AGREEMENT IS NOT CONFIDENCE: TRAJECTORY AGREEMENT DEGRADES AS TRAINING CONVERGES

Anonymous authors
Paper under double-blind review

ABSTRACT

When a model's logits were never logged, a cheap substitute for confidence is how often the checkpoints saved during training agree with the final prediction. The same signal drives a published selective-prediction method that weights late disagreement most heavily. We show that this signal carries less information the closer the checkpoints are to the end of training. Two elementary bounds make this precise. The calibration error of trajectory agreement is pinned to within δ_W of the model's error rate, where δ_W is the mean disagreement in the checkpoint window. Its AUROC for error detection is at most one half plus the largest class-conditional probability that some checkpoint in the window disagrees with the final prediction. As the window approaches the final checkpoint, agreement tends to the constant 1: its calibration error tends to the error rate and its AUROC to one half. On 3D visual grounding (ScanRefer and Nr3D, four training runs) and CIFAR-10 with ResNet-18 (three seeds), moving a fixed-size window later raises calibration error and lowers AUROC in 13 of 14 run-by-window-size cells, with bootstrap intervals excluding zero; in the last-checkpoint window, 89–94% of errors receive agreement 1. Temperature-scaled softmax and a three-seed deep ensemble on the same rows are calibrated (ECE 0.031 and 0.041 against 0.228 for agreement, on ScanRefer), so the failure belongs to the construction, not to the model. For the published method, late weighting helps where accuracy keeps improving and hurts where it stalls, and its late limit falls to AUROC 0.51–0.54 in every run.

1 INTRODUCTION

A confidence score is only as useful as the information it can carry about errors. Many deployed systems do not expose the object that carries it best. Identifier-token 3D language models answer with an object identifier token, and the distribution over scene objects is not part of their output (Huang et al., 2024); black-box language models are queried through sampled text (Wang et al., 2023; Manakul et al., 2023; Xiong et al., 2024). The common response is a logit-free proxy: agreement among several predictions stands in for the probability of being right.

The cheapest source of several predictions is a single training run. Saving the model's predictions at every checkpoint costs no extra training, and the fraction of checkpoints that agree with the final prediction looks like a confidence. Selective Prediction via Training Dynamics (SPTD) turns this into a method: it rejects test points whose intermediate predictions disagree with the final one, weighting disagreement by $(t/T)^k$ so that late checkpoints count most (Rabanser et al., 2025). The intuition is sound for early training. Easy examples are learned first and stay learned (Toneva et al., 2019; Baldock et al., 2021), so instability separates hard inputs from easy ones.

The intuition fails at the end of training. Once the optimiser has settled, consecutive checkpoints make nearly the same predictions on every input, right or wrong. Agreement with the final prediction then approaches 1 everywhere, and a signal that is 1 everywhere says nothing about errors. We call this *convergence-induced saturation* and study it as a property of the construction, independent of any particular model.

Contributions.

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 bounds (Section 3).** For a window W of checkpoints with mean disagreement δ_W and a final model with error rate q, the calibration error of fixed-final agreement lies in $[\,|q-\delta_W|,\ q+\delta_W\,]$, and its error-detection AUROC is at most $\frac{1}{2}$ plus the largest class-conditional probability that the window disagrees with the final prediction. The calibration bound also holds for binned ECE, and neither bound needs an assumption on the model.
- **Saturation along the trajectory (Section 5.2).** On 3D visual grounding and CIFAR-10, sliding a fixed-size window toward the end of training raises calibration error and lowers AUROC in 13 of 14 run-by-window-size cells, with 95% intervals excluding zero; the fourteenth has the same sign.
- **The construction, not the model (Section 5.1).** On the same predictions, temperature-scaled softmax and a three-seed probability ensemble are calibrated, while within-trajectory agreement is miscalibrated by the same amount in three independent runs.
- **A published score (Section 5.3).** At its original CIFAR-10 setting, SPTD's late weighting improves AUROC by 0.016–0.020 on CIFAR-10, where accuracy keeps rising, and lowers it on all four 3D grounding runs, where accuracy stalls; its late limit is near chance in all seven runs.

Agreement is not useless. Added to native softmax it lowers out-of-fold log loss slightly (Section 5.1). But it is a feature, not a confidence, and its value falls as training converges. Our 3D experiments use one grounding model family with ground-truth object proposals; CIFAR-10 with ResNet-18 provides a second architecture and domain (Section 4).

2 Setting

Trajectory scores. A training run produces checkpoints $\theta_1, \ldots, \theta_T$ with predictions $f_t(x)$; f_T is the deployed model. Its correctness is $Y(x) = \mathbb{1}[f_T(x) = y(x)]$ and its error rate is $q = \Pr(Y = 0)$. For a window $W \subseteq \{1, \ldots, T-1\}$ of $K = |W|$ checkpoints, *fixed-final agreement* is

$$A_W(x) = \frac{1}{K} \sum_{t \in W} \mathbb{1}[f_t(x) = f_T(x)], \qquad \delta_W = \mathbb{E}[1 - A_W(x)], \tag{1}$$

so δ_W is the window's mean disagreement with the final prediction. Agreement attaches a score to the final prediction and never changes it, so every comparison below is on identical predictions. SPTD (Rabanser et al., 2025, Alg. 1) ranks inputs by the weighted disagreement $g_k(x) = \sum_{t<T} (t/T)^k \mathbb{1}[f_t(x) \neq f_T(x)]$. At $k = 0$ it orders inputs exactly as A_W with $W = \{1, \ldots, T-1\}$; as $k \to \infty$ it orders them as A_W with $W = \{T-1\}$.

Metrics. For a score $S \in [0, 1]$, write $\eta(S) = \mathbb{E}[Y \mid S]$. The calibration error is $\mathrm{CE}(S) = \mathbb{E}|\eta(S) - S|$, split into overconfidence $O = \mathbb{E}[(S - \eta(S))_+]$ and underconfidence $U = \mathbb{E}[(\eta(S) - S)_+]$. Agreement takes at most $K + 1$ values, so CE is estimated exactly on its support; with $K \leq 9$ this equals ten-bin equal-width ECE (Naeini et al., 2015), which we report for continuous scores. Error detection uses AUROC with $1 - S$ as the score for an error, ties counted as one half (Hendrycks & Gimpel, 2017), and AURC is the area under the risk–coverage curve (Geifman & El-Yaniv, 2017).

