary-pre-trained DeiT3 classifier (Touvron et al., 2022) proposes candidate labels, geometric layout rules group and order the boxes into lines, and a base decoder selects label sequences using classifier and six-gram language-model (LM) scores with weights selected on the development (DEV) set. A compositionally pre-trained $N{=}100$ SVTRv2* then verifies candidate labels on their own crops, adding full CTC probability with a DEV-selected weight (Figure 5 in Appendix D). Inference uses no reference boxes, lines, or text; the classifier uses no synthetic-line pre-training. This modular inference system differs from the direct tablet-side image-to-sequence model studied in concurrent work (Anonymous, 2026).

5 Evaluation Protocol

5.1 Evaluation Sets and Annotation Budgets

Training and selection (Ur III, MaiCuBeDa Hilprecht). For line recognition, Stage-B samples N tablets from a pool of 214 Ur III tablets. A tablet-disjoint 50-tablet development (DEV) set (309 CTC lines, 342 VLM lines) selects learning rates and early stopping. Because the VLM's mapping also scores numerals, its in-period, cross-period, and HT line counts are 718, 1,254, and 1,408 versus 670, 1,243, and 1,320 for CTC (Appendix B). Primary line evaluation excludes lines with fewer than three in-vocabulary signs; whole-side evaluation keeps all annotated in-inventory signs.

Automatic-system budgets and selection. The detector trains on 450 tablets and validates on 50, DeiT3 adapts on the 214-tablet pool, and the recognizers on a 100-tablet subset. Of the 50 DEV tablets, 43 are in detector training and 7 in detector validation, while DEV also selects the system's checkpoints and decoding weights. These selections are therefore conditional on a detector that has seen DEV; the evaluation tablets are outside detector training, detector validation, and DEV (Appendix D.1).

Three primary evaluation sets are disjoint from recognizer training and DEV at the tablet level and from Stage-A at the CDLI-text level (§5.3). Sets (2) and (3) were not used to tune or select the primary recognizers; the generalization claims rest on them.

(1) *In-period (reused; Ur III)*: 100 tablets (198 tablet-side images, "stems"), 670 lines, same period and collection as training; these lines featured in earlier in-period analyses and were inspected during development, so we label them *reused* rather than pristine.

(2) *Cross-period*: 192 MaiCuBeDa tablets (1,243 lines) span six non-Ur III periods, up to ∼500 years from training (Appendix F). HS 1032 lacks CDLI mapping and period metadata. The set was assembled without inspecting model outputs. Cross-period transfer is relative to the Ur III adaptation period: the crop dictionary spans multiple periods, with most crop-period labels unknown, so Stage-A is not period-disjoint.

(3) *Cross-collection*: MaiCuBeDa-HT (Ahlborn & Homburg, 2025), the Haft Tappeh collection of Middle-Elamite-period Akkadian tablets (c. 1500–1400 BC), annotated by a different campaign. The CTC evaluator covers 1,320 lines on 305 tablets. Of the annotated tokens, 18.5% are outside our inventory and are excluded from image and label, as in the other sets under the oracle protocol. HT differs from training in site, language, period, scribal tradition, and annotation campaign (a different imaging batch under the same MSII rendering technology).

Frozen manifests fix all three sets (cross-period and HT hashes in Appendix C). The 670-line in-period set instantiates only 147 of the 749 labels (93 with ≥5 occurrences), the cross-period set 212, and the cross-collection set 115; performance covers these instantiated subsets, not the full inventory. The 285 types in Stage-A text cover 100/99.95/96.7% of in-period/cross-period/HT evaluation tokens.

5.2 Comparisons and Model Selection

Primary protocol (matched-protocol ablation). For each recognizer and annotation budget $N \in \{25, 100\}$, we train one arm with Stage-A pre-training and one without it. Both arms use the same Stage-B tablets, seeds $\{42, 123, 456\}$, batch size, epoch budget, and patience (CTC: 100 epochs/15; PaddleOCR-VL: 40/10). Only the learning rate may differ: each arm selects it by seed-42 DEV LCS Recall from a fixed grid, then applies it to all seeds. Appendix C gives the grids and the VLM's reuse of previously DEV-selected rates after text-disjoint retraining. Without Stage-A, CTC models start randomly and the base VLM receives a fresh LoRA adapter. Analysis and cross-period evaluation plans were locked before primary evaluation (Reproducibility Statement).

5.3 Text-Overlap Controls

All primary results (§6.1), strict subsets, and automatic sequence-model results use *text-disjoint* Stage-A models (§4.1). **Text-level deduplication does not remove recurring formulae.** Stage-A contains none of the 341 CDLI texts linked to DEV, in-period, or cross-period tablets (one cross-period tablet has no CDLI mapping). It also contains none of the 282 HT texts, because it draws only on Ur III texts. Yet 48.6% of in-period lines still recur verbatim because administrative formulae appear across tens of thousands of texts. Cross-period and HT lines recur far less (Table 4).

We therefore recompute every contrast on a *strict* subset requiring *both* conditions: a line shares *no 4-gram* with any Stage-A line *and* its exact sequence never occurs in the Stage-A corpus (Table 5, Appendix A). Both are needed because a short line (<4 signs, 45% of the in-period set) has no 4-gram and would pass a 4-gram filter even when it recurs verbatim; the exact-unseen condition removes these. The in-period strict subset is 43% of lines (n=290 in the annotation audit); a second computation on the SER records uses raw sign names and reports its own subset sizes (Appendix A).

5.4 Metrics and Uncertainty

Primary and secondary metrics. Our primary metric is **Sign Error Rate** (SER↓), the corpus-level sign-token Levenshtein distance

$$\mathrm{SER} = \frac{\sum_i (S_i + D_i + I_i)}{\sum_i |y_i|},$$

where S_i, D_i, and I_i count substitutions, deletions, and insertions against reference y_i; unlike LCS Recall, SER penalizes insertions. *Oracle-line SER* scores saved outputs on the reconstructed lines; PaddleOCR-VL outputs are normalized and vocabulary-filtered beforehand. *Automatic whole-side SER* concatenates all output rows of a side without spatial masking, retaining unknown and unmatched output tokens. Its fixed reference has 3,280 in-inventory signs; 903 annotated regions whose readings map outside the inventory (e.g., numeral notations) are excluded from the reference only, so outputs produced there can still incur errors. It therefore measures a closed-inventory

target, not full transcription, and is not comparable to oracle-line SER (Appendix D.5). **Exact-Line Accuracy**, LCS Recall, and length diagnostics are defined in Appendix A.

Uncertainty and multiple comparisons. Primary ablation contrasts compare 3-seed means using paired tablet-level bootstrap (B=10,000, percentile 95% CI), resampling the same tablets across fixed runs. The intervals condition on fixed training runs: seeds are not resampled, and Stage-A retraining variance is excluded. The 95% CIs are unadjusted for multiple comparisons. All 18 contrasts in Tables 1 and 13 have CIs excluding zero. The strict in-period attested-order contrast does not (§6.2).

6 Results

6.1 Recognition, Adaptation, and Transfer

Absolute recognition. With oracle sign localization and closed-vocabulary, stitched inputs, adaptation on 100 real Ur III tablets brings the models to 17.9–24.8% SER in-period, 38.4–41.9% cross-period, and 66.1–73.6% cross-collection (Table 1; qualitative examples in Figures 3 and 4). Exact-line accuracy falls from 42–57% in-period to 1–3% cross-collection. The reported HT SER

Table 1: **Synthetic pre-training benefit at N=100.** SER↓ (%) after Stage-B adaptation on 100 tablets with and without Stage-A pre-training on the synthetic pairs (§4.1); N=0 is Stage-A alone. GT-localized, closed-vocabulary inputs; means over seeds 42/123/456. Δ = without − with Stage-A, with paired-tablet 95% CIs conditional on fixed runs. Line counts differ by model (§5.1), so cross-model values are descriptive. N=25: Table 13; per-seed values: Appendix E.

Set	Model	N=0	with St.-A	no St.-A	Δ (pp) [95% CI]
In-period (Ur III)	SVTRv2*	33.6	21.5	72.5	+50.9 [48.5, 53.3]
	CRNN	37.5	24.8	73.1	+48.3 [45.9, 50.6]
	PaddleOCR-VL	32.8	17.9	45.8	+27.9 [26.0, 29.9]
Cross-period	SVTRv2*	40.7	38.4	90.8	+52.4 [49.3, 54.9]
	CRNN	48.3	41.9	90.5	+48.6 [45.7, 50.9]
	PaddleOCR-VL	41.1	41.2	76.2	+35.0 [32.5, 37.0]
Cross-collection	SVTRv2*	70.4	71.2	97.8	+26.7 [24.9, 28.6]
	CRNN	72.5	73.6	97.5	+23.9 [22.1, 25.6]
	PaddleOCR-VL	69.6	66.1	90.1	+23.9 [22.3, 25.6]

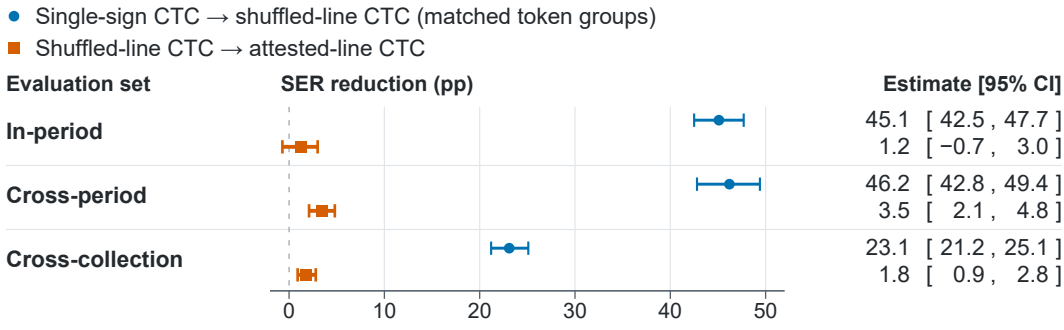


Figure 2: **Training-format contrasts** within SVTRv2* (N=100; 3-seed Stage-B mean). Positive favors the second format; bars: paired-tablet 95% CIs given fixed Stage-A checkpoints. Values are on strict-subset lines (exact-unseen, no shared 4-gram). Canvas, target length, and alignment co-vary in the first contrast; co-occurrence and order in the second. They do not causally decompose Table 1 (Table 19).

applies only to in-inventory signs. Another 18.5% of annotated HT tokens are excluded from its reference and input, so the result does not measure recognition of the full collection.

Value of synthetic pre-training. Removing Stage-A raises SER by 24–52 pp at N=100 (Table 1). The gaps are smallest for PADDLEOCR-VL and on HT, where absolute SER remains high for every model. For PADDLEOCR-VL, whose without-Stage-A arm keeps the strong pre-trained base with a fresh LoRA adapter, removing Stage-A still costs 24–35 pp. At N=25 the gaps are larger (up to 68 pp; Table 13); there the no-Stage-A CRNN collapses to near-total error (SER $\approx$ 100–102%), and completing the full 100-epoch schedule still leaves it weak (Appendix K).

The gain survives a converged control and a strict text subset. For SVTRv2*, the Stage-A advantage remains large against a 300-epoch from-scratch baseline: +46.7/+50.7/+25.2 pp at N=100 (Appendix E). On strict text subsets, the removal costs stay within $\sim$2 pp of the all-line costs across all nine N=100 contrasts (Table 5).

Adaptation and transfer. Without target tablets (N=0), Stage-A models reach 33–72% SER. Adapting on 100 Ur III tablets cuts in-period SER by 12–15 pp, to 18–25%. Off-domain reductions are smaller: -0.2 to $+6.3$ pp cross-period and -1.1 to $+3.5$ pp cross-collection. Further Ur III supervision mostly buys in-distribution refinement, while Stage-A already supplies most off-domain capability at N=0 (the composed corpus covers nearly all evaluation tokens; for the VLM, partly from its pre-trained base). Cross-period performance is heterogeneous across the six periods (Appendix F). Cross-model differences are descriptive: initialization, capacity, token mapping, and Stage-A budget differ (Appendix E).

Strong per-sign classifier baseline. A dictionary-trained DeiT3-medium classifier (Touvron et al., 2022), adapted on the same 100 tablets, classifies each GT sign crop in line order. Evaluation lines, labels, OOV rule, and strict subsets match SVTRv2* exactly. Given true sign boundaries, DeiT3 performs better than SVTRv2* on the two off-domain sets (Table 2); on strict in-period lines its SER is 22.2% versus 25.4% ($\Delta = -3.2$ pp, 95% CI $[-5.5, -0.8]$). The compositional pre-training result therefore does not establish that sequence decoding is preferable to classifying each localized sign; with predicted boundaries, SVTRv2* serves as a verifier (§6.3). Different architectures, initializations, and objectives make this a pipeline comparison, not an isolated test of sequence context (Appendix J).

Table 2: **Strong per-sign comparison under the same oracle protocol.** N=100, three-seed SER (%); Δ = per-sign $-$ SVTRv2*; paired-tablet 95% CIs condition on fixed checkpoints. Strict subsets and exact-line scores: Table 22.

