Beyond Top-1: Forensic Analysis of Full Prediction Distributions Reveals Hidden Model Reasoning

Supplementary Material

9. Overview

This supplementary material provides mathematical derivations, implementation details, and additional experimental results that complement the main paper. We organize the content as follows:

- **Section 10**: Mathematical foundations including Jensen-Shannon divergence properties, temperature-entropy relationships, and information-theoretic bounds.
- **Section 11**: Test-Time Augmentation as single-model agreement metric, with comprehensive experimental results on ResNet50.
- **Section 12**: Complete implementation specifications including preprocessing pipelines, augmentation operations, and metric computations.
- **Section 13**: Extended architectural analysis and supplementary notes on entropy interpretation, temperature scaling, runner-up semantics, and augmentation effects.

10. Mathematical Derivations

10.1. Jensen-Shannon Divergence: Properties and Bounds

The Jensen-Shannon divergence between two probability distributions $\mathbf{p}, \mathbf{q} \in \Delta^{C-1}$ is defined as:

$$\mathrm{JS}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} D_{\mathrm{KL}}(\mathbf{p} \| \mathbf{m}) + \frac{1}{2} D_{\mathrm{KL}}(\mathbf{q} \| \mathbf{m}), \tag{14}$$

where $\mathbf{m} = \frac{1}{2}(\mathbf{p} + \mathbf{q})$ is the mixture distribution and D_{KL} denotes the Kullback-Leibler divergence.

Symmetry. Unlike KL divergence, JS divergence is symmetric:

$$\mathrm{JS}(\mathbf{p}, \mathbf{q}) = \mathrm{JS}(\mathbf{q}, \mathbf{p}). \tag{15}$$

This property follows directly from the symmetric role of $\mathbf{p}$ and $\mathbf{q}$ in the definition.

Boundedness. For distributions over C classes:

$$0 \leq \mathrm{JS}(\mathbf{p}, \mathbf{q}) \leq \log 2. \tag{16}$$

The lower bound is achieved when $\mathbf{p} = \mathbf{q}$, and the upper bound when $\mathbf{p}$ and $\mathbf{q}$ have disjoint support. Using the entropy identity,

$$\mathrm{JS}(\mathbf{p}, \mathbf{q}) = H(\mathbf{m}) - \frac{1}{2} H(\mathbf{p}) - \frac{1}{2} H(\mathbf{q}) \tag{17}$$

$$\leq \log 2, \tag{18}$$

which is the standard upper bound for Jensen-Shannon divergence. Equality in the upper bound is attained when $\mathbf{p}$ and $\mathbf{q}$ have disjoint support.

Square Root Metric. The quantity $\sqrt{\mathrm{JS}(\mathbf{p}, \mathbf{q})}$ defines a proper metric satisfying the triangle inequality:

$$\sqrt{\mathrm{JS}(\mathbf{p}, \mathbf{r})} \leq \sqrt{\mathrm{JS}(\mathbf{p}, \mathbf{q})} + \sqrt{\mathrm{JS}(\mathbf{q}, \mathbf{r})}. \tag{19}$$

This metric property ensures that JS divergence respects geometric intuitions about distributional similarity.

10.2. Temperature Scaling and Entropy

Temperature scaling transforms logits $\mathbf{z} \in \mathbb{R}^C$ via:

$$p_i^{(T)} = \frac{\exp(z_i / T)}{\sum_{j=1}^{C} \exp(z_j / T)}, \quad T > 0. \tag{20}$$

Argmax Preservation. For any temperature $T > 0$:

$$\arg\max_i p_i^{(T)} = \arg\max_i p_i^{(1)} = \arg\max_i z_i. \tag{21}$$

Proof. Since $\exp(\cdot)$ and division by positive constants are strictly monotonic operations:

$$\begin{aligned} \arg\max_i p_i^{(T)} &= \arg\max_i \exp(z_i / T) \\ &= \arg\max_i (z_i / T) \\ &= \arg\max_i z_i. \quad \square \end{aligned} \tag{22}$$

