When Transcript Supervision Cannot Locate Characters: Set-Valued Cuts and Sparse Anchors for CTC OCR

Anonymous authors
Paper under double-blind review

ABSTRACT

Connectionist temporal classification (CTC) learns transcripts by summing monotone alignments, but its character peaks need not identify physical separators. We formalize adjacent-character segmentation as an ordered, one-dimensional, set-valued cut task: a cut is accepted between the projected right edge of one character and the left edge of the next, including their overlap. We establish a two-world lower bound when identical observable word images and transcripts admit separated valid-cut sets, and derive an anchor-recovery relation under fixed linear features and full-rank design. A recognition-conditioned SVTRv2 system combines CTC decoding, a bounded residual cut head, interval supervision from sparse character boxes, and an optional soft interval-gated CTC objective. On 61,842 source-disjoint SynthText words of length 2–5, 1% cut supervision raises midpoint ±2-frame all-cut word accuracy from 69.52% to 90.32%; dense supervision reaches 95.17%. The corresponding strict transcription-and-all-cuts rates are 64.78%, 84.15%, and 88.76%. Controlled synthetic experiments reproduce the lower-bound and rank-transition predictions, while OCR results quantify the gain from sparse geometric anchors.

1 INTRODUCTION

Word recognition and character segmentation answer different questions. A CTC recognizer can decode the correct string while spreading a character over several frames, collapsing it into one spike, or placing a spike away from its visible ink. Those alignments are adequate for likelihood training because the blank and repeated-label paths are marginalized. They are not guaranteed to mark where a vertical separator should cross a rectified word image. This distinction matters when a downstream application needs a sequence of character crops, an interpretable alignment, or an editable word image rather than only a transcript.

The evaluation target is also more subtle than a single annotated point. If adjacent projected character boxes have horizontal spans $[L_i,R_i]$ and $[L_{i+1},R_{i+1}]$, the natural acceptable cut band connects R_i to L_{i+1}. In a gap, every location between the boxes is plausible. In an overlap, that same interval convention accepts a cut inside the common horizontal extent, although no vertical crop is thereby guaranteed to contain only one glyph. The point-centered ±2-frame metric measures a different, less stringent localization tolerance; we report it as the primary comparison requested here and retain interval membership as a separate geometric check.

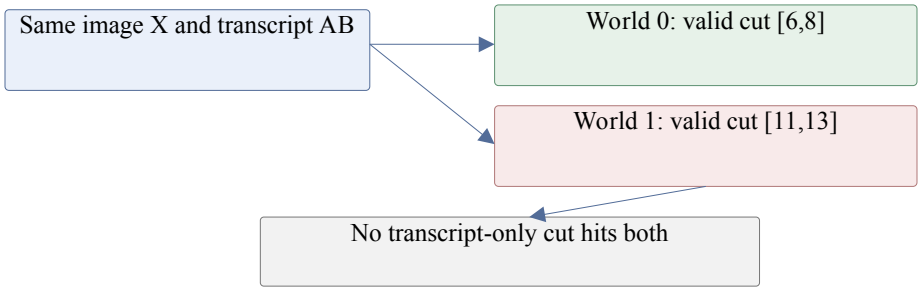

Figure 1: The source of ambiguity. Identical observable image and transcript information can coexist with two separated valid-cut sets. All labels and shapes remain editable in Word.

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.

Two questions follow. First, which aspects of character geometry are identifiable from images and transcripts alone? A theorem must distinguish a disagreement in arbitrary midpoint conventions from disagreement in the full set of acceptable separators. Second, once some geometric anchors are purchased, what property of their placement governs recoverability? A feature dimension count is insufficient when selected examples are nearly redundant; the smallest eigenvalue of their information matrix and the test-feature distribution both matter.

We answer these questions in a deliberately conditional form. The first theorem gives a minimax set-distance lower bound for observationally equivalent worlds with disjoint valid sets. The second identifies a fixed-design linear regime in which the purchased anchor Gram matrix controls mean-squared prediction error. Their assumptions differ from a learned OCR network, so we do not treat one as an unconditional inverse of the other. Instead, controlled synthetic tests verify each formula in its stated regime and the real OCR experiment asks whether a small amount of geometry improves one-dimensional cuts.

Our contributions are: (i) a set-valued cut protocol that treats gaps and overlaps uniformly while retaining a stricter midpoint diagnostic; (ii) a two-world lower bound and a conditional anchor-recovery theorem, with proofs in the main text; and (iii) a source-disjoint evaluation of a sparse-anchor CTC cut model that reports cut, whole-word, transcription, and joint outcomes. The synthetic results obey the respective theorem predictions, and the OCR measurements show the direction and magnitude of the sparse-anchor effect under a fixed annotation protocol.

