TABLETVLM: END-TO-END CUNEIFORM TABLET TRANSCRIPTION WITH A SPATIALLY GROUNDED VISION LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Transcription of cuneiform tablets requires solving spatial localization, reading-order reconstruction, and sign recognition across substantial variations in image quality, tablet condition, and scribal style. Existing approaches address these challenges through discrete pipeline stages, requiring different datasets, evaluations, and models, or attempt direct end-to-end training. These difficulties are compounded by the low-resource nature of cuneiform and by existing datasets that are often incomplete, inconsistently annotated, or derived from heterogeneous sources. We propose a single vision-language model that directly transcribes cuneiform tablets without explicit sign detection supervision or reading-order reconstruction and investigate how a training curriculum with supervision at different visual and sequential scales can be used to improve transcription performance. Among the four curricula tested, the model with the best trade-off between transcription and localization performance achieves a mean tablet-side sign-error rate of 27.24% on 4,805 held-out tablet sides, while its peak cross-attention falls within the corresponding ground-truth sign region for 78.27% of 879 held-out annotated signs. More broadly, our results suggest that end-to-end VLMs can learn both accurate sequence prediction and spatial grounding in low-resource OCR settings with limited and uneven supervision.

1 INTRODUCTION

Cuneiform tablets preserve a large textual record of the ancient Near East, but much of this material remains difficult to access computationally without expert transcription. We focus on texts from the Third Dynasty of Ur (Ur III), whose large administrative corpus provides a comparatively well-defined setting for studying automated cuneiform transcription. Improving transcription from tablet images could expand the amount of searchable text available for both philological research and large-scale computational analysis.

1.1 CHALLENGES

Transcription of cuneiform tablets is difficult due to the complexity and variance in tablets. Signs are formed by pressing a stylus into wet clay, introducing three-dimensional warping not normally encountered in OCR or HTR. As opposed to type-print, cuneiform sign appearance varies by scribal style, and spacing between signs and lines are highly irregular. Roughly four millennia of aging have also subjected Ur III tablets to erosion, cracking, or discoloration which obscure signs and introduce similar visual noise.

Sign classes also have an extreme long-tailed distribution. In our unified dataset, 471 sign classes appear 3,160,375 times, but the top 10% of classes (47 of 471) account for 68.28% of those occurrences, while 67 classes (14.23%) appear fewer than ten times (Figure 1). This makes it difficult for models to learn rare signs and generalize effectively.

Spatial annotations are uneven as well. On the eBL tablet sides represented by Sign Snippets (Lewenstein et al., 2026) retained for this study, raw boxes align with 56.0% of visible transcribed sign positions, falling to 30.7% after the data cleaning and sign mapping (Section 3.1; Fig-

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.

ure 2). This means a large portion of signs on annotated tablet photos still lack bounding boxes, limiting both the supervision and reliable evaluation of segmentation- or detection-based sign localization models.

Current approaches often decompose localization, recognition, and sequencing as separate steps. This increases control and interpretability, but sparse sign boxes leave visible signs unlabeled during detector training (Che et al., 2026), and errors in localization then propagate to recognition and sequencing. Separately optimized stages also do not naturally share learned representations (Williams et al., 2023). On the other hand, direct end-to-end training on full tablet images avoids box-level targets (Kim et al., 2022), but introduces substantial visual and sequential variability. Our experiments show that direct training yields substantially poorer transcription accuracy than training with intermediate supervision, highlighting the complexity of the task.

1.2 Approach

We investigate a single vision-language model that jointly learns visual sign recognition and sequence generation to transcribe complete tablet-side images, with multiple training curricula used to evaluate how different sources of supervision support these capabilities. Our model consists of an ImageNet-initialized SwinV2-Tiny backbone, a projection layer, and a three-layer Transformer decoder initialized from scratch, which uses self- and cross-attention to autoregressively generate cuneiform text. We use labeled individual sign images, full tablet-side images, and tablet text sequences. Following the synthetic data generation introduced in concurrent work (Anonymous, 2026), we also construct synthetic training examples by combining visual sign instances with text sequences, providing intermediate supervision between isolated sign recognition and full tablet transcription. We compare curricula that incorporate different combinations of sign-level learning, text-only warm-up, and synthetic line and tablet-side training before real tablet-side transcription. These comparisons examine whether intermediate training stages improve transcription accuracy and spatial alignment over simpler training schedules.

1.3 Results & Contributions

Our model demonstrates that a single vision-language model can learn both transcription and localization without explicit sign detection supervision. On 4,805 held-out tablet sides, the model with the best trade-off between transcription and localization performance achieves a mean side SER of 27.24% and macro sign F1 of 52.24%, with attention peaks inside the correct sign region for 78.27% of 879 annotated signs.

Our contributions are threefold: (1) we develop a unified vision-language model for direct cuneiform tablet-side transcription; (2) we investigate multiple training curricula to assess how different sources and sequences of supervision support learning within this architecture; and (3) we evaluate transcription on 4,805 held-out tablet-side images and spatial alignment through a separate localization evaluation using annotated sign regions.

2 Related Work

Cuneiform sign detection and localization. One line of cuneiform OCR research focuses on locating and identifying individual signs in tablet photographs or renderings (Dencker et al., 2020; St"otzner et al., 2023; Cobanoglu et al., 2024). These methods provide explicit spatial predictions but require additional processing to recover ordered transcriptions. Incomplete annotations also complicate evaluation, as correct detections of unannotated signs are counted as false positives.

Cuneiform sign recognition. Isolated recognition methods classify cropped sign images (Verwimp et al., 2025) or align prototypes with photographed signs, using synthetic examples to improve rare-class recognition (Mikulinsky et al., 2025). These methods learn sign identity but do not evaluate localization, reading order, or complete tablet transcription.

Cuneiform OCR pipelines. Modular pipelines combine explicit detection, recognition, and ordering stages. DeepScribe uses separately trained sign detection, classification, and line detection (Williams et al., 2023); other approaches combine tablet-side extraction, DETR, and heuristic reading-order reconstruction (Che et al., 2026), or SAM2 segmentation with a Vision Transformer

classifier (Poudel et al., 2026). These pipelines provide interpretable intermediate outputs, but allow upstream errors to propagate and lack joint optimization for the final transcription objective.