3 Why late agreement cannot be a confidence

Proposition 1 (Calibration is pinned to the error rate)**.** *For any window W,*

$$\mathrm{CE}(A_W) = |q - \delta_W| + 2\min(O, U), \qquad |q - \delta_W| \ \leq\ \mathrm{CE}(A_W)\ \leq\ q + \delta_W. \tag{2}$$

Both inequalities also hold for binned ECE with any binning.

The identity follows from $\mathrm{CE} = O + U$ and $\mathbb{E}[A_W] - \mathbb{E}[Y] = O - U = q - \delta_W$. The lower bound drops $2\min(O, U)$. The upper bound is Jensen's inequality followed by $|Y - A_W| \leq (1 - Y) + (1 - A_W)$. Proofs are in Appendix A.

Proposition 1 says that trajectory agreement can be calibrated in the large only for a window whose mean disagreement equals the error rate, $\delta_W = q$. This is the within-trajectory analogue

of the generalisation–disagreement equality that holds approximately for independently trained networks (Jiang et al., 2022). Nothing ties a particular window to that point. Along a trajectory, δ_W falls from its early value toward zero as the window moves later, and the lower bound $q - \delta_W$ rises toward q.

Proposition 2 (Discrimination is capped by disagreement mass)**.** *Let P_y be the law of a score S given $Y = y$. Then* $\mathrm{AUROC}(S) \le \frac{1}{2} + \mathrm{TV}(P_0, P_1)$. *If S is constant on the event $E_W = \{f_t(x) = f_T(x)\ \forall t \in W\}$, which holds for A_W and for g_k restricted to W, then*

$$\mathrm{AUROC}(S) \;\le\; \tfrac{1}{2} + \max_y \Pr(E_W^c \mid Y = y) \;\le\; \tfrac{1}{2} + \frac{K\,\delta_W}{\min(q, 1-q)}. \tag{3}$$

Corollary 1 (Late limit)**.** *If the window checkpoints converge in prediction to the final model, $\delta_W \to 0$, then $\mathrm{CE}(A_W) \to q$ and $\mathrm{AUROC}(A_W) \to \frac{1}{2}$ for any fixed K. The limit is the constant score 1, whose calibration error is the error rate.*

Two consequences follow. First, the damage in the limit is proportional to the error rate: small for an image classifier at 94% accuracy, large for a 3D grounding model at 46% or 30%. The proxy is worst on the tasks where abstention matters most. Second, the direction of change along a trajectory depends on the sign of $q - \delta_W$. If early windows disagree more than the model errs ($\delta_W > q$), agreement is underconfident there and calibration first improves as the window moves later, then degrades once $\delta_W < q$. If $\delta_W < q$ from the start, calibration only degrades. Both patterns appear below. The bounds are not loose in the regime that matters: in the last-checkpoint window on ScanRefer seed 42, $q = 0.538$ and $\delta_W = 0.056$, so Proposition 1 confines calibration error to $[0.482, 0.594]$ (measured: 0.501), and Proposition 2 with $\max_y \Pr(E_W^c \mid Y = y) = \max(0.021, 0.087)$ caps AUROC at 0.587 (measured: 0.533).

4 Experimental setup

3D visual grounding. The model selects the referred object among ground-truth ScanNet object proposals (Dai et al., 2017) given a text query. It is a GeoClassifier: frozen Uni3D object features (Zhou et al., 2024), a frozen DistilBERT query encoder (Sanh et al., 2019), a one-layer graph network over nearest-neighbour objects, FiLM fusion, and one logit per proposal (3.8M trained parameters; AdamW, cosine schedule with one warm-up epoch, 10 epochs, label smoothing 0.1; Appendix B). We train three seeds on ScanRefer (Chen et al., 2020) and one run on Nr3D (Achlioptas et al., 2020), and store per-epoch predictions for every validation query ($T = 10$). Evaluation uses ScanRefer validation (9,508 queries, 141 scenes; final accuracy 0.462, 0.459, 0.450) and Nr3D validation (8,584 queries, 130 scenes; accuracy 0.298). Native confidence is the maximum softmax probability over proposals, recomputed from exported logits with 100% prediction parity against the stored final dumps. Accuracy saturates early: on ScanRefer seed 42 it rises by 1.3 points after epoch 4 while agreement with the final prediction rises from 0.66 to 0.94 at epoch 8.

CIFAR-10. ResNet-18 (He et al., 2016) on CIFAR-10 (Krizhevsky, 2009), three seeds, 80 epochs of SGD with step decays at epochs 40 and 60, test predictions stored every epoch ($T = 80$; final accuracy 0.942–0.943). Accuracy here keeps improving for most of training (0.81–0.83 before the first decay, 0.94 at the end), which gives a contrasting regime.

Controls and inference. Temperature scaling (TS) (Guo et al., 2017) is fitted out of fold on scene-grouped folds and never changes a prediction. The three-seed probability ensemble averages softmax distributions over identical candidate sets (Lakshminarayanan et al., 2017). All intervals are 95% paired bootstrap intervals over scenes for 3D (1,000 resamples) and over examples for CIFAR-10. *Fixed-K windows* slide $K \in \{3, 5\}$ consecutive non-final checkpoints across the run, holding K and the score grid constant; "early" and "late" are the first and last such window. On CIFAR-10 we slide within the last 30 non-final epochs (49–78), where accuracy has plateaued; the full-trajectory result is reported as a contrast.

Table 1: Confidence scores attached to the same final predictions (ScanRefer seed 42, Nr3D). The ensemble changes predictions (accuracy 0.469). Agreement uses all nine non-final epochs ("all") or only epoch 8 ("last"). Lower ECE, Brier and AURC and higher AUROC are better.

