A Metacognitive Framework for Test-Time Adaptation: Taxonomy and Research Challenges

Dario Loi[1,*], **Gabriele Onorato**[1], **Antonella Coviello**[1], **Shahriar Rifat**[1], **Nathaniel D. Bastian**[2], **Francesco Restuccia**[1]

[1]Northeastern University
[2]Pathfinder Applied Research, LLC
[*]loi.d@northeastern.edu

Abstract

In Test-Time Adaptation (TTA), when out-of-distribution samples interleave unpredictably with corrupted in-distribution samples, adaptation without corrective measures is prone to significant performance decrease. Existing protocols rely on unreliable cues (e.g., predictive entropy) which are unable to distinguish corrupted in-distribution data from semantically novel open-set samples. To address this issue, we formalize TTA through a two-level metacognitive framework that decouples task execution from control, facilitating a unified taxonomy that categorizes existing TTA protocols across three core axes, i.e., *which knowledge* is monitored, *which part* of the models are targeted for adaptation, and *how* adaptation occurs. We provide evidence that robust corrective measures in TTA require monitoring grounded in structured internal representations to reliably estimate model competence. Lastly, we frame monitoring as a resource-allocation problem, examine neuromorphic substrates for bounded self-regulation, and outline key open challenges.

Code — https://github.com/Restuccia-Group/Metacog-TTA

Introduction

Deep Neural Networks (DNNs) attain strong performance across perception tasks such as image classification, object detection, and semantic segmentation (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016; Dosovitskiy et al. 2021), provided that inference is performed on samples drawn from the same distribution as the training data. This assumption is frequently violated in practice. Sensor noise, compression artifacts or changes in acquisition conditions can shift the input distribution, degrading model accuracy. This has been shown to occur even for shifts that a human observer would regard as minor (Hendrycks and Dietterich 2019; Recht et al. 2019). The gap between In-Distribution (ID) and Out-of-Distribution (OOD) performance has motivated a body of work on improving the resilience of DNNs (Madry et al. 2018; Hendrycks et al. 2020; Gulrajani and Lopez-Paz 2021).

Test-Time Adaptation (TTA) can be used to recover model performance during inference, requiring no access to labels or training data (Wang et al. 2021; Schneider et al. 2020). Successful adaptation requires that the test stream is only degraded by covariate-shifted ID data, which rarely holds in real-world settings. Niu et al. (2023) removed this assumption by allowing the test stream to contain OOD samples, label imbalance and small batch sizes. In such regimes, unconditional updates lead to a decrease in performance. Recent re-evaluations have also confirmed that current methods degrade or diverge once the stream is noisy, temporally correlated or open-set (Zhao et al. 2023; Press et al. 2023; Gong et al. 2023).

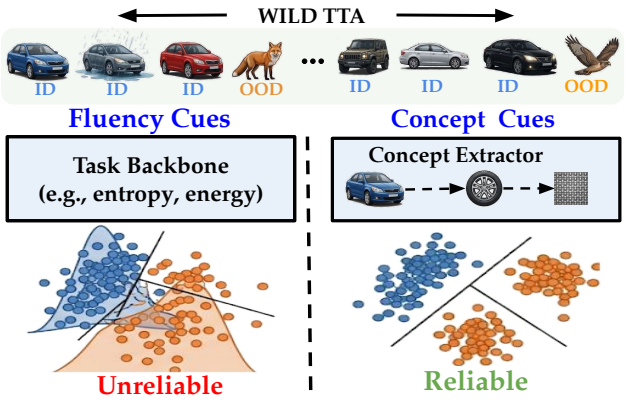


Figure 1: Standard vs Concept Cues in Wild Test-Time Adaptation (WTTA)

The Need for Metacognition. An effective solution to the TTA problem requires a model that can assess prediction trustworthiness and self-regulate its behavior. This self-directed capacity, the monitoring and control of a cognitive process rather than its execution, is what cognitive psychology calls metacognition (Flavell 1979). Nelson and Narens (1990) formalize the metacognitive model as a two-level architecture in which an object level performs the task and a meta level holds a model of the object level, monitors its state, and issues control signals that modify it. Current TTA can already be framed through this lens. Specifically, the DNN is the *object level*, the adaptation procedure is the *meta level*.

The ineffectiveness of TTA in OOD regimes can be attributed to poor object-level modeling, as current methods use a meta-level estimate of task correctness. Entropy, confidence margins, uncertainty, and free energy all serve as such cues, es-

[*]Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

timating the *Feeling of Rightness* (Thompson, Prowse Turner, and Pennycook 2011; Ackerman and Thompson 2017). These estimates are comparable to experience-based cues (Koriat 1997; Koriat and Levy-Sadot 1999), heuristics for task performance based on the *fluency* of computation, where effortless processing or confident answers are taken to be a proxy for accuracy. In TTA, models can exhibit high fluency despite being incorrect in their prediction, rendering single-cue monitoring brittle in scenarios that require robustness.