End-to-end image-to-sequence OCR. Recent OCR models increasingly map document images to ordered text or structured sequences using vision-language models. HunyuanOCR (Hunyuan Vision Team et al., 2025) follows a unified image-to-sequence approach, and PaddleOCR-VL (Cui et al., 2025) combines a separate layout and reading-order model with a VLM for element recognition. These models typically rely on large quantities of paired image-text supervision. Cuneiform presents a more data-constrained setting in which such supervision is limited and heterogeneous, motivating our investigation of how different sources of supervision support full tablet-side transcription.

Synthetic and compositional supervision. Synthetic training data provides a way to construct paired image-sequence supervision when manually aligned examples are scarce. StackMix (Shonenkov et al., 2021) generates labeled handwritten sequences by recombining smaller visual fragments, demonstrating that supervision available at a local scale can be composed into longer recognition examples. More recently, Li et al. (2026) construct realistic full-page document supervision by composing document elements and layouts, using synthetic data to address the scarcity of end-to-end document annotations. Closely related concurrent work (Anonymous, 2026) applies this principle to cuneiform by arranging isolated real sign crops according to tablet text sequences to produce synthetic line-level supervision. We utilize synthetic line and tablet images to bridge the gap between sign images and tablet images.

Progressive and curriculum-based training. Curriculum learning organizes training so that simpler tasks are learned before more difficult ones (Bengio et al., 2009). Recent document-recognition systems have adopted similar strategies: Li et al. (2026) combine synthetic full-page supervision with progressive learning for end-to-end document parsing, while FireRed-OCR (Wu et al., 2026) trains a VLM through successive stages that move from document perception toward structured sequence generation. Our curriculum follows the same general principle but increases difficulty along spatial and sequential dimensions, requiring increasingly complex recognition, localization, and reading-order reconstruction while retaining a single model throughout training.

3 Dataset

3.1 Data Sources and Normalization

We build a unified Ur III dataset with sign crops from Sign Snippets (Lewenstein et al., 2026), sign annotations from the electronic Babylonian Library (eBL) (electronic Babylonian Library), and tablet photographs, catalog metadata, and ASCII Text Format (ATF) transcriptions from the Cuneiform Digital Library Initiative (CDLI) (CDLI contributors, 2026). These sources provide sign-level and tablet-level supervision. We normalize their labels to a shared Unicode cuneiform vocabulary.

Data Cleaning & Filtering We apply strict cleaning and filtering criteria using online catalog definitions and manual inspection. For Sign Snippets, we retain consistent labels that map to Unicode cuneiform, identify source tablet using CDLI IDs, and obtain individual crops from annotated sign regions (Appendix A.2). For CDLI, we match photographs and ATF records by catalog identity, parse side and line structure, and automatically extract obverse and reverse images from composite photographs, discarding cases without two eligible regions (Appendix A).

Sign Annotation Coverage Sign annotations cover only part of the visible text in the eBL images. Across 206 artifacts containing 14,610 visible transcribed sign positions, annotation coverage is 56.0% in the raw metadata and 30.7% after filtering. Figure 2 shows coverage before and after filtering for the 178 artifacts with nonzero coverage; 28 zero-coverage artifacts are omitted from the plot but included in the overall rates. Coverage is measured by matching boxes to transcription positions, so unmatched positions may also reflect outdated transcription paths (Appendix A.2).

Synthetic Data Following the synthetic data generation procedure introduced in concurrent work (Anonymous, 2026), we construct synthetic line and tablet-side training examples from these

independently available resources, bridging isolated sign recognition and full tablet-side transcription (Appendix A.5).

Target The target is an ordered sequence of Unicode signs with line, column, and uncertainty markers. The training inventory contains 471 signs and 8 special tokens. Unicode mapping details are in Appendix A.1. Figure 1 shows the sign-class distribution across the unified manifest.

3.2 Dataset Splits and Provenance

We use CDLI P-numbers to enforce source-tablet isolation. The transcription test set contains 2,485 tablets, approximately 10% of the cleaned CDLI tablet pool. A separate localization evaluation holdout contains 67 of 133 eligible eBL tablets with sign annotations. After excluding both sets from the training pool, we assign 10% of the remaining source P-numbers to validation by deterministic SHA-256 ranking with seed 13. Synthetic examples use donors from their assigned split. No source tablet appears in more than one of the training, validation, or test splits, and validation and test set vocab is fully covered by the train set (Appendix A.6).

Table 1: Dataset composition after normalization, source-level splitting, and provenance filtering.

Component	Train	Validation	Test
Real sign crops	4,520	444	–
Synthetic lines	140,981	4,437	–
Synthetic tablet sides	17,623	555	–
Real CDLI tablet sides	38,920	4,256	4,805
Total	**202,044**	**9,692**	**4,805**

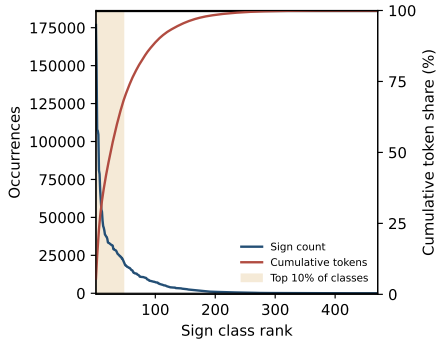


Figure 1: Sign occurrences by class rank (left axis) and cumulative percentage of tokens (right axis). Shading marks the top 10% of classes; synthetic examples repeat source text.

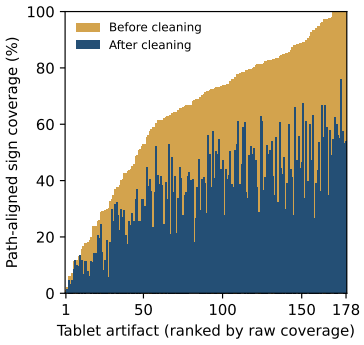


Figure 2: Path-aligned coverage for 178 artifacts, sorted by raw coverage. Gold and blue show coverage before and after cleaning; 28 zero-coverage artifacts are omitted.