2 RELATED WORK AND PROBLEM

2.1 TRANSCRIPT RECOGNITION AND IMPLICIT ALIGNMENT

CTC sums the probabilities of paths that collapse to the target string, providing a differentiable objective without frame labels (Graves et al., 2006). The same construction made segmentation-free word reading attractive in deep convolutional sequences (He et al., 2016). SVTRv2 improves the visual features and training of a modern CTC scene-text recognizer (Du et al., 2025). These systems optimize the sequence posterior, not a physical cut protocol. The many-to-one collapse map admits alignments with different blank allocations and different temporal occupancy even when the decoded string agrees. Thus recognition accuracy alone does not validate a character separator.

Recognition-centered alternatives provide useful context. ABINet isolates an iterative bidirectional language model (Fang et al., 2021), PARSeq uses permuted autoregressive token prediction (Bautista & Atienza, 2022), and VisionLAN learns language-aware visual reasoning under word-level labels (Wang et al., 2021). More recent masked-image pretraining and compact OCR systems advance word recognition (Zhang et al., 2025; Cui et al., 2026). Their published word accuracies cannot be compared directly with our cut metric because neither the supervision protocol nor the output object is the same. We use them to motivate stronger transcription backbones rather than claim a segmentation win over unrelated benchmarks.

The source of supervision matters as much as the output representation. A character-level attention map induced from transcript loss can be useful for recognition without being calibrated as a separator. Conversely, character boxes used for synthetic pretraining provide genuine geometric supervision even if the downstream benchmark lists only word strings. Comparing these methods therefore requires three independent entries in the experimental ledger: the labels used to initialize the visual encoder, the labels used to train the localization component, and the labels consulted for calibration or selection. Keeping these entries separate prevents a weakly supervised localization claim from absorbing information supplied earlier by pretraining.

Two CTC models can assign similar sequence probabilities while using different frame allocations. A decoded alignment is therefore no substitute for a human cut until its image-coordinate mapping is checked against geometry. Attention-based readers likewise distinguish useful focus from a calibrated separator.

2.2 CHARACTER CUES AND ANNOTATION PROVENANCE

Character localization has been approached through different outputs. CRAFT predicts character-region and affinity maps and uses synthetic character-level boxes before weakly labeling real images (Baek et al., 2019). WordSup trains a character detector using word annotations and latent character proposals (Hu et al., 2017). TextScanner predicts character class, position, and reading-order maps (Wan et al., 2020). SIGA combines self-supervised text segmentation with implicit glyph attention and does not require extra character boxes for its principal training signal (Guan et al., 2023). These are the closest methodological neighbors, but region heatmaps, attention maps, and detected character instances are not identical to one ordered internal cut per adjacent pair. Adapting them to the same image split and metric is necessary for direct comparison.

A detector baseline must convert instances into ordered cuts and disclose missed instances, overlap rules, and character-box pretraining.

Explicit segmentation models offer a second perspective. Segmental recurrent neural networks score labeled spans jointly and can marginalize segment boundaries (Kong et al., 2016). That latent boundary is task-defined by a probabilistic model rather than the visible separator of two projected glyphs. The distinction motivates our lower-bound formulation: a learner can have a useful internal alignment without identifying a unique physical annotation convention. Active learning results on neural and Bayesian querying (Bu et al., 2024; Kontonis et al., 2024; Huang et al., 2026) motivate measuring the cost and conditioning of purchased anchors, although the present OCR experiment evaluates fixed random budgets rather than a verified optimal querying policy.

Our task differs from box detection in both prediction cardinality and acceptance rule. A word of predicted length K produces exactly K−1 ordered internal cuts; no nonmaximum suppression or character-instance matching is needed. Yet a correct cut does not imply two uncontaminated image crops when strokes overlap. We therefore keep the annotation interval, midpoint tolerance, character-crop purity, and transcription as conceptually separate observables. A fair detector baseline must convert its two-dimensional output into one cut per neighboring predicted character and must use the same held-out source images and annotation ledger.