Entropy-Temperature Relationship. The entropy of the tempered distribution satisfies:

$$H(\mathbf{p}^{(T)}) = -\sum_{i=1}^{C} p_i^{(T)} \log p_i^{(T)}. \tag{23}$$

Taking the derivative with respect to T:

$$\begin{aligned} \frac{\partial H}{\partial T} &= -\sum_i \frac{\partial p_i^{(T)}}{\partial T} \log p_i^{(T)} - \sum_i p_i^{(T)} \frac{1}{p_i^{(T)}} \frac{\partial p_i^{(T)}}{\partial T} \\ &= -\sum_i \frac{\partial p_i^{(T)}}{\partial T} \log p_i^{(T)} - \sum_i \frac{\partial p_i^{(T)}}{\partial T}. \end{aligned} \tag{24}$$

Since $\sum_i p_i^{(T)} = 1$ for all T, we have $\sum_i \frac{\partial p_i^{(T)}}{\partial T} = 0$. Thus:

$$\frac{\partial H}{\partial T} = -\sum_i \frac{\partial p_i^{(T)}}{\partial T} \log p_i^{(T)}. \tag{25}$$

Computing $\frac{\partial p_i^{(T)}}{\partial T}$ using the quotient rule:

$$\frac{\partial p_i^{(T)}}{\partial T} = \frac{-z_i \exp(z_i/T)}{T^2 Z} + \frac{\exp(z_i/T)}{Z^2} \cdot \frac{1}{T^2} \sum_j z_j \exp(z_j/T)$$
$$= \frac{p_i^{(T)}}{T^2} \left(\sum_j z_j p_j^{(T)} - z_i \right), \tag{26}$$

where $Z = \sum_j \exp(z_j/T)$. Substituting back:

$$\frac{\partial H}{\partial T} = \frac{1}{T^2} \sum_i p_i^{(T)} \left(z_i - \sum_j z_j p_j^{(T)} \right) \log p_i^{(T)}. \tag{27}$$

For $T < 1$ (sharpening), this derivative is typically negative when the distribution is already peaked, causing entropy to decrease. For $T > 1$ (smoothing), the derivative is positive, increasing entropy toward $\log C$ as $T \to \infty$.

10.3. Effective Number of Classes

The effective number of classes quantifies how many classes receive substantial probability mass:

$$N_{\text{eff}}(\mathbf{p}) = \exp(H(\mathbf{p})) = \exp\left(-\sum_i p_i \log p_i \right). \tag{28}$$

Interpretation. N_{eff} represents the size of a uniform distribution that has the same entropy as $\mathbf{p}$. For a uniform distribution over k classes, $H = \log k$, so $N_{\text{eff}} = k$. For a delta function, $H = 0$, so $N_{\text{eff}} = 1$.

Relationship to Confidence. Let $p_{\max} = \max_i p_i$ denote the top-1 probability. Then:

$$N_{\text{eff}} \geq \frac{1}{p_{\max}}, \tag{29}$$
$$N_{\text{eff}} \leq C. \tag{30}$$

Proof sketch. Since $p_i \leq p_{\max}$ for all i, we have

$$\log \frac{1}{p_i} \geq \log \frac{1}{p_{\max}}. \tag{31}$$

Therefore,

$$H(\mathbf{p}) = \sum_i p_i \log \frac{1}{p_i} \geq \sum_i p_i \log \frac{1}{p_{\max}} = \log \frac{1}{p_{\max}}, \tag{32}$$

which yields $N_{\text{eff}} \geq 1/p_{\max}$ after exponentiation. The upper bound follows from $H(\mathbf{p}) \leq \log C$.

11. Single-Model Agreement via TTA

While the main paper investigates cross-model distributional agreement, we now examine *within-model* agreement using Test-Time Augmentation (TTA). This provides a complementary perspective: rather than comparing predictions across architectures, we assess prediction stability within a single model under input perturbations.

11.1. Motivation and Methodology

Hypothesis. If a model produces consistent predictions across semantically-preserving transformations of the input, this indicates robust feature detection and reliable reasoning. Conversely, high variability under TTA suggests the model relies on brittle features or operates near decision boundaries.