	ScanRefer (n=9,508, acc. 0.462)				Nr3D (n=8,584, acc. 0.298)			
Score	ECE	Brier	AUROC	AURC	ECE	Brier	AUROC	AURC
Native max-softmax	0.071	0.191	0.795	0.310	0.186	0.256	0.592	0.642
Temperature-scaled softmax	**0.031**	0.186	0.793	0.313	**0.063**	**0.213**	**0.603**	**0.637**
3-seed probability ensemble	0.041	**0.177**	**0.813**	**0.286**	–	–	–	–
Agreement, all (K=9)	0.228	0.253	0.764	0.344	0.323	0.342	0.585	0.652
Agreement, last (K=1)	0.501	0.501	0.533	–	0.670	0.670	0.510	–

5 RESULTS

5.1 AGREEMENT AGAINST STANDARD CONFIDENCE ON THE SAME PREDICTIONS

Table 1 compares scores attached to identical predictions. Agreement over all nine non-final epochs has ECE 0.228 on ScanRefer and 0.323 on Nr3D, nearly all of it overconfidence (signed bias 0.223 and 0.320). Equal-mass binning gives the same ordering (agreement 0.223 and 0.320, temperature scaling 0.034 and 0.066). The reliability diagrams in Figure 3 (Appendix C) show agreement below the diagonal everywhere above confidence 0.2, and the last-checkpoint window placing 94% of queries at confidence 1, where accuracy is 0.48 on ScanRefer and 0.30 on Nr3D. Out-of-fold temperature scaling of the same model's logits reaches 0.031 and 0.063; the gap to agreement is 0.197 $[0.176, 0.208]$ on ScanRefer and 0.260 $[0.244, 0.271]$ on Nr3D. Agreement also ranks errors worse than native softmax on both datasets. The queries it marks as certain ($A_W = 1$, 26.1% of ScanRefer) are wrong 21.5% of the time; the same number of queries with the highest softmax confidence are wrong 18.0% of the time. The one place agreement ranks better is the easy subset: on the official single-candidate split it reaches AUROC 0.868 against 0.843 for softmax (error rate among its certain queries 2.8% against 3.6%), while on the multiple-candidate split, which is 81% of queries, it trails (0.709 against 0.747). Using only the last checkpoint makes every number worse: ECE approaches the error rates (0.538 and 0.702) and AUROC approaches 0.5, as Corollary 1 predicts. In that window 91.3% of ScanRefer errors (89.2–92.5% across the three seeds) and 93.6% of Nr3D errors receive agreement exactly 1.

Independent runs and ensembles. Correlated checkpoints are the obvious suspect, so we trained two further ScanRefer seeds with identical configurations. Within-trajectory agreement has ECE 0.228, 0.215 and 0.220 in the three runs, with mean agreement 0.67–0.69 against accuracy 0.45–0.46. The three runs disagree with each other substantially: all three final models pick the same object on only 52.2% of queries. Averaging their softmax distributions gives ECE 0.041, better than any single seed (0.070–0.087), with a difference to seed-42 agreement of -0.187 $[-0.202, -0.169]$. Miscalibration therefore replicates across independent trajectories and disappears under a valid probability construction on the same rows. Three-seed vote shares take only three values and are not calibrated (ECE 0.330); we report them only descriptively. Two further GeoClassifier variants with four stored epochs each show the same pattern (agreement ECE 0.215 and 0.197 against softmax 0.042 and 0.044; Appendix C).

Agreement is a feature, not a confidence. On ScanRefer (seed 42), a scene-grouped out-of-fold logistic model that adds agreement and its product with softmax to native softmax lowers log loss by 0.0125 $[0.0069, 0.0210]$ (200 scene-bootstrap resamples) and Brier score by 0.0057 relative to Platt-scaled softmax alone. The signal is real. It has to be mapped through a fitted model before it means anything as a probability, and Section 5.4 shows that one global mapping is not enough.

5.2 SATURATION ALONG THE TRAJECTORY

Figure 1a follows one ScanRefer run. From the earliest window (epochs 0–2) to the latest (6–8), mean agreement rises from 0.52 to 0.87, calibration error from 0.17 to 0.41, and AUROC falls

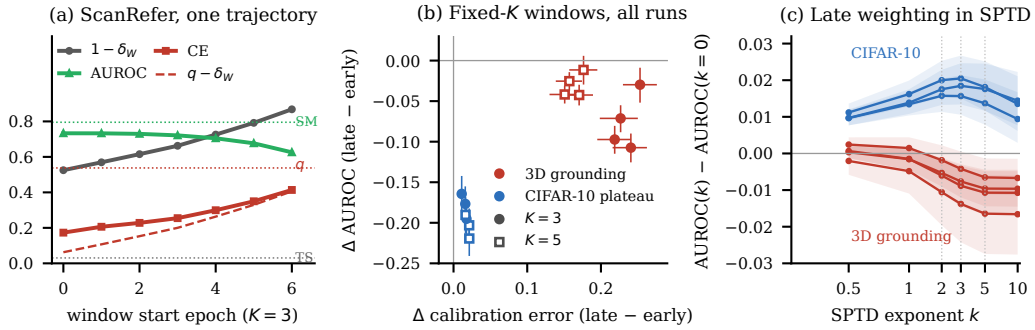


Figure 1: **(a)** One ScanRefer trajectory, fixed-final agreement over sliding windows of $K{=}3$ epochs. As the window moves later, mean agreement $1-\delta_W$ rises, calibration error (CE) climbs onto the floor $q-\delta_W$ of Proposition 1, and AUROC falls; dotted lines mark the error rate q, native softmax AUROC (SM) and temperature-scaled softmax ECE (TS). **(b)** Late minus early window for every run and K, with 95% bootstrap intervals; CIFAR-10 windows lie on the accuracy plateau. All 14 points lie in the lower-right quadrant; one interval (Nr3D, $K{=}5$) crosses zero in AUROC. **(c)** SPTD's weighting $(t/T)^k$: AUROC change relative to uniform weighting, seven runs with 95% bands. Dotted verticals mark the published values $k \in \{2, 3, 5\}$.

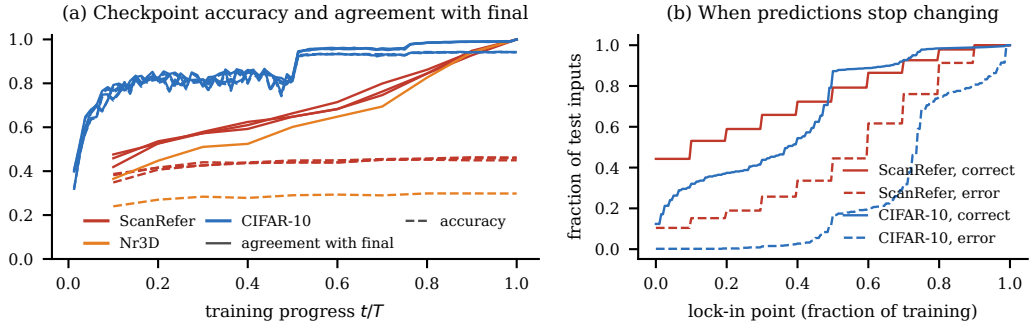


Figure 2: **(a)** Accuracy of each checkpoint (dashed) and its agreement with the final prediction (solid), against training progress, for all seven runs. On 3D grounding accuracy flattens by $t/T \approx 0.5$ while agreement keeps rising to 1; on CIFAR-10 accuracy jumps at each learning-rate decay. **(b)** Lock-in point: the earliest checkpoint after which the prediction never changes again, for inputs the final model gets right (solid) and wrong (dashed); ScanRefer and CIFAR-10, seed 42.

from 0.73 to 0.63. The floor $q-\delta_W$ of Proposition 1 rises from 0.06 to 0.41 and calibration error converges onto it: in late windows almost all miscalibration is calibration-in-the-large.