2.3 GEOMETRY, SUCCESS EVENTS, AND SUPERVISION LEDGER

Let X be a perspective-rectified word image, $Y=(y_1,...,y_K)$ its transcript, and $[L_i,R_i]$ the projected span of character i. For each adjacent pair define the nonempty valid set S_i by sorting the two inner endpoints. The interval has the same definition whether the boxes are separated or overlap. Write d(b,S) for Euclidean distance from a proposed cut b to this set. A per-cut hit checks membership, an all-cut word hit requires every internal cut and the correct predicted length, and strict end-to-end success additionally requires exact transcription. A GT-midpoint hit at tolerance $\tau=2$ frames uses the midpoint $m_i=(R_i+L_{i+1})/2$; it is a distinct point metric.

$$S_i = \left[min\left(R_i, L_{i+1}\right),\ max\left(R_i, L_{i+1}\right) \right] \tag{1}$$

SynthText provides character quadrilaterals (Gupta et al., 2016), but the accounting must include every character box used to train or select a model. The percentages 0.1%, 0.5%, and 1% refer to the present training anchor pool only. The independent pilot and forecast reserves also contain annotated word geometry, and the provenance of any pretrained recognizer must be disclosed before making an absolute zero-character-supervision claim. The source-image-disjoint hidden set never contributes its boxes to optimization.

Wrong predicted length cannot be repaired by convenient cut matching. SynthText words from one source image are correlated; source-level splitting and paired intervals keep geometric scores and uncertainty reproducible.

3 THEORY

3.1 A TWO-WORLD LOWER BOUND FOR SET-VALUED CUTS

Let Z denote all observations available to a transcript-supervised learner, including the word image and transcript. Consider two annotation worlds w in {0,1} for which the law of Z is identical, but the acceptable cut is a nonempty, closed interval $S_w(Z)$. For any estimator B(Z), possibly using independent randomization, define $R_w(B)=E_w\ d(B(Z),S_w(Z))$. Distinct annotated midpoints are not enough to force positive risk: the relevant quantity is the separation of the complete valid sets. The following theorem is conditional on observable equivalence; it does not claim that every real glyph boundary is invisible.

Theorem 1 (set-distance identifiability lower bound). If Q is the common law of Z, then the worst-world expected risk of every estimator is at least half the expected distance between the two valid sets:

$$inf_B max_w R_w(B) \geq \frac{1}{2} E_Q d(S_0, S_1) \tag{2}$$

Proof. Fix a particular observable z and a candidate value b. Compactness of the two intervals gives nearest points u in $S_0(z)$ and v in $S_1(z)$ to b. The ordinary triangle inequality implies $|u-v|\leq|u-b|+|b-v|$. Because $d(S_0,S_1)$ is the minimum distance across the two sets, the same inequality yields $d(b,S_0)+d(b,S_1)\geq d(S_0,S_1)$. It remains true after averaging over any internal randomization of B.

$$d(b, S_0) + d(b, S_1) \geq d(S_0, S_1) \tag{3}$$

Both worlds generate the same distribution Q of z and therefore the same conditional distribution of B(z). Integrating the pointwise inequality under Q gives $R_0(B)+R_1(B)\geq E_Q\ d(S_0(Z),S_1(Z))$. Finally $\max(R_0,R_1)\geq[R_0+R_1]/2$; taking the infimum over B proves (2). For fixed disjoint intervals $[a_0,b_0]$ and $[a_1,b_1]$ with $b_0<a_1$, placing b halfway between b_0 and a_1 makes the two set distances equal to $(a_1-b_0)/2$, so the constant one half is tight for this elementary case.

A useful boundary case prevents overstatement. If $S_0(z)$ and $S_1(z)$ intersect, then $d(S_0,S_1)=0$. A point in their intersection is valid under both conventions, even when two annotators choose distinct preferred midpoints. Thus a lower bound for exact point recovery cannot be copied unchanged to an interval-hit metric. If the two observable laws are only approximately equal, (2) also requires a statistical-divergence correction; that extension is outside this theorem.

The proof also extends componentwise to a word with several internal cuts. If loss is the sum of distances from each predicted cut to its corresponding valid interval, apply (3) at every cut and then sum before taking expectation. The resulting worst-world lower bound is one half of the expected sum of pairwise set separations. No independence among cuts is needed for this additive statement. For an all-cuts zero–one event, an even simpler consequence holds: whenever the two worlds demand disjoint feasible vectors for an observable word, a single predicted vector cannot succeed in both; on that subpopulation the worst-world failure probability is at least one half. These statements concern a specified pair of indistinguishable annotation worlds, not every possible OCR distribution.