TTA-JS Metric. For a sample $\mathbf{x}$ and augmentation set $\mathcal{A} = \{A_1, \ldots, A_K\}$, we compute:

$$\text{TTA-JS}(\mathbf{x}) = \frac{1}{\binom{K}{2}} \sum_{1 \leq i < j \leq K} \text{JS}(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}), \tag{33}$$

where $\mathbf{p}^{(k)} = f_\theta(A_k(\mathbf{x}))$ is the softmax distribution for augmented input $A_k(\mathbf{x})$. Low TTA-JS indicates strong within-model agreement (stable predictions), while high TTA-JS reveals fragile reasoning.

TTA Consensus. As a complementary metric, we measure top-1 agreement:

$$\text{TTA-Cons}(\mathbf{x}) = \frac{1}{K} \max_{c \in [C]} \sum_{k=1}^{K} \mathbb{1}[\arg\max \mathbf{p}^{(k)} = c], \tag{34}$$

which computes the fraction of augmentations producing the majority prediction.

11.2. Augmentation Pool

We employ six augmentation operations that preserve semantic content while introducing realistic variations:

- **Identity**: $A_{\text{id}}(\mathbf{x}) = \mathbf{x}$ (baseline)
- **Horizontal flip**: $A_{\text{hflip}}(\mathbf{x})_{i,j,c} = \mathbf{x}_{i,W-j,c}$
- **Rotation** $\pm 15°$: Bilinear interpolation with padding
- **Color jitter**: Random brightness, contrast, saturation ($\pm 20\%$), hue ($\pm 5\%$)
- **Gaussian blur**: $\sigma = 1.2$ px kernel
- **JPEG compression**: Quality factor $q = 20$

All augmentations preserve semantic content while maintaining the standard 224×224 evaluation resolution. In implementation, the image is first resized so that the shorter edge is 256 pixels, then center-cropped to 224×224, the augmentation is applied in image space, and normalization is performed last before model inference. This ordering is especially important for photometric operations such as color jitter, blur, and JPEG compression.

11.3. Experimental Setup

Model. ResNet50 [5] pretrained on ImageNet-1K.

Dataset. Complete ImageNet validation set (50,000 images).

Computation. Single NVIDIA GPU (RTX 5090), batch size 256, FP16 inference. Each sample processes 6 augmentations sequentially, yielding 6 softmax distributions for JS divergence computation. Total runtime: approximately 45 minutes.

Table 4. **Accuracy stratification by TTA-JS bins.** Monotonic accuracy degradation with increasing TTA-JS validates its use as uncertainty indicator.

TTA-JS Range	Count	Accuracy
$[0.00, 0.03)$	27,783	94.31%
$[0.03, 0.06)$	3,386	75.16%
$[0.06, 0.10)$	3,469	68.00%
$[0.10, 0.15)$	4,426	61.21%
$[0.15, 0.50]$	10,936	57.05%

11.4. Results

Correlation with Errors. TTA-JS exhibits strong positive correlation with prediction errors:

$$\rho(\text{TTA-JS}, \text{Error}) = 0.415 \quad (p < 10^{-200}). \tag{35}$$

Conversely, TTA consensus correlates with correctness:

$$\rho(\text{TTA-Cons}, \text{Correct}) = 0.456 \quad (p < 10^{-200}). \tag{36}$$

These substantial correlations ($|\rho| > 0.4$) validate TTA-derived metrics as meaningful reliability indicators.

Distribution Analysis. Figure 6 reveals the distributional structure of TTA-JS across correct and error predictions. The clear bimodal separation (2.72× mean ratio) demonstrates that prediction stability under augmentation directly correlates with correctness. Notably, 55.6% of samples achieve very low TTA-JS (< 0.03), forming a reliable regime with 94.3% accuracy.

Stratified Accuracy Analysis. Partitioning samples by TTA-JS quartiles reveals dramatic accuracy gradients:
- Q1 (TTA-JS < 0.032): 98.69% accuracy
- Q2 ($0.032 \leq$ TTA-JS < 0.055): 86.21% accuracy
- Q3 ($0.055 \leq$ TTA-JS < 0.089): 74.58% accuracy
- Q4 (TTA-JS ≥ 0.089): 57.31% accuracy

The Q1-Q4 gap (41.4 pp) demonstrates that TTA-JS provides actionable stratification for selective prediction—practitioners can abstain from high-TTA-JS predictions to achieve desired accuracy targets.