Figure 2 shows why. On ScanRefer, queries the final model gets right lock in at a median epoch of 1; wrong-instance and wrong-class errors lock in at a median epoch of 6 and switch more than twice as often (3.9 and 3.7 against 1.7 switches). The separation between the groups lives in the first two thirds of training. By epoch 8 both have stopped moving, and a window placed there cannot tell them apart.

Figure 1b collects early-to-late changes across runs. For the three ScanRefer seeds, calibration error rises by 0.15–0.24 and AUROC falls by 0.025–0.107 at both window sizes, all intervals excluding zero (seed 42, $K{=}3$: $+0.241$ $[0.216, 0.262]$ and -0.107 $[-0.125, -0.090]$). Nr3D replicates at $K{=}3$ ($+0.253$; AUROC -0.030 $[-0.052, -0.009]$). At $K{=}5$ its calibration error rises by 0.176 but the AUROC change, -0.012 $[-0.030, 0.006]$, has an interval that includes zero; this is the one cell of 14 that does not reach significance. Table 2 lists every cell. A reference-free variant, the modal share within the window, saturates in the same way ($K{=}3$: calibration error $+0.205$, AUROC -0.143).

Table 2: Early versus late fixed-K windows (late minus early), 95% paired bootstrap intervals. 3D: scene bootstrap, windows over epochs 0–8, final epoch 9. CIFAR-10: example bootstrap, windows within the plateau (epochs 49–78), final epoch 79.

Run	K	Δ calibration error	Δ AUROC
ScanRefer seed 42	3	$+0.241$ $[0.216, 0.262]$	-0.107 $[-0.125, -0.090]$
	5	$+0.170$ $[0.150, 0.189]$	-0.042 $[-0.055, -0.030]$
ScanRefer seed 7	3	$+0.219$ $[0.195, 0.241]$	-0.097 $[-0.115, -0.080]$
	5	$+0.151$ $[0.130, 0.170]$	-0.042 $[-0.053, -0.031]$
ScanRefer seed 13	3	$+0.227$ $[0.198, 0.250]$	-0.071 $[-0.088, -0.055]$
	5	$+0.157$ $[0.139, 0.174]$	-0.025 $[-0.037, -0.014]$
Nr3D	3	$+0.253$ $[0.230, 0.276]$	-0.030 $[-0.052, -0.009]$
	5	$+0.176$ $[0.157, 0.195]$	-0.012 $[-0.030, +0.006]$
CIFAR-10 seed 42	3	$+0.016$ $[0.011, 0.020]$	-0.177 $[-0.198, -0.155]$
	5	$+0.021$ $[0.017, 0.025]$	-0.203 $[-0.224, -0.182]$
CIFAR-10 seed 7	3	$+0.018$ $[0.012, 0.021]$	-0.195 $[-0.216, -0.173]$
	5	$+0.021$ $[0.017, 0.025]$	-0.219 $[-0.241, -0.198]$
CIFAR-10 seed 13	3	$+0.011$ $[0.007, 0.015]$	-0.164 $[-0.188, -0.142]$
	5	$+0.016$ $[0.012, 0.021]$	-0.190 $[-0.213, -0.168]$

The CIFAR-10 contrast. Over the full CIFAR-10 trajectory, calibration error of agreement *falls* from early to late windows (e.g. -0.41 for seed 42, K=3) while AUROC still falls (-0.085 to -0.141). This is the first case of Proposition 1: early checkpoints are much worse than the final model, so early windows disagree more than the model errs ($\delta_W > q$), and agreement over all 79 non-final epochs is underconfident (mean 0.877 against accuracy 0.942). Moving later removes that underconfidence first. On the accuracy plateau, where $\delta_W < q$ throughout, the 3D pattern returns in all three seeds and both window sizes: calibration error rises by 0.011–0.021 toward $q \approx 0.058$, and AUROC falls by 0.16–0.22, all intervals excluding zero. In the last-epoch window AUROC is 0.53–0.54 and 91.5–93.2% of errors have agreement 1.

5.3 A PUBLISHED TRAJECTORY SCORE

SPTD's weighting depends only on training progress t/T, so the weight profile is the same whether a run stores 10 checkpoints or 1,600; our runs sample it more coarsely than the original experiments, which used T between 800 and 2,200. The original paper sets k per dataset (k=2 on CIFAR-10 and CIFAR-100, 3 and 5 on two others), reports that $k \in [1, 3]$ works best, and notes that increasing k too far lowers accuracy at medium coverage (Rabanser et al., 2025). Figure 1c and Table 3 sweep k on all seven runs. The sign of the effect depends on the regime. On CIFAR-10, late weighting helps: at k=2, the original CIFAR-10 setting, AUROC improves by 0.016, 0.018 and 0.020 over uniform weighting, which is consistent with the original paper and brings SPTD level with native softmax (0.927–0.930 against 0.919–0.932). On 3D grounding the same setting hurts: at k=2 AUROC drops by 0.006, 0.005 and 0.002 on the three ScanRefer seeds and 0.011 on Nr3D, with intervals excluding zero in three of the four runs, and at k=5 by 0.007–0.016 with every interval excluding zero. At k=1 no run changes by more than 0.005 and every interval includes zero. SPTD stays below native softmax on every 3D run at every k. In the limit $k \to \infty$, where Proposition 2 binds, SPTD's AUROC is 0.51–0.54 in all seven runs.