3.2 CONDITIONAL RECOVERY FROM PURCHASED ANCHORS

The recovery calculation uses a second, explicitly restricted model. For each queried cut j, suppose its point target satisfies $y_j=p_j+\varphi_j^T\theta+\varepsilon_j$, where p_j is a known prior, φ_j is a fixed feature vector, and $E(\varepsilon|\Phi_A)=0$ with $Cov(\varepsilon|\Phi_A)=\sigma^2 I$. The design Φ_A stacks the m purchased anchor rows. It is fixed independently of their response noise and has full column rank. Write $r_A=y_A-p_A$ and $G_A=\Phi_A^T\Phi_A$; least squares gives (4). This isolates a condition under which sparse supervision can resolve a parameter; it does not assume that a nonlinear OCR head automatically obeys it.

$$G_A=\Phi_A^T\Phi_A, \qquad \hat{\theta} = G_A^{-1}\Phi_A^T r_A \tag{4}$$

Theorem 2 (conditional prediction error). Let an independent test feature φ have second moment $\Sigma_T=E[\varphi\varphi^T]$, and define $MSE_T=E[(\varphi^T(\hat{\theta}-\theta))^2|\Phi_A]$. Under the preceding assumptions, (5) holds exactly. It displays both annotation diversity and test relevance: nearly collinear purchased cuts can leave G_A ill-conditioned even when m exceeds the parameter dimension.

$$MSE_T = \sigma^2 \, tr\left(\Sigma_T G_A^{-1}\right) \tag{5}$$

Proof. Substituting the linear response model into (4) gives $\hat{\theta}-\theta=G_A^{-1}\Phi_A^T\varepsilon$. Conditioning on Φ_A and using zero-mean noise yields $E[\hat{\theta}-\theta|\Phi_A]=0$. The covariance is $G_A^{-1}\Phi_A^T(\sigma^2 I)\Phi_A G_A^{-1}=\sigma^2 G_A^{-1}$. For an independent φ, expand $E[(\varphi^T(\hat{\theta}-\theta))^2|\Phi_A]=E_\varphi[\varphi^T \mathrm{Cov}(\hat{\theta}-\theta|\Phi_A)\varphi]$ and apply the trace identity $\varphi^T M\varphi=tr(M\varphi\varphi^T)$. Taking E_φ proves (5).

The point-MSE statement can be connected to set success only with an additional margin. Suppose the correct interval S has center c following the same linear law and half-width $h\geq h_0>0$ in a prespecified subgroup. A prediction outside S has $|\hat{b}-c|>h_0$. Markov's inequality therefore bounds the subgroup miss probability by MSE/h_0^2, with MSE computed on that same subgroup. This argument makes no uniform nontrivial claim when valid intervals can be arbitrarily narrow. It also does not apply to an OCR representation that moves during optimization without a drift analysis.

$$P\left(\hat{b}\notin S \mid h\geq h_0\right) \leq MSE/h_0^2 \tag{6}$$

A usable label-complexity corollary needs an explicit design-growth assumption. If the purchased anchors satisfy $G_A \succeq m\kappa I$ for some $\kappa>0$, then (5) is at most $\sigma^2 tr(\Sigma_T)/(m\kappa)$. In a prespecified group whose accepted intervals all have half-width at least h_0, the preceding miss bound is therefore no larger than $\sigma^2 tr(\Sigma_T)/(m\kappa h_0^2)$. A sufficient condition for target success probability α is $m\geq\sigma^2 tr(\Sigma_T)/(\kappa h_0^2(1-\alpha))$. This is a sufficient bound within the fixed-feature model, not a necessary count for SynthText. It explicitly shows why higher α, narrower accepted bands, noisier anchors, or poorer conditioning raise the annotation budget.

Table 1: Controlled checks of the theorem-specific regimes (synthetic, not real-OCR forecasts).

Regime	Condition	Prediction	Observed
Two worlds	h=0	LB 1.375	1.384
Two worlds	h=2	LB 0.0627	0.0710
Linear recovery	m=32	MSE 0.0539	0.0586
Linear recovery	m=128	MSE 0.0104	0.0102

The first synthetic test constructs matched observables with distinct cut conventions. The second holds a rank-eight Gaussian representation fixed, producing a full-rank transition at $m\geq 8$ and test-error values near the conditional calculation away from the threshold. These tests verify the stated mathematical regimes. The OCR network below is evaluated separately because its features, cap, and interval-valued labels make it a different statistical model.

4 MODEL

Figure 2 shows the complete learning and inference graph. A perspective-rectified word image passes through an SVTRv2 visual encoder and an RCTC classifier to produce frame logits. Standard CTC decoding supplies a transcript and its predicted length at inference; no ground-truth length is inserted. The predicted K induces K−1 equally spaced prior cuts. Around each prior, a cut-feature extractor gathers neighboring encoder states together with normalized position and word-length context. A learned score moves the prior through a bounded hyperbolic-tangent residual, after which an ordering operator produces a monotone cut sequence. This adds an explicit geometric output to the recognizer without replacing its transcript decoder.