Our Position. To maintain adaptation performance in adversarial conditions, adaptation must move beyond monolithic fluency metrics. Robust TTA requires monitoring that is more faithful than unreliable metrics like entropy, the information-theoretic analogue of the metacognitive *Feeling of Rightness*, in order to remain robust in the presence of OOD data. In this work, we divide metacognitive monitoring into two subcategories: *correctness* monitoring, assessing whether a prediction is erroneous; and *competence* monitoring, determining whether a task falls within the model's operational domain. We provide a metacognition-oriented taxonomy of the existing TTA literature, and a unifying template with which to represent any TTA protocol. To evaluate future metacognitive adaptation protocols, we extend the existing object-level evaluation procedures with meta-level metrics that measure the metacognitive efficiency of monitoring mechanisms, independently of the underlying object-level. We propose a class of concept-based readouts (Alain and Bengio 2017; Kim et al. 2018) to act as interpretable deliberate cues. Additionally, we empirically validate the impact of deliberate cues on the robustness and stability of adaptation protocols, admitting up to $22.1\times$ fewer OOD samples for adaptation, while recovering up to 72% of the achievable gain under an ideal OOD detector. Finally, we frame metacognitive monitoring as a resource-allocation problem and discuss event-driven neuromorphic computing as a promising substrate for bounded self-regulation, outlining the architectural challenges required to close the loop between monitoring and adaptive control.

Background: Metacognitive Adaptation

Metacognition originates in clinical and cognitive psychology, where it denotes the monitoring and regulation of ongoing cognitive operations (Flavell 1979; Ackerman and Thompson 2017), most often learning tasks. The concept of metacognition has acquired diverse meanings across psychology (Wells and Matthews 1996), education (Veenman, Van Hout-Wolters, and Afflerbach 2006) and Artificial Intelligence (AI) (Cox 2005). According to the working definition, which closely follows Nelson and Narens (1990), we split any adaptive learning system into two (as shown in Figure 2): an object level that executes the task, and a meta level that maintains an internal model of task progress to issue steering control signals.

Automatic and deliberate knowledge. Beyond monitoring and control, the original framework by Flavell (1979) incorporates a third core component, metacognitive knowledge and experiences, which together form the diagnostic *cues* the meta level leverages to guide regulation. The nature of this information partitions metacognitive processes into *automatic* and *deliberate* (Shea et al. 2014; Thompson 2009). Automatic monitoring relies on fast, experience-based heuristics that act as a proxy for task performance without the need for external validation; one such example is the *Feeling of Rightness* (Thompson 2009), a subjective sense of correctness induced automatically by the *fluent* completion of a task. In TTA, automatic monitoring corresponds to cues extracted from a model's outputs, such as predictive entropy (Wang et al. 2021). Conversely, deliberate monitoring relies on deeper object-level introspection (Shea et al. 2014). Adaptation protocols realize deliberate monitoring via structured comparisons across transformed views, teacher networks, or ensemble agreements (Wang et al. 2022; Nguyen et al. 2023). This automatic–deliberate dichotomy extends symmetrically to control, distinguishing unconditioned loss optimization, such as entropy minimization, from reference-anchored constraints, such as consistency and invariance regularization.

Test-Time Adaptation (TTA). The rapid expansion of TTA has prompted several complementary efforts to taxonomize the literature. Liang, He, and Tan (2024) group different TTA methods by their adaptation regime, differentiating protocols by domain, batch and online adaptation settings. Wang et al. (2024) focus on continual adaptation, where weights are adapted in non-stationary conditions such as distribution shifts. A parallel line of work organizes the field by the test stream conditions (Zhao et al. 2023; Press et al. 2023; Marsden, Döbler, and Yang 2024) under which each protocol remains stable. In this work, we instead organize the field through the structural lens of the two-level metacognitive loop (Nelson and Narens 1990). Under this formulation, the *object level* is the feedforward network executing inference on the test stream, while the *meta level* is the supervisory adaptation algorithm that monitors task execution and issues steering signals to modify model state.

The object level. Let $f_\theta : \mathcal{X} \to \Delta^C$ denote a parameterized DNN mapping inputs to the probability simplex over C target classes, trained on a source distribution P_S. In our metacognitive adaptation framework, this represents the *object level*. At time step t, the network receives an unlabeled batch $\mathcal{B}_t = \{x_{t,1}, \ldots, x_{t,n}\} \sim P_T$, computes posterior distributions $f_{\theta_t}(x_{t,i})$, and assigns predictions $\hat{y}_{t,i} = \arg\max_{c\in\{1,\ldots,C\}} [f_{\theta_t}(x_{t,i})]_c$. Governed entirely by its fixed parameterization θ_t, the object level possesses no intrinsic mechanism to detect distribution shift or regulate its internal state. For instance, under standard covariate shift, where $P_T(y \mid x) = P_S(y \mid x)$ but $P_T(x) \neq P_S(x)$, the ground-truth labeling rule remains invariant while input marginals shift (Schneider et al. 2020; Liang, He, and Tan 2024), thus degrading inference performance.

The minimal meta level. Any mechanism that aims to restore performance by detecting distribution shift and regulating the object level constitutes the *meta level* of our abstract protocol. Early TTA protocols (Wang et al. 2021, 2022) realize this by optimizing a self-supervised surrogate loss ℓ over incoming test batches

$$\mathcal{L}_t(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(x_{t,i}; \theta\big), \quad \theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_t(\theta_t). \quad (1)$$

Under this formulation, updates are applied without evaluating sample trustworthiness or prediction validity. Subsequent

methods modify the definition of ℓ using either *automatic cues* derived from a single forward pass, such as entropy (Grandvalet and Bengio 2004; Wang et al. 2021; Niu et al. 2022, 2023; Zhang et al. 2025), pseudo-labels, and Maximum Softmax Probability (MSP) (Gong et al. 2023) or *deliberate cues* computed against an explicit reference, such as consistency across augmented views (Wang et al. 2022; Döbler, Marsden, and Yang 2023; Nguyen et al. 2023) and self-supervised pretext tasks (Liu et al. 2021; Sun et al. 2020).

