Knowledge-Based Systems

CATE: Consensus-Aware Calibration for Test-Time Prompt Tuning via Energy Anchoring

--Manuscript Draft--

Manuscript Number:	KNOSYS-D-25-10058R1
Article Type:	Full Length Article
Keywords:	Test-time Prompt Tuning; Confidence Calibration; Energy-based; Text Similarity
Abstract:	Test-Time Prompt Tuning (TPT) enables vision-language models (VLMs) to adapt to new domains without labeled data by minimizing marginal entropy (MEM) over augmented views. However, MEM inherently sharpens the output distribution without correcting its underlying uncertainty, leading to inflated confidence that does not reflect true reliability under distribution shift, thus resulting in a significant calibration mismatch between predicted confidence and actual accuracy. Moreover, existing TPT methods rely on softmax-based view selection, which favors high-confidence but visually redundant samples, sacrificing feature diversity that is crucial for robust adaptation. To address these limitations, we introduce CATE—a consensus-aware test-time prompt tuning framework via energy anchoring—that leverages inter-view consensus as a reliable unsupervised signal. CATE identifies a stable anchor class through energy-based voting across low-uncertainty views, then aggregates semantically consistent predictions to form a consensus set. To calibrate adaptation, each view is weighted by both energy-based confidence and cross-modal similarity, enabling entropy minimization that prioritizes both reliability and semantic alignment. Experiments on fine-grained and distribution-shifted benchmarks show that CATE achieves significantly better calibration than existing methods, while maintaining competitive accuracy. Our results highlight the importance of incorporating visual uncertainty and consensus signals into test-time adaptation.

Response Letter to ` Knowledge-Based Systems ` Submission

Paper ID: KNOSYS-D-25-10058
Paper Title: CATE: Consensus-Aware Calibration for Test-Time Prompt Tuning via Energy Anchoring

Response to the Editor

Editor's Comment: Please be aware that, as recommended by the referees, kindly use the English Language Editing service available from Elsevier's Author Services. In your cover letter please provide a point-by-point response to each comment.

Author Response:

Thank you for your valuable recommendation and for the opportunity to improve our manuscript. We have made all efforts to address the concerns regarding content, language, and experimentation in the revised version.

Regarding the language, we have taken your advice very seriously. We have conducted a comprehensive proofreading and linguistic refinement of the entire manuscript to ensure it meets the high publication standards of the journal. The revised text has been thoroughly polished to guarantee clarity, correctness, and professional academic tone.

Regarding the revision, we have refined the manuscript to ensure comprehensiveness and have expanded our experimental section to provide a more robust validation of our proposal. As requested, a detailed point-by-point response to each comment from the referees is provided below. We hope these revisions significantly enhance the quality and impact of our work.

Response to Reviewer #1

Comment # 1:

The contribution currently appears incremental compared to existing works such as C-TPT and O-TPT. Please emphasize more clearly what is fundamentally novel in your method and add a subsection explaining how CATE differs conceptually and practically from prior TPT calibration strategies.

Author Response:

We thank the reviewer for the insightful comment regarding the novelty and positioning of our contribution. We agree that, without sufficient clarification, the distinction between CATE and prior TPT calibration methods such as C-TPT and O-TPT may appear incremental.

In response, we have substantially revised the Introduction to explicitly articulate the fundamental conceptual shift introduced by CATE. In particular, we clarify that prior works primarily intervene at the objective level via regularization, while still relying on confidence-based view selection, whereas CATE reformulates TPT from a selection-level perspective, identifying biased view selection as the root cause of miscalibration. We further formalize this issue as the Overconfidence-Redundancy Dilemma and motivate energy-anchored inter-view consensus as a principled alternative.