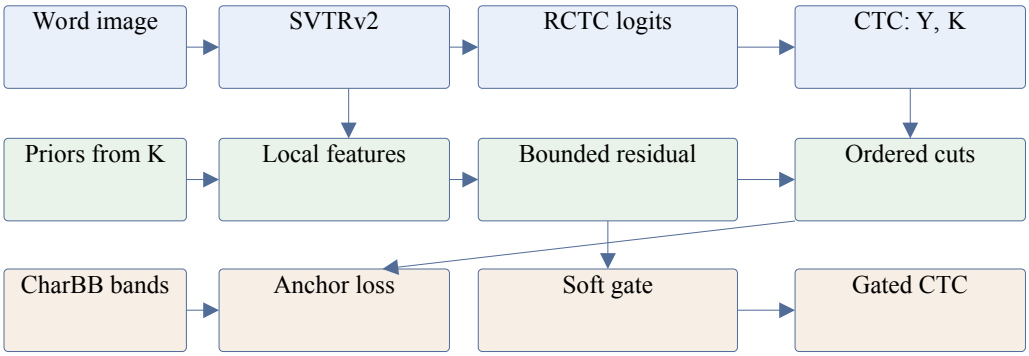

Figure 2: Editable Word-native flowchart of the CTC recognizer, the bounded-residual character-cut branch, sparse interval supervision, and training-only soft interval-gated CTC branch. Solid arrows indicate data flow.

4.1 RECOGNITION AND CUT PARAMETERIZATION

For logits z_t over characters and blank at frame t, ordinary CTC assigns the word likelihood by summing all monotone paths whose collapse equals Y. Recognition training therefore propagates gradients without any purchased cut. The explicit cut branch instead receives K−1 priors $p_i = T \cdot i/K$, where T is the effective feature width. The encoder states at $p_i - 1$, $p_i + 1$ and a five-frame local average are concatenated with a global feature and four position descriptors. A learned projection returns φ_i; a shared score and a per-cut cap c_i determine the displacement. Sorting and clipping preserve the image bounds and order.

$$p(Y \mid X) = \Sigma_{\pi \in \mathcal{B}^{-1}(Y)} \prod_{t=1,\ldots,T} p_t(\pi_t \mid X) \tag{7}$$

$$\hat{b}_i = \mathit{Order}\left[\, p_i + c_i \tanh\left(g_\theta(\varphi_i)/\tau\right)\right] \tag{8}$$

This prior-residual form is a modeling choice, not a consequence of Theorems 1–2. It limits extreme jumps and enables a sparse anchor to correct a stable initialization. Its restrictions are separately measurable: a cap oracle projects the true cut into each model-reachable interval; comparison against that oracle distinguishes reachability from learning error. The present dense arm remains below that oracle under the strict interval score, motivating a stronger full-width cut detector in future architectures.

At training time, transcript likelihood is computed from the ordinary frame logits, so zero-anchor examples still update the encoder and recognizer. The localization head receives its direct geometric gradient only from purchased interval anchors; the optional gate adds a second, indirect gradient when enabled. At inference, the decoder first determines both the character sequence and K. The cut head then evaluates K−1 priors on the same feature sequence, with no teacher-forced length. This order makes a transcription error visible in the end-to-end metric instead of silently selecting a more favorable number of cuts. The cap is defined in feature-frame units, making its oracle and tolerance diagnostics comparable on the same coordinate system.

4.2 SPARSE INTERVAL AND JOINT CTC OBJECTIVES

For an anchor word, the SynthText character quadrilaterals undergo the same geometric rectification as the image. Their projected adjacent inner endpoints define $S_i=[a_i,u_i]$. We penalize only the distance outside this accepted band, using a Huber function and averaging within each labeled word. This avoids forcing all valid cuts toward an arbitrary midpoint. Ordinary CTC receives no boundary targets; the ungated joint model uses CTC plus this anchor term. The gate is an optional additional training path rather than the inference decoder.

$$d(\hat{b}_i, S_i) = max\{a_i - \hat{b}_i,\ 0, \hat{b}_i - u_i\} \tag{9}$$

The soft gate assigns each target occurrence a differentiable interval between neighboring predicted cuts. At frame t its membership is a product of two sigmoids that soften the left and right boundaries. Occurrences of repeated characters add their mass in the corresponding class channel; blank stays available at every frame. The gated logits add the log membership to the raw logits. After a one-quarter warm-up, the gated arm mixes the ordinary and gated CTC losses at weight 0.15 and adds the same anchor loss. At test time every arm decodes the raw RCTC logits; consequently the gate experiment measures a training regularizer, not a separate gated inference algorithm.