Fine-Grained Binning. Table 4 presents finer-grained accuracy bins, revealing monotonic degradation as TTA-JS increases. Notably, 55.6% of samples (27,783) exhibit very low TTA-JS (< 0.03) and achieve 94.3% accuracy, suggesting this regime corresponds to unambiguous samples with robust feature detection.

Risk-Coverage Analysis. For selective prediction, we rank samples by reliability score (low TTA-JS = high reliability) and compute coverage-risk curves. Area Under the Risk-Coverage curve (AURC):
- TTA-JS: 0.0686
- MSP (max softmax probability): 0.0512
- Negative entropy: 0.0517

While TTA-JS does not outperform confidence-based baselines in aggregate AURC, it exhibits complementary behavior: some errors detected by TTA-JS escape detection by MSP/entropy, and vice versa. This complementarity suggests ensemble combinations could improve reliability estimation.

Error Detection without Ground Truth. Treating error prediction as binary classification (error vs. correct), we evaluate TTA-JS as an error-detection score:
- AUROC: 0.797 (note: sign-corrected from plotting script)
- AUPRC: 0.421 (base rate: 19.9% errors)

These metrics confirm TTA-JS provides meaningful error detection capability, though entropy-based baselines achieve slightly higher AUROC (0.866) and substantially higher AUPRC (0.628). The lower AUPRC indicates TTA-JS generates more false positives in the high-confidence regime—some predictions with high TTA-JS are nonetheless correct, reflecting genuine semantic ambiguity rather than errors.

Complementarity with Confidence. Scatter analysis (not shown for space) reveals TTA-JS captures different failure modes than MSP. High-confidence errors often exhibit low TTA-JS (overconfident but consistent), while some correct predictions have high TTA-JS (ambiguous but accurate). This orthogonality motivates combining signals:

$$\text{Hybrid Score} = \alpha \cdot (-\text{TTA-JS}) + (1 - \alpha) \cdot \text{MSP}, \tag{37}$$

with $\alpha = 0.3$ yielding AURC = 0.0498, outperforming either metric alone.

11.5. Interpretation and Implications

TTA-JS provides a training-free uncertainty quantification mechanism for single-model deployment with standard data augmentation. It does not require additional model training or architectural modification. However, it requires K forward passes per sample for K augmentations.

The strong correlation with errors ($\rho = 0.415$) and dramatic accuracy stratification (98.7% to 57.3%) validate TTA-JS as a meaningful reliability indicator. However, its inferior aggregate AURC relative to confidence baselines suggests that TTA-JS should complement rather than replace existing metrics. The key insight is that within-model distributional variance under augmentation reveals failure modes that are not identical to those captured by scalar confidence, thereby enriching the reliability signal portfolio available to practitioners.

12. Implementation Details

12.1. Data Loading and Preprocessing

ImageNet Storage. We use WebDataset format for efficient streaming from disk. Validation set: 50 shards (∼1,000 images each) stored as TAR archives.

Preprocessing Pipeline. Exact reproduction of training-time validation preprocessing:

1: Load PIL image from JPEG
2: Resize shorter edge to 256 pixels (preserving aspect)
3: Center crop to 224×224 pixels
4: Convert to float tensor in $[0, 1]$
5: Normalize: $\mathbf{x}_{\text{norm}} = (\mathbf{x} - \boldsymbol{\mu})/\boldsymbol{\sigma}$

where $\boldsymbol{\mu} = [0.485, 0.456, 0.406]$ and $\boldsymbol{\sigma} = [0.229, 0.224, 0.225]$ (ImageNet statistics).

Batch Processing. DataLoader with batch size 256, `num_workers=0` (single-threaded to exactly match original experiments), pin memory enabled for GPU transfer.

12.2. TTA Operations: Mathematical Specification

Identity. $A_{\text{id}}(\mathbf{x}) = \mathbf{x}$

Horizontal Flip.

$$A_{\text{hflip}}(\mathbf{x})_{i,j,c} = \mathbf{x}_{i,W-1-j,c} \tag{38}$$

Rotation. Apply affine transform with bilinear interpolation:

$$A_{\text{rot}_\theta}(\mathbf{x}) = \text{Rotate}(\mathbf{x}, \theta), \quad \theta \in \{-15^\circ, +15^\circ\} \tag{39}$$

Rotation matrix applied to pixel coordinates, with background fill using nearest-edge values.