Set	Per-sign DeiT3	SVTRv2*	Δ (pp) [95% CI]
In-period	21.0	21.5	-0.5 [-1.8, 0.9]
Cross-period	27.8	38.4	-10.6 [-12.2, -8.9]
Cross-collection	62.4	71.2	-8.8 [-9.9, -7.7]

6.2 Contributions of Line Format and Attested Text Structure

Which parts of the synthetic-data construction help? We train the same CTC architecture (SVTRv2*) on single signs, on shuffled multi-sign lines, and on lines following attested text. These Stage-A models form a separate checkpoint family (the attribution family): the arms share the renderer and match token exposure and optimizer updates, but not compute, and then follow the locked Stage-B protocol. Shuffling permutes tokens within each split and re-cuts lines, changing co-occurrence as well as order. Construction, per-arm results, and all-lines contrasts are in Appendix G.

Training-format effects within SVTRv2*. The largest tested step is from single signs to multi-sign lines. A parent-matched control renders the same shuffled token groups as separate one-sign images ("singleton-matched"). At N=100, shuffled-line pre-training lowers strict-subset SER by 45.1, 46.2, and 23.1 pp in-period, cross-period, and cross-collection, respectively (Figure 2). Thus grouping

alone does not explain the gain, although canvas, target length, and CTC alignment still co-vary. This compares CTC pre-training formats, not per-sign classification (§6.1).

Attested text structure adds a smaller, domain-dependent benefit. Compared with shuffled recomposition, attested co-occurrence and order jointly reduce strict-subset SER by $+1.2$ $[-0.7, 3.0]$ pp in-period and by $+3.5/+1.8$ pp cross-period/cross-collection, respectively. The in-period interval crosses zero; the off-domain intervals remain positive (Figure 2, Table 19). Each arm uses one Stage-A seed, so the intervals exclude Stage-A training variance.

6.3 Exploratory Modular End-to-End Recognition

In the exploratory whole-side comparison, the full modular system reaches **38.05%** SER, versus **68.92%** for the strongest original geometric route (Table 3), a **30.87 pp** difference (mean of three runs that vary the classifier and verifier seeds; paired 95% CI $[28.10, 33.88]$). This comparison evaluates complete systems; it does not isolate the effect of any one module. The original baselines apply compositionally pre-trained $N{=}100$ recognizers to YOLO-grouped crops. They share its detector weights, but proposals, crop selection, ordering, decoding, and component budgets differ. All systems are scored on the same 198 sides of the 100 reused in-period tablets; the reference excludes 903 out-of-inventory regions, but every output is scored (§5.4). DEV-based selections are conditional on detector exposure (§5.1).

Table 3: **Original geometric-route baselines and full system.** Mean whole-side SER over three recognition-seed runs on the 100 reused in-period tablets; all outputs are scored. 95% CIs condition on fixed models. Budgets and decoding differ (Table 8; Appendix D).

System	Runs	SER (%↓)	95% CI
Geometric + CRNN	3	72.22	[68.64, 76.12]
Geometric + SVTRv2*	3	70.78	[67.26, 74.63]
Geometric + PaddleOCR-VL	3	68.92	[65.42, 72.81]
Full system	3	**38.05**	[35.82, 40.43]

Holding the other components fixed within each run (detector, classifier, candidate pool, layout, LM, and decoder weights), removing the SVTRv2* verifier raises mean SER from 38.05% to 39.88%. Verification reduces substitutions and insertions, with more deletions and slightly lower LCS Recall (Table 12; per-pair results in Appendix D). Separately, an independent SVTRv2* on the base decoder's selected crops and order scores 46.39% (Table 9).

7 Conclusion and Limitations

Composing real sign crops according to text-corpus sequences supplies pre-training supervision without paired real-line images: under oracle localization, removing it costs 24–52 pp SER across three recognizers and three domains. Within SVTRv2*, the multi-sign line format matters more than attested sign order and co-occurrence. Target-tablet annotation mainly helps in-period, and cross-period and cross-collection recognition remain difficult. Given true sign boundaries, a per-sign classifier is better off-domain than SVTRv2*; the in-period full-set difference is inconclusive. With predicted localization, the exploratory system reduces whole-side SER from 68.92% to 38.05% versus the best original geometric baseline. For Assyriology, composed supervision could help produce searchable sign sequences for digitized tablets and align them with existing editions, subject to expert checking.

Limitations. Primary evaluation uses GT-localized, fixed-vocabulary stitches of MaiCuBeDa renderings; collection shifts vary site, language, and campaign, not imaging domain. The multi-period dictionary may expose styles of unseen tablets/texts (§5.1). Automatic evidence uses reused in-period data, conditional component selection, and a fixed-inventory reference excluding some annotated regions. Recognition of out-of-inventory signs and numeral notations, and evaluation under new imaging conditions, remain open.

AI USE STATEMENT

Generative AI tools were used to assist with code implementation, experiment orchestration, refinement of experimental designs, analysis and verification of results, literature search, and drafting of exposition. All AI-assisted code and analyses were reviewed and tested by the authors, every quantitative result reported here was produced by the authors' own experimental pipeline, and all research questions, claims, and conclusions are the authors' own. We take responsibility for the final content of this work.

ETHICS STATEMENT

We use the publicly released MaiCuBeDa Hilprecht (Mara & Homburg, 2023; Homburg et al., 2025) and MaiCuBeDa-HT (Ahlborn & Homburg, 2025) corpora (CC BY-SA 4.0) and CDLI texts (CDLI contributors, 2026); isolated-sign exemplars come from the eBL annotated sign-image dataset (Lewenstein et al., 2026) (CC BY-NC 4.0) and an Old Babylonian sign dataset (Smidt et al., 2026) (CC BY 4.0). No human subjects were involved; the work aims to accelerate scholarly access to cuneiform sources, and outputs are sign-name sequences that require expert review before use.

REPRODUCIBILITY STATEMENT

Locked plan. The primary-analysis protocol was fixed before any primary-protocol evaluation: a written analysis plan (2026-08-08) set the LR grids, epoch budgets, patience, seeds $\{42, 123, 456\}$, the DEV-only LR-selection rule, and the Holm family, and a locked evaluation plan (2026-08-13) recorded the cross-period manifest. These are internally generated artifacts whose hashes identify the exact versions used (Appendix C), not an externally time-stamped registration. **Text-disjoint retraining.** After the audit of §5.3 the Stage-A corpus was deduplicated, all three Stage-A models were retrained, and the Stage-B runs were repeated as described in §5. Training-run details, compute, artifact hashes, deviations from bitwise replay, and the HT manifest construction are in Appendices C and B. **Review materials.** An accompanying anonymous records-only supplement provides selected per-line records and portable scripts for the current Stage-A ablation, strict-subset SER, within-SVTRv2* attribution, converged-scratch control, and per-sign comparison, plus side-level records and scoring scripts for the automatic whole-side comparison and additional development controls (Appendix K). Its README maps each comparison to the supplied records, splits, label mappings, available configurations, and commands. Six of the 108 main-ablation run$\times$set records—the without-Stage-A CRNN N=25 runs for seeds 123/456 on three sets—remain explicit failure entries scored by the empty-output convention; no model output is synthesized for them. The package also supplies the Stage-A text manifest and an independent overlap check. It excludes raw images, model weights, and retired matched-line automatic diagnostics. Missing historical run metadata, including the per-sign baseline's gaps (Appendix J), remains identified as missing; packaging-time inventories do not reconstruct contemporaneous records. Third-party material retains its source attribution and applicable terms. Upon publication, we will release the broader code and analysis materials described in Appendix C.

REFERENCES

Lukas Ahlborn and Timo Homburg. MaiCuBeDa HT: Mainz cuneiform benchmark dataset for the Haft Tappeh collection. heiDATA, https://doi.org/10.11588/DATA/8TCR5C, 2025. Version 1.1, CC BY-SA 4.0.

Anonymous. End-to-end cuneiform tablet transcription with a spatially grounded vision-language model. In *Submitted to The Fifteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uYq3VIKeJ3. under review.

Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. What if we only use real datasets for scene text recognition? Toward scene text recognition with fewer labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3113–3122, 2021.

Raphael Baena, Syrine Kalleli, and Mathieu Aubry. General detection-based text line recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Bartosz Bogacz and Hubert Mara. Digital Assyriology—Advances in visual cuneiform analysis. *ACM Journal on Computing and Cultural Heritage*, 15(2):38:1–38:22, 2022. doi: 10.1145/3491239.

Bartosz Bogacz, Maximilian Klingmann, and Hubert Mara. Automating transliteration of cuneiform from parallel lines with sparse data. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 615–620, 2017. doi: 10.1109/ICDAR.2017.106.

Kai Brandenbusch, Eugen Rusakov, and Gernot A. Fink. Context aware generation of cuneiform signs. In *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 65–79. Springer, 2021. doi: 10.1007/978-3-030-86549-8_5.

CDLI contributors. Cuneiform Digital Library Initiative. https://cdli.earth, 2026. Online resource, accessed 2026.

Wentao Che, Esteban Garcés Arias, Asim Niaz, Andreas Bender, and Enrique Jiménez. Automated sign detection across the Electronic Babylonian Library: A large-scale dataset and end-to-end cuneiform OCR pipeline. *arXiv preprint arXiv:2606.22608*, 2026.

Danlu Chen, Aditi Agarwal, Taylor Berg-Kirkpatrick, and Jacobo Myerston. CuneiML: A cuneiform dataset for machine learning. *Journal of Open Humanities Data*, 9:30, pp. 1–9, 2023. doi: 10.5334/johd.151.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024.

Zijian Chen, Tingzhu Chen, Wenjun Zhang, and Guangtao Zhai. OBI-Bench: Can LMMs aid in study of ancient script on oracle bones? In *International Conference on Learning Representations (ICLR)*, 2025.

Yan Hon Michael Chung and Donghyeok Choi. Fine-tuning vision–language models as OCR systems for low-resource languages: A case study of Manchu. *Computational Humanities Research*, 2:e20, 2026. doi: 10.1017/chr.2026.10042.

Yunus Cobanoglu, Luis Sáenz, Ilya Khait, and Enrique Jiménez. Sign detection for cuneiform tablets. *it – Information Technology*, 66(1):28–38, 2024. doi: 10.1515/itit-2024-0028.

Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza. Benchmarking large language models for handwritten text recognition. *arXiv preprint arXiv:2503.15195*, 2025.

Cheng Cui, Yubo Zhang, Ting Sun, Xueqing Wang, Hongen Liu, Manhui Lin, Yue Zhang, Tingquan Gao, Changda Zhou, Jiaxuan Liu, Zelun Zhang, Jing Zhang, Jun Zhang, and Yi Liu. PP-OCRv5: A specialized 5M-parameter model rivaling billion-parameter vision-language models on OCR tasks. *arXiv preprint arXiv:2603.24373*, 2026.

Tobias Dencker, Pablo Klinkisch, Stefan M. Maul, and Björn Ommer. Deep learning of cuneiform sign detection with weak supervision using transliteration alignment. *PLOS ONE*, 15(12):e0243039, 2020. doi: 10.1371/journal.pone.0243039.

Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. SVTR: Scene text recognition with a single visual model. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 884–890, 2022.

Yongkun Du, Zhineng Chen, Hongtao Xie, Caiyan Jia, and Yu-Gang Jiang. SVTRv2: CTC beats encoder-decoder models in scene text recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20147–20156, 2025. doi: 10.1109/ICCV51701.2025.01874.

Shai Gordin, Morris Alper, Avital Romach, Luis Saenz Santos, Naama Yochai, and Roey Lalazar. CuReD: Deep learning optical character recognition for cuneiform text editions and legacy materials. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pp. 130–140, 2024. doi: 10.18653/v1/2024.ml4al-1.14.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 369–376, 2006. URL https://www.cs.toronto.edu/~graves/icml_2006.pdf.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2123/.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 690–696, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-2121/.

Timo Homburg, Lukas Ahlborn, Kai-Christian Bruhn, and Hubert Mara. The Mainz cuneiform benchmark dataset series: Sign annotations of 3D rendered tablets. *Journal of Open Archaeology Data*, 13:22, pp. 1–7, 2025. doi: 10.5334/joad.172.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026.

Antonia Karamolegkou, Nicolas Angleraud, Benoît Sagot, and Thibault Clérice. Reading or guessing? Visual grounding failures of vision-language models for OCR in ancient Greek editions. *arXiv preprint arXiv:2605.27750*, 2026.

Amir Hossein Kargaran, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. GlotOCR Bench: OCR models still struggle beyond a handful of Unicode scripts. *arXiv preprint arXiv:2604.12978*, 2026.

Maria A. Levchenko. Evaluating LLMs for historical document OCR: A methodological framework for digital humanities. In *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities*, pp. 75–85, 2025. URL https://aclanthology.org/2025.lm4dh-1.7/.