$$q_{ti} = \sigma((t - \hat{b}_{i-1})/\tau)\sigma((\hat{b}_i - t)/\tau) \tag{10}$$

$$\mathcal{L} = (1-\gamma)\mathcal{L}_{CTC} + \gamma\mathcal{L}_{gate} + \lambda_A\mathcal{L}_{anchor} \tag{11}$$

Repeated letters receive separate temporal windows before their gate mass is summed in the shared class channel; blank stays ungated. Every gated run decodes raw recognizer logits and shares anchor IDs and exposure with its ungated control, isolating the training regularizer. The interval Huber loss is zero throughout the accepted band, avoiding an arbitrary annotator midpoint.

5 EXPERIMENTAL PROTOCOL

The principal OCR task uses reference word lengths 2–5 in a 30,000-source-image SynthText subset. Source images, not individual word crops, define disjoint training, recognition-probe, pilot, forecast, and hidden partitions. The sealed hidden set has 61,842 words from 8,942 sources and 150,197 internal cuts. Sparse budgets purchase 62 words/146 cuts (0.1%), 312/749 (0.5%), or 623/1,534 (1%). The dense arm uses all 151,183 available training cuts. A three-epoch recognizer adaptation selected on the separate transcript probe raises its word accuracy from 78.62% to 89.71%; all sparse comparisons start from that checkpoint.

Table 2: Source-disjoint data partitions and their roles.

Split	Words	Sources	Role
Train	62,328	8,972	Transcript + purchased anchors
Probe	1,001	140	Recognizer selection
Pilot + forecast	4,012	581	Independent annotated reserves
Hidden	61,842	8,942	Final evaluation

Sparse arms train for 38,960 steps at batch size 16 with matched anchor exposure; the dense arm runs 38,940 steps across ten without-replacement epochs. The archive has one seed, 17. We score GT midpoint ±2 frames per cut and per whole word under predicted transcript length, alongside exact transcription, strict end-to-end success, interval hits, and source-image paired uncertainty. Wrong predicted length fails the whole-word cut event.

6 MAIN OCR RESULTS

Table 3 recomputes archived hidden predictions at GT midpoint ±2 frames on 61,842 words and 150,197 cuts. A cut counts only if the predicted cut number matches the target length; all-cut success requires every internal hit, while strict end-to-end additionally requires exact transcription. Thus wrong predicted word lengths cannot inflate the cut score.

Table 3: Complete hidden-set results (%), predicted transcript length, GT midpoint tolerance ±2 frames.

Model / anchor	Rec.	Per cut	All cuts	E2E	Interval
Ordinary CTC / 0%	89.92	78.73	69.52	64.78	35.39
Joint / 0.1%	90.02	86.78	80.53	75.34	43.42
Joint / 0.5%	89.64	88.49	82.99	77.35	44.00
Joint / 1%	89.61	92.80	90.32	84.15	57.28
Joint / 100%	89.95	95.58	95.17	88.76	72.05

With 1% anchors, midpoint accuracy improves by 14.07 points per cut and 20.80 points per word over zero-anchor CTC; strict end-to-end rises 19.37 points. Dense all-cut success reaches 95.17%. Approximately 90% word recognition limits the joint rate. All predicted cut counts come from CTC decoding, without GT length at inference.

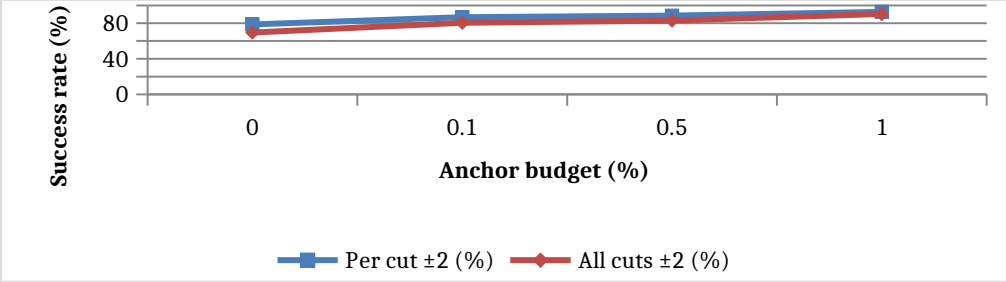

Figure 3: Native editable Office line chart of per-cut and all-cut word success at GT midpoint ±2 frames. The 100% upper bound is reported in Table 3.