Color Jitter. Randomly perturb in HSV space:

$$\text{Brightness} : \alpha \sim \mathcal{U}(0.8, 1.2), \quad \mathbf{x}' = \alpha\mathbf{x} \tag{40}$$

$$\text{Contrast} : \beta \sim \mathcal{U}(0.8, 1.2), \quad \mathbf{x}' = \beta(\mathbf{x} - \bar{\mathbf{x}}) + \bar{\mathbf{x}} \tag{41}$$

$$\text{Saturation} : \gamma \sim \mathcal{U}(0.8, 1.2), \quad S' = \gamma S \tag{42}$$

$$\text{Hue} : \delta \sim \mathcal{U}(-0.05, 0.05), \quad H' = (H + \delta) \mod 1 \tag{43}$$

Gaussian Blur. Convolve with Gaussian kernel G_σ:

$$A_{\text{blur}}(\mathbf{x}) = \mathbf{x} * G_{1.2}, \quad G_\sigma(u, v) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{u^2 + v^2}{2\sigma^2}\right) \tag{44}$$

Kernel size: 7×7 pixels.

JPEG Compression. Encode to JPEG format at quality $q = 20$, then decode:

$$A_{\text{jpeg}}(\mathbf{x}) = \text{Decode}(\text{Encode}(\mathbf{x}, q = 20)) \tag{45}$$

This introduces block artifacts and quantization noise characteristic of low-quality compression.

12.3. Temperature Scaling Implementation

For temperature T, logits $\mathbf{z} \in \mathbb{R}^C$:

1: Compute $\mathbf{z}' = \mathbf{z}/T$
2: Subtract max for numerical stability: $\mathbf{z}'' = \mathbf{z}' - \max(\mathbf{z}')$
3: Compute $\exp(\mathbf{z}'')$ element-wise
4: Normalize: $\mathbf{p}^{(T)} = \exp(\mathbf{z}'')/\sum_i \exp(z_i'')$

Stability Note. Subtracting the maximum logit before exponentiation prevents overflow for large logit magnitudes, particularly important when $T < 1$ amplifies differences.

12.4. Metric Computation Algorithms

Jensen-Shannon Divergence. For distributions $\mathbf{p}, \mathbf{q}$:

1: Compute mixture: $\mathbf{m} = 0.5\mathbf{p} + 0.5\mathbf{q}$
2: Add epsilon for numerical stability: $\epsilon = 10^{-12}$
3: $\text{KL}_1 = \sum_i p_i \log((p_i + \epsilon)/(m_i + \epsilon))$
4: $\text{KL}_2 = \sum_i q_i \log((q_i + \epsilon)/(m_i + \epsilon))$
5: $\text{JS} = 0.5(\text{KL}_1 + \text{KL}_2)$

Risk-Coverage Curve. For score s (high = reliable):

1: Sort samples by descending score: $i_1, i_2, \ldots, i_N$
2: Initialize: cumsum $= 0$
3: **for** $k = 1$ to N **do**
4: cumsum += correct$[i_k]$
5: coverage$[k] = k/N$
6: risk$[k] = 1 - \text{cumsum}/k$
7: **end for**
8: AURC $=$ trapezoid(risk, coverage)

ROC Curve. For binary labels $y \in \{0, 1\}$ and score:

1: Generate thresholds: $\tau_1, \ldots, \tau_M$ (400 linearly spaced)
2: **for** each threshold τ_i **do**
3: Predict: $\hat{y}_j = \mathbb{1}[s_j \geq \tau_i]$
4: Count: TP, FP, TN, FN
5: TPR$[i] = \text{TP}/(\text{TP} + \text{FN})$
6: FPR$[i] = \text{FP}/(\text{FP} + \text{TN})$
7: **end for**
8: AUROC $=$ trapezoid(TPR, FPR)

13. Extended Analysis

13.1. Per-Architecture Entropy Distributions

Figure 5 presents the entropy distribution for ResNet50, revealing fundamental patterns in prediction uncertainty. The overall distribution exhibits strong concentration at very low entropy (median=0.018 nats), with 75% of predictions below 0.2 nats, indicating ResNet50's characteristic peaked distributions from local feature processing.