Monitoring under stream contamination. In practical deployments, test stream degradation frequently deviates from the pure covariate shift assumption, degrading unconstrained adaptation protocols (Zhao et al. 2023; Press et al. 2023; Gong et al. 2023). To formalize these settings, Niu et al. (2023) introduced WTTA, modeling the test stream as a mixture

$$P_t = (1 - \pi_t)\widetilde{P}_t + \pi_t Q_t, \tag{2}$$

where $\widetilde{P}_t$ is the covariate-shifted ID component sharing the training label space $\mathcal{Y}$, Q_t is an OOD component whose label set is disjoint from $\mathcal{Y}$, and $\pi_t \in [0, 1]$ is an unobserved, time-varying contamination rate. Beyond open-set contamination, TTA protocols also face (i) label distribution drift ($P_T(y) \neq P_S(y)$) (Niu et al. 2023; Marsden, Döbler, and Yang 2024), (ii) temporal correlation between consecutive samples (Gong et al. 2022; Yuan, Xie, and Li 2023), and (iii) very small batches ($n \to 1$), which invalidate batch-normalization updates (Niu et al. 2023; Zhang, Levine, and Finn 2022). The mixture in Equation (2) poses a monitoring problem that automatic cues cannot solve: corrupted ID inputs should drive adaptation, whereas semantically novel OOD inputs should be excluded from it, yet both can produce high predictive entropy $H(p)$. Automatic cues estimate correctness rather than domain competence (Thompson 2009), and distribution shift makes predictions overconfident (Guo et al. 2017; Snoek et al. 2019). Entropy gates therefore admit confidently misclassified samples from Q_t, whose gradients drive the network toward collapse (Niu et al. 2023; Zhang et al. 2025; Han, Na, and Hwang 2025).

A Unified Meta-Level Formulation. To systematically compare existing protocols, we express the metacognitive loop as a general discrete-time dynamical system. We divide the system's internal state into two components, a static state ϕ, fixed prior to deployment (e.g., initial parameters θ_0, source normalization statistics, Fisher importances (Niu et al. 2022), class prototypes (Li et al. 2023b) or static concept dictionaries) and a dynamic state s_t which can be updated online by the TTA protocol (e.g., replay buffers, running statistics, or teacher model weights). At each time step t, a monitoring map m evaluates the object level's state to produce knowledge K_t, which controls subsequent adaptation and inference

$$\begin{aligned}
K_t &= m(\mathcal{B}_t, \theta_t, s_t, \phi), \\
n_t &= \nu(\mathcal{I}_t, \theta_t, \phi), \\
L_t(\theta) &= \frac{1}{|\mathcal{I}_t|} \sum_{x \in \mathcal{I}_t} w_t(x)\ell(x; \theta, \phi, s_t, n_t) + \lambda_t \Omega(\theta, \phi), \\
\theta_{t+1} &= A_t \odot \theta_0 + (\mathbf{1} - A_t) \odot \mathcal{U}_{\eta_t}(\theta_t, L_t), \\
s_{t+1} &= \Sigma(s_t, \theta_{t+1}, \mathcal{B}_t, K_t),
\end{aligned} \tag{3}$$

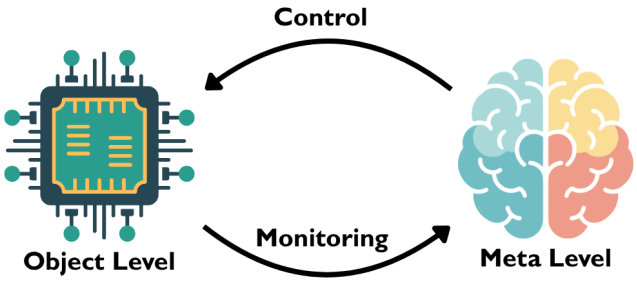


Figure 2: The two levels of Metacognitive Adaptation.

$$\hat{Y}_t = g(\mathcal{B}_t, \theta_t, s_t, \phi, K_t) \in (\mathcal{Y} \cup \{\perp\})^n.$$

In this formulation, the active set $\mathcal{I}_t \subseteq \mathcal{B}_t \cup \mathcal{M}_t$ draws instances from both the current batch and an episodic memory $\mathcal{M}_t \subseteq s_t$, with sample-wise weights $w_t(x)$ accommodating selective filtering or explicit entropy maximization (Gao, Zhang, and Liu 2024). The surrogate objective L_t penalizes parameter divergence from ϕ via regularizer Ω scaled by λ_t. The normalization context $n_t = \nu(\mathcal{I}_t, \theta_t, \phi)$ allows for cross-sample dependencies in the surrogate loss, such as those introduced by normalization layers. Parameter optimization is performed by an update operator $\mathcal{U}_{\eta_t}$ (e.g., gradient descent or sharpness-aware minimization (Foret et al. 2021)) restricted to a designated parameter subspace $\Theta_{\text{ad}} \subseteq \Theta$, while the vector $A_t \in [0, 1]^{|\theta_0|}$ controls parameter interpolation toward θ_0, stabilizing adaptation over long time horizons. Finally, the decision function g outputs predictions over the extended label space $\mathcal{Y} \cup \{\perp\}$, providing an explicit mechanism for abstention. Canonical unmonitored protocols correspond to the degenerate case where control variables are completely decoupled from K_t. Setting $\mathcal{I}_t = \mathcal{B}_t$, $w_t = 1$, $\lambda_t = 0$, $A_t = \mathbf{0}$, $g = \text{argmax}$, $\nu = (\mu_{\mathcal{I}_t}, \sigma^2_{\mathcal{I}_t})$, with Σ accumulating running normalization statistics via exponential moving averaging, and adapting over the affine normalization subspace $\Theta_{\text{ad}} = \Theta_{\text{affine}}$, recovers Tent (Wang et al. 2021). More broadly, TTA protocols are categorized by how they instantiate Equation (3) along three primary axes: ❶ *which knowledge* is extracted by m and optimized by ℓ (automatic vs. deliberate); ❷ *what loci* are adapted among Θ_{ad}, Σ, and g; and ❸ *how updates are regulated* through the conditioning of $(\mathcal{I}_t, w_t, \eta_t, \lambda_t, A_t)$ on K_t. The resulting taxonomy is shown in Tables 1 and 2.