Checkpoint density changes the size of these effects but, in our runs, not their sign (Table 7, Appendix C). Keeping every second, fourth or eighth CIFAR-10 checkpoint, counting back from the final one, shrinks the k=2 gain from 0.016–0.020 at T=80 to 0.007–0.008 at T=10, with every interval still excluding zero; at the same T=10 all four 3D runs lose AUROC. On 3D, density works the other way: going from T=5 to T=10 enlarges the k=2 loss in all four runs (from between -0.001 and -0.006 at T=5 to between -0.002 and -0.011 at T=10). Denser 3D trajectories were not stored, so we cannot test SPTD's original density there.

The effects at finite k are small, and we do not read them as a failure of SPTD. They show where its design assumption holds, and Figure 2b shows the difference between the regimes. On CIFAR-10, correct predictions lock in at a median of 37.5% of training and errors at 74%; 19.3% of errors are still changing in the last tenth of training against 1.1% of correct predictions, so late disagreement

Table 3: SPTD weighting sweep: AUROC for error detection at each k, and the change from $k{=}0$ with a 95% paired bootstrap interval at $k{=}2$ (the original setting for CIFAR-10 and CIFAR-100) and $k{=}5$ (the largest setting in the original paper). $k{=}0$ ranks like agreement over all non-final checkpoints; $k{=}\infty$ uses only the last one. Native softmax AUROC for reference.

Run	softmax	$k{=}0$	1	2	3	5	∞	$\Delta_{k=2}$	$\Delta_{k=5}$
ScanRefer s42	0.795	0.764	0.763	0.758	0.755	0.753	0.533	-0.006 $[-0.011, -0.001]$	-0.011 $[-0.017, -0.005]$
ScanRefer s7	0.787	0.770	0.769	0.765	0.763	0.761	0.526	-0.005 $[-0.009, -0.001]$	-0.010 $[-0.014, -0.004]$
ScanRefer s13	0.802	0.772	0.773	0.770	0.768	0.765	0.540	-0.002 $[-0.006, +0.002]$	-0.007 $[-0.011, -0.001]$
Nr3D	0.592	0.585	0.580	0.574	0.571	0.568	0.510	-0.011 $[-0.020, -0.002]$	-0.016 $[-0.027, -0.006]$
CIFAR-10 s42	0.923	0.911	0.924	0.927	0.926	0.924	0.541	$+0.016$ $[+0.011, +0.020]$	$+0.014$ $[+0.008, +0.020]$
CIFAR-10 s7	0.932	0.913	0.926	0.930	0.931	0.930	0.532	$+0.018$ $[+0.013, +0.023]$	$+0.018$ $[+0.012, +0.025]$
CIFAR-10 s13	0.919	0.909	0.926	0.929	0.930	0.927	0.534	$+0.020$ $[+0.015, +0.025]$	$+0.018$ $[+0.012, +0.025]$

singles out errors, and early disagreement, dominated by inputs the final model later gets right, is noise. On 3D grounding, correct predictions lock in almost immediately and errors by the middle of training, so the separating disagreement is early, and weighting it down loses information. In both regimes the last checkpoints carry little disagreement of any kind, which is why the late limit fails everywhere.

5.4 A SECONDARY CASE: AGREEMENT INVENTS A SUBGROUP GAP

A calibration audit often compares subgroups. In 3D grounding the natural split is candidate multiplicity: whether the scene holds one or several objects of the target class. Counting same-class candidates among the scored proposals under the ScanRefer 18-class map (chosen before looking at results), the ECE gap between multiple- and single-candidate queries is 0.224 $[0.187, 0.244]$ under agreement and 0.027 $[-0.006, 0.046]$ under native softmax. The absolute-ECE gap under temperature scaling and the ensemble is negative (-0.066, -0.044), but that reflects underconfidence on single-candidate queries, not reversed overconfidence: every score is more overconfident on multiple-candidate queries in signed bias, and agreement is the only score overconfident in both groups (signed bias 0.038 and 0.257). Agreement adds about 0.065 of signed-bias gap beyond softmax, far less than the 0.20 its absolute ECE suggests. The native-softmax gap depends on the ontology (it is 0.080 under raw proposal labels), while the agreement gap is large under every ontology. Nr3D is less clear: the agreement gap is 0.080–0.086 under all three ontologies, but the softmax gap ranges from 0.018 $[-0.020, 0.051]$ under raw proposal labels to 0.068 $[0.020, 0.103]$ under the 18-class map, so there agreement clearly amplifies the gap only under raw labels. On the official ScanRefer single/multiple split, one global Platt or isotonic map cannot repair agreement: it cuts overall ECE to 0.049 or 0.009 but inflates single-candidate ECE from 0.041 to 0.22 or 0.20, because the multiple-candidate group is 4.4 times larger. Split-specific maps fix both groups (Appendix C). An audit that used agreement as its confidence would have reported a subgroup failure that the model's own probabilities do not show.

6 RELATED WORK

Training dynamics as a signal. Per-example training dynamics identify hard, ambiguous or mislabelled *training* examples: forgetting events (Toneva et al., 2019), dataset cartography (Swayamdipta et al., 2020), the area under the margin (Pleiss et al., 2020) and prediction depth (Baldock et al., 2021). These methods use labels and aim at data diagnosis rather than test-time confidence. SPTD (Rabanser et al., 2025) carries the idea to test points and selective prediction, and is the closest method to the construction we analyse; our results qualify rather than contradict it. Temporal ensembling uses a running average of predictions across epochs as a training target (Laine & Aila, 2017).

Ensembles from one run. Snapshot and checkpoint ensembles (Huang et al., 2017; Chen et al., 2017), fast geometric ensembling (Garipov et al., 2018) and SWAG (Maddox et al., 2019) build predictive distributions from a single trajectory, usually with cyclic or high learning rates designed to visit diverse solutions. Checkpoints along one standard trajectory are functionally much closer to each other than independently initialised networks (Fort et al., 2019), which matches the saturation

we measure. Our agreement construction differs from these methods in that it scores the final prediction by agreement instead of averaging probabilities; the ensemble control in Section 5.1 shows that averaging probabilities across independent runs does not suffer the failure.

Disagreement and calibration. Disagreement between independently trained networks tracks test error when the ensemble is calibrated in a class-aggregated sense (Jiang et al., 2022), and agreement across models predicts out-of-distribution accuracy (Baek et al., 2022). Proposition 1 gives the per-window, within-trajectory version: calibration in the large requires $\delta_W = q$. Calibration of modern networks and its estimation are studied by Guo et al. (2017), Naeini et al. (2015), Kumar et al. (2019), Nixon et al. (2019), Ovadia et al. (2019) and Minderer et al. (2021); label smoothing, used by our 3D model, also changes softmax calibration (Müller et al., 2019). Selective prediction and error detection follow Geifman & El-Yaniv (2017) and Hendrycks & Gimpel (2017).