Or Lewenstein, Daniel López, Cyrill Dankwardt, Mays Fadhil Alrawi, Louisa Grill, Brian Mak, Albert Setälä, Fiammetta Gori, Aino Hätinen, Felix Rauchhaus, Zsombor Földi, and Enrique Jiménez. A large-scale dataset of annotated cuneiform sign images for digital palaeography. *Journal of Open Humanities Data*, 12:53, pp. 1–8, 2026. doi: 10.5334/johd.503.

Yunhao Liang, Ruixuan Ying, Bo Li, Hong Li, Kai Yan, Feiteng Fang, Qingwen Li, Min Yang, Satoshi Okamoto, Zhe Cui, and Shiwen Ni. Visual merit or linguistic crutch? A close look at DeepSeek-OCR. *arXiv preprint arXiv:2601.03714v2*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Hubert Mara and Timo Homburg. MaiCuBeDa Hilprecht – Mainz cuneiform benchmark dataset for the Hilprecht collection. heiDATA, https://doi.org/10.11588/DATA/QSNIQ2, 2023.

Rachel Mikulinsky, Morris Alper, Shai Gordin, Enrique Jiménez, Yoram Cohen, and Hadar Averbuch-Elor. ProtoSnap: Prototype alignment for cuneiform signs. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.00129.

PaddlePaddle Team. PaddleOCR-VL-1.6. https://huggingface.co/PaddlePaddle/PaddleOCR-VL-1.6, 2026. Model release, accessed 2026; technical report arXiv:2606.03264.

Vittorio Pippi, Konstantina Nikolaidou, Silvia Cascianelli, George Retsinas, Giorgos Sfikas, Rita Cucchiara, and Marcus Liwicki. Quo vadis handwritten text generation for handwritten text recognition? In *ICCV Workshops*, 2025. arXiv:2508.09936.

Utsav Poudel, Rasik Bhattarai, Siddhartha Pathak, Raghavendra Ramacharna, and Gaurav Jaswal. Zero-shot SAM2 segmentation and vision transformer-based recognition of Elamite cuneiform symbols from degraded tablet images. *arXiv preprint arXiv:2608.18544*, 2026.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Eugen Rusakov, Kai Brandenbusch, Denis Fisseler, Turna Somel, Gernot A. Fink, Frank Weichert, and Gerfrid G. W. Müller. Generating cuneiform signs with cycle-consistent adversarial networks. In *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing (HIP@ICDAR)*, pp. 19–24, 2019. doi: 10.1145/3352631.3352632.

Gowtham Senthil, K. Nandhakumar, and Gorthi Rama Krishna Sai Subrahmanyam. Handwritten Hindi word generation to enable few instance learning of Hindi documents. In *International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5. IEEE, 2020. doi: 10.1109/SPCOM50965.2020.9179634.

Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2017. doi: 10.1109/TPAMI.2016.2646371.

Alex Shonenkov, Denis Karachev, Maxim Novopoltsev, Mark Potanin, and Denis Dimitrov. StackMix and blot augmentations for handwritten text recognition. *arXiv preprint arXiv:2108.11667*, 2021.

Gustav Ryberg Smidt, Anna Goddeeris, Olivier Goossens, Katrien De Graef, Hendrik Hameeuw, and Eli Verwimp. Old Babylonian cuneiform signs from Akkadian documentary tablets. Zenodo, https://doi.org/10.5281/zenodo.18335468, 2026.

Mohamed Ali Souibgui, Ali Furkan Biten, Sounak Dey, Alicia Fornés, Yousri Kessentini, Lluís Gómez, Dimosthenis Karatzas, and Josep Lladós. One-shot compositional data generation for low resource handwritten text recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2563–2571, 2022. doi: 10.1109/WACV51458.2022.00262.

Ernst Stötzner, Timo Homburg, Jan Philipp Bullenkamp, and Hubert Mara. R-CNN based polygonal wedge detection learned from annotated 3D renderings and mapped photographs of open data cuneiform tablets. In *Eurographics Workshop on Graphics and Cultural Heritage (GCH)*, pp. 47–56, 2023a. doi: 10.2312/gch.20231157.

Ernst Stötzner, Timo Homburg, and Hubert Mara. CNN based cuneiform sign detection learned from annotated 3D renderings and mapped photographs with illumination augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1680–1688, 2023b.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *European Conference on Computer Vision (ECCV)*, 2022.

Eli Verwimp, Gustav Ryberg Smidt, Hendrik Hameeuw, and Katrien De Graef. Signs of the past, patterns of the present: On the automatic classification of Old Babylonian cuneiform signs. *arXiv preprint arXiv:2507.13959*, 2025.

Edward C. Williams, Grace Su, Sandra R. Schloen, Miller C. Prosser, Susanne Paulus, and Sanjay Krishnan. DeepScribe: Localization and classification of Elamite cuneiform signs via deep learning. *ACM Journal on Computing and Cultural Heritage*, 18(2):31:1–31:32, 2025. doi: 10.1145/3716850.

Appendix guide. Primary-result details cover overlap controls and strict subsets (§A, Table 5), data provenance (§B), training and compute (§C, Table 7), per-seed results (§E), period breakdowns (§F), and the per-sign baseline (§J). Appendix D gives the automatic pipeline, component budgets, and whole-side results. Appendix G gives the primary attribution study followed by a separate, stronger global-shuffle control. The VLM rank and image-sensitivity probes (§H) and natural-line diagnostic (§I) use earlier text-exposed checkpoints and their stated auxiliary metrics. CRNN stopping-condition and VLM DEV checks appear in Appendix K.

A Overlap Audit and Overlap-Controlled Subsets: Full Tables

Auxiliary metrics (definitions). **Exact-Line Accuracy** is the fraction of lines with $S_i+D_i+I_i = 0$, i.e., whose predicted sign sequence equals the reference. **LCS Recall** measures ordered sign recovery: $\sum_i \mathrm{LCS}(\hat{y}_i^V, y_i)/\sum_i |y_i|$. Here LCS is the longest common subsequence and $\hat{y}_i^V$ is the prediction projected to the 749-label vocabulary. Counts are pooled across lines. This metric gives partial credit but does not penalize insertions. For the N=100 with-Stage-A runs in Table 1, we compute $\sum_i |\hat{y}_i|/\sum_i |y_i|$ per seed and average over three seeds. Across the nine recognizer×set cells, this ratio ranges from 0.974 (CRNN, cross-collection) to 1.001 (SVTRv2*, cross-collection). There is no systematic over-generation, but insertions and deletions can offset each other in the length ratio. SER therefore remains the primary metric.

Table 4: Overlap between the text-disjoint Stage-A corpus and each evaluation set. The last column is the *strict*-subset fraction (no shared 4-gram *and* no verbatim match); the no-4-gram-only fractions are higher (in-period 72%). In-period percentages are from the annotation-audit script (671 lines); the SER-record pipeline counts 670 lines and reports its own subset sizes separately (§A). (Before deduplication: 99/100 in-period and 47/50 development texts were in Stage-A, with 56.2% and 42.4% verbatim lines. Deduplication removed 859 lines, 0.22%, from 146 texts of the 388,174 composed lines, leaving 387,315.)

Evaluation set	Tablets	Text in Stage-A	Verbatim lines	Strict subset
Development (Ur III; selection)	50	0/ 50	34.6%	49.8%
In-period (Ur III; reused)	100	0/100	48.6%	43.2%
Cross-period	192	0/191†	9.0%	84.4%
Cross-collection (HT)	315	0/282‡	1.2%	98.0%

†HS 1032 has no CDLI mapping. ‡315 manifest tablets contain 282 CDLI P-numbers; primary CTC scoring loads qualifying lines from 305 of these tablets.

Table 5: **Strict-subset SER ablation** (with vs. without Stage-A), 3-seed corpus-SER mean over strict lines only. Strict = exact-unseen ∩ no-shared-4-gram against the Stage-A corpus, computed on each model's loaded SER records. The PADDLEOCR-VL evaluator loads 718/1,254/1,408 lines across sets, versus 670/1,243/1,320 for CTC, so its strict n is larger. These are the same line sets as in Table 1. ΔSER = $\text{SER}_{\text{no}} - \text{SER}_{\text{yes}}$; Exact = exact-line accuracy (%). The strict removal cost reproduces the all-line Δ of Table 1 within $\sim$2 pp on all nine N=100 contrasts (max 1.9 pp); at N=25 it departs more, by up to $\sim$5.8 pp in-period (SVTRv2*) and less off-domain, as the strict subset drops the easier formulaic lines. Source: `compute_strict_ablation_ser.py` (records only; no retrain).

				SER (%↓)		Exact (%↑)	
Set	Model	N	n	with	no	with	no
In-period	SVTRv2*	25	291	33.0	95.1	22.5	0.0
	SVTRv2*	100	291	25.4	76.2	34.9	0.6
	CRNN	25	291	38.3	100.4	16.5	0.0
	CRNN	100	291	29.4	76.1	26.6	0.1
	PADDLEOCR-VL	25	386	31.9	78.7	26.9	2.4
	PADDLEOCR-VL	100	386	20.5	50.4	48.7	12.4
Cross-period	SVTRv2*	25	1050	39.3	98.2	14.3	0.0
	SVTRv2*	100	1050	40.0	91.7	14.4	0.1
	CRNN	25	1050	46.2	100.3	8.1	0.0
	CRNN	100	1050	43.9	91.5	10.4	0.0
	PADDLEOCR-VL	25	1071	40.9	94.0	14.4	0.0
	PADDLEOCR-VL	100	1071	43.0	77.3	12.0	0.7
Cross-collection	SVTRv2*	25	1293	70.0	97.8	2.1	0.0
	SVTRv2*	100	1293	71.4	98.0	1.7	0.0
	CRNN	25	1293	72.7	101.8	1.1	0.0
	CRNN	100	1293	73.8	97.5	1.1	0.0
	PADDLEOCR-VL	25	1385	65.2	97.7	3.0	0.0
	PADDLEOCR-VL	100	1385	66.3	90.1	2.6	0.0

Table 6: **Auxiliary LCS-Recall companion to the SER text-exposure audit** (§5.3; the primary strict-subset SER result is in Table 5). Overlap-controlled ablation at N=100 (3-seed mean): LCS Recall and exact-line accuracy (%) with vs. without Stage-A on the *strict* subset of lines that both share no 4-gram with any Stage-A line *and* have no exact verbatim match in it (the exact-unseen condition removes short verbatim lines that a 4-gram filter alone would admit; §5.3), next to the all-line LCS Recall gap. n = strict-subset lines (subsets defined on the CTC token mapping; on HT the PADDLEOCR-VL also scores the two subset lines the CTC evaluator fails to load, hence n=1,295); "all lines" = the all-line LCS Recall gap in this auxiliary metric. N=25 strict LCS Recall gaps: in-period +56.1/+59.6/+41.4, cross-period +54.0/+52.4/+47.8, cross-collection +26.9/+27.7/+27.9 pp (SVTRv2*/CRNN/PADDLEOCR-VL).

			with Stage-A		no Stage-A		ΔLCS (pp)	
Set	Model	n	LCS	Exact	LCS	Exact	strict	all lines
In-period	SVTRv2*	290	75.3	35.1	28.5	0.6	+46.8	+46.7
	CRNN	290	71.6	26.7	28.5	0.1	+43.1	+44.0
	PADDLEOCR-VL	290	78.3	44.7	51.4	11.3	+26.9	+26.7
Cross-period	SVTRv2*	1049	60.4	14.5	13.5	0.1	+46.9	+47.7
	CRNN	1049	56.8	10.5	11.9	0.0	+44.9	+45.8
	PADDLEOCR-VL	1049	57.2	11.5	24.8	0.6	+32.4	+33.0
Cross-collection	SVTRv2*	1293	30.0	1.7	5.9	0.0	+24.1	+24.1
	CRNN	1293	27.7	1.1	5.3	0.0	+22.4	+22.5
	PADDLEOCR-VL	1295	34.5	2.4	12.5	0.0	+22.0	+22.0

Strict-subset definitions and counts. Two strict computations appear in this paper, differing only in tokenization and line universe, not in rule.

(i) Annotation audit. The audit (`audit_dedup.py` → `AUDIT_DEDUP.json`) normalizes sign tokens and parses 671 in-period lines (≥3 signs); 290 (43.2%) are strict—no verbatim match *and* no shared 4-gram with the Stage-A corpus. This computation underlies the auxiliary LCS view (Table 6, n=290).

(ii) SER evaluation records. The SER strict ablation (Table 5) and the parent-matched contrast (§6.2) use the per-line SER records with raw sign names. Of the 670 loaded in-period CTC lines, 480 (72%) share no 4-gram, 345 (51%) never recur verbatim, and 291 satisfy both conditions. The no-4-gram set includes 189 short (<4-sign) verbatim recurrences. The PADDLEOCR-VL evaluator loads more lines (718 in-period), giving a larger strict subset (Table 5).