❶ **Which knowledge.** We classify both the surrogate loss ℓ and the monitoring mapping m by their informational grounding: *automatic* (A) when evaluated strictly on single-pass forward outputs through predictive entropy (Wang et al. 2021), Dirichlet evidence (Zhang et al. 2025) or free energy (Chung et al. 2025); *deliberate* (D) when evaluated against an explicit offline (ϕ) or online (s_t) reference such as source normalization statistics, prototype banks or teacher representations (Wang et al. 2022; Yuan, Xie, and Li 2023; Li et al. 2023b). Because a protocol's monitoring cue does not dictate the nature of its optimization objective, we report ℓ and m independently in Table 1.

❷ **What is adapted.** Control signals must target specific model components when performing TTA. Surveyed pro-

Protocol	Stream	❶ Update (ℓ)	❶ Trigger (m)	❷ Adapted	❸ When
BN-Adapt (Schneider et al. 2020)	B	–	–	stats	–
Tent (Wang et al. 2021)	–	entropy^{A}	–	affine, stats	–
MEMO (Zhang, Levine, and Finn 2022)	B	marginal entropy^{D}	–	all, stats	T
CoTTA (Wang et al. 2022)	C	teacher on views^{D}	source confidence^{D}	all, stats	S
RDumb (Press et al. 2023)	C	entropy^{A}	entropy^{A}; redundancy^{D}	affine, stats	FWT
EATA (Niu et al. 2022)	C	entropy^{A}; Fisher anchor^{D}	entropy^{A}; redundancy^{D}	affine, stats	FW
SAR (Niu et al. 2023)	BIOX	sharpness-aware entropy^{A}	entropy^{A}; loss average^{A}	affine	FE
SoTTA (Gong et al. 2023)	O	sharpness-aware entropy^{A}	max. prob.^{A}; balance^{D}	affine, stats	FM
NOTE (Gong et al. 2022)	BI	entropy on memory^{A}	source statistics^{D}; balance^{D}	stats, affine	M
RoTTA (Yuan, Xie, and Li 2023)	CI	teacher^{D}	entropy^{A}; age and balance^{D}	affine, stats	WM
DeYO (Lee et al. 2024)	BIX	entropy^{A}	entropy^{A}; label shift^{D}	affine, stats	FW
OWTTT (Li et al. 2023b)	O	prototypes; alignment^{D}	prototype distance^{D}	all, stats	FM
PeTTA (Hoang, Vo, and Do 2024)	CI	teacher; anchor to θ_0^{D}	divergence^{D}; memory^{A}	affine, stats	MG
ROID (Marsden, Döbler, and Yang 2024)	BCI	SLR^{A}; consistency^{D}	certainty^{A}; diversity^{D}	affine, stats, out	FWS
LAME (Boudiaf et al. 2022)	I	batch Laplacian^{D}	–	out	–
REM (Han, Na, and Hwang 2025)	BCIX	ranked masked views^{D}	–	affine	–
MECTA (Hong et al. 2023)	C	host entropy^{A}	statistics divergence^{D}	affine, stats, sel	G
GALA (Sahoo et al. 2025)	BC	entropy^{A}	gradient alignment^{D}	affine, sel	F
PALM (Maharana, Zhang, and Guo 2025)	C	entropy^{A}	gradient magnitude^{A}	affine, sel	G
LAW (Park et al. 2024)	CI	host entropy^{A}	layer Fisher trace^{D}	affine, sel	G

Table 1: Taxonomy of adaptation protocols; symbols defined in Table 2. Cues marked as automatic (A) or deliberate (D). **Stream**: B degenerate batch, C dynamic domain, I temporally correlated or imbalanced, O open-set contamination, X mixed domains.

tocols deploy updates across four main loci: (i) updating running statistics without gradient computation (Schneider et al. 2020; Lim et al. 2023) (ii) restricting gradient updates strictly to affine normalization parameters (Wang et al. 2021; Niu et al. 2022, 2023; Song et al. 2023; Hong et al. 2023) (iii) updating all model parameters, which may require periodic resets or partial restoration toward θ_0 to prevent divergence or collapse (Zhang, Levine, and Finn 2022; Wang et al. 2022) (iv) leaving weights untouched and adjusting output assignments directly by refining logits (Boudiaf et al. 2022), adjusting class priors (Marsden, Döbler, and Yang 2024), or querying a cache (Karmanov et al. 2024). While most protocols fix these loci statically at build time, a distinct class of methods treats the trainable subspace as a dynamic, cue-conditioned function $\Theta_{\text{ad}}(K_t)$ (designated sel in Table 1). These methods dynamically select target layers via a number of cues such as statistical divergence (Hong et al. 2023), gradient trajectory alignment (Sahoo et al. 2025), posterior gradient magnitude (Maharana, Zhang, and Guo 2025) or accumulated Fisher traces (Park et al. 2024).