Logit-free confidence. For language models, consistency across sampled outputs (Wang et al., 2023; Manakul et al., 2023; Kuhn et al., 2023), verbalised confidence (Tian et al., 2023; Xiong et al., 2024) and self-evaluation (Kadavath et al., 2022) replace token probabilities. These proxies sample from one final model rather than from a trajectory, so our bounds do not apply to them directly. They share the structural question we ask: what a proxy's calibration error and AUROC can be at best, given the variability of the predictions it aggregates.

7 Discussion and limitations

Is this not obvious? That consecutive checkpoints agree more as training settles is expected. What is not obvious from that fact alone is how much it costs and where. Proposition 1 says the cost in the limit is the whole error rate, so the same construction that is nearly harmless on CIFAR-10 at 94% accuracy is badly wrong on a grounding task at 30–46%, which is where one would want to abstain. The published method we examined weights late disagreement most, on the stated grounds that late disagreement is most indicative of an error (Rabanser et al., 2025). Its authors already observed that very large k hurts. Our bound explains why, and our results agree with the argument in one regime and contradict it in another, where even the original CIFAR-10 setting lowers AUROC. Figure 2b shows which quantity decides between the two regimes: whether errors are still changing late in training while correct predictions are not. That quantity can be measured on a validation set before a trajectory score is trusted. The bounds also give a check that needs labels only to estimate the error rate: δ_W is computed from stored predictions alone, so with q estimated on a labelled validation set, a candidate window with $q - \delta_W$ large is miscalibrated by at least that amount for any model (Proposition 1), and one with $K\delta_W$ small relative to $\min(q, 1-q)$ cannot rank errors well (Proposition 2). A window with $\delta_W \approx q$ passes this check but is not thereby calibrated, because the term $2\min(O, U)$ remains; the check rules windows out, it does not certify them. Finally, the subgroup analysis in Section 5.4 shows the practical failure mode. An audit that reads agreement as confidence reports a disparity that the model's own probabilities do not have.

What to use instead. When the logits exist, out-of-fold temperature scaling was the best calibrated score we tested, and it needs one checkpoint. When several independent runs exist, averaging their probabilities was both calibrated and the best ranker. When only a trajectory of discrete predictions exists, agreement should be recalibrated on labelled data separately for each regime of interest (Table 4), the window should be chosen by its measured disagreement rather than by its position, and the error rate should be reported alongside, since it is what the calibration error approaches.

Limitations. The 3D results come from one model family with ground-truth proposals; a detector front end would add errors that agreement might track differently. An independent 3D architecture could not be run in our environment, and CIFAR-10 with ResNet-18 is the external control. Our trajectories store one checkpoint per epoch, 10 for 3D and 80 for CIFAR-10, far fewer than the original SPTD experiments. Because SPTD's weights depend only on t/T, the comparison is on the same profile; thinning the stored trajectories shows that finer sampling enlarges the CIFAR-10 gain, so it could change the magnitude of the finite-k effects on 3D as well. Per-epoch probabilities were not stored, so we could not compare against averaging probabilities over the same checkpoints, as snapshot or FGE-style ensembles do; the only probability-averaging control is the three-seed ensemble.

The construction assumes discrete predictions: regression outputs would need a tolerance for agreement, and agreement among outputs sampled from the final model, as in self-consistency decoding, is a different source of multiple predictions that we do not study. Three-seed vote agreement is too coarse to evaluate as a confidence. The subgroup analysis depends on how multiplicity is counted, which is why we treat it as secondary. The bounds in Section 3 are elementary; their value is that they locate the failure in the construction and predict its limit.

8 Conclusion

Agreement between a model's checkpoints and its final prediction is informative early in training and uninformative at the end. Its calibration error is pinned to the error rate within the window's mean disagreement, and its discrimination is capped by that disagreement, so a window close to the final checkpoint scores almost every error as certain. Across two 3D grounding datasets and CIFAR-10 the degradation appears wherever the window is moved later on a plateau, it survives independent training runs, and it vanishes under temperature scaling and probability ensembles on the same predictions. Trajectory agreement is a feature that can be calibrated, not a confidence that can be read off.

AI use statement

We used generative AI tools (large language model coding and writing assistants) for the following purposes. Writing: drafting the paper and editing text. Retrieval: searching for related work and retrieving bibliographic metadata, which we checked against the published versions. Execution: writing and refactoring analysis scripts, including the SPTD weighting sweep (Section 5.3), and running them on stored model outputs. Mathematics: drafting the proofs of Propositions 1 and 2 (Appendix A), which the authors checked. Every numerical result in the paper is computed by the analysis scripts in the supplementary material, which have unit tests. No data or result was generated by a generative model itself. The authors reviewed all AI-assisted text, code and proofs and take full responsibility for the content.

Reproducibility statement

All analyses run on CPU from stored per-epoch prediction dumps and exported final-checkpoint logits. Definitions are in Section 2, proofs in Appendix A, model and training configurations in Appendix B, and the remaining tables in Appendix C. The supplementary material contains the analysis scripts with their tests, the SPTD sweep and its checkpoint-thinning variant, the bootstrap code, and the JSON outputs from which every number in the paper is read.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In *European Conference on Computer Vision (ECCV)*, pp. 422–440, 2020.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and Zico Kolter. Agreement-on-the-Line: Predicting the Performance of Neural Networks under Distribution Shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Robert J. N. Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep Learning Through the Lens of Example Difficulty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language. In *European Conference on Computer Vision (ECCV)*, pp. 202–221, 2020.