Both computations draw strict membership from the same Stage-A corpus (93,440 *distinct* sign sequences after deduplication, from 387,315 training instances); its 285 sign types cover 100/99.95/96.7% of in-period/cross-period/HT evaluation tokens. The CTC subset sizes are 290/1,049/1,293 (annotation audit, ≥3-sign line universe) vs. 291/1,050/1,293 (SER-record ablation, all loaded CTC lines), the in-period/cross-period differences coming from tokenization and the line universe; the PADDLEOCR-VL SER-record subsets are larger on its own line universe (Table 5). Each analysis retains its own filtering pipeline and reports its subset size. The strict-subset SER ablation (Table 5) keys strict membership on the unique per-line identifier (tablet-side stem and line id); the main table (Table 1) and the parent-matched contrast (Figure 2) do not use that subset key—they aggregate the per-line records directly. Verbatim recurrence is driven by administrative formulae: the receipt formula `ŠU BA TI` alone is 58 of the 326 recurring in-period lines and occurs 10,762 times in the Stage-A corpus.

Auxiliary LCS view (Table 6). The strict-subset LCS gaps track the SER conclusion: within 1 pp of the all-line LCS Recall gaps at N=100 and within 5 pp at N=25. On the strict subset the no-Stage-A in-period LCS Recall is 28.5/28.5/51.4% (SVTRv2*/CRNN/PADDLEOCR-VL)—low but not zero—against with-Stage-A 75.3/71.6/78.3%. With-Stage-A absolute in-period LCS Recall is up to 6 pp higher on the full set than on the strict subset, a descriptive difference consistent with formula recurrence rather than a measured causal contribution.

Earlier and text-disjoint checkpoint comparison. We compare the earlier text-exposed checkpoints against the text-disjoint retrain under the identical in-period protocol. In auxiliary LCS Recall, the differences are mixed and small: SVTRv2* scores 0.3–0.7 pp *lower* (71.9/78.3 vs. 72.2/79.0 at N=25/100), CRNN 1.9–4.0 pp higher (71.1/77.9 vs. 67.1/76.0), and CTC N=0 moves by ≤2 pp. The comparison absorbs Stage-A run variation and, for CRNN, a different DEV-selected LR, so it describes cross-generation differences (≤4 pp) and does not isolate a text-exposure effect. The VLM's cells move similarly (N=25/100: 73.7/83.4 text-exposed vs. 74.7/82.7 text-disjoint); its N=0 change is −2.0 pp.

B CROSS-COLLECTION (MAICUBEDA-HT) MANIFEST AND CROP-DICTIONARY PROVENANCE

MaiCuBeDa-HT (Ahlborn & Homburg, 2025) releases, for the MSII rendering of each tablet side, W3C sign annotations (transliteration, line number, character index, polygon or box, CDLI P-number link) and cropped images of the annotated lines. We checked each released line image against the final bounding box of its cumulative line annotation, in the sign boxes' pixel frame. Of 2,842 lines, 2,832 match to ≤1 px; we drop the 10 mismatches and 36 lines without an image. We then express each sign box relative to its line image and use the same 4-px-padded crop and stitching function as for Ur III. Padding is clipped at the line-image border. Signs are ordered by character index; tokens outside the 749-label inventory (18.5%) are dropped from image and label under each family's token mapping, exactly as on the Ur III sets (the VLM mapping additionally maps numeral signs such as `1(disz)`, hence 1,408 vs. 1,322 manifest lines). Two of the 1,322 CTC manifest lines fail to load, leaving 1,320 scored lines; this is separate from the mapping difference. The frozen manifest (315 tablets, 282 CDLI P-numbers, none of which occurs in the Stage-A corpus) is hashed below. The

released per-sign `char/` images are fixed 250×250 context windows rather than sign boxes and were not used.

Crop-dictionary provenance and leakage audits. The 199,243-image dictionary comprises 146,296 crops from the eBL annotated sign-image dataset (Lewenstein et al., 2026), 35,947 from an expert-annotated Old Babylonian sign dataset (Smidt et al., 2026) (photorealistic renderings), and 17,000 augmentation variants of rare-sign crops. The variants are produced by `build_clean_dict.py` for under-represented sign types by resampling existing real crops under rotation, scaling, brightness/contrast, blur, and additive noise—standard geometric/photometric augmentation, not a skeleton-alignment or generative model; they add sample variety for existing classes only. Where the eBL metadata (Lewenstein et al., 2026) records a script period, the labels are: Unknown 144,221; Neo-Babylonian 494; Neo-Assyrian 376; Middle Assyrian 254; Old Assyrian 235; Middle Babylonian 206; Ur III 128; Parthian 108; Late Babylonian 100; Hellenistic 90; Persian 36; Presargonic 27; Fara 16; Sargonic 5. The OB source (Smidt et al., 2026) is entirely Old Babylonian; augmented crops inherit unlabeled origins. For the tablet-overlap audit, we parsed 10,780 unique tablet/museum IDs from crop filenames. We compared these against all MaiCuBeDa tablets used here: 214 training, 50 development, 100 in-period, and 192 cross-period tablets (556 unique HS numbers), plus the 315-tablet HT manifest. Both numeric HS-number and normalized full-string matching find *no overlap*. The dictionary includes 308 other Hilprecht tablets; the separation is at tablet level. The three empty inventory entries (`du`, `šar`$_2$, `600`) held only MaiCuBeDa Hilprecht crops, which were removed when the dictionary was built; as noted in §4.1, token normalization never requests them, and none occurs in the Stage-A manifest or any evaluation reference. The sign detector's training data (450 train / 50 val tablets) shares no tablet with the 100 in-period test tablets, which are exactly its held-out split; the detector pool does, however, contain the Stage-B pool, DEV, and cross-period tablets—the automated evaluation (§6.3) therefore runs only on the in-period test tablets.

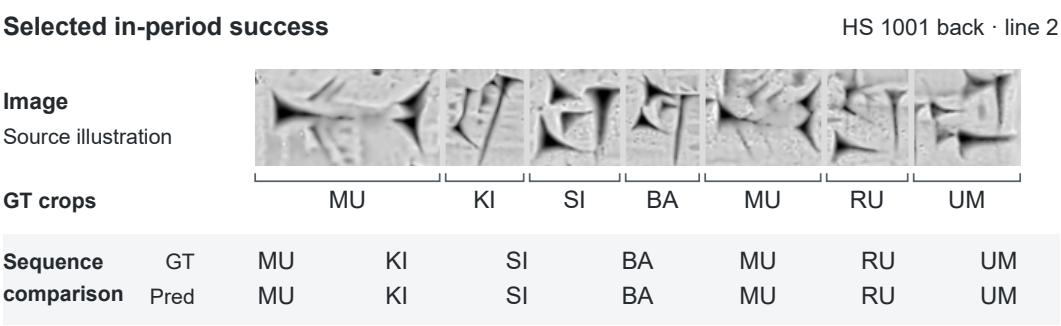
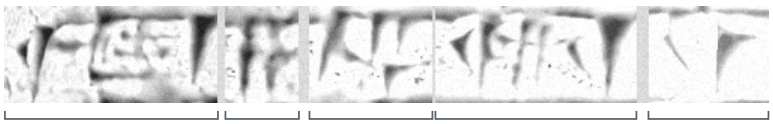


Figure 3: **Selected successful in-period example** (text-disjoint SVTRv2*, $N=100$, seed 42; HS 1001 back, line 2; strict subset). GT and prediction match. The GT-crop stitch illustrates the source, not the inference tensor. Brackets locate GT crops; sequence rows show edit alignment, not image positions. Rates: Table 1; strict subset: §5.3; oracle/automatic and cross-collection examples: Figure 4.

Paired in-period line HS 193 front · line 4

a Oracle · GT localization



Sequence comparison	GT	U_3	ZA	HA	AR	KI
	Pred	U_3	ZA	HA	TA	KI

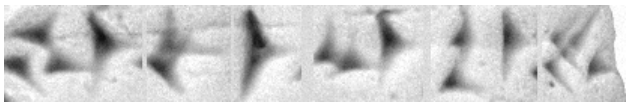
b Independent SVTRv2* on base-selected crops



Sequence comparison	GT	U_3	ZA	HA	AR	KI
	Pred	IGI	ZA	HA	–	KI

Separate cross-collection example HT 05-13-113 front · line 5

c Oracle · GT localization



Sequence comparison	GT	BA	NU	DIŠ	AN	GIŠ	BU
	Pred	MUNUS	NU	ŠI	AN	ŠU	$ŠUM_2$

Brackets identify GT crops or numbered automatic crops.

Sequence rows show edit alignment. Underline = error; – = gap.

Figure 4: **Localization and cross-collection examples** (text-disjoint SVTRv2*, N=100, seed 42). **(a,b)** HS 193 front, line 4: oracle input and the exact automatic model input formed from four base-selected boxes. Panel **(b)** shows independent SVTRv2* recognition on base-selected crops, not the full system. **(c)** An oracle-localized HT example, HT 05-13-113 front, line 5. Brackets identify GT crops in (a,c) and numbered automatic crops in (b). GT is used only for display and evaluation; sequence rows show sign-token edit alignment, not image positions (ties: diagonal, deletion, insertion). These selected cases do not represent overall rates (Tables 1 and 9).

C Training and Compute Details

Evaluation determinism. Within each matched comparison and input protocol, checkpoints are evaluated on fixed images (fixed augmentation seed), and CTC decoding is verified against the training-time evaluator on CPU for each evaluation batch.

Reproduction entry points. The anonymous records supplement's `verify_current.py` is the single entry point for primary and strict-subset SER, attribution, converged-scratch and per-sign

comparisons, development controls, and the unified automatic comparison. It checks hashes and recomputes edit distances, LCS, aggregates, and paired tablet intervals from saved records using Python and NumPy. `SOURCE_PROVENANCE.json` maps the supplied records to their sources; `SHA256SUMS.txt` binds the package files. This verifies saved outputs, without reproducing neural training or inference.

Training-run disclosures. The VLM's two-stage execution is detailed in §C.1. Stage-B VLM runs use gradient checkpointing and a per-process CUDA memory cap so that two runs share one GPU; both are memory-only (the loss is bitwise identical to the plain path; gradients agree to within GPU nondeterminism).

Artifact hashes. Analysis plan SHA-256 `1f80a787...`; locked evaluation plan `27e7c053...`; cross-period manifest `e09d11d1...`; deduplicated Stage-A manifest `5450bd27...`; HT manifest `f6e632bc....` The file name, purpose, and full 64-character hash of each of these five artifacts are listed in the supplement's `ARTIFACT_INDEX.json`. These hashes identify the original internal artifacts.

Compute. All runs used a single NVIDIA RTX 5090 (32 GB; CUDA 12.8). A primary Stage-B fine-tune takes minutes for the CTC models (9 and 17 min at N=100, seed 42, for SVTRv2* and CRNN) and $\sim$5.6 h for the VLM (LoRA; run two per GPU under a 0.36 per-process memory fraction). These are measured per-run times from the training logs, not a whole-project total; a complete accounting of Stage-A pre-training and detector-training compute was not preserved and is not reported.

Table 7: **Training configuration per run type**, read from the per-run records, training scripts, logs, and LoRA adapter configs (not generic defaults; sources and VLM Stage-A boundaries are given in the text below). Optimizer is AdamW throughout; weight decay is 10^{-4} for the CTC runs, while the VLM run record stores none. Gradient-norm clip is 5.0 (CTC) / 1.0 (VLM). CTC line images are height-normalized to 64 px (aspect preserved) with worker-level augmentation (inter-sign gaps, noise). "Budget" is the epoch cap, "Stopping" gives the DEV early-stopping patience, and "Ckpt sel." is the checkpoint-selection rule; LRs marked "DEV-sel" are chosen per model$\times N$ on DEV, while Stage-A pre-training LRs are the script constants. All Stage-B/ladder/scratch runs use 3 seeds {42,123,456}; Stage-A pre-training uses a single seed.

Run type	LR	Scheduler	Batch	Budget	Stopping	Ckpt sel.
Stage-A SVTRv2*	3×10^{-4}	warmup 1k, cosine (0.05 floor)	32	30 ep	patience 5	best Val-SER
Stage-A CRNN	3×10^{-4}	cosine (T=30)	32	30 ep	patience 5	best Val-SER
Stage-B CTC (SVTRv2*, CRNN)	DEV-sel	cosine (T=100)	8	$\leq$100 ep	patience 15	best DEV-LCS Recall
Stage-B ladder / parent-matched (SVTRv2*)	DEV-sel	cosine (T=100)	8	$\leq$100 ep	patience 15	best DEV-LCS Recall
Scratch (SVTRv2*, no St.-A)	3×10^{-4}	cosine (T=300)	8	300 ep	off (converged)	best DEV-LCS Recall
Stage-B VLM (LoRA r16/α32)	DEV-sel	cosine	8$\times$1	40 ep	patience 10	best val LCS Recall

Configuration sources and boundaries. The CTC values are read from each run's `result.json`; the VLM Stage-B values from its `config.json` (batch 8, accumulation 1, 40 epochs, cosine), with early stopping at patience 10 read from `train.log` and best-val-LCS-Recall selection from `best_meta.json` (its `config.json` stores neither). The DEV-selected LR grids are $\{5\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$ (CTC) and $\{5\times10^{-5}, 10^{-4}, 2\times10^{-4}\}$ (VLM), selected by development-set LCS Recall at seed 42; the VLM reused its original DEV-selected values after the text-disjoint retraining, while the CTC grids were fully re-run. The ladder/parent-matched rows are the Stage-B fine-tunes on those arms; the ladder *Stage-A* pre-training runs are fixed at 30 epochs / final checkpoint

(Appendix G), not the DEV-best rule of the table. The VLM LoRA adapter is $r16/\alpha32$, dropout 0.05, q/k/v/o projections (5.49 M trainable parameters, 0.60%), base `PaddlePaddle/PaddleOCR-VL-1.6`.