❸ How to adapt. To prevent performance collapse, updates should be performed *selectively*, according to some adaptation trigger. At one extreme, updates can be *unconditional*, firing on every batch. Adaptation becomes *conditional* when knowledge K_t regulates control variables in Equation (3). This conditioning manifests through four primary mechanisms: (i) filtering batch elements or scaling sample weights (w_t) within the active set $\mathcal{I}_t$ (ii) regulating admission and replay policies within memory buffers $\mathcal{M}_t \subseteq s_t$ (iii) dynamically scaling learning rates η_t or regularization penalties λ_t (iv) triggering rollback vectors A_t to pull model weights toward reference state θ_0 upon detecting task performance collapse.

Reading the taxonomy. A systematic examination of Table 1 reveals that current adaptation protocols are fundamentally constrained by a representational bottleneck in meta-level monitoring. Along ❶, deliberate knowledge is disproportionately allocated to loss construction rather than update gating: while several protocols optimize deliberate objectives, their triggers mostly rely on forward-pass scalar heuristics or endogenous, drifting state summaries rather than fixed pre-deployment references, resulting in circular self-reinforcement of biases, which drives overconfidence and risks conflating corrupted ID samples with semantic OOD shifts. In ❷, adaptation loci remain largely static, only a few emerging dynamic protocols condition the adaptation subspace, but they do so based on heuristics rather than the semantic content of the underlying representations. Lastly, under ❸, supervisory regulation rarely exercises selective abstention ($g \in \mathcal{Y} \cup \{\perp\}$) to account for epistemic uncertainty. Because automatic cues cannot localize the underlying causes of computational failure across these three axes, closing the monitoring gap requires establishing a deliberate knowledge channel (Shea et al. 2014) grounded in structured, inspectable internal representations via concept-based interpretability.

Concept-Based Interpretability for Metacognitive Monitoring

Addressing the limitations of current TTA protocols requires designing a deliberate monitoring channel (Koriat 1997; Shea et al. 2014), thus mitigating the degradation of the adaptation signal in the presence of OOD contamination (see Equation (2)). Logit-derived heuristics have been shown to conflate ID data corrupted by covariate shifts with OOD data sampled from entirely different semantic classes (Guo et al.

	Entry	Component
❶ *Which knowledge*		
A	Automatic	ℓ, m
D	Deliberate	ℓ, m, ϕ, s_t
❷ *What is adapted*		
stats	Normalization statistics	Σ
affine	Affine normalization parameters	Θ_{ad}
all	All parameters	Θ_{ad}
out	Output assignments	g
sel	Subset chosen at run time by cue	$\Theta_{\text{ad}}(K_t)$
❸ *How the update is regulated*		
F	Restricted to samples passing cue threshold	$\mathcal{I}_t$
W	Weighted sample-wise by cue	w_t
M	Computed on samples held in memory	Σ, $\mathcal{I}_t$
G	Scaled by step size, penalty or forget rate	η_t, λ_t, Σ
E	Reset to reference when cue triggers	$A_t = \mathbf{1}$
S	Partially reverted every step	A_t
T	Fully reverted on a fixed schedule	$A_t = \mathbf{1}$

Table 2: Vocabulary of Metacognitive Adaptation.

2017; Snoek et al. 2019). In contrast, cues extracted from the internal feature representations of the model retain evidence of domain support, and can function as a competence monitoring channel (Lee et al. 2018). Recent methods such as OWTTT (Li et al. 2023b), which discards samples exceeding a distance threshold to class prototypes extracted from the classification head, have begun exploring feature-space filtering, but they remain tied to coarse, class-level geometry. In contrast, concept-based factorization techniques (Kim et al. 2018; Fel et al. 2023; Bhusal et al. 2025) have emerged as a principled mechanism to extract compact representations of ID semantics learned by the model, projecting high-dimensional model activations onto an explicit concept dictionary $\mathbf{W}_s \in \phi$ learned *a priori*. Deliberate monitoring built on interpretable semantic cues could enhance meta-level decisions, while being unconstrained from the classification head's latent space (Ghorbani et al. 2019; Fel et al. 2023).

Position Motivation. To motivate our position of semantic concepts as deliberate cues, we formalize our feature-based monitoring channel. Let f_θ be a model trained on the source (ID) distribution P_S, let $h_\theta(\mathbf{x}) \in \mathbb{R}^d$ denote the representation of input $\mathbf{x} \sim P_S$ extracted from an intermediate layer of f_θ. Concept extraction yields a static dictionary $\mathbf{W}_s \in \mathbb{R}^{k \times d}$ consisting of k basis vectors that are semantically representative of the ID data (Fel et al. 2023; Bhusal et al. 2025). At test time t, given an unlabeled batch $\mathcal{B}_t = \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,n}\} \sim P_T$, we obtain intermediate activations $\mathbf{Z}_t = [h_{\theta_0}(\mathbf{x}_{t,1}), \ldots, h_{\theta_0}(\mathbf{x}_{t,n})]^\top \in \mathbb{R}^{n \times d}$, we approximate these activations through linear decomposition (Park, Choe, and Veitch 2024) in the form $\mathbf{Z}_t \approx \mathbf{U}_t \mathbf{W}_s$, where $\mathbf{U}_t \in \mathbb{R}^{n \times k}_{\geq 0}$ are nonnegative concept activation coefficients (Fel et al. 2023; Bhusal et al. 2025). The reconstruction residual can be computed as