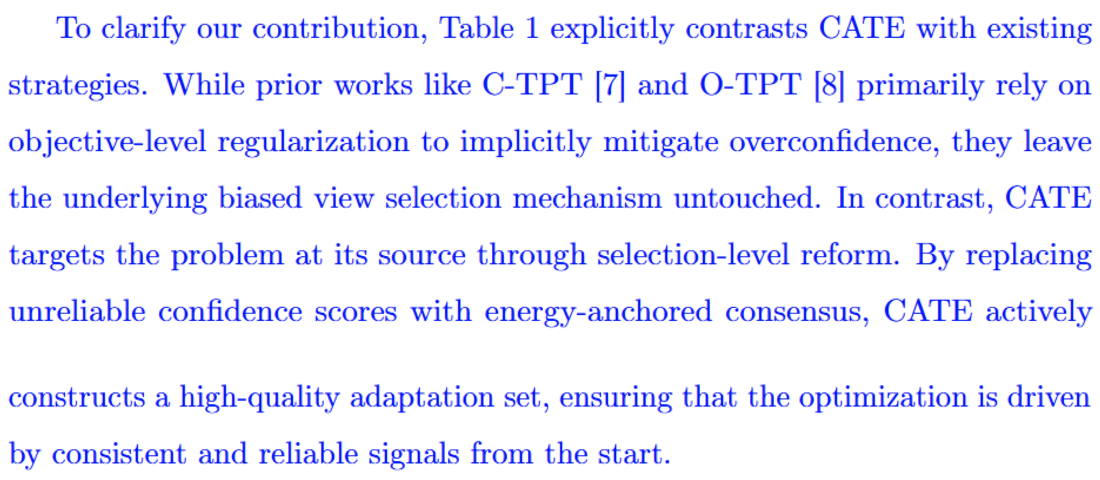
To make this distinction explicit and accessible, we additionally introduce a comparative table that contrasts CATE with existing methods in terms of point of intervention and selection mechanism, highlighting that CATE represents a qualitatively different strategy rather than an incremental modification.

Action taken:

•We rewrote and expanded the Introduction to emphasize the core conceptual novelty of CATE as a selection-level reform of TPT, rather than an objective-level regularization.

•We introduced Table 1 (Method Comparison) to explicitly contrast CATE with C-TPT and O-TPT, clarifying their conceptual and practical differences.

•We added explicit discussion distinguishing confidence-based filtering from energy-anchored consensus selection, and formalized the motivation through the Overconfidence-Redundancy Dilemma.



To clarify our contribution, Table 1 explicitly contrasts CATE with existing strategies. While prior works like C-TPT [7] and O-TPT [8] primarily rely on objective-level regularization to implicitly mitigate overconfidence, they leave the underlying biased view selection mechanism untouched. In contrast, CATE targets the problem at its source through selection-level reform. By replacing unreliable confidence scores with energy-anchored consensus, CATE actively constructs a high-quality adaptation set, ensuring that the optimization is driven by consistent and reliable signals from the start.

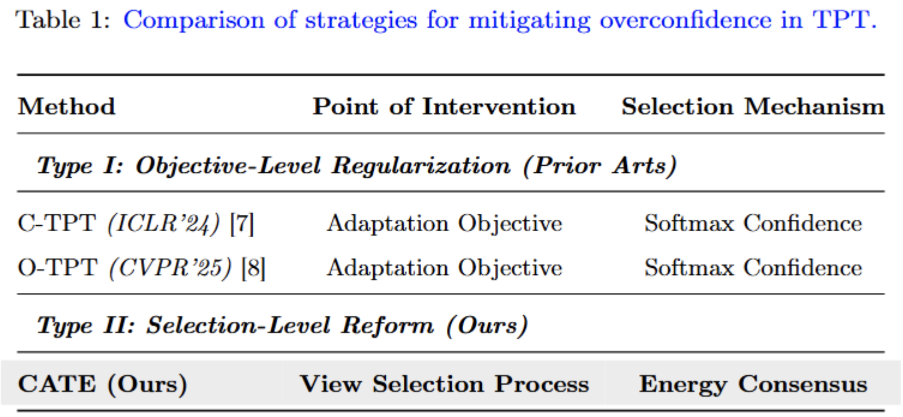


Table 1: Comparison of strategies for mitigating overconfidence in TPT.