C.1 Recognizer definitions and serialization

CTC architectures. The retained implementations use grayscale $64 \times W$ tensors in $[0, 1]$. Both convolutional stems have four 3×3, stride-1, padding-1 convolutions with channels $1 \to 64 \to 128 \to 256 \to 256$, each followed by batch normalization, ReLU, and max pooling. Pool sizes/strides are $(2, 2), (2, 2), (2, 1), (2, 1)$, yielding height 4 and width $\lfloor W/4 \rfloor$. CRNN averages over height, then applies two bidirectional LSTM layers with 256 hidden units per direction and inter-layer dropout 0.2; a $512 \to 750$ linear head produces CTC logits (3,975,534 trainable parameters). For SVTRv2*, a learned query attends over the four height positions independently at each column (256 dimensions, four heads), followed by layer normalization. Sinusoidal positional encoding precedes six residual mixing/FFN blocks in the order `LLGLLG`: L is a kernel-3 depthwise 1D convolution with GELU; G is four-head self-attention. Both use pre-layer normalization, and each FFN is $256 \to 512 \to 256$ with GELU and dropout 0.1; global-attention dropout is also 0.1. Final layer normalization and a $256 \to 750$ head give 3,532,398 parameters. There is no semantic-guidance module.

Cropping, batching, and decoding. Real-line inputs use 4-pixel-padded sign boxes, Lanczos height resizing, and corner-median background normalization to intensity 220 (invert first if the estimated background is below 80), followed by 2–8-pixel gaps and Gaussian noise with $\sigma = 6$. CTC minibatches are right-padded with zero. Stage-A CTC losses use each image's $\lfloor W/4 \rfloor$ valid length; Stage-A SVTRv2* also masks padded positions in global attention. The retained Stage-B implementations differ: SVTRv2* uses valid CTC lengths but passes no attention padding mask, whereas CRNN uses the full padded output length for each batch member. These conventions are shared by the with/without-Stage-A arms within each model. Primary CTC loss uses mean reduction and `zero_infinity=True`; greedy decoding collapses consecutive repeated classes, then removes blank. The separately controlled attribution renderer and parent-weighted objective are specified in Appendix G.

Inventory and canonical mapping. Class order is the sorted 747 dictionary keys followed by `X` and `lacuna`; blank has index 749. Of those keys, 744 have crop images. The portable supplement's `core/labels/inventory.json` and `token_mapping.py` preserve the class order and alias map. Mapping handles bracketed reconstructions as `lacuna`, illegible `X`, editorial markers, ASCII transliteration spellings, numeric subscripts, case, and aliases before inventory lookup. Both mappings reject pure numeral strings; CTC additionally rejects numeral-prefix expressions such as `1(disz)`, whereas the VLM can retain them when mapped to an inventory entry. The VLM also rejects normalized strings longer than 24 characters. Thus the model families have different scored line universes.

VLM prompting and output. The model and its processor/tokenizer are loaded from `PaddlePaddle/PaddleOCR-VL-1.6`. A user message contains the RGB image and the following fixed prompt, also used outside Ur III:

> Transliterate this Ur III Sumerian cuneiform line. Output only space-separated sign names (e.g. BA AN KI). Do not explain. Do not translate. Do not output natural language.

The processor's chat template supplies the generation prefix. Training appends tokenized, space-separated sign names and EOS; prompt positions have loss label -100, so only response tokens contribute. Oracle evaluation uses `do_sample=False` and at most 64 new model-token IDs. Its historical code decodes `output[prompt_length:-1]`, unconditionally omitting the last generated ID, then applies the saved parser: remove recognized response prefixes, split at whitespace or `-/,;`, normalize signs, and retain only inventory entries. Oracle SER is computed after this parsing. Automatic evaluation scores full outputs (Appendix D).

Stage-A records and reproduction limits. The CTC logs record 347,322 training and 39,993 validation pairs, split by P-number from the 387,315-pair corpus; selected epochs are 30 (SVTRv2*) and 23 (CRNN). The VLM's initial executed configuration has 126,000 training records, batch 8, seed 42, LR 2×10^{-4}, and a three-epoch schedule. The selected lineage continues the saved step-42,000 adapter for 5,250 batch-8 updates on 42,000 records, with fresh AdamW/cosine states at

2.3×10^{-5}; this is not uninterrupted replay of that initial schedule. No per-run sampled-ID manifest or optimizer/RNG continuation state was retained; the run records do not pin a processor/tokenizer revision. VLM weight decay 0.01 is specified by the retained training code but not its run configuration. The records supplement supplies saved outputs, label mappings, split identifiers, and a minimal Stage-A text manifest for independent overlap checks; it excludes images, weights, and training/inference runtimes. The missing run states prevent exact replay of historical training.

D Automatic Whole-Side Recognition

D.1 Inputs, supervision, and frozen components

The proposed modular pipeline maps a complete tablet-side image and its surface type (front, back, or edge) to an ordered sequence of canonical sign labels. Reference boxes, reference lines, and reference text are not inputs. Table 8 separates the supervision of its components. The detector uses 450 training and 50 validation tablets; these contain, respectively, 43 and 7 of the recognizer DEV tablets, but none of the 100 evaluation tablets. The automatic experiment is therefore restricted to the in-period evaluation set: the detector pool also contains cross-period tablets.

Table 8: **Supervision and selection of automatic components.** Tablet counts belong to individual components and are not additive: the $N{=}100$ recognizer set is contained in the $N{=}214$ classifier pool. Evaluation tablets are separate from adaptation and detector training/validation tablets. DEV was seen by the detector and used for component/decoder selection. Inference takes an image and surface type, without reference boxes, lines, or text.

Component	Direct supervision	Selection and inference role
YOLO26n	450 training and 50 validation tablets with sign boxes	43 DEV tablets in training; 7 in validation. All routes share weights; geometric and base-selected routes use different detection settings.
DeiT3	Dictionary classification pre-training; adaptation on 214 tablets (7,582 usable characters from 210 tablets)	Seeds 123/42/456; the same DEV checkpoint-selection rule. Seed 123 supplies base-selected controls; all three run the final pipeline. Absent from geometric baselines.
Six-gram LM	332,876 deduplicated CDLI text lines; 71,929 P-numbers	Excludes 341 DEV/evaluation-associated P-numbers. Scores base-decoder/final candidates; absent from geometric baselines.
Sequence recognizers	Compositional Stage-A; $N{=}100$ adaptation, seeds 42/123/456	Same DEV-selected checkpoints decode geometric or base-selected inputs without DeiT3 labels.
Candidate verifier	Reuses $N{=}100$ SVTRv2*, seeds 42/123/456; no extra adaptation tablets	Final pipeline only: full CTC probability on each candidate's own stitch; $\lambda = 0.25$ selected using historical DEV ignore scoring.

YOLO26n (Jocher et al., 2026) produces proposals at confidence 0.05, NMS IoU 0.90, image size 960, and at most 1,000 detections. Nonfinite or nonpositive-area boxes are rejected and numeric duplicates collapsed. A fixed detector calibrator supplies proposal probabilities. Geometric line hypotheses and a conservative, training-derived column rule determine output order. For detected two-column layouts, the supplied surface type gives left-to-right column order on fronts and right-to-left order on backs; for other surfaces, including edges, geometric row order is retained. These rules do not read reference annotations.

The DeiT3-medium classifier (Touvron et al., 2022) uses dictionary-classification pre-training, followed by adaptation on the 214-tablet pool (seeds 123/42/456). Of these tablets, 210 contribute 7,582 usable in-inventory characters, including characters without a line label. AdamW uses learning rate 10^{-5}, weight decay 10^{-4}, batch size 32, and a 30-epoch cosine schedule and patience 8; strict-improvement DEV LCS selection retains epochs 26, 25, 20, respectively. The classifier uses RGB sign crops with zero extra padding at inference and the same 749-label inventory. This component does not use synthetic-line pre-training. Predicted signs retain their supporting detector crops.

D.2 ORIGINAL GEOMETRIC-ROUTE BASELINES

The original baselines use YOLO proposals, geometric line grouping, and the text-disjoint, with-Stage-A $N{=}100$ CRNN, SVTRv2*, or PADDLEOCR-VL checkpoints (seeds 42/123/456). The detector weights are identical to those of the full system, but the original settings are confidence 0.20 followed by a 0.25 filter, image size 640, NMS IoU 0.70, and at most 300 detections. Grouping sorts boxes by center-y, splitting at successive gaps above $\max(0.5\,\text{median box height}, 8)$ pixels, and orders boxes by center-x within a group. There is no DeiT3-based crop selection or language-model reranking.

The completed baseline evaluation retains all 198 sides and all 1,106 predicted clusters, without reference-line matching. The original minimum of three boxes leaves 183 clusters with empty outputs; the other 923 are decoded. A shared seed-42 render cache uses the original stitching recipe. Previously matched nonempty box groups are reproduced (637 groups), but expanding traversal to all clusters changes random draws: this is a protocol-aligned re-evaluation, not a bitwise replay of historical matched-line scores. CTC decoding is greedy. The original VLM has a 128-new-token limit, compared with 64 on base-selected crops. One original VLM output (seed 123) reaches its limit and is retained; full continuations and unknown tokens are scored. These settings and the original DEV-selected checkpoints were fixed before the completed evaluation; the detector threshold had previously been examined on this historically reused evaluation set. The comparison is consequently not a new blind test or a controlled estimate of changing only the input front end.

D.3 INDEPENDENT RECOGNITION ON BASE-SELECTED CROPS

The independent CRNN, SVTRv2*, and PADDLEOCR-VL comparisons reuse the text-disjoint, with-Stage-A $N{=}100$ checkpoints for seeds 42/123/456. They receive the same crops and order selected by the *base* per-sign decoder, not by the candidate verifier, and do not receive DeiT3's sign labels. Stitches use height 64, crop padding 4, gaps of 2–8 pixels, the original noise recipe, and a deterministic seed per proposal path. Input hashes agree across all nine runs; the VLM additionally converts grayscale to RGB.

The independent recognizers process all 95 DEV sides (50 tablets) and 198 evaluation sides (100 tablets). Their 696 and 1,186 proposed rows, including 18 and 78 empty rows, are retained. No output rows are dropped by matching to reference lines. CTC recognizers retain greedy decoding. VLM outputs retain unknown tokens and trailing content after canonical sign-name normalization; none of the three evaluation runs reaches the 64-new-token generation limit. Oracle-line results use the vocabulary-filtered protocol in Appendix C.1.

D.4 CANDIDATE CONSTRUCTION AND SEQUENCE VERIFICATION

The base pool combines the top five visual classes, beam size 64, and up to 16 hypotheses per generator, using 17 fixed weight vectors with either no language model or a six-gram modified Kneser–Ney model, queried with KenLM (Heafield, 2011; Heafield et al., 2013). The LM uses the text-disjoint CDLI Ur III corpus (CDLI contributors, 2026): deduplicating exact sequences within each P-number leaves 332,876 lines and 1,528,710 tokens from 71,929 P-numbers. Sources exclude all 341 P-numbers linked to DEV, the in-period evaluation set, or the cross-period test; Stage-B training texts remain eligible. Exact strings can still recur across distinct tablets.

An empty candidate is always available. Candidate generation uses $T_{\text{gen}} = 1.55$; reranking recomputes visual log probabilities from raw logits at $T_{\text{rank}} = 3$, without changing the pool. The base score is

$$\begin{aligned} s_0(h) = \sum_{t \in h} \log p_{T_{\text{rank}}}(t) + \lambda_d D(h) + \lambda_\ell \log P_{\text{LM}}(h) \\ + \lambda_o O(h) + b|h| + a\,\mathbf{1}[|h| > 0], \end{aligned} \tag{1}$$

where D sums calibrated detector log odds, O penalizes overlapping successive proposals, and λ_d, λ_ℓ, λ_o, b, and a are the DEV-selected base-decoder weights.