$$\mathbf{R}_t = \mathbf{Z}_t - \mathbf{U}_t \mathbf{W}_s, \tag{4}$$

capturing the amount of feature information left unexplained by the basis vectors in $\mathbf{W}_s$. In shifted test distribution, object-level representations drift outside of the subspace spanned by $\mathbf{W}_s$, thus spiking $\|\mathbf{R}_t\|$. The resulting cue is entirely independent of the labeling rules $P(y|x)$, and can therefore function as an effective deliberate cue for competence monitoring. To decouple the effect of $\mathbf{Z}_t$'s magnitude on the cue, we adopt the *normalized residual*

$$\bar{\mathbf{R}}_t = \frac{\|\mathbf{R}_t\|}{\|\mathbf{Z}_t\|} \tag{5}$$

as our deliberate concept-based monitoring cue.

Metrics for metacognition. The effectiveness of TTA protocols is largely measured in terms of their capacity to recover or exceed the object-level's pre-deployment accuracy on test data subject to degradations (Liang, He, and Tan 2024). Even though the existing literature has established the importance of monitoring quality for robust TTA, especially in adversarial settings (Niu et al. 2023), direct evaluation of the metacognitive effectiveness of said monitoring channels is largely missing. As such, an ideal meta level would be one that is capable of closely tracking object-level competence, as well as determining the samples that warrant adaptation from the ones that risk degrading the quality of the model. These capabilities are ideally independent of the object level's performance, and must therefore be evaluated separately. To this end, we propose a set of metrics to evaluate the effectiveness of a meta level in a TTA protocol. First, we measure *correctness* monitoring using Signal Detection Theory (Galvin et al. 2003; Fleming and Lau 2014). For a decision threshold $\tau \in \mathbb{R}$ and monitoring score c, oriented so that larger c indicates greater belief in object-level correctness, the Type-II Hit Rate $\text{HR}_2(\tau)$ is given as

$$P(c \geq \tau \,|\, \mathbb{I} = 1) = \frac{\sum_{i=1}^{n} \mathbb{I}\,(c_i \geq \tau \wedge \hat{y}_i = y_i)}{\sum_{i=1}^{n} \mathbb{I}(\hat{y}_i = y_i)}, \tag{6}$$

representing the probability that the meta level is highly confident given that the object-level's prediction is correct. We then define the Type-II False Alarm Rate $\text{FAR}_2(\tau)$ as

$$P(c \geq \tau \,|\, \mathbb{I} = 0) = \frac{\sum_{i=1}^{n} \mathbb{I}\,(c_i \geq \tau \wedge \hat{y}_i \neq y_i)}{\sum_{i=1}^{n} \mathbb{I}(\hat{y}_i \neq y_i)}. \tag{7}$$

By integrating the parametric curve $(\text{FAR}_2(\tau), \text{HR}_2(\tau))$ across all thresholds $\tau \in (-\infty, +\infty)$, we obtain metacognitive sensitivity:

$$\text{AUROC}_2 = \int_0^1 \text{HR}_2\big(\text{FAR}_2^{-1}(u)\big)\, du, \tag{8}$$

the Area Under the Type-2 ROC Curve, where $\text{AUROC}_2 = 1.0$ corresponds to perfect correctness monitoring, and $\text{AUROC}_2 = 0.5$ indicates chance-level correctness monitoring. Furthermore, to evaluate *competence* monitoring, given $\text{HR}_{sem}(\tau) = P(c \geq \tau \,|\, \mathbf{x} \in \mathcal{X}_{\text{ID}})$ and $\text{FAR}_{sem}(\tau) = P(c \geq \tau \,|\, \mathbf{x} \in \mathcal{X}_{\text{OOD}})$ we propose semantic discrimination:

$$\text{AUROC}_{sem} = \int_0^1 \text{HR}_{sem}\big(\text{FAR}_{sem}^{-1}(u)\big)\, du, \tag{9}$$

measuring the separation between ID and OOD data according to the meta level's score c. Similarly, given $\text{HR}_{corr}(\tau) =$

$P(c \geq \tau \mid \mathbf{x} \sim P_S)$ and $\text{FAR}_{corr}(\tau) = P(c \geq \tau \mid \mathbf{x} \sim \tilde{P}_T)$, where $\tilde{P}_T$ is the test data distribution subject to covariate shift corruption, we propose corruption sensitivity:

$$\text{AUROC}_{corr} = \int_0^1 \text{HR}_{corr}\big(\text{FAR}_{corr}^{-1}(u)\big)du, \qquad (10)$$

which measures the meta-level's response to covariate shift.