Method	Point of Intervention	Selection Mechanism
Type I: Objective-Level Regularization (Prior Arts)		
C-TPT *(ICLR'24)* [7]	Adaptation Objective	Softmax Confidence
O-TPT *(CVPR'25)* [8]	Adaptation Objective	Softmax Confidence
Type II: Selection-Level Reform (Ours)		
CATE (Ours)	**View Selection Process**	**Energy Consensus**

Comment # 2:

The paper lacks theoretical depth in motivating why energy-guided consensus should reliably improve calibration. The revision should include either theoretical justification (even light analytical reasoning or propositions) or an explicit discussion of the method's theoretical limitations.

Author Response:

We thank the reviewer for pointing out the need for stronger theoretical motivation behind energy-guided consensus in improving calibration. We agree that, beyond empirical evidence, it is important to provide analytical reasoning to explain *why* energy constitutes a more reliable signal than softmax confidence under distribution shift.

In response, we have added a dedicated **Theoretical Analysis** subsection that provides light but principled analytical justification for our design choices. Rather than claiming a full probabilistic guarantee, this analysis focuses on two fundamental properties that directly motivate CATE: (i) **density sensitivity** and (ii) **gradient stability under perturbation**. These properties clarify why energy-guided consensus is inherently more robust than confidence-based selection when constructing adaptation sets at test time.

Action taken:

• We added a new subsection "Theoretical Analysis" (Section 4.3) that provides analytical reasoning supporting energy-guided consensus.

• We show that softmax confidence can be expressed as a biased surrogate of energy, explaining its saturation and failure under distribution shift, while energy remains sensitive to overall logit mass.

• We further analyze the gradient structure of the energy function, demonstrating that energy induces a probability-weighted aggregation of class gradients, leading to cancellation effects and improved stability under view-level perturbations.

• This analysis directly motivates the robustness of energy-based anchor selection and explains its effectiveness for calibration, without overstating theoretical guarantees.

4.3. Theoretical Analysis

We analyze the properties of energy versus softmax confidence to clarify why energy provides a more reliable signal for anchor selection under distribution shift. We formalize two key distinctions—density sensitivity and gradient stability—that directly motivate the design of CATE.

(1) Confidence as a biased energy surrogate. Let $f_k(\mathbf{x})$ denote the logit of class k, and define the free energy as $E(\mathbf{x}) = -\log \sum_{k=1}^{K} \exp(f_k(\mathbf{x}))$. Let $f^{\max}(\mathbf{x}) = \max_k f_k(\mathbf{x})$ be the maximum logit and $\mathcal{C}(\mathbf{x}) = \max_k p_k(\mathbf{x})$ be the softmax confidence. We observe the following identity:

$$\log \mathcal{C}(\mathbf{x}) = f^{\max}(\mathbf{x}) - \log \sum_{k=1}^{K} \exp(f_k(\mathbf{x})) = f^{\max}(\mathbf{x}) + E(\mathbf{x}). \tag{16}$$

Eq. (16) reveals that confidence is effectively an energy-based quantity biased by the dominant logit $f^{\max}(\mathbf{x})$. When the logit gap is large, $\mathcal{C}(\mathbf{x})$ saturates to 1 regardless of the absolute logit mass. In contrast, $E(\mathbf{x})$ aggregates contributions from all logits, providing a non-saturating summary of the input density. This structural difference explains why confidence fails to distinguish high-energy OOD samples from in-distribution data, whereas energy remains sensitive.

(2) Energy as a smoothed gradient aggregate. Let $p_k(\mathbf{x}) = \mathrm{softmax}(f(\mathbf{x}))_k$. By direct differentiation, the gradient of the energy function is:

$$\nabla_{\mathbf{x}} E(\mathbf{x}) = -\sum_{k=1}^{K} p_k(\mathbf{x}) \, \nabla_{\mathbf{x}} f_k(\mathbf{x}). \tag{17}$$