The adopted extension uses an existing compositionally pre-trained $N{=}100$ SVTRv2* (seeds 42/123/456). Its adaptation tablets are a subset of the classifier's 214-tablet pool. Each candidate h is scored on a stitch x_h of its own selected detector crops, rather than the shared base-selected

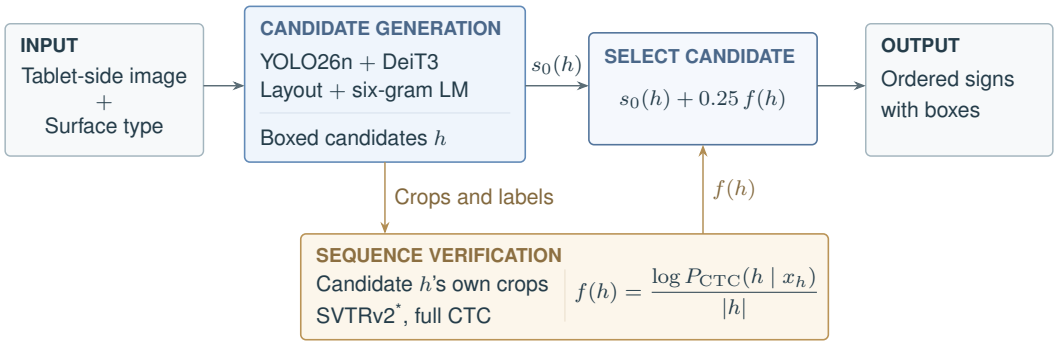


Figure 5: **Modular end-to-end recognition with sequence verification.** YOLO uses 450 training/50 validation tablets, DeiT3 a 214-tablet adaptation pool, and the verifier an included 100-tablet subset. The verifier scores each candidate's labels on its own crops with full CTC probability; it creates no boxes or reference alignment. The weight 0.25 was selected using historical DEV ignore scoring and frozen; final evaluation retains all outputs. Empty candidates have $f(h) = 0$. Inference takes only a side image and surface type. Independent recognizers instead share the selected crops without classifier labels (Table 9).

stitch used by the independent recognizers:

$$s(h) = s_0(h) + \lambda f(h), \qquad f(h) = \frac{\log P_{\mathrm{CTC}}(h \mid x_h)}{|h|}. \tag{2}$$

The feature is zero for an empty candidate. We use the full CTC sequence probability (Graves et al., 2006), not the greedy-path probability. Scoring neither creates boxes nor aligns outputs to reference text or CTC emission times. Figure 5 shows this inference path.

Only $\lambda \in \{0, 0.1, 0.25, 0.5, 1, 2, 4\}$ is selected for this extension. In each of the existing five DEV tablet folds, base-decoder weights remain fold-specific and λ is selected on the other 40 tablets. Under the *historical annotated-ignore DEV objective*, conditional out-of-fold errors fall from 534 to 529 over 1,769 reference signs. The all-DEV choice is $\lambda = 0.25$. With the existing 300-epoch, without-Stage-A N=100 SVTRv2* control, the same procedure selects $\lambda = 0$ in every fold and on all DEV. This development check is conditional on the detector's exposure to DEV and the DEV-selected recognition checkpoints. The weight remains fixed when evaluation changes to full-output scoring.

D.5 Reference construction, scoring, and uncertainty

All systems use one fixed reference on the same 100 tablets and 198 sides: 3,280 in-inventory signs, with 903 other annotated regions excluded because the fixed mapper does not assign their readings to the 749-label inventory (387 are numeral notations; the rest are readings such as *babbar* and *suen*). Every parseable character accepted by the fixed mapper contributes, including short rows and characters without a `Line` field. Within each logical `(Column, Line)` group, complete, positive, unique `Charindex` values determine order: a logical line can wrap across physical rows, so sorting all characters by x is insufficient (Homburg et al., 2025). Ambiguous indices and missing-line assignments use the frozen annotation-geometry fallback while preserving indexed sequences. Numeric column labels determine reference column order; logical groups within a column follow geometric vertical order. These reference-construction rules are not inference inputs.

Predicted rows are concatenated in their predicted order, retaining every output token. There is no reference-row matching, spatial ignore masking, or deletion of outputs near excluded annotations. SER is $\sum_{\text{sides}}(S + D + I)/\sum_{\text{sides}}|Y|$, where $|Y|$ is the length of a side's reference. Outputs corresponding to the 903 excluded regions can therefore incur errors, even when they depict real signs. This is a fixed closed-inventory task, not a full-inventory scholarly transcription score.

Intervals use 10,000 percentile bootstrap resamples of whole tablets, seed 20260923, with one shared sampling matrix for all systems and seeds. Each row reports the arithmetic mean of three corpus-level recognition-seed SERs; per-seed values are in Table 10. For the complete pipeline and its ablation, the

fixed classifier/verifier seed pairs are $(123, 42)$, $(42, 123)$, and $(456, 456)$; each classifier regenerates its own candidates. The two added classifiers use the original training recipe and DEV selection rule. Both Stage-A initializations, detector, LM, and decoding parameters remain fixed; there is no decoder retuning. These are recognition-seed repeats, not independent retraining of every system component. Table 11 reports all three pairs and their sample standard deviations. The intervals condition on these models, excluding retraining and selection uncertainty. Paired differences are computed at full precision before rounding.

D.6 Overall gains, stronger controls, and module ablation

The main comparison (Table 3) separates the original geometric baselines from the final pipeline. Full-system SER reductions are 34.17 pp relative to CRNN (paired 95% CI $[31.38, 37.08]$), 32.73 pp relative to SVTRv2* (paired 95% CI $[30.08, 35.61]$), and 30.87 pp relative to PaddleOCR-VL (paired 95% CI $[28.10, 33.88]$). The main text uses the strongest of these original baselines, PaddleOCR-VL, to summarize the overall gain; all three are reported.

Table 9: **Strong controls using base-selected crops and order.** All systems use the same full-output scoring target as Table 3. The three recognizers receive identical base-selected crop images and order, without classifier labels, and report three-seed means. These inputs come from the original classifier seed 123; the full system averages three recognition-seed pairs. These are complete-system comparisons; 95% CIs condition on fixed models and budgets differ (Table 8).

System	Runs	SER (%↓)	95% CI
CRNN	3	55.81	[52.68, 59.39]
SVTRv2*	3	46.39	[43.92, 49.02]
PaddleOCR-VL	3	47.01	[44.20, 50.18]
Full system	3	**38.05**	[35.82, 40.43]

As stronger controls, independent recognizers decode crops and order already selected by the original seed-123 base per-sign decoder. The full system reduces SER by 8.34 pp relative to SVTRv2*, the lowest observed SER among these three controls (paired 95% CI $[7.05, 9.68]$; Table 9). This comparison uses an improved input selection procedure and is distinct from the original geometric baseline comparison.

Table 10: **Per-seed automatic recognition.** Each run covers 198 sides and 3,280 reference signs using a DEV-selected, text-disjoint, with-Stage-A $N{=}100$ checkpoint. Images and order are shared within each route, not across routes. The geometric route retains 1,106 clusters (183 too small to decode); the base-selected route retains 1,186 rows (78 empty). No output is selected by reference-line matching. Seeds vary the recognizer, not all components. VLM generation caps are 128 and 64 tokens, respectively.

Recognizer	Seed	SER (%↓)	S	D	I
Original geometric route					
CRNN	42	71.71	1175	175	1002
CRNN	123	71.89	1175	176	1007
CRNN	456	73.05	1184	180	1032
SVTRv2*	42	69.51	1048	174	1058
SVTRv2*	123	71.86	1104	169	1084
SVTRv2*	456	70.98	1109	168	1051
PADDLEOCR-VL	42	68.02	1050	173	1008
PADDLEOCR-VL	123	69.42	1013	164	1100
PADDLEOCR-VL	456	69.33	1046	159	1069
Base-selected crops and order					
CRNN	42	55.24	1044	260	508
CRNN	123	55.03	1020	275	510
CRNN	456	57.16	1074	242	559
SVTRv2*	42	45.67	842	363	293
SVTRv2*	123	46.83	879	360	297
SVTRv2*	456	46.68	890	351	290
PADDLEOCR-VL	42	46.92	857	318	364
PADDLEOCR-VL	123	46.86	846	305	386
PADDLEOCR-VL	456	47.26	862	300	388

Table 11: **Full-pipeline recognition-seed repeats.** Pairs were fixed before the two additional evaluations; the first is the original run. Each pair covers all 198 sides. DeiT3 uses the same $N{=}214$, 7,582-character adaptation recipe; SVTRv2* uses $N{=}100$. Detector, LM, both Stage-A initializations, and decoder settings are fixed. Each classifier regenerates its own candidates. Δ is the matched SER reduction from verification, in percentage points.

DeiT3 seed	SVTRv2* seed	Without verifier	Full SER (%)	Δ (pp)
123	42	39.76	37.59	2.16
42	123	39.73	38.11	1.62
456	456	40.15	38.45	1.71
Mean		39.88	38.05	1.83
Sample SD		0.24	0.43	—

The matched ablation (Table 12) reduces mean edit errors from 1308.0 to 1248.0, with fewer substitutions and insertions but more deletions. Mean output length changes from 3210.3 to 3048.3 tokens; mean LCS recovery is 2356.7 versus 2348.3 of 3,280 reference signs, and mean exact-side counts are 3.67 versus 3.00 of 198. The original $(123, 42)$ pair scores 37.59% SER and supplies the qualitative automatic example. The comparison supports use of the existing sequence model for candidate selection, not a general architecture advantage or an isolated causal effect of synthetic pre-training.

D.7 REPRODUCTION AND INTERPRETATION

All 18 recognizer runs (nine original-route and nine base-selected-crop controls) retain their original DEV-selected checkpoints and complete outputs. Predictions, reference records, shared-input hashes, and tablet-bootstrap arrays are provided for saved-record verification. A raw-image DEV replay

Table 12: **Sequence-verification ablation, averaged over three recognition-seed pairs.** Full-output whole-side SER uses 3,280 fixed-inventory signs on 100 reused tablets. The two rows use the same three candidate pools; only verification is removed. Counts are per-run means. SER falls by 1.83 percentage points (paired-tablet 95% CI [1.15, 2.53]). Edit categories do not identify the responsible module.

System	SER (%↓)	S	D	I	Errors	Output tokens
Without verifier	39.88	633.0	372.3	302.7	1308.0	3210.3
Full system	38.05	528.3	475.7	244.0	1248.0	3048.3

taking only an image and surface metadata reproduced the selected candidate scores and tokens. Record replay does not retrain the models or regenerate their neural predictions.

The evaluation set has been inspected historically and is not a new blind test. Current edit categories do not separately identify the contributions of detection, crop selection, and line grouping. The present comparison measures complete frozen systems under a shared evaluation target, with different component supervision and decoding.

E Small-Budget Results and Per-Seed Values

Table 13: **Stage-A benefit with 25 adaptation tablets.** SER↓ (%), three-seed means; protocol and model-specific line universes as in Table 1. Δ = SER without − with Stage-A, with conditional paired-tablet 95% CIs. †Without-Stage-A CRNN seeds 123/456 selected no checkpoint and enter the means as empty outputs (SER=100%); seed 42 has a decoded, near-blank checkpoint.

Set	Model	with St.-A	no St.-A	Δ (pp) [95% CI]
In-period	SVTRv2*	28.4	96.3	+67.9 [65.6, 70.2]
	CRNN†	33.6	100.5	+66.9 [64.5, 69.3]
	PaddleOCR-VL	26.0	76.3	+50.2 [47.9, 52.5]
Cross-period	SVTRv2*	38.0	97.9	+59.9 [56.3, 62.8]
	CRNN†	44.6	100.3	+55.8 [52.1, 58.7]
	PaddleOCR-VL	39.0	93.3	+54.3 [51.4, 56.7]
Cross-collection	SVTRv2*	70.0	97.8	+27.8 [25.9, 29.8]
	CRNN†	72.4	101.8	+29.4 [27.5, 31.4]
	PaddleOCR-VL	64.9	97.7	+32.8 [31.2, 34.5]

At N=0, PaddleOCR-VL and SVTRv2* are within 1 pp SER on every set; CRNN trails by up to 7.6 pp. At N=100, the ordering varies by domain. These are descriptive comparisons: pre-training, capacity, token mapping, and Stage-A budget differ. Tables 14 and 15 list every run behind Tables 1 and 13: Table 14 gives the DEV-selected learning rate of each arm, and Table 15 the per-seed SER. As in §5, the primary test for each contrast is whether its per-comparison 95% tablet-bootstrap CI excludes zero (all 18 do). Zero of the 10,000 bootstrap resamples reversed the sign; this empirical count is not a calibrated p-value bound, and we do not rely on it in place of the CI. **Without-Stage-A convergence.** Most without-Stage-A CTC runs with a DEV-selected checkpoint early-stopped on DEV well before the 100-epoch budget (best epochs 1–66; the two failed CRNN N=25 runs below record best epoch 0, a sentinel for "no checkpoint selected"); the CRNN N=100 runs stopped latest (best epochs 87/82/97; two of the three reached the 100-epoch budget). As a converged control we retrain the SVTRv2* N=100 no-Stage-A arm from scratch for 300 epochs (early stop off, cosine over the full budget; training loss saturates near zero and DEV LCS peaks by epoch 244–267 under this schedule); it narrows the removal cost by only 1.5–4.2 pp—to +46.7 [44.5, 48.9], +50.7 [47.7, 53.1], and +25.2 [23.4, 27.0] pp in-, cross-period, and cross-collection (3-seed mean, tablet bootstrap 95% CI). The SVTRv2* advantage thus persists under the tested 300-epoch budget. For without-Stage-A CRNN at N=25, seeds 123/456 kept DEV LCS at 0. Training stopped at epoch 15 with loss ≈4.2

and no selected checkpoint. Their records are explicit failures scored as empty outputs (all deletions, SER = 100%). Seed 42 selected epoch 46 of 61 (loss 3.25); its near-blank output scores 101.5% in-period. This is an optimization failure under the fixed budget and grid, and the affected Table 13 rows are marked (†). The Stage-A effect is thus measured at a fixed downstream annotation and Stage-B budget; the conclusions rest on the N=100 contrasts (24–52 pp SER), not on the degenerate cells. An additional fixed-100-epoch control revisits these runs in Appendix K.