Empirical Validation. We evaluate the quality of our proposed concept-based monitoring channels, comparing their performance with fluency-based monitoring on WTTA conditions. We employ a frozen ResNet-34 backbone (He et al. 2016) pretrained on ImageNet (Deng et al. 2009). To fit our concept dictionary $\mathbf{W}_s$, we extract $\mathbf{Z} \in \mathbb{R}^{n \times 512}$, the last-layer activations, in line with the FACE protocol adopted by Bhusal et al. (2025). For our ID dataset, we use Imagenette (Howard 2019), a 10-class subsample of ImageNet. We simulate the effect of covariate shift on ID samples by degrading the test set with four different types of corruption: Gaussian noise, blur, fog and contrast (See Appendix Table 4). For the OOD data, we use samples from the NINCO (Bitterwolf, Müller, and Hein 2023) dataset, specifically curated to guarantee no data overlap with the large corpus of ImageNet. First, we evaluate the metacognitive sensitivity, semantic discrimination and corruption sensitivity of automatic and deliberate cues on the experimental setup described above. Semantic

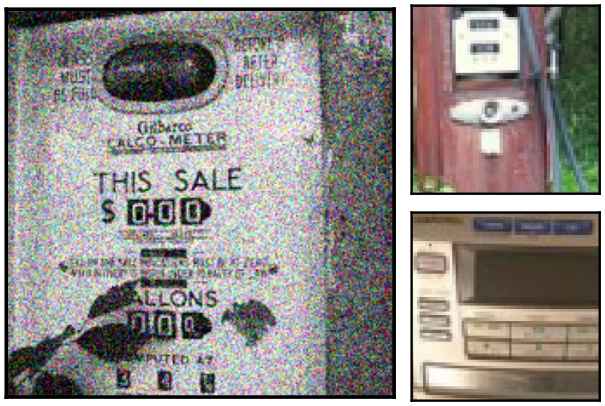


Figure 3: (*Left*) Imagenette gas pump under Gaussian noise, severity 3; (*Right*) the top two most relevant concepts extracted by FACE (Bhusal et al. 2025).

discrimination is measured against OOD samples subject to the same corruptions as the ID samples, so that competence cannot be determined by detecting corruptions. We fit $\mathbf{W}_s$ on 2000 ID samples, and repeat the experiment for 15 seeds. For each seed, we take the worst measurement across each corruption. We measure on two automatic cues, entropy and MSP, and four deliberate cues: normalized residuals from FACE (Bhusal et al. 2025), CRAFT (Fel et al. 2023), PCA low-rank decomposition and OWTTT(Li et al. 2023b) class prototypes. Remaining details are deferred to the Appendix, results shown in Table 3. Automatic cues emerge as competent correctness monitors but show poor competence monitoring performance. Deliberate cues show substantially higher competence monitoring capabilities, and maintain competitive correctness monitoring performance, concept and prototype-based approaches outperform PCA by incorporating more information from the object-level in their meta-level model, as shown in Figure 3.

Monitor	**AUROC**$_{\text{sem}}$ ↑	**AUROC**$_{\text{corr}}$ ↑	**AUROC**$_2$ ↑
Automatic cues			
Entropy	$0.609_{\pm.002}$	$0.831_{\pm.002}$	$0.854_{\pm.001}$
MSP	$0.602_{\pm.002}$	$0.808_{\pm.002}$	$\mathbf{0.867}_{\pm.001}$
Deliberate cues			
OWTTT	$\mathbf{0.935}_{\pm.001}$	$0.928_{\pm.001}$	$0.862_{\pm.002}$
Concept residual ***(ours)***			
FACE	$0.908_{\pm.005}$	$0.947_{\pm.002}$	$0.822_{\pm.002}$
CRAFT	$0.912_{\pm.002}$	$\mathbf{0.950}_{\pm.001}$	$0.831_{\pm.004}$
PCA	$0.857_{\pm.070}$	$0.863_{\pm.024}$	$0.727_{\pm.032}$

Table 3: Metacognitive cue evaluation. mean ± std over 15 seeds; best in **bold**; ↑ means higher is better.

Deliberate cues as test-time gates. Having established deliberate cues as effective competence monitors, we now show their performance on a partial WTTA scenario where the test stream is progressively corrupted by OOD samples with proportion $\pi \in [0, 1)$. To leverage each cue's metacognitive capabilities, we devise a gate mechanism that excludes OOD samples from the update. Let $c_{t,i}$ denote the monitoring score of Equations (8) to (10) evaluated on $x_{t,i}$, monitoring yields knowledge K_t and active set

$$K_t = \{c_{t,i}\}_{i=1}^n, \qquad \mathcal{I}_t = \{x_{t,i} \in \mathcal{B}_t : c_{t,i} \geq \tau_t\}, \qquad (11)$$

adaptation is performed by minimization over surrogate loss $\ell(x;\theta) = H\big(f_{\theta_t}(x)\big)$ as in Tent (Wang et al. 2021)

$$L_t(\theta) = \frac{1}{|\mathcal{I}_t|}\sum_{x \in \mathcal{I}_t} H\big(f_{\theta_t}(x)\big),$$
$$\theta_{t+1} = \theta_t - \eta \nabla_{\Theta_{\text{ad}}} L_t(\theta_t). \qquad (12)$$

This corresponds to Equation (3) with $w_t{=}1$, $\lambda_t{=}0$, $A_t{=}\mathbf{0}$, $g = \text{argmax}$, $\nu = (\mu_{\mathcal{I}_t}, \sigma^2_{\mathcal{I}_t})$, with Σ accumulating running normalization statistics via exponential moving averaging, and $\Theta_{\text{ad}} = \Theta_{\text{affine}}$; K_t regulates the update by populating the active sets $\mathcal{I}_t$, which corresponds to the F mechanism in Table 2. We consider three candidate scores $c_{t,i}$ conditioning our gate

$$c^{\text{ent}}_{t,i} = -H\big(f_{\theta_t}(x_{t,i})\big), \qquad (13)$$
$$c^{\text{res}}_{t,i} = -\bar{\mathbf{R}}_{t,i} \text{ with } \mathbf{Z}_t = h_{\theta_0}(\mathcal{B}_t), \qquad (14)$$
$$c^{\text{pro}}_{t,i} = \max_k \cos\big(h_{\theta_0}(x_{t,i}), \mathbf{P}_k\big), \qquad (15)$$

where $\mathbf{P} \in \phi$ are the class prototypes of Li et al. (2023b).