4 Model Design

Our model follows an encoder-projector-decoder architecture for autoregressive cuneiform transcription. The encoder extracts spatial visual features from the tablet image, and the projector maps these features into the decoder's embedding space. The decoder then generates a sequence of Unicode cuneiform signs and structural markers, conditioning each output token on the image tokens and previously generated tokens.

Encoder The encoder uses SwinV2-Tiny pretrained on ImageNet-1K (Liu et al., 2022). Its four stages progressively reduce spatial resolution while increasing feature dimensionality, producing a

Table 2: Swin feature dimensions (height × width × channels) with standard downsampling and with Stage 4 spatial downsampling disabled.

Swin stage	Standard Swin	Stage 4 at /16
Input	$512 \times 1024 \times 3$	$512 \times 1024 \times 3$
Patch embedding	$128 \times 256 \times 96$	$128 \times 256 \times 96$
Stage 1	$128 \times 256 \times 96$	$128 \times 256 \times 96$
Stage 2	$64 \times 128 \times 192$	$64 \times 128 \times 192$
Stage 3	$32 \times 64 \times 384$	$32 \times 64 \times 384$
Stage 4	$16 \times 32 \times 768$	$32 \times 64 \times 768$

final feature map at $1/32$ of the input height and width. In preliminary experiments, the model was unable to converge during tablet-side training, suggesting that reduced spatial resolution may obscure the detail needed to recognize individual signs.

We therefore disable spatial downsampling in the fourth Swin stage during Phases 3–5. This preserves feature maps at $1/16$ of the input height and width, yielding four times as many spatial positions as the standard configuration while retaining 768 feature channels. Table 2 compares the feature dimensions, and Appendix B.1 explains the modification.

Projector & Decoder Separate projectors map 768-channel features through widths $768 \rightarrow 384 \rightarrow 192$ for lines and $768 \rightarrow 192 \rightarrow 192$ for sides. The spatial grid is represented as a sequence of visual tokens. With Stage 4 spatial downsampling disabled, the full grid contains 2,048 visual tokens, each with 192 dimensions:

$$32 \times 64 \times 768 \;\rightarrow\; 32 \times 64 \times 192 \;\rightarrow\; 2048 \times 192.$$

The decoder is a three-layer autoregressive Transformer with a hidden dimension of 192, three attention heads, and a feed-forward dimension of 768. Each layer contains causal self-attention over previous tokens, cross-attention over the projected image tokens, and a feed-forward network.

A linear output layer maps each decoder hidden state to a score for every vocabulary token, sharing its weights with the token embedding matrix. The vocabulary is constructed from the training transcriptions and includes Unicode cuneiform signs, the structural tokens `<LINE>`, `<COLUMN>`, `<ILLEGIBLE>`, and `<GAP>`, and model-control tokens. Autoregressive generation begins with `<BOS>` and stops when the model predicts `<EOS>` or reaches the 512-token generation limit. This limit corresponds to the 99.17th percentile across the whole CDLI transcription population.

The model contains 30.67 million parameters, of which 29.75 million are used for tablet-side transcription, including 27.58 million in the encoder and 1.97 million in the decoder.

5 Training

We study four training curricula assembled from the same five fixed phase definitions. Each curriculum is an independent run, model weights are carried between phases within a run. The four curricula are:

Curriculum	Phase sequence
Full curriculum	$1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$
Without synthetic transcription	$1 \rightarrow 2 \rightarrow 5$
Text warm-up followed by real sides	$2 \rightarrow 5$
Real sides only	5

All curricula use the same dataset manifest and split, a vocabulary of 479 tokens, and random seed 13. Synthetic-line and synthetic-tablet training sets are each capped at 10,000 randomly selected examples to prevent overfitting, while their validation sets remain uncapped. Each run starts with an ImageNet-pretrained SwinV2-Tiny encoder and randomly initialized decoder and projection layers. Training follows the phase definitions in Table 3, except that curricula $2 \rightarrow 5$ and 5 train the patch embedding and all Swin stages during Phase 5.

Table 3: Training phases with representative training examples, trained components, and frozen Swin encoder stages.

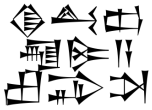
Phase	Data	Visual	Objective	Trained components	Frozen stages
1	Isolated sign crops		Sign classification	Encoder, classifier	None
2	Training transcriptions (text only)		Causal next-token prediction	Decoder (cross-attention bypassed)	Patch embedding; Stages 1–4
3	Synthetic lines		Autoregressive line transcription	Encoder, decoder, line projector	Patch embedding; Stages 1–3
4	Synthetic tablet sides		Autoregressive tablet-side transcription	Encoder, decoder, tablet-side projector	Patch embedding; Stages 1–2
5	Real tablet sides		Autoregressive tablet-side transcription	Encoder, decoder, tablet-side projector	None for $2 \rightarrow 5$, 5; otherwise patch embedding and Stages 1–2

Table 4: Fixed settings for each phase when included in a curriculum.

Phase	Epochs	Batch size	Relative Spatial Resolution
1: Sign recognition	10	12	/32
2: Text warm-up	1	32	—
3: Synthetic lines	10	32	/16
4: Synthetic tablet sides	10	6	/16
5: Real tablet sides	25	9	/16

Image preparation and targets. Images are converted to RGB, resized while preserving aspect ratio, and centered on a 512×1024 canvas. Maximum content bounds are 256×256 pixels for sign crops, 128×1024 for lines, and 512×512 for tablet sides. The remaining canvas is filled with the image's median color, and pixel values are normalized using ImageNet channel means and standard deviations. A visual attention mask excludes padded regions.

Optimization. Epoch counts and batch sizes are listed in Table 4. Each phase uses a fresh AdamW optimizer with a constant learning rate of 10^{-4}, weight decay of 0.01, and gradient clipping at norm 1.0. Training uses BF16 precision on an NVIDIA GeForce RTX 5070 Ti. We compute teacher-forced validation metrics after each image-training epoch and report the final Phase-5 checkpoints for curriculum comparisons.