Table 14: DEV-selected learning rates (with / without Stage-A) per model and N, chosen by seed-42 development LCS Recall and applied to all three seeds and all evaluation sets (the rate does not vary by set). These are the LRs behind the per-seed SER in Table 15.

Model	N	LR with Stage-A	LR without Stage-A
SVTRv2*	25	10^{-4}	10^{-4}
SVTRv2*	100	3×10^{-4}	3×10^{-4}
CRNN	25	10^{-4}	3×10^{-4}
CRNN	100	3×10^{-4}	3×10^{-4}
PADDLEOCR-VL	25	5×10^{-5}	2×10^{-4}
PADDLEOCR-VL	100	2×10^{-4}	2×10^{-4}

Table 15: Per-seed **SER** (%↓) for the 18 primary contrasts (seeds 42/123/456); in-period = the reused set of §5.1. DEV-selected LRs are in Table 14. Δ = the SER cost of Tables 1 and 13 (rounded once from full precision, so it may differ from the difference of the two rounded seed means by ≤0.1 pp); the 3-seed means reproduce Tables 1 and 13. †Without-Stage-A CRNN N=25: seeds 123/456 kept DEV LCS at 0 and stopped at epoch 15 without a selected checkpoint. These failures are scored as empty outputs (all deletions, 100.0%). Seed 42 decodes near-blank outputs scoring 101.5/100.9/105.4%. Per-seed auxiliary LCS is in the released per-line records.

Set	Model	N	with Stage-A (s42/s123/s456)	without (s42/s123/s456)	Δ
In-period	SVTRv2*	25	28.5 / 28.5 / 28.1	97.2 / 94.9 / 96.7	+67.9
In-period	SVTRv2*	100	20.5 / 22.4 / 21.6	65.3 / 79.2 / 72.9	+50.9
In-period	CRNN†	25	32.9 / 33.6 / 34.4	101.5 / 100.0 / 100.0	+66.9
In-period	CRNN	100	24.3 / 24.6 / 25.5	72.2 / 77.7 / 69.3	+48.3
In-period	PADDLEOCR-VL	25	26.0 / 26.3 / 25.7	77.1 / 74.2 / 77.4	+50.2
In-period	PADDLEOCR-VL	100	17.6 / 17.7 / 18.4	45.8 / 47.7 / 44.0	+27.9
Cross-period	SVTRv2*	25	38.2 / 38.3 / 37.6	98.7 / 96.0 / 99.0	+59.9
Cross-period	SVTRv2*	100	37.6 / 38.4 / 39.2	87.3 / 93.0 / 92.1	+52.4
Cross-period	CRNN†	25	44.0 / 44.4 / 45.3	100.9 / 100.0 / 100.0	+55.8
Cross-period	CRNN	100	41.8 / 42.2 / 41.8	90.7 / 91.6 / 89.2	+48.6
Cross-period	PADDLEOCR-VL	25	38.8 / 39.3 / 38.8	93.7 / 93.1 / 92.9	+54.3
Cross-period	PADDLEOCR-VL	100	41.7 / 40.9 / 41.1	75.7 / 76.5 / 76.5	+35.0
Cross-collection (HT)	SVTRv2*	25	70.1 / 70.0 / 69.8	99.0 / 96.6 / 97.7	+27.8
Cross-collection (HT)	SVTRv2*	100	71.1 / 71.3 / 71.1	97.3 / 98.3 / 97.9	+26.7
Cross-collection (HT)	CRNN†	25	72.1 / 72.1 / 73.0	105.4 / 100.0 / 100.0	+29.4
Cross-collection (HT)	CRNN	100	74.6 / 72.9 / 73.3	97.6 / 98.2 / 96.5	+23.9
Cross-collection (HT)	PADDLEOCR-VL	25	65.4 / 64.7 / 64.7	97.3 / 97.4 / 98.6	+32.8
Cross-collection (HT)	PADDLEOCR-VL	100	66.2 / 66.8 / 65.4	90.5 / 90.4 / 89.3	+23.9

F PER-PERIOD BREAKDOWN OF THE CROSS-PERIOD SET

The cross-period set spans six attested periods of very different size and difficulty; Table 16 reports per-period tablet and token counts and N=100 performance.

Table 16: Cross-period performance is heterogeneous: the largest period (ED IIIb) sits 13–16 pp below the micro average. Per-period LCS Recall (%) at N=100 with Stage-A (3-seed mean, aggregated within period); token counts use the CTC serialization (the VLM is scored on its own records, which add 11 lines). Small cells warrant caution: ED IIIa has 3 tablets and 48 tokens; the unattributed tablet (HS 1032) lacks a period label in the source metadata. Approximate dates: ED IIIb ∼2500–2340 BC; ED IIIa and Old Akkadian also precede Ur III, whereas Old Assyrian and Old Babylonian follow it. The range extends up to ∼500 years from Ur III.

Period	Tablets	Tokens	SVTRv2*	CRNN	PADDLEOCR-VL
Old Assyrian	44	2054	65.7	62.3	63.4
Old Babylonian	13	1614	69.4	64.7	60.2
ED IIIb	114	1584	46.5	44.1	46.6
Early Old Babylonian	7	662	72.1	70.1	73.5
Old Akkadian	10	157	56.9	51.4	56.5
ED IIIa	3	48	33.3	30.6	40.3
Unattributed	1	38	81.6	81.6	89.7
Macro avg. (6 periods)	—	—	57.3	53.9	56.7
Micro avg. (all)	192	6157	62.0	58.7	59.3

G ATTRIBUTION STUDY: PROTOCOL, SEQUENCE-ORDER CONTROL, AND SEED STABILITY

This appendix gives the full attribution protocol, the all-lines N=100 SER reductions (Table 18), the parent-matched construction, and the separate global-shuffle sensitivity analysis.

Table 17: **Three-arm attribution study** (SVTRv2*, per-arm SER↓ %; SVTRv2* is our SVTRv2-inspired CTC variant, §4.2). Per-arm absolute SER on each set for the isolated-sign, shuffled-order, and attested-order Stage-A arms. Lifting isolated-sign supervision to multi-sign line format gives the largest tested training-format gain on every set. N=100 values are 3-seed Stage-B means; the isolated-sign arm is an initialization/task-transfer control. Corresponding all-lines reductions with CIs are in Table 18; the strict, parent-matched contrasts in Figure 2.

	in-period		cross-period		cross-collection	
Stage-A arm	N=0	N=100	N=0	N=100	N=0	N=100
Isolated-sign (control)	95.4	61.8	97.5	82.3	99.0	94.6
Shuffled-order line	36.4	25.6	39.2	38.6	68.3	71.2
Attested-order line	32.8	21.2	39.2	34.4	68.2	69.5

Table 18: All-lines N=100 SER reductions (pp) [95% CI] for the three-arm study (SVTRv2* attribution family; 3-seed Stage-B mean; paired tablet bootstrap, conditional on the trained Stage-A checkpoints). These are the all-lines contrasts corresponding to Table 17; the strict, parent-matched contrasts are in Figure 2.

Reduction	in-period	cross-period	cross-collection
isolated→shuffled	36.2 [34.7, 37.7]	43.7 [40.5, 46.7]	23.5 [21.6, 25.4]
shuffled→attested	4.4 [3.3, 5.5]	4.2 [3.0, 5.2]	1.7 [0.8, 2.7]
isolated→attested	40.6 [38.9, 42.3]	47.9 [44.3, 51.1]	25.2 [23.3, 27.1]

Primary three-arm protocol (Table 17). All three arms share one on-the-fly renderer and corpus (text-disjoint: 387,315 lines, 1,778,825 (sign, crop) tokens, 285 types, 146,784 distinct crops; height 64, gaps 2–8 px, vertical jitter ±3, Gaussian noise σ=6, brightness 0.90–1.10, background value 220,

real-background padding). *Attested-order* renders each source line in corpus order; *shuffled-order* permutes all (sign, crop) tokens within each P-number split and re-cuts them into lines of the original lengths; *isolated-sign* renders each token as its own one-sign image. A *parent* is the group of tokens the loss treats as one unit for grouping and weighting: in the line arms one parent is one rendered line, and in the isolated-sign arm the tokens of a parent are rendered as separate one-sign images but still weighted together as that parent. Every arm uses a batch of 32 parents per optimizer step with a parent-grouped CTC loss (mean over the tokens of a parent, then mean over the 32 parents), so per-parent weighting is identical across arms; in the isolated-sign arm each parent group yields $\approx$4.59 one-sign images, so that arm sees $\approx$147 images per step versus 32 line-images for the line arms. Token exposure and optimizer updates per epoch are matched; total compute is *not* claimed equal. Each arm trains for a fixed 30 epochs (no early stopping; the epoch-30 checkpoint is used), Stage-A seed 42, AdamW (LR 3×10^{-4}, weight decay 10^{-4}, 1000-step warmup then cosine), 10% P-number validation split. Stage-B then follows the locked protocol (LR chosen from $\{5\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$ by seed-42 DEV LCS, applied to seeds 123/456; $N \in \{25, 100\}$; SVTRv2* only). The all-lines reduction CIs (Table 18) and the strict CIs (Figure 2) are tablet bootstraps conditional on these single Stage-A checkpoints and therefore exclude Stage-A training variance; we do not provide a separate estimate of that variance here. The two-seed stability check below concerns only the separate global-shuffle control family (a distinct, more aggressive perturbation), not the primary Stage-A ladder.

Design boundary and parent-matched control. The isolated-to-shuffled contrast also changes token grouping: the shuffled arm re-groups tokens into new lines, whereas the isolated-sign arm keeps single tokens. So on its own that step only evaluates the *bundled* multi-sign line-format change (canvas, target length, CTC alignment, grouping, and multi-sign exposure jointly), not stitching in isolation. We therefore add a *parent-matched* control that renders the *same* shuffled groups as isolated one-sign images, holding grouping fixed (isolated vs. multi-sign line-format inputs; canvas, length, and CTC training still co-vary): singleton-matched $\rightarrow$ shuffled still lowers SER by +45.1/+46.2/+23.1 pp on the strict subset in-period, cross-period, and cross-collection, respectively (3-seed mean, tablet bootstrap; §6.2). The gain therefore persists when token grouping is held fixed.

Table 19: **Strict-subset attribution: full values** plotted in Figure 2 (SVTRv2* attribution family; N=100; 3-seed Stage-B mean; SER-record strict subset, exact-unseen $\cap$ no-shared-4-gram; paired tablet bootstrap 95% CI, conditional on the Stage-A checkpoints). SER reductions (pp); positive = improvement. The first contrast is parent-matched (grouping held fixed); the second is not; the point differences telescope through the shared shuffled arm, but their CIs cannot simply be added. They do not provide a causal decomposition of the main Stage-A effect.

Contrast	in-period	cross-period	cross-collection
singleton-matched $\rightarrow$ shuffled	45.1 [42.5, 47.7]	46.2 [42.8, 49.4]	23.1 [21.2, 25.1]
shuffled $\rightarrow$ attested	1.2 [−0.7, 3.0]	3.5 [2.1, 4.8]	1.8 [0.9, 2.8]

Table 20: **Global-shuffle order control (SVTRv2*, LCS Recall %).** A stronger perturbation than the within-split shuffle of the primary attribution study (Table 17): Stage-A on the real-order text-disjoint corpus vs. the same corpus with (sign, crop) pairs *globally* shuffled. N=25/100 are 3-seed Stage-B means (locked protocol; DEV-selected LRs 10^{-4}/3×10^{-4} for both), N=0 the Stage-A models alone. Δ = real − shuffled, 95% tablet-bootstrap CI (seed stability below; N=0 Δ: single checkpoints, no CI). Real order is worth $\sim$4 pp in-period; under this stronger perturbation the off-domain effect is smaller and seed-dependent (cf. the consistent in-/cross-period benefit under the milder within-split shuffle, Table 17).