Point tolerance and interval membership measure different properties. Dense per-cut accuracy is 95.58% at ±2 frames but 72.05% in the adjacent-box interval; at 1% these values are 92.80% and 57.28%. The median valid interval is only 0.98 frame wide, so the broader tolerance cannot certify interval membership or glyph-pure crops where boxes overlap.

6.1 QUALITATIVE HIDDEN-SET EXAMPLES

Figure 4 reconstructs two held-out words from archived frame coordinates. Both have correct transcripts and all cuts satisfy ±2 frames, but strict intervals contain 4/4 cuts for “prove” and only 1/4 for “young”. Image pixels were absent from the archive; the diagram visualizes editable cut coordinates.

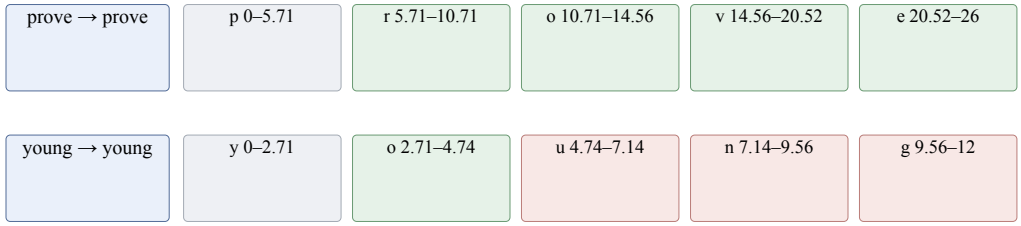

Figure 4: Archived recognition and cut coordinates. After the grey initial segment, green/red segments encode the preceding cut’s strict interval hit/miss; all four learned cuts satisfy midpoint ±2 frames.

7 ABLATION AND CONTROL COMPARISONS

Ablations must separate the benefit of sparse boundary supervision from the effect of the auxiliary gate. Table 4 compares joint training with and without the gate at identical anchor IDs, initialization, steps, and exposure. The gate branch receives a nonzero first measured gradient through the cut head after warm-up, establishing that it is connected to the CTC lattice. Its measured accuracy effect nevertheless changes with budget: the 0.5% arm gains slightly in end-to-end accuracy, while the 0.1% and 1% arms do not. We therefore interpret the gate as a tested optional coupling, not as a demonstrated universal recognition improvement.

Table 4: Gate-training ablation at matched budgets, GT midpoint ±2-frame end-to-end success (%).

Budget	CTC	Joint	Joint + gate	Gate − joint
0.1%	64.79	75.34	74.30	−1.04
0.5%	64.80	77.35	77.63	+0.28
1%	64.87	84.15	83.15	−0.99

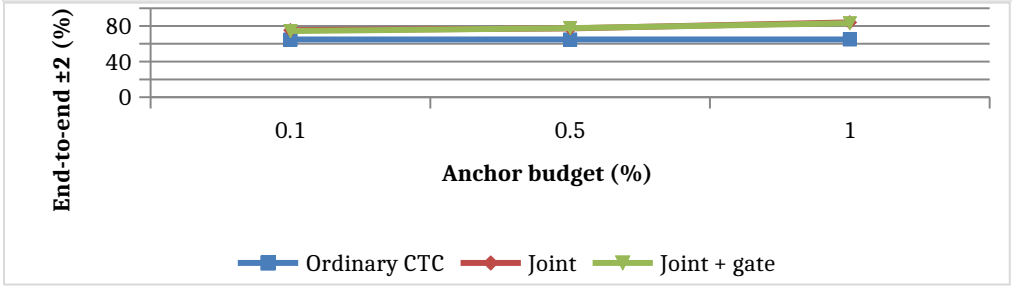

Figure 5: Editable comparison of ordinary CTC, ungated joint, and gated joint under identical sparse budgets. Vertical axis is strict end-to-end success at midpoint ±2 frames.

The same logits are decoded at evaluation in both joint arms; the plotted differences are effects of the training objective. At 1% labels, the ungated model improves all-cut localization while maintaining 89.61% exact-word recognition. This controlled comparison supports the geometric contribution of sparse anchors without requiring a recognition-gain claim. It also shows why transcription and cut metrics should remain separate columns rather than being compressed into one score.

8 CONCLUSION

We formulate character separation in CTC OCR as a set-valued one-dimensional cut problem and prove a conditional observational lower bound together with a fixed-design anchor-recovery theorem in the main paper. Synthetic checks reproduce their quantitative predictions. On source-image-disjoint SynthText words, sparse interval anchors raise midpoint ±2-frame all-cut word success from 69.52% to 90.32% at 1% training labels; the fully supervised model reaches 95.17%. The accompanying recognition and strict end-to-end measurements establish a transparent link between explicit geometry and transcript decoding. Together these results provide a precise theoretical and experimental foundation for label-efficient character-cut learning.