6 Evaluation

We evaluate transcription by free generation and compare curricula using the final Phase-5 checkpoints (epoch 25). Localization is evaluated on a separate annotated holdout by measuring cross-attention alignment with ground-truth sign boxes while supplying the reference transcription.

Sign Error Rate After removing structural, uncertainty, and model-control tokens, we compute overall SER as the total minimum edit count divided by the total reference sign count, and mean tablet-side SER as the average of these ratios across tablet sides.

$$\text{SER} = \frac{\sum_{i=1}^{N}(S_i + D_i + I_i)}{\sum_{i=1}^{N} L_i}, \qquad \text{Mean side SER} = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i + D_i + I_i}{L_i},$$

where N is the number of tablet sides, L_i is the reference sign count, and S_i, D_i, I_i are substitution, deletion, and insertion counts under a minimum-edit alignment.

Macro Sign F1 We match predicted and reference sign sequences using a minimum-edit alignment, then compute F1 for each sign class and average across classes. This gives rare and frequent classes equal weight. A class that is never predicted receives an F1 of zero.

$$F_{1,c} = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \qquad \text{Macro } F_1 = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} F_{1,c},$$

where $\mathcal{C}$ contains the sign classes present in the reference set, and TP_c, FP_c, and FN_c count correctly matched, incorrectly predicted, and missed instances of class c, respectively.

Localization We measure localization using final-layer cross-attention averaged over heads while supplying the reference transcription. Peak in box measures how often the model's strongest attention falls inside the correct sign's box, while mass in box measures how much of its total attention falls inside that box.

$$\text{Peak in box} = \frac{100}{N}\sum_{i=1}^{N}\mathbf{1}\left[\arg\max_{p} A_i(p) \in B_i\right], \qquad \text{Mass in box} = \frac{100}{N}\sum_{i=1}^{N}\sum_{p} A_i(p)\, w_i(p),$$

where N is the number of scored signs, $A_i(p)$ is normalized attention at image pixel p, B_i is the annotated sign box, and $w_i(p)$ is the fraction of pixel p covered by B_i. Peak in box measures how often the maximum-attention pixel center lies inside the sign box, while mass in box measures the attention assigned to that box, with higher values indicating stronger alignment.

7 Results and Discussion

Table 5: Test results for phase-5 epoch-25 checkpoints from the four runs. All runs use the same 4,805-side transcription test set and 879-position localization holdout. Training cost and structural diagnostics are reported in Appendix C

Phase sequence	Overall SER	Macro F1	Mean side SER	Peak in box %	Mass in box %
$1 \to 2 \to 3 \to 4 \to 5$	30.46	**54.04**	26.19	76.79	48.29
$1 \to 2 \to 5$	30.83	52.24	27.24	**78.27**	**50.00**
$2 \to 5$	**29.24**	53.64	**25.93**	45.51	23.29
5	105.20	0.96	108.06	1.48	3.30

Transcription The results suggest that the specific curriculum matters much less than whether the model is given any pretraining before Phase-5. The three curriculum-trained models all end up in a narrow performance range, with SER around 30% and macro F1 around 50%. Their Phase-5 validation losses and teacher-forced SERs also converge to nearly the same values by the end of training. This indicates that the earlier phases mainly change the model's initialization and optimization trajectory, while continued training on the shared Phase-5 objective pulls the models toward a similar final solution. In other words, once the model has learned how to extract useful information from the image, the common final objective appears to dominate the exact path used to get there.

The contrast with the Phase-5-only model makes this much clearer. Training directly on Phase 5 produces an SER of 105.20% and a macro F1 of only 0.96%, despite the training and validation

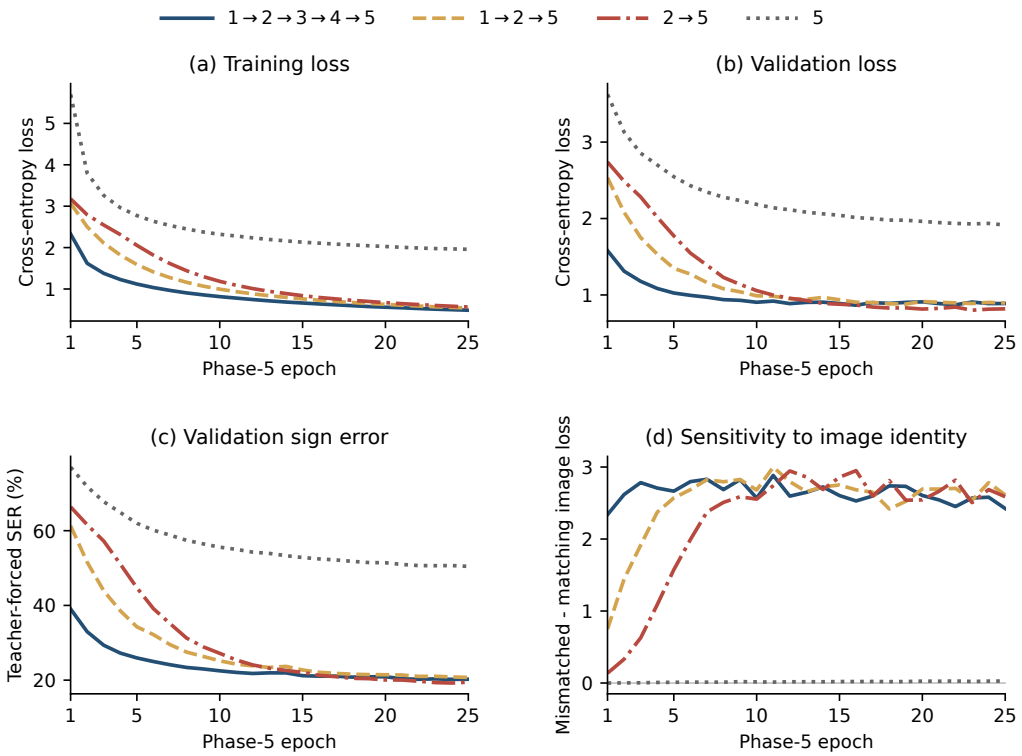


Figure 3: Phase-5 training curves for the four curricula: (a) training loss, (b) validation loss, (c) teacher-forced validation SER, and (d) the validation loss increase when images are mismatched with their transcriptions. A larger loss increase indicates greater sensitivity to image identity.

losses steadily decreasing. More importantly, its image-identity sensitivity remains essentially zero, which means it is not using the image tokens to determine the output. This strongly suggests that the model has learned to predict primarily from previously supplied target tokens only. The curriculum phases therefore seem to serve primarily as a mechanism for preventing this language-only shortcut and providing a more favorable initialization before full tablet-side training begins.