	N=0			N=25			N=100		
Set	real	shuf.	Δ	real	shuf.	Δ [CI]	real	shuf.	Δ [CI]
In-period (Ur III)	66.8	62.5	+4.3	72.2	68.6	+3.7 [2.0, 5.4]	79.0	74.9	+4.1 [2.9, 5.3]
Cross-period	59.7	61.1	−1.4	62.4	64.6	−2.1 [−3.4, −0.9]	62.0	62.5	−0.5 [−1.7, 0.7]
Cross-collection	30.7	33.2	−2.4	31.5	34.4	−2.9 [−3.8, −1.9]	30.2	29.8	+0.4 [−0.4, 1.3]

Separate global-shuffle control. This separate sensitivity analysis is distinct from the primary within-split shuffle of Table 17. From the text-disjoint corpus (387,315 lines, 1,778,825 sign tokens, 285 types, 146,784 distinct crop images) we list every (sign, crop) pair in corpus order, apply one global permutation (seed 42), and re-cut the permuted list into lines of the original lengths, attaching to each line the source line's P-number. Verified equalities with the real-order corpus: line count, token count, per-sign counts, crop multiset, (sign, crop)-pair multiset, and line-length distribution are identical. Each shuffled line keeps a source-line P-number *label* only for bookkeeping; because tokens are permuted *globally across splits*, this does not preserve the original per-line (sign, crop) membership of the validation split (unlike the primary within-split shuffle of Table 17, which is unaffected). The renderer (height 64, gaps 2–8 px, jitter, noise, brightness) and the Stage-A/Stage-B training configurations are unchanged. 2,243 shuffled lines (0.58%) coincide with some real line; 2/3/4-gram overlap with the real corpus is 51.6/7.9/0.4%, and the shuffled corpus has 1.8/5.1/4.5× more distinct 2/3/4-grams. Manifest SHA-256 `d835d53f. . . .` **Seed stability of the global-shuffle control.** We trained a second Stage-A model of each kind (seed 123, same corpora) and evaluated all four at N=0. LCS Recall (%) for real-order / shuffled models, at seeds 42 and 123, is: in-period 66.8/62.5 and 66.9/63.2 (Δ = +4.3, +3.7); cross-period 59.7/61.1 and 61.2/61.0 (−1.4, +0.2); cross-collection 30.7/33.2 and 32.9/32.6 (−2.4, +0.2). Real-order seeds differ by 0.1/1.5/2.1 pp across the three sets; shuffled seeds differ by 0.7/0.1/0.5 pp. The in-period advantage is stable and larger than the observed seed variation. Off-domain effects are smaller and seed-dependent. This global shuffle is stronger than the primary within-split shuffle, which changes co-occurrence and order jointly. The primary study finds a small all-lines N=100 SER benefit (Table 17); its strict in-period CI crosses zero (§6.2). These are complementary sensitivity analyses. Two seeds do not estimate Stage-A variance or bound the text-structure effect, and the primary study's Stage-A variance remains unmeasured. Stage-B CIs in Table 20 condition on the seed-42 Stage-A checkpoints.

H VLM DIAGNOSTICS: FULL TABLES

This appendix collects two PADDLEOCR-VL diagnostics: an image-sensitivity probe (Table 21) and the LoRA-rank check. (The cross-generation text-exposure comparison is in Appendix A.)

Table 21: Image-sensitivity results on the 670-line in-period set (earlier PADDLEOCR-VL Stage-A checkpoint; the three Stage-B checkpoints are summarized in the text). Permuted: images randomly reassigned (derangement, seed=99). Blank: solid gray (pixel 128). Noise: Gaussian $\mathcal{N}(128, 50^2)$. ChgRate: % of lines where prediction differs from Correct. The harness re-stitches the input images itself (fresh gap/noise draw), so Correct-row LCS Recall can differ slightly from the same legacy checkpoints' scores.

Checkpoint	Condition	LCS	Exact%	ChgRate%
PADDLEOCR-VL Stage-A (N=0)	Correct	73.26	41.49	0.0
	Permuted	7.29	1.49	98.5
	Blank	6.39	0.75	99.1
	Noise	6.92	0.60	99.3

The three Stage-B checkpoints evaluated (r=16 at N=100 and N=200; r=64 at N=100) behave the same way: permuted, blank, and noise inputs change 98.7–100% of predictions and collapse LCS Recall to 0.71–7.93% (Correct rows: 75.13–81.90%).

LoRA rank. At N=100 under matched tablet subsets (seeds 123 and 456), we see no systematic rank effect across the two controlled seeds: r=16/32/64 reach auxiliary LCS Recall of 72.27 ± 0.82, 72.25 ± 1.11, and 71.95 ± 0.74% (mean±std over the two seeds). With only two seeds this is a small stability check, not an equivalence test. The seed-42 runs (76.07, 81.94, 82.72%) reused the original production tablet subset, so their larger spread is confounded with subset selection.

I NATURAL-LINE DIAGNOSTIC

As a diagnostic, we also trained SVTRv2* on *natural-line* images: contiguous rectangular crops of tablet lines with a 15% margin, rather than sign-level stitched composites. The natural-line recognizer includes an OOV class (749 semantic labels + OOV + blank = 751-class output) to handle signs outside the inventory, and initializes from the earlier text-exposed Stage-A checkpoint with the CTC head expanded by an OOV row.

At N=214 tablets, the natural-line model achieves 48.31% aggregate LCS Recall (OOV signs included in GT) on a 906-line natural-line evaluation split; removing OOV from both prediction and GT gives 45.98%; 21.6% of its GT tokens are OOV. The stitched formulation reaches 79% LCS Recall on the same N=214 tablets, but this is a *motivating diagnostic, not a controlled comparison*: the two use different evaluation lines and OOV treatment, and the natural-line model initializes from the earlier text-exposed Stage-A checkpoint. It bears on the input-format choice (§3.1) and is not evidence that end-to-end learning cannot work.

J STRONG PER-SIGN CLASSIFIER BASELINE

Because the oracle protocol supplies GT sign boxes, we compare SVTRv2* against a strong per-sign classifier that keeps the protocol but replaces sequence decoding with per-crop classification (main-text summary in §6.1).

Configuration and sources. **Architecture / initialization:** ImageNet-pre-trained DeiT3-medium (`deit3_medium_patch16_224.fb_in22k_ft_in1k`, ImageNet-22k→1k) with a linear head over 749 labels (sorted dictionary signs + `X` + `lacuna`). It uses no MAE, multitask head, or earlier cuneiform weights. **Stage-A:** dictionary classification on 199,243 crops (`dictionary_clean.json`, tablet-disjoint from all test sets). The run config records 193,266 training and 5,977 validation crops (`val_frac`=0.03). Training uses AdamW, LR 10^{-4}, weight decay 0.05, cosine scheduling, and at most 30 epochs with patience 6. Best validation per-sign accuracy is 0.860 at epoch 25. **Stage-B:** adaptation uses the main table's 100-tablet `train_n100` split and seeds 42/123/456. As for SVTRv2*, checkpoints maximize LCS Recall on the 50-tablet DEV set. The saved DEV curves confirm patience 8. No per-run training manifest or config was saved; the split and AdamW LR 10^{-5} come from the retained implementation, with the LR unconfirmed by a run record.

Inference / scoring: classify each GT crop and concatenate predictions in line order. SER, exact-line accuracy, and strict membership use the same GT, OOV rule, and `(stem,line_id)` keys as SVTRv2*. The evaluated line sets match exactly: 670/1,243/1,320 full lines and 291/1,050/1,293 strict lines. Saved predictions record checkpoint paths but no checkpoint hashes. A later hash inventory accompanies the supplement; it does not establish a contemporaneous checkpoint-to-output binding. CIs are paired tablet bootstraps (B=10,000, percentile 95%), conditional on fixed checkpoints, with $\Delta = \text{SER}_{\text{per-sign}} - \text{SER}_{\text{SVTRv2*}}$. Architecture, initialization, and objective differ, so this comparison does not isolate sequence context.

Table 22: **Strong per-sign classifier vs. SVTRv2* under the oracle protocol** (N=100, 3-seed mean). SER↓ and exact-line↑ (%); $\Delta = \text{SER}_{\text{per-sign}} - \text{SER}_{\text{SVTRv2*}}$ (negative = classifier lower), paired tablet-bootstrap 95% CI conditional on the fixed checkpoints. Per-sign top-1 accuracy: in-period 79.0, cross-period 72.2, cross-collection 37.6.

Set	Subset	$\text{SER}_{\text{p-s}}$	$\text{SER}_{\text{SVTRv2*}}$	Δ (pp) [95% CI]	$\text{Ex}_{\text{p-s}}$	$\text{Ex}_{\text{SVTRv2*}}$
In-period	full	21.0	21.5	−0.5 [−1.8, 0.9]	44.9	47.1
	strict	22.2	25.4	−3.2 [−5.5, −0.8]	39.6	34.9
Cross-period	full	27.8	38.4	−10.6 [−12.2, −8.9]	25.9	16.4
	strict	28.3	40.0	−11.7 [−13.2, −10.0]	25.3	14.4
Cross-collection	full	62.4	71.2	−8.8 [−9.9, −7.7]	3.6	1.7
	strict	62.5	71.4	−8.8 [−10.0, −7.7]	3.5	1.7

On full lines the in-period difference is small with a CI that includes zero; on the in-period strict subset and on both off-domain sets the intervals favor the per-sign classifier (CIs excluding zero). With sign boundaries given, the sequence training gain therefore does not establish an advantage over per-sign classification. The inconclusive full-set in-period result does not establish equivalence. Automatic recognition is evaluated separately in Appendix D.

K Development Checks

These checks retain the primary checkpoint family and table entries. They examine the stopping condition of small-budget CRNN training and the DEV checkpoint and output-parsing sensitivity of the VLM. They do not reselect primary checkpoints by evaluation-set performance.

K.1 CRNN small-budget control without early stopping

We rerun the without-Stage-A $N{=}25$ CRNN for all three original seeds from random initialization, using the same data, augmentation, learning rate 3×10^{-4}, batch size 8, AdamW weight decay 10^{-4}, and 100-epoch cosine schedule. Only patience is extended from 15 to 100, so every run completes 100 epochs. Checkpoints still maximize DEV LCS; none is selected using evaluation-set SER. The original failed-run records remain unchanged in Table 13.

The selected epochs are 88, 71, and 89 for seeds 42, 123, and 456. The two formerly blank runs begin producing correct DEV signs after the old stopping point. This shows that the epoch-15 failures were not permanent failures to learn. However, the decoded models remain weak (Table 23); longer training does not rescue the small-budget scratch recognizer or overturn the pre-training benefit.

Table 23: **Additional $N{=}25$ CRNN control.** Three-seed mean corpus SER (%). Original no-Stage-A includes the two declared empty-output failures; Fixed 100 decodes three actual DEV-selected checkpoints. The With St.-A column gives the unchanged primary results. The last column is Fixed 100 minus Stage-A, with paired-tablet 95% CI.

Domain	Original no St.-A	Fixed 100	With St.-A	Gain [95% CI]
In-period	100.51	99.44	33.63	65.81 [63.41, 68.11]
Cross-period	100.31	100.90	44.56	56.35 [53.20, 59.00]
Cross-collection	101.82	105.03	72.42	32.61 [30.62, 34.73]

K.2 VLM validation checkpoint and output-parsing checks

The pilot compares the original DEV-selected $N{=}100$, seed-42 PaddleOCR-VL adapter with steps 1000/2000/3000/4000 on 342 DEV lines (1,597 signs), fixing augmentation (seed 12345), prompt, and deterministic generation, and saving all raw responses. Unfiltered edit counts are 278/300/284/292/276. Step 4000 improves both SER (17.41% to 17.28%, two errors) and LCS. This does not show conflicting metric rankings: training-time validation used changing augmentation draws.

The historical PaddleOCR-VL evaluator normalizes sign names and removes tokens outside the inventory before scoring. Its saved-token SER is not unfiltered generation SER. For these five DEV checkpoints, explicitly retaining unknown output tokens leaves the total edit counts unchanged, although some individual alignments differ. This check is limited to these five DEV checkpoints. Original oracle-line results retain their saved-token scoring protocol.

The symmetric check covers three seeds with/without Stage-A (Table 24), using identical images, references, prompts, and generation settings. Scoring normalizes known signs but retains unknown tokens and response prefixes. Seed-42 Stage-A matches the pilot. The latest archived adapters are step 3000 for Stage-A seed 456 and 4000 otherwise, not necessarily final epochs.

Changes are two to five errors over 1,597 signs. Unfiltered SER and LCS agree for all six pairs; with the historical parser, without-Stage-A seed 123 has one fewer error but lower LCS at the later point.

Table 24: **Symmetric VLM DEV checkpoint check.** Fixed identical inputs, $N{=}100$, three seeds per arm. "Selected" is the original DEV-LCS checkpoint; "later" is the latest archived periodic adapter (step 3000 for Stage-A seed 456, 4000 otherwise). Both SER columns retain unknown output tokens; ΔE is later minus selected edit count. This is DEV sensitivity, not a new test result.

Stage-A	Seed	Selected SER (%)	Later SER (%)	ΔE
Yes	42	17.41	17.28	-2
Yes	123	17.16	16.84	-5
Yes	456	17.60	17.72	$+2$
No	42	41.58	41.70	$+2$
No	123	43.64	43.77	$+2$
No	456	42.20	42.52	$+5$

Filtering changes at most three errors (0.19 pp) across these twelve DEV outputs. These DEV checks leave the primary checkpoints and three-domain results unchanged.