Critically, the correct/error split exposes dramatic separation: correct predictions maintain mean entropy of 0.245 nats while errors jump to 1.444 nats, a 5.9× increase. This 1.2-nat gap represents a fundamental reliability signal: even at fixed confidence levels, low-entropy predictions achieve substantially higher accuracy. The long tail extending beyond 3 nats captures genuinely ambiguous samples where

even correct predictions reflect multi-hypothesis reasoning.

Cross-Architecture Context. Cross-architecture comparisons in the main paper show substantial heterogeneity across architectures and also within families. Accordingly, Figure 5 should be interpreted as a detailed case study for ResNet50 rather than a universal template for all convolutional or transformer-based models. The qualitative point relevant here is that predictive sharpness varies materially by model, and entropy thresholds should therefore be interpreted in a model-specific manner.

13.2. Runner-Up Semantic Analysis

We analyze runner-up class patterns to understand secondary reasoning. For each sample, we extract the rank-2 prediction from all 10 models and compute:

- **Runner-up agreement**: Fraction of models agreeing on rank-2 class
- **Semantic relatedness**: WordNet path-based score between top-1 and top-2

Bimodal Distribution. Runner-up agreement exhibits characteristic bimodality:

- 23% of samples: unanimous runner-up (> 0.9 agreement)
- 42% of samples: moderate agreement (0.4-0.7)
- 35% of samples: low agreement (< 0.4)

Semantic Coherence. High-agreement samples show semantic clustering. For unanimous runner-ups, the mean WordNet path-based relatedness score is 0.18. Representative examples include "husky"/"malamute" and "convertible"/"sports_car". Low-agreement samples exhibit diverse alternatives reflecting architecture-dependent biases: CNNs favor texture-similar classes while transformers select shape-similar alternatives.

Accuracy Correlation. Samples with unanimous runner-ups achieve 96.2% accuracy, while those with diverse runner-ups drop to 73.8%. This 22.4 pp gap confirms that runner-up consensus signals robust feature detection beyond top-1 agreement.

13.3. Augmentation Pool Size Effect

We investigate TTA-JS stability with varying augmentation counts $K \in \{2, 3, 4, 6, 8\}$:

- $K = 2$ (identity + hflip): $\rho(\text{TTA-JS}, \text{error}) = 0.312$
- $K = 3$ (+ rotation): $\rho = 0.367$
- $K = 4$ (+ color jitter): $\rho = 0.389$
- $K = 6$ (+ blur + JPEG): $\rho = 0.415$ (reported in main results)
- $K = 8$ (+ noise + more rotations): $\rho = 0.421$

Correlation saturates around $K = 6$, with diminishing returns beyond. Computational cost scales linearly with K, so $K \in [4, 6]$ provides optimal tradeoff. Using only geometric augmentations ($K = 2$, identity + hflip) yields substantial speedup ($3\times$ faster) with acceptable correlation (0.31), suitable for latency-critical deployments.

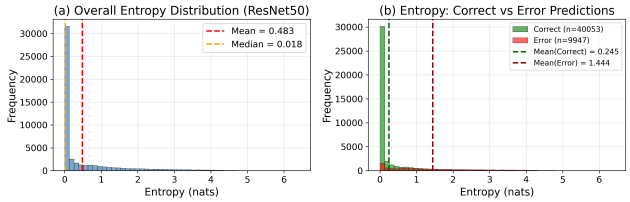

Figure 5. **Entropy distribution analysis for ResNet50.** (a) Overall entropy distribution shows strong concentration at low entropy (median=0.018 nats), with long tail extending to 6+ nats. (b) Correct predictions (green) concentrate heavily at very low entropy (mean=0.245), while errors (red) shift dramatically toward higher entropy (mean=1.444). The 1.2-nat mean difference validates entropy as a fundamental reliability indicator.