Localization The main difference between the earlier curriculum stages is in localization performance. The $1 \rightarrow 2 \rightarrow 5$ run performs best, achieving 78.27% peak attention in box and 50.00% attention mass in box. This suggests that sign-level training can substantially improve the model's ability to localize relevant visual regions. In contrast, the $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$ curriculum achieves slightly lower localization scores of 76.79% peak attention in box and 48.29% attention mass in box. One possible explanation is that training on synthetic images introduces a domain mismatch that makes localization on real tablet images more difficult.

Performance Ceiling and Error Sources The results indicate that there may be a shared performance ceiling imposed by several factors common to all three curricula. In addition to transcription complexity, performance may also be limited by the shared encoder-decoder architecture, autoregressive error propagation during inference, and imperfect localization of relevant visual regions. Because all successful curricula ultimately use the same Phase-5 data, model architecture, objective, and decoding procedure, these factors would affect each run similarly. Once sufficient sequence grounding has been established, such shared bottlenecks may dominate the remaining error and limit the benefit of additional curriculum stages.

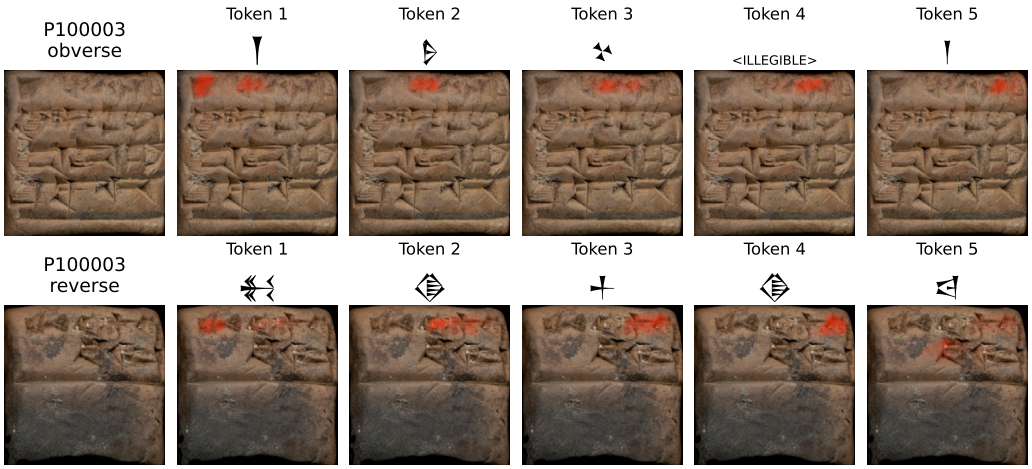


Figure 4: Generated-token attention for the full-curriculum epoch-25 model on P100003. Each row shows the original side and the first five generated tokens, including structural markers. Red overlays show final-layer cross-attention averaged across heads, conditioned on generated prefixes and scaled to each token's peak (FP32 greedy decoding; opacity 0.65). These qualitative examples complement the reference-conditioned evaluation.

8 Conclusion

We developed a single vision-language model for cuneiform tablet transcription and evaluated curricula spanning sign recognition, text modeling, and real tablet-side transcription. The signs $\rightarrow$ text $\rightarrow$ real-sides curriculum (1–2–5) achieves a strong balance between transcription and localization, with a mean tablet-side SER of 27.24%, a macro sign F1 of 52.24%, and peak cross-attention within the corresponding ground-truth sign region for 78.27% of annotated signs.

Our results demonstrate that full tablet transcription and spatial grounding can be learned within a single image-to-text model without explicit bounding-box supervision. This provides an alternative to conventional multi-stage cuneiform OCR pipelines, in which localization, recognition, and sequencing are optimized separately and errors can propagate between stages. The curriculum experiments further suggest that intermediate supervision at simpler visual and sequential scales can shape capabilities that are difficult to learn from full tablet images alone, although the contribution of individual synthetic stages remains unresolved. Future work should therefore isolate the effects of synthetic line and tablet-side training, investigate stronger mechanisms for spatial grounding, and evaluate whether the same curriculum principles transfer across periods, scripts, and other low-resource document transcription settings.

Overall, these findings support single-model image-to-sequence transcription as a viable alternative to multi-stage OCR pipelines for tasks that require joint visual recognition, spatial reasoning, and sequence prediction in low-resource domains.

Ethics Statement

The study uses cuneiform images and transcriptions associated with CDLI, eBL, and Sign Snippets. Their provenance and reuse conditions should be respected when distributing derived data or model artifacts. Autoregressively generated transcriptions can hallucinate and should be treated as aids to scholarly work rather than authoritative editions of cultural heritage material.

Reproducibility Statement

Section 3.1 and Appendix A describe data sources, filtering, normalization, source-level splits, and synthetic rendering. Sections 5 and 6 specify the model, phase schedule, checkpoint selection, metrics, and localization holdout. Appendix B records implementation and artifact identities. The reported comparisons use the fixed manifests and seed described in Section 5.

AI Use Statement

We used OpenAI Codex to assist with manuscript structure and editing, LaTeX tables, dataset figure code, and discussion of the reported comparisons. It was also used to research existing techniques, related works, and brainstorm potential research ideas. We take responsibility for the final text, figures, code, and numerical claims.

References

Anonymous. Anonymous concurrent submission, 2026. Under review.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009. doi: 10.1145/1553374.1553380.

CDLI contributors. Cuneiform Digital Library Initiative (CDLI), 2026. URL https://cdli.earth/.

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, 2026. URL https://arxiv.org/abs/2606.22608. arXiv:2606.22608.

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.