Hugh Chen, Scott Lundberg, and Su-In Lee. Checkpoint Ensembles: Ensemble Methods from a Single Training Process. arXiv preprint arXiv:1710.03282, 2017.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective. arXiv preprint arXiv:1912.02757, 2019.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot Ensembles: Train 1, get M for free. In *International Conference on Learning Representations (ICLR)*, 2017.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-Scene: Bridging 3D Scene and Large Language Models with Object Identifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J. Zico Kolter. Assessing Generalization of SGD via Disagreement. In *International Conference on Learning Representations (ICLR)*, 2022.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know. arXiv preprint arXiv:2207.05221, 2022.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified Uncertainty Calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Samuli Laine and Timo Aila. Temporal Ensembling for Semi-Supervised Learning. In *International Conference on Learning Representations (ICLR)*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the Calibration of Modern Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When Does Label Smoothing Help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring Calibration in Deep Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying Mislabeled Data using the Area Under the Margin Ranking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Stephan Rabanser, Anvith Thudi, Kimia Hamidieh, Adam Dziedzic, Israfil Bahceci, Akram Bin Sediq, Hamza Sokun, and Nicolas Papernot. Selective Prediction via Training Dynamics. *Transactions on Machine Learning Research*, 2025.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring Unified 3D Representation at Scale. In *International Conference on Learning Representations (ICLR)*, 2024.

A PROOFS

Proposition 1. Let $S = A_W$ and $\eta(S) = \mathbb{E}[Y \mid S]$. By definition $\mathrm{CE} = \mathbb{E}|\eta(S) - S| = O + U$ and $O - U = \mathbb{E}[S - \eta(S)] = \mathbb{E}[S] - \mathbb{E}[Y] = (1 - \delta_W) - (1 - q) = q - \delta_W$. Hence $\mathrm{CE} = |O - U| + 2\min(O, U) = |q - \delta_W| + 2\min(O, U) \geq |q - \delta_W|$. For the upper bound, Jensen's inequality gives $\mathbb{E}|\mathbb{E}[Y - S \mid S]| \leq \mathbb{E}|Y - S|$, and $|Y - S| \leq |Y - 1| + |1 - S| = (1 - Y) + (1 - S)$, so $\mathrm{CE} \leq q + \delta_W$. For a binned estimator with bins B_b, $\mathrm{ECE} = \sum_b \Pr(B_b)\,|\mathbb{E}[Y - S \mid B_b]|$; the triangle inequality over bins gives $\mathrm{ECE} \geq |\mathbb{E}[Y - S]| = |q - \delta_W|$, and Jensen within each bin gives $\mathrm{ECE} \leq \mathbb{E}|Y - S| \leq q + \delta_W$. □

Proposition 2. Orient S so that larger values indicate an error, and let $X \sim P_0$ (errors) and $X' \sim P_1$ (correct) be independent. With ties counted as one half, $\mathrm{AUROC} = \frac{1}{2} + \frac{1}{2}D$ with $D = \Pr(X > X') - \Pr(X < X')$. Define $h(s) = \Pr(X' < s) - \Pr(X' > s) \in [-1, 1]$. Then $D = \mathbb{E}_{P_0}[h]$, and $\mathbb{E}_{P_1}[h] = 0$ because two independent draws from P_1 are exchangeable. Therefore $D = \mathbb{E}_{P_0}[h] - \mathbb{E}_{P_1}[h] \leq (\sup h - \inf h)\,\mathrm{TV}(P_0, P_1) \leq 2\,\mathrm{TV}(P_0, P_1)$, which gives the first inequality. If S equals a constant s^* on E_W, both P_0 and P_1 place mass at least $1 - \varepsilon_y$ on s^*, where $\varepsilon_y = \Pr(E_W^c \mid Y = y)$. For any set B, if $s^* \in B$ both probabilities lie in $[1 - \varepsilon_y, 1]$, and otherwise both are at most ε_y; hence $\mathrm{TV} \leq \max_y \varepsilon_y$. By the union bound, $\varepsilon_y \leq \sum_{t \in W} \Pr(f_t \neq f_T \mid Y = y) = K\,\delta_W^{(y)}$ with $\delta_W^{(y)} = \mathbb{E}[1 - A_W \mid Y = y]$. Finally $\delta_W = (1 - q)\,\delta_W^{(1)} + q\,\delta_W^{(0)}$ gives $\delta_W^{(0)} \leq \delta_W / q$ and $\delta_W^{(1)} \leq \delta_W / (1 - q)$. □

Corollary 1. Proposition 1 gives $|\mathrm{CE}(A_W) - q| \leq \delta_W$ and Proposition 2 gives $\frac{1}{2} \leq \mathrm{AUROC} \leq \frac{1}{2} + K\delta_W / \min(q, 1 - q)$ when $\mathrm{AUROC} \geq \frac{1}{2}$ (for $\mathrm{AUROC} < \frac{1}{2}$ apply the same argument to $-S$). Both converge as $\delta_W \to 0$. □

B SETUP DETAILS

3D grounding model. GeoClassifier with frozen Uni3D features of ground-truth object segments (1,024 dimensions) and frozen DistilBERT (`distilbert-base-uncased`) text features with attention pooling; hidden size 256; one graph layer with four attention heads over a k=2 nearest-neighbour object graph with geometric edge features; FiLM fusion of semantic and geometric object features; a scene-context token; a relational refiner (weight 0.7). Training uses AdamW (learning rate 4×10^{-5}, weight decay 0.01, gradient clipping at 1.0), batch size 8, cosine decay to 1% of the peak after one warm-up epoch, 10 epochs, label smoothing 0.1, mixed precision on Apple MPS. The three ScanRefer seeds (42, 7, 13) use byte-identical configurations apart from the seed; every stored artifact is hash-verified. Nr3D uses the same configuration on the Nr3D split of ReferIt3D.

CIFAR-10. ResNet-18 adapted to 32×32 inputs, SGD with momentum 0.9, learning rate 0.1, weight decay 5×10^{-4}, batch size 128, standard crop-and-flip augmentation, learning rate divided by 10 at epochs 40 and 60, 80 epochs, seeds 42, 7 and 13. Test predictions are stored after every epoch.

Temperature scaling and recalibration. Five scene-grouped folds (seed 0); the temperature is fitted by multiclass negative log-likelihood on training scenes and applied unchanged to held-out scenes (fold temperatures 1.09–1.13 on ScanRefer and 1.30–1.34 on Nr3D). Platt and isotonic recalibration of agreement (Zadrozny & Elkan, 2002) use the same folds.