This indicates that the energy gradient is a probability-weighted average of all class-wise logit gradients. Applying the triangle inequality yields:

$$\|\nabla_{\mathbf{x}} E(\mathbf{x})\| \leq \sum_{k=1}^{K} p_k(\mathbf{x}) \|\nabla_{\mathbf{x}} f_k(\mathbf{x})\| \leq \max_k \|\nabla_{\mathbf{x}} f_k(\mathbf{x})\|. \tag{18}$$

By comparison, the local variation of confidence is dominated by a single class direction: $\nabla_{\mathbf{x}} f^{\max}(\mathbf{x}) = \nabla_{\mathbf{x}} f_{y^*}(\mathbf{x})$. Eqs. (17)–(18) formalize that energy responds to input perturbations through an aggregate of multiple logit directions. This probability-weighted averaging induces a cancellation effect among conflicting class gradients, rendering energy-based ranking locally smoother and empirically more stable across augmented views than the mode-seeking confidence score.

Comment # 3:

The use of energy as the core uncertainty signal is motivated intuitively but not deeply analyzed. Please expand on the rationale, provide a comparison with alternative uncertainty measures, and consider adding ablation experiments highlighting when energy works well or fails under strong domain shifts.

Author Response:

Thank you for this valuable suggestion. We agree that, beyond intuitive motivation, the choice of energy as the core uncertainty signal should be supported by clearer analytical rationale and contrasted against alternative uncertainty measures. In the revision, we strengthen this aspect in two ways.

First, we provide a dedicated theoretical analysis to clarify why energy is a more reliable anchor signal than softmax confidence under distribution shift. Our analysis formalizes (i) confidence as a biased surrogate of energy due to its dependence on the dominant logit and its saturation behavior, and (ii) the stability advantage of energy via its probability-weighted gradient aggregation (smoother response under perturbations). This offers a principled explanation for why energy-guided consensus construction is more stable than confidence-driven selection.

Second, we include an ablation study to validate the necessity of key components in CATE, including the energy-based consensus selection and the per-view reweighting design. The ablation results empirically confirm that replacing energy-based selection with confidence-based selection leads to the largest degradation, especially in calibration, demonstrating that energy is the primary driver of the improvement.

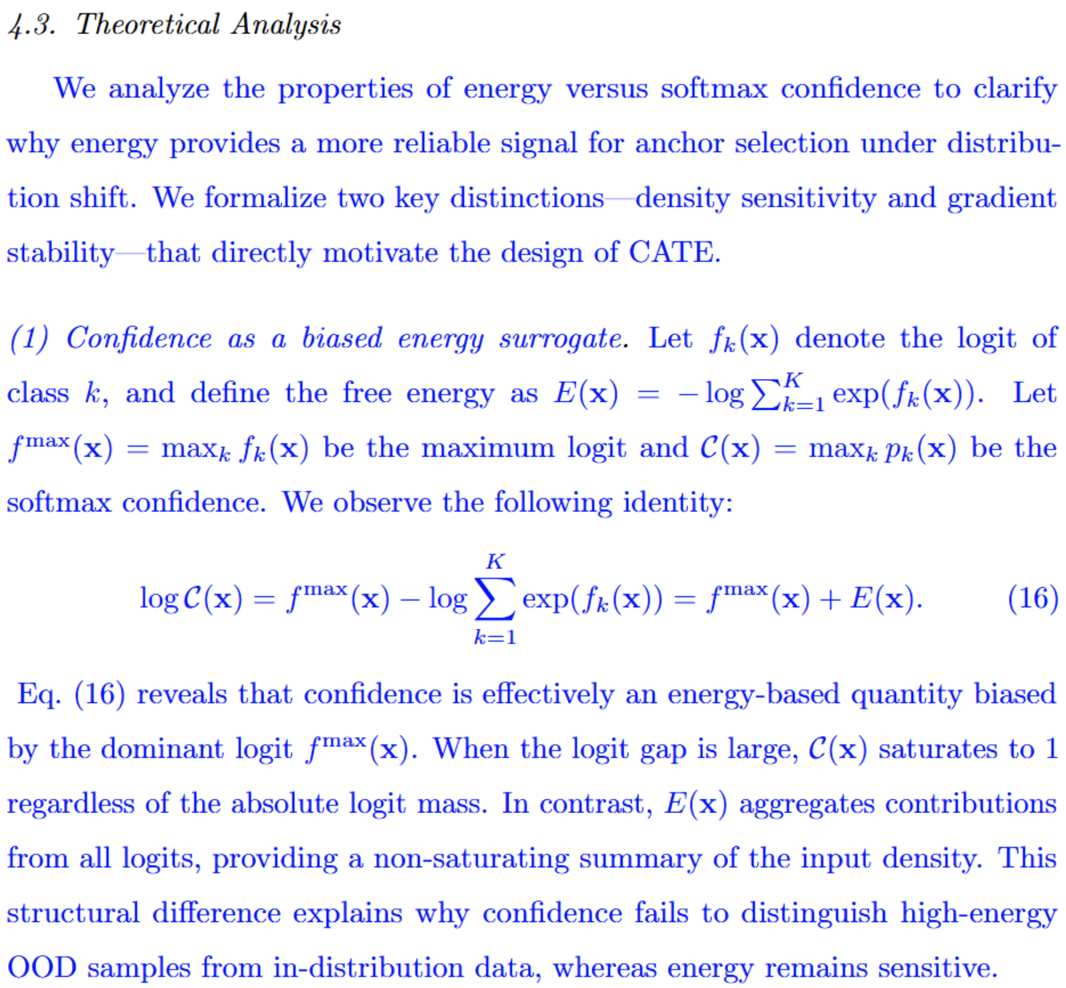
Regarding “when energy works well or fails under strong domain shifts,” we explicitly acknowledge this limitation in our newly added limitations discussion: under extreme shifts, low energy does not guarantee correctness and anchor errors may still occur. We highlight this as a meaningful future direction, e.g., adaptive thresholding and more robust/differentiable selection mechanisms.