Cheng Cui et al. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model, 2025. URL https://arxiv.org/abs/2510.14528. arXiv:2510.14528.

Tobias Dencker, Pablo Klinkisch, Stefan M. Maul, and Bj"orn 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.

electronic Babylonian Library. electronic Babylonian Library (eBL). https://www.ebl.lmu.de/.

Hunyuan Vision Team et al. HunyuanOCR technical report, 2025. URL https://arxiv.org/abs/2511.19575. arXiv:2511.19575.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2111.15664.

Or Lewenstein et al. A large-scale dataset of annotated cuneiform sign images for digital palaeography. *Journal of Open Humanities Data*, 12:53, 2026. doi: 10.5334/johd.503. URL https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.503.

Gengluo Li et al. Towards real-world document parsing via realistic scene synthesis and document-aware training, 2026. URL https://arxiv.org/abs/2603.23885. arXiv:2603.23885.

Ze Liu, Han Hu, Yutong Lin, et al. Swin transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.09883.

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*, 2025. URL https://openreview.net/forum?id=XHTirKsQV6.

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, 2026. URL https://arxiv.org/abs/2608.18544. arXiv:2608.18544.

Alex Shonenkov, Denis Karachev, Maxim Novopoltsev, Mark Potanin, and Denis Dimitrov. StackMix and blot augmentations for handwritten text recognition, 2021. URL https://arxiv.org/abs/2108.11667. arXiv:2108.11667.

Ernst St"otzner, Timo Homburg, and Hubert Mara. CNN based cuneiform sign detection learned from annotated 3D renderings and mapped photographs with illumination augmentation, 2023. URL https://arxiv.org/abs/2308.11277. ICCV Workshops.

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, 2025. URL https://arxiv.org/abs/2507.13959. arXiv:2507.13959.

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, 2023. URL https://arxiv.org/abs/2306.01268. arXiv:2306.01268.

Hao Wu et al. FireRed-OCR technical report, 2026. URL https://arxiv.org/abs/2603.01840. arXiv:2603.01840.

A Dataset construction details

A.1 Unicode normalization

We construct the Unicode mapping by extending the `Cdli2cuneiform` workflow[1], which maps transliterated sign names and values to Unicode cuneiform characters. Candidate mappings are aggregated from Akkademia, the Oracc Sign List, CDLI, CuneiML, Nuolenna, Text-Fabric, and Wikidata. Candidate outputs are normalized by removing spaces and retaining only valid cuneiform code points. For each source entry, the most frequent valid Unicode output is selected. Ties are resolved pseudorandomly from sorted candidates using seed 0. Reading labels are subsequently lowercased, and duplicate labels are reduced using the same seeded procedure.

This procedure provides a deterministic mapping for model training but does not constitute a philological resolution of disagreements between sign lists. The saved mapping artifact is therefore treated as the canonical mapping used throughout the experiments. The tokenizer adds <LINE>, <COLUMN>, <ILLEGIBLE>, and <GAP> as structural tokens and <PAD>, <BOS>, <EOS>, and <UNK> as model-control tokens. Validation and transcription test sets contain 402 and 403 distinct sign characters, respectively, with no sign classes outside the training inventory.

[1] https://github.com/HarrisonLiruiHuang/Cdli2cuneiform

A.2 SIGN SNIPPETS FILTERING

We retain Sign Snippets examples assigned to the Ur III period and use Electronic Babylonian Literature (eBL) resources to recover their source images and catalog identifiers. After lowercasing and trimming `signName` and `value`, we remove bracketed sign names and names equal to `unknown`, `struct`, or `x`. We also remove rows whose two nonempty labels differ as strings and rows whose sign name cannot be mapped to Unicode. String inequality is a filtering heuristic, not evidence that the annotation is incorrect: different readings can map to the same sign.

Mappings between sign images and Unicode labels were visually inspected using a custom Streamlit interface. eBL identifiers were subsequently mapped to CDLI P-numbers and linked to their corresponding source tablets. These source identities are retained throughout dataset generation to prevent sign crops originating from held-out tablets from entering training examples.

Reapplying these filters to the saved full Sign Snippets metadata export begins with 158,946 records. Filtering removes 4,681 bracketed sign names, 755 `unknown`/`struct`/`x` entries, and 53,095 unequal name–value pairs, leaving 100,415 metadata records before Ur III and final source-eligibility filtering. Of the unequal pairs, 50,050 map to the same Unicode under our mapping. The final study manifest contains 4,964 retained Ur III sign crops across the training and validation splits.

The coverage comparison is restricted to tablets represented by retained sign rows in the train/validation manifest; it does not estimate coverage for all eBL tablets. The denominator counts visible sign positions in eBL text lines, excluding broken-away signs, gaps of unknown length, and structural markers. Path-aligned coverage counts distinct positions with a box at the same eBL transcription path; stale paths may lower this estimate. Boxes without an identifiable side or a visible-sign denominator are omitted from the corresponding ratios.

A.3 CDLI ATF PARSING

CDLI catalog records provide ATF transcriptions and metadata, while composite photographs are retrieved by P-number. We restrict the study dataset to Ur III tablets and match photographs to transcription records using these identities. A paired example requires an available photograph, a non-empty side transcription, and period metadata.

We parse ATF side directives for the obverse, reverse, left, right, top, and bottom surfaces and normalize recognized spelling and punctuation variants in surface names. Document headers, comments, and other non-transcription metadata are removed. Editorial line numbers, including primed, ranged, and subdivided labels, are stripped, and indented continuation text is joined to its preceding line.

Transliterations are segmented at hyphens and determinative braces while retaining the enclosed determinative text. Editorial punctuation, inline status annotations, and allograph suffixes are removed before readings are lowercased. Tokenization preserves parenthesized numeric expressions and pipe-delimited compound sign names while separating compound operators outside these groups.

Parsed rows are separated by `<LINE>`, while column directives are represented using `<COLUMN>`. Bracketed spans and isolated `x` readings are mapped to `<ILLEGIBLE>`. ATF status markers explicitly describing illegibility are likewise represented by `<ILLEGIBLE>`, whereas broken, damaged, effaced, missing, or trace-only content is represented by `<GAP>`. Normalized sign readings are then converted to Unicode, with unmapped labels represented by `<ILLEGIBLE>`. We retain tablet-side targets containing between 1 and 100 Unicode cuneiform signs, excluding structural markers from the sign count.