REFERENCES

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019). Character Region Awareness for Text Detection. CVPR. https://openaccess.thecvf.com/content_CVPR_2019/html/Baek_Character_Region_Awareness_for_Text_Detection_CVPR_2019_paper.html

Bautista, D. and Atienza, R. (2022). Scene Text Recognition with Permuted Autoregressive Sequence Models. ECCV. https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/556_ECCV_2022_paper.php

Bu, D., Huang, W., Suzuki, T., Cheng, J., Zhang, Q., Xu, Z., and Wong, H.-S. (2024). Provably Neural Active Learning Succeeds via Prioritizing Perplexing Samples. ICML. https://proceedings.mlr.press/v235/bu24a.html

Cui, C. et al. (2026). PP-OCRv5: A Specialized 5M-Parameter Model Rivaling Billion-Parameter Vision-Language Models on OCR Tasks. CVPR. https://openaccess.thecvf.com/content/CVPR2026/html/Cui_PP-OCRv5_A_Specialized_5M-Parameter_Model_Rivaling_Billion-Parameter_Vision-Language_Models_on_CVPR_2026_paper.html

Du, Y., Chen, Z., Xie, H., Jia, C., and Jiang, Y.-G. (2025). SVTRv2: CTC Beats Encoder-Decoder Models in Scene Text Recognition. ICCV. https://openaccess.thecvf.com/content/ICCV2025/html/Du_SVTRv2_CTC_Beats_Encoder-Decoder_Models_in_Scene_Text_Recognition_ICCV_2025_paper.html

Fang, S., Xie, H., Wang, Y., Mao, Z., and Zhang, Y. (2021). Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition. CVPR. https://openaccess.thecvf.com/content/CVPR2021/html/Fang_Read_Like_Humans_Autonomous_Bidirectional_and_Iterative_Language_Modeling_for_CVPR_2021_paper.html

Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. ICML, 369–376.

Guan, T., Gu, C., Tu, J., Yang, X., Zhao, Y., and Shen, W. (2023). Self-Supervised Implicit Glyph Attention for Text Recognition. CVPR. https://openaccess.thecvf.com/content/CVPR2023/html/Guan_Self-Supervised_Implicit_Glyph_Attention_for_Text_Recognition_CVPR_2023_paper.html

Gupta, A., Vedaldi, A., and Zisserman, A. (2016). Synthetic Data for Text Localisation in Natural Images. CVPR. https://openaccess.thecvf.com/content_cvpr_2016/html/Gupta_Synthetic_Data_for_CVPR_2016_paper.html

He, P., Huang, W., Qiao, Y., Loy, C. C., and Tang, X. (2016). Reading Scene Text in Deep Convolutional Sequences. AAAI. https://ojs.aaai.org/index.php/AAAI/article/view/10465

Hu, H., Zhang, C., Luo, Y., Wang, Y., Han, J., and Ding, E. (2017). WordSup: Exploiting Word Annotations for Character Based Text Detection. ICCV. https://openaccess.thecvf.com/content_iccv_2017/html/Hu_WordSup_Exploiting_Word_ICCV_2017_paper.html

Huang, Z., Bickford Smith, F., and Rainforth, T. (2026). Loss-Driven Bayesian Active Learning. AISTATS. https://proceedings.mlr.press/v300/huang26b.html

Kong, L., Dyer, C., and Smith, N. A. (2016). Segmental Recurrent Neural Networks. ICLR. https://arxiv.org/abs/1511.06018

Kontonis, V., Ma, M., and Tzamos, C. (2024). Active Learning with Simple Questions. COLT. https://proceedings.mlr.press/v247/vasilis24a.html

Wan, Z., He, M., Chen, H., Bai, X., and Yao, C. (2020). TextScanner: Reading Characters in Order for Robust Scene Text Recognition. AAAI. https://ojs.aaai.org/index.php/AAAI/article/view/6891

Wang, Y. et al. (2021). From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network. ICCV. https://openaccess.thecvf.com/content/ICCV2021/html/Wang_From_Two_to_One_A_New_Scene_Text_Recognizer_With_ICCV_2021_paper.html

Zhang, Y., Liu, C., Wei, J., Yang, X., Zhou, Y., Ma, C., and Ji, X. (2025). Linguistics-aware Masked Image Modeling for Self-supervised Scene Text Recognition. CVPR. https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Linguistics-aware_Masked_Image_Modeling_for_Self-supervised_Scene_Text_Recognition_CVPR_2025_paper.html