Threshold Estimation. To obtain the threshold τ, we use the online estimation method proposed by Li et al. (2023b), which selects the threshold minimizing intra-cluster variance over the resulting ID/OOD split realised on the scores observed before the current batch, to avoid contamination. In Figure 4, we show the effect of our mechanism across the three gates shown in Equations (13) to (15), together with ungated adaptation

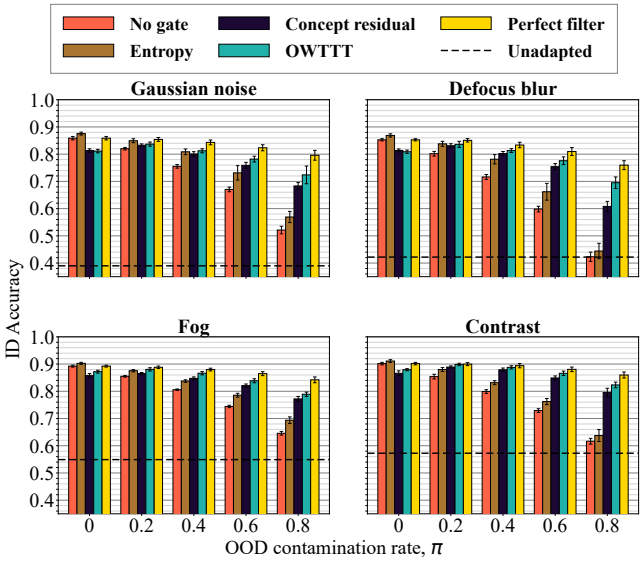


Figure 4: Clean ID accuracy after 200 adaptation steps, across various contamination rates π. Deliberate cues show increased resilience to OOD and closely recover the perfect filter's accuracy. Automatic cues degrade to near-unadapted performance.

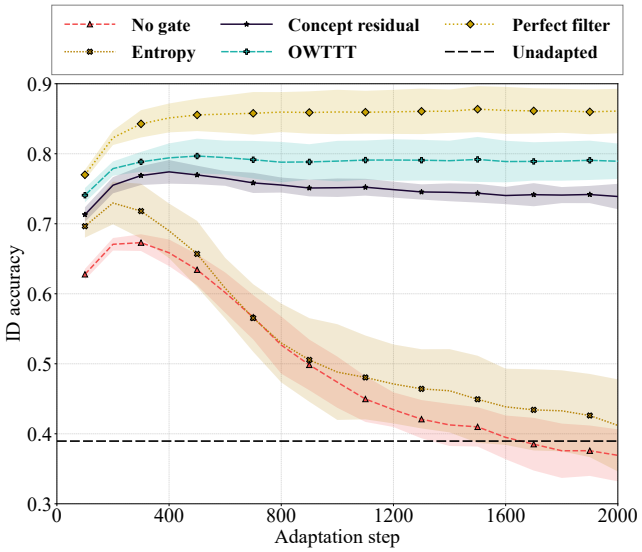


Figure 5: Clean ID accuracy on a longer horizon task (2000 steps, Gaussian noise, severity 3). Deliberate cues prevent collapse by filtering OOD samples out.

and a perfect filter that has access to ground truth membership labels as controls. Deliberate cues show resilience to stream contamination across all four corruptions, almost recovering the perfect filter's ideal adaptation performance. Remarkably, entropy shows no measurable difference with respect to ungated adaptation, underlining the necessity of deliberate cues for robust WTTA adaptation.

Preventing long-horizon collapse. Lastly, we highlight the effect of our protocol on preventing long range adaptation collapse. We fix $\pi = 0.6$ on the Gaussian Noise corruption scenario and run adaptation for 2000 steps. We repeat the experiment across 10 random seeds, reporting accuracy on the uncontaminated ID data, results shown in Figure 5. Automatic gating decays monotonically similarly to ungated adaptation; meanwhile deliberate gating maintains consistent performance across the whole time horizon, and closely matches the performance of the ideal filter. Moreover, while empirical results show that OWTTT (Li et al. 2023b) maintains higher performance over long adaptation horizons compared to our concept residual proposal, our method can explicitly quantify *how much* and *which* specific concepts drive the adaptation process (Figure 3), thus rendering these representational shifts in a human understandable space (Fel et al. 2023; Bhusal et al. 2025). In the end, our concept residual proposal grounds into the metacognitive framework via inspectable and deliberate knowledge (Shea et al. 2014), thereby ensuring that the meta-level control signals remain an auditable cognitive process (Nelson and Narens 1990), allowing practitioners to verify the exact semantic rationale behind the model's self-regulation process.

Neuromorphic Metacognitive Monitoring

Our taxonomy and empirical analysis indicate a trade-off in metacognitive adaptation. Automatic cues are inexpensive and continuously available but confuse prediction correctness with model competence, while deliberate cues provide a more reliable estimate of whether the current input is in the model's operational domain but require a structured inspection of internal representations