A.4 CDLI TABLET-SIDE EXTRACTION

Although the transcription parser recognizes multiple tablet surfaces, automatic image extraction is restricted to obverse and reverse sides. Composite photographs are first converted to grayscale for region detection. To identify photographs with the expected black background, we inspect a square patch at each corner with side length equal to the larger of 16 pixels or 8% of the shorter image dimension. An image is retained only when at least 95% of pixels in every corner have grayscale intensity at most 50 on a 0–255 scale.



Figure 5: Example synthetic tablet side assembled from five synthetic lines. The sign crops come from real tablets, but the composition does not represent a physical tablet.

Pixels with intensity greater than 50 are treated as foreground, and 8-connected components are extracted using OpenCV. Components are retained when their foreground area occupies 0.8–70% of the complete image and their bounding-box width-to-height ratio lies between 0.10 and 12.0. The two largest eligible components are selected by foreground area. Assuming the standard CDLI composite layout, the upper component is assigned to the obverse and the lower component to the reverse; photographs with fewer than two eligible components are discarded.

Each selected component is cropped from the original color photograph and stored as a JPEG at quality 95. The foreground mask is used only to determine the crop boundary, so black pixels falling within the bounding rectangle are retained. Only successfully extracted sides with corresponding normalized transcription records are included in the manifest.

A.5 Synthetic rendering

Synthetic lines. Within each split and period, we divide CDLI transcriptions at line and column boundaries and retain sequences of 5–10 signs for which eligible sign crops are available. We sample a real crop for each sign and compose the crops in transcription order into a labeled synthetic line.

Synthetic tablet sides. Within each split and period, we interleave lines from different source tablets and group four to eight lines into a synthetic tablet side. These compositions do not reconstruct the discourse or layout of a physical tablet; their target sequences are joined with `<LINE>` tokens.

When duplicate tablet-side transcription records exist, we retain the record containing the largest number of Unicode signs, with stable ordering used to resolve ties. Lines containing remaining control tokens or signs without an eligible crop in the same split and period are discarded. For every sign in a retained sequence, a crop is sampled from the same split and period using seed 0 and sorted candidate lists. Crops may be reused between examples, and each eligible source line generates one synthetic image.

Each crop is converted to grayscale and intensity-normalized using its 5th and 95th percentiles, clipping the result to $[0, 1]$. Each normalized pixel intensity u deterministically scales the base clay color RGB $(176, 137, 96)$ by $0.55 + 0.70u$; constant-intensity crops use the base clay color directly. Crops are resized to a height of 128 pixels using Lanczos interpolation while preserving aspect ratio. They are placed horizontally in transcription order with up to 8 pixels of overlap, using linear blending at crop boundaries. Synthetic lines are stored as RGB JPEGs at quality 95 with chroma subsampling disabled.

Synthetic lines are shuffled within each source tablet, and source-tablet groups are interleaved using seed 0. The ordered lines are partitioned into groups of four to eight, subject to a nominal maximum

of 100 signs. Because individual lines contain at most 10 signs, an eight-line composition contains at most 80 signs. Lines are stacked vertically without additional resizing, horizontally centered, and separated by 8-pixel gaps with 12 pixels of outer padding on the same RGB $(176, 137, 96)$ background. Synthetic sides are stored as JPEGs at quality 95 with chroma subsampling disabled.

A.6 Source assignment and dataset accounting

The fixed transcription test set is inherited from a reference manifest and is not resampled. Every test P-number is excluded from training and validation, including other tablet sides, sign crops, and transcription donors from that tablet. After removing test and localization-holdout identities, 6,470 of 64,699 candidate source identities are assigned to validation and 58,229 to training. Only 4,398 validation identities occur in retained examples because final synthetic inclusion also requires eligible crops and transcriptions; 10% describes source assignment, not the final row proportion.

Validation assignment follows SHA-256 ranking but skips a candidate if removing it would eliminate the last training representative of a period or sign within a real-image level and period. It therefore preserves training coverage rather than taking an unconstrained hash-ranked split.

Each synthetic line records the P-number supplying its transcription and every sign crop; each synthetic side records the union of its constituent lines' identities. An example is retained only if all donors belong to its split and period. Decoder text warm-up uses only training-split transcriptions. The final manifest has zero pairwise source-identity overlap across training, validation, and transcription test splits.

Across all components, the training, validation, and test splits contain 58,229, 4,398, and 2,485 unique recorded source P-numbers, respectively. These counts include transcription and sign-image donors and are not counts of photographed tablet sides. Considering only real CDLI tablet-side examples, the corresponding counts are 20,117, 2,198, and 2,485. The rebuilt manifest preserves all 4,964 eligible real sign crops and all 43,176 real tablet sides from the training/validation pool. It contains 145,418 synthetic lines and 18,178 synthetic sides across those splits; the frozen transcription test set contains 4,805 sides.

In the full curriculum, synthetic training pools are capped at 10,000 lines and 10,000 synthetic sides, while all 4,520 training sign crops and 38,920 real training tablet sides are used. The image-training pool therefore contains 63,440 distinct examples before repeated epochs. Text-only decoder warm-up uses the uncapped training transcription pool. Manifest totals are not the number of examples observed in every curriculum stage.

Before phase caps, the training sign crops cover 167 sign classes and each synthetic training pool covers 166; real training sides cover 470 of the combined 471 classes. The intermediate image sources therefore cover only part of the output vocabulary.

A.7 Label coverage and uncertainty

The training manifest contains 2,855,151 cuneiform sign-token occurrences, compared with 173,947 in validation and 131,277 in the transcription test set. These totals include repeated textual content appearing in synthetic lines and synthetic tablet sides.

Table 6 summarizes uncertainty markers for the real CDLI tablet-side subset. Percentages are computed over cuneiform signs together with the four structural-marker classes, excluding model-control tokens. Because `<ILLEGIBLE>` represents both damaged or unknown content and readings that could not be mapped to Unicode, its frequency should not be interpreted solely as a mapping-failure rate.