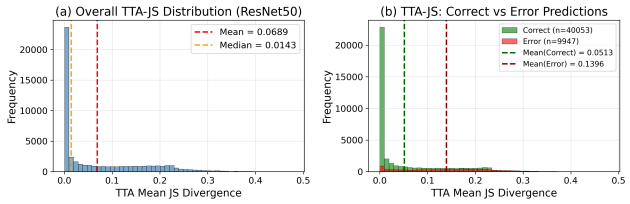

Figure 6. **TTA-JS distribution reveals prediction stability patterns.** (a) Overall TTA-JS distribution (median=0.014) shows most predictions are stable under augmentation, with 55.6% achieving very low JS (< 0.03). (b) Correct predictions exhibit $2.72\times$ lower TTA-JS than errors (0.051 vs. 0.140), demonstrating that within-model agreement strongly signals reliability. The clear separation between correct and error distributions validates TTA-JS as a training-free single-model uncertainty metric based on multiple test-time augmentations.

14. Conclusion

This supplementary material provides mathematical foundations, implementation specifications, and selected additional analyses that support the main paper's findings. Key contributions are as follows:

- Formal properties of JS divergence and temperature scaling used in the paper
- A comprehensive TTA-JS experiment showing within-model agreement as a complementary uncertainty signal
- Complete implementation details enabling reproduction of the reported procedures
- Supplementary qualitative analyses and deployment-oriented interpretation

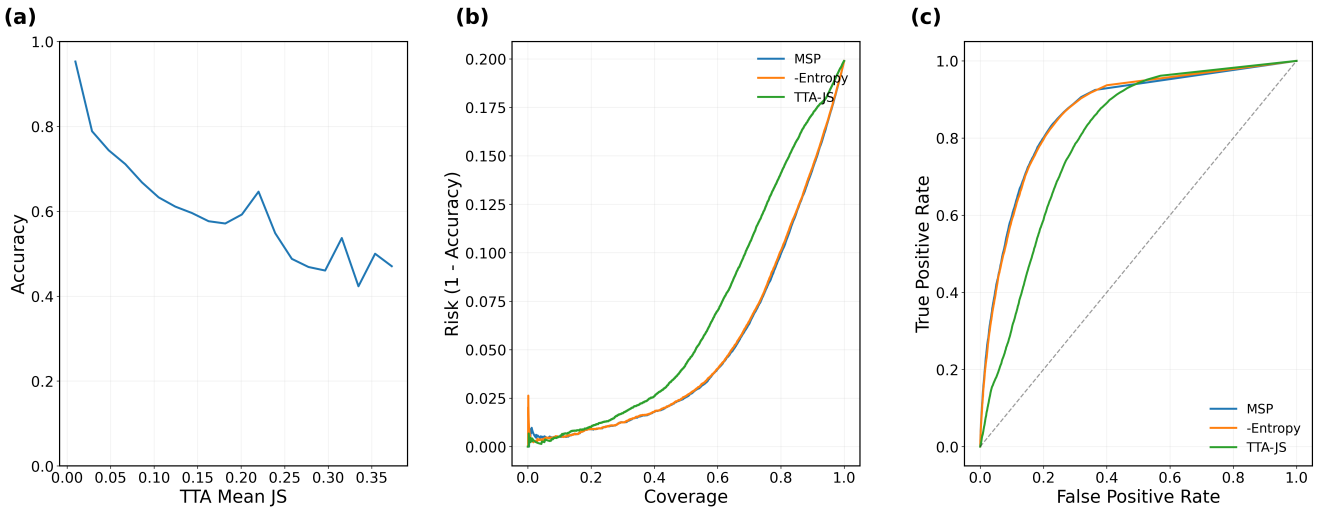

Figure 7. **Test-Time Augmentation as single-model uncertainty signal.** (a) Accuracy decreases monotonically with TTA-JS, from 98.7% (low JS) to 57.3% (high JS). (b) Risk-coverage curves show TTA-JS provides competitive selective prediction (AURC=0.069), though MSP (0.051) and entropy (0.052) achieve slightly better aggregate performance. (c) Error detection via ROC curves (AUROC=0.797) demonstrates that TTA-JS captures reliability signals, though confidence-based baselines reach higher AUROC (0.866). The key finding: TTA-JS detects complementary failure modes invisible to scalar confidence, justifying hybrid approaches. All results on ResNet50, ImageNet validation set (N=50,000).