C ADDITIONAL TABLES

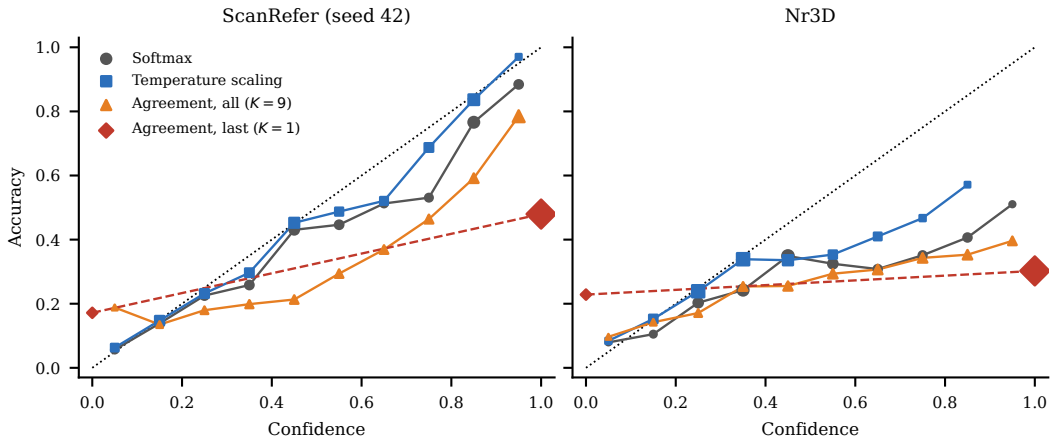


Figure 3: Reliability diagrams (10 equal-width bins; marker area proportional to bin mass) for the final predictions of ScanRefer seed 42 and Nr3D. Softmax and out-of-fold temperature scaling follow the diagonal more closely than agreement over all nine non-final epochs. Agreement in the last-checkpoint window takes only the values 0 and 1 and puts 94% of the mass at confidence 1.

Table 4: Out-of-fold recalibration of agreement across the candidate-multiplicity split (ScanRefer seed 42, official Unique/Multiple labels; 1,759 and 7,749 queries).

Method	All ECE	Single-candidate ECE	Multiple-candidate ECE	Log loss
Raw agreement	0.228	0.041	0.272	2.544
Platt, global	0.049	0.222	0.068	0.583
Isotonic, global	0.009	0.197	0.048	0.577
Platt, split-specific	0.018	0.030	0.023	0.546
Isotonic, split-specific	0.010	0.037	0.008	0.547

Table 5: Signed bias (mean confidence minus accuracy) by multiplicity group, ScanRefer seed 42, proposal-set ScanRefer-18 counts.

Score	abs. ECE gap	signed-bias gap	bias, single	bias, multiple
Native softmax	+0.027	+0.155	−0.059	+0.095
Temperature-scaled	−0.066	+0.176	−0.121	+0.055
Probability ensemble	−0.044	+0.166	−0.105	+0.060
Agreement (all)	+0.224	+0.219	+0.038	+0.257

Earlier variants. Two further GeoClassifier variants trained for four epochs on ScanRefer (one close to the main configuration, one with a two-layer graph network, relational edge features and no scene-context token) have agreement ECE 0.215 and 0.197 against native softmax 0.042 and 0.044 on all 9,508 queries, with 100% prediction parity. With four checkpoints the agreement grid is $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$, so we treat these as a coarse check.

Table 6: Other scores computed from the final logits (ranking by the raw score; ECE after scene-grouped out-of-fold Platt scaling). Entropy ranks errors slightly better than max-softmax on Nr3D; none of these scores uses the trajectory.

	ScanRefer		Nr3D	
Score	AUROC	ECE (Platt)	AUROC	ECE (Platt)
Max-softmax	0.795	–	0.592	–
Top-two margin	0.779	0.026	0.539	0.007
Negative entropy	0.788	0.043	0.624	0.019
Negative energy	0.311	0.625	0.457	0.003

Table 7: SPTD under thinned trajectories: every m-th checkpoint kept, counting back from the final one, so the retained T checkpoints span the same training horizon. AUROC at $k{=}0$ and the change at $k{=}2$ and $k{=}5$ with 95% paired bootstrap intervals (1,000 resamples; unit scene for 3D, example for CIFAR-10). The full-density rows repeat Table 3.

Run	T	AUROC, $k{=}0$	Δ, $k{=}2$	Δ, $k{=}5$
ScanRefer s42	10	0.764	-0.006 $[-0.011, -0.001]$	-0.011 $[-0.017, -0.005]$
	5	0.740	-0.005 $[-0.010, -0.001]$	-0.005 $[-0.010, -0.001]$
ScanRefer s7	10	0.770	-0.005 $[-0.009, -0.001]$	-0.010 $[-0.014, -0.004]$
	5	0.740	-0.004 $[-0.008, -0.001]$	-0.004 $[-0.008, -0.001]$
ScanRefer s13	10	0.772	-0.002 $[-0.006, +0.002]$	-0.007 $[-0.011, -0.001]$
	5	0.747	-0.001 $[-0.004, +0.002]$	-0.001 $[-0.004, +0.002]$
Nr3D	10	0.585	-0.011 $[-0.020, -0.002]$	-0.016 $[-0.027, -0.006]$
	5	0.578	-0.006 $[-0.013, +0.002]$	-0.006 $[-0.013, +0.002]$
CIFAR-10 s42	80	0.911	$+0.016$ $[+0.011, +0.020]$	$+0.014$ $[+0.008, +0.020]$
	40	0.905	$+0.017$ $[+0.012, +0.022]$	$+0.013$ $[+0.006, +0.020]$
	20	0.893	$+0.014$ $[+0.008, +0.020]$	$+0.009$ $[+0.002, +0.016]$
	10	0.864	$+0.007$ $[+0.002, +0.012]$	$+0.005$ $[-0.001, +0.011]$
CIFAR-10 s7	80	0.913	$+0.018$ $[+0.013, +0.023]$	$+0.018$ $[+0.012, +0.025]$
	40	0.909	$+0.017$ $[+0.012, +0.022]$	$+0.015$ $[+0.009, +0.022]$
	20	0.900	$+0.013$ $[+0.008, +0.019]$	$+0.009$ $[+0.003, +0.016]$
	10	0.882	$+0.008$ $[+0.003, +0.012]$	$+0.005$ $[-0.001, +0.011]$
CIFAR-10 s13	80	0.909	$+0.020$ $[+0.015, +0.025]$	$+0.018$ $[+0.012, +0.025]$
	40	0.904	$+0.019$ $[+0.014, +0.025]$	$+0.016$ $[+0.009, +0.024]$
	20	0.893	$+0.016$ $[+0.011, +0.023]$	$+0.013$ $[+0.006, +0.021]$
	10	0.865	$+0.008$ $[+0.003, +0.014]$	$+0.007$ $[+0.001, +0.014]$