Action taken:

• Added a theoretical justification section: We added a new subsection “Theoretical Analysis” (Section 4.3) that provides analytical reasoning for energy anchoring, including (i) the confidence–energy identity and saturation behavior, and (ii) gradient stability properties of energy.

• Added ablation experiments: We added a new subsection “Ablation Studies” (Section 5.3) with a dedicated table (Table 6) evaluating the effectiveness and necessity of the energy-based consensus selection and the dual-component weighting mechanism.

- Added explicit discussion of limitations under strong shifts: We added a “Limitations and Future Directions” section (Section 6) clarifying that energy may still produce confident errors under extreme distribution shifts, and outlining directions to address such cases.



4.3. Theoretical Analysis

We analyze the properties of energy versus softmax confidence to clarify why energy provides a more reliable signal for anchor selection under distribution shift. We formalize two key distinctions—density sensitivity and gradient stability—that directly motivate the design of CATE.

(1) Confidence as a biased energy surrogate. Let $f_k(\mathbf{x})$ denote the logit of class k, and define the free energy as $E(\mathbf{x}) = -\log \sum_{k=1}^{K} \exp(f_k(\mathbf{x}))$. Let $f^{\max}(\mathbf{x}) = \max_k f_k(\mathbf{x})$ be the maximum logit and $\mathcal{C}(\mathbf{x}) = \max_k p_k(\mathbf{x})$ be the softmax confidence. We observe the following identity:

$$\log \mathcal{C}(\mathbf{x}) = f^{\max}(\mathbf{x}) - \log \sum_{k=1}^{K} \exp(f_k(\mathbf{x})) = f^{\max}(\mathbf{x}) + E(\mathbf{x}). \qquad (16)$$

Eq. (16) reveals that confidence is effectively an energy-based quantity biased by the dominant logit $f^{\max}(\mathbf{x})$. When the logit gap is large, $\mathcal{C}(\mathbf{x})$ saturates to 1 regardless of the absolute logit mass. In contrast, $E(\mathbf{x})$ aggregates contributions from all logits, providing a non-saturating summary of the input density. This structural difference explains why confidence fails to distinguish high-energy OOD samples from in-distribution data, whereas energy remains sensitive.