A.8 Localization holdout details

The 67-tablet localization holdout was reserved from 133 tablets containing labeled sign coordinates. All holdout tablets have resolved CDLI identities, including documented joins, stored in an exclusion ledger applied before synthetic generation and train–validation assignment. The exclusions cover real tablet sides, sign crops, synthetic components, and transcription donors.

Table 6: Label uncertainty in real CDLI tablet-side transcriptions.

Split	Signs	Illegible	Share	Gap	Share
Train	1,084,771	76,007	5.68%	2,244	0.17%
Validation	118,849	7,933	5.43%	234	0.16%
Test	131,277	9,142	5.65%	264	0.16%

The complete frozen localization population contains 283 artifacts, of which 279 were processed and four remain unresolved. The 67-tablet subset contains 178 side records. Ninety-nine sides have no eligible localization targets, leaving 79 sides with 882 eligible annotated positions; three out-of-vocabulary positions are excluded for evaluation of the final full-curriculum model, leaving 879 scored positions. No retained localization side exceeds the model context limit.

Localization uses separate eBL side crops. Verified sign anchors identify a unique connected photograph region; explicit surface annotations are used when available. The region is expanded to retain complete sign boxes. Stale transcription paths, changed readings, conflicting or ambiguous assignments, unresolved crops, and sides with transcription issues are excluded from scoring.

Attention is evaluated in FP32 with the canonical reference sequence, preserving out-of-vocabulary positions as <UNK>. Final-layer head-averaged attention is bilinearly upsampled to the input canvas, cropped to the resized image content, and renormalized to sum to one. Peak-in-box uses the maximum pixel's center; mass-in-box weights boundary pixels by fractional box overlap. Both metrics average equally over the 879 scored positions. Uniform attention over image content yields a mean mass-in-box of 3.58%.

Forty-four localization-holdout P-numbers occur in the transcription test set, corresponding to 83 transcription-test sides. These sides remain in the frozen transcription benchmark, and none of their known source identities occur in the final training or validation manifest. Exclusions apply to datasets rebuilt after the localization reservation; earlier-model localization measurements are diagnostic unless their training provenance establishes the same isolation. Held-out claims are associated with the exact training manifest of each checkpoint.

Source-level isolation controls overlap through known catalog identities and documented joins, but does not guarantee independence from unidentified catalog aliases or external image-encoder pretraining. Tablet-level separation also does not imply sequence-level uniqueness: recurring expressions can appear on distinct tablets, and identical textual subsequences are not removed across splits.

B Model implementation and reproducibility

B.1 Disabling Stage 4 spatial downsampling

Standard SwinV2 patch merging concatenates the feature vectors from each non-overlapping 2×2 neighborhood, then applies a learned linear reduction and normalization. At the entrance to Stage 4, this maps an $H \times W \times 384$ feature map to $(H/2) \times (W/2) \times 768$.

During Phases 3-5, we replace neighborhood concatenation with four channelwise copies of the feature vector at each spatial position. For a vector $x_{h,w} \in \mathbb{R}^{384}$, the modified operation is

$$y_{h,w} = \mathrm{LayerNorm}(W_{\mathrm{merge}}[x_{h,w}; x_{h,w}; x_{h,w}; x_{h,w}]) \,,$$

where W_{merge} is the pretrained linear reduction from 1,536 to 768 channels. The reduction and normalization retain their pretrained parameters at initialization. This preserves the $H \times W$ spatial grid while producing the 768 channels expected by the Stage 4 transformer blocks. For a 512×1024 input, the final grid is therefore 32×64 rather than 16×32 (Table 2). Phase 1 retains standard patch merging.

B.2 Projection, position embeddings, and training

The feature adapters and projection layers use LayerNorm, a linear map, and dropout. Each visual pathway has learned row and column embeddings initialized to zero; the decoder has learned

sequence-position embeddings. Phase 1 classifies mean-pooled encoder features with a separate head. Valid visual tokens are packed before decoder attention. Last-epoch weights carry between phases, while validation-best checkpoints are retained separately.

B.3 CHECKPOINTS AND DECODING

All reported curriculum comparisons use the checkpoint from epoch 25 of Phase 5. Evaluation uses PyTorch `2.11.0+cu128` and batch-one greedy decoding in BF16, suppressing `<PAD>` and `<BOS>` after initialization. Each experiment retains its dataset manifest, vocabulary, training history, and evaluation summaries, including checkpoint hashes and decoding settings.

B.4 DATA VERSIONS AND PROVENANCE

The reported curricula share an identical dataset manifest and evaluation tokenizer. SHA-256 fingerprints of the manifest and tokenizer, including normalization settings, are recorded with the evaluation results. The Sign Snippets dataset has a December 2025 Zenodo archive (`https://doi.org/10.5281/zenodo.17949595`), distinct from the 2026 article (Lewenstein et al., 2026); its image crops are distributed under CC BY-NC 4.0. Localization uses a fixed eBL annotation export dated 11 September 2026, cached transcriptions and photographs, and recorded holdout membership and source exclusions.

C TRAINING COST AND SUPPLEMENTARY DIAGNOSTICS

Table 7 reports saved Phase-5 diagnostics. GPU-hours sum completed training calls through epoch 25 across all included phases on one RTX 5070 Ti. They exclude separate validation, evaluation, checkpoint writing, preparation, and external pretraining, and are distinct from total run wall time.

Structural micro-F1 pools true positives, false positives, and false negatives for the four structural markers after unit-cost alignment retaining signs and markers; matches require identical marker types with no positional tolerance. Generation-limit counts identify sides reaching 512 tokens without `<EOS>`. These diagnostics capture structure and termination differences excluded from sign-only metrics.

Table 7: Training cost and free-generation diagnostics on the 4,805 test sides. Structural micro-F1 is a percentage; the final column counts sides.

Curriculum	GPU-hours	Structural micro-F1	Generation limit
$1 \to 2 \to 3 \to 4 \to 5$	7.765	55.43	65
$1 \to 2 \to 5$	6.766	64.42	34
$2 \to 5$	8.618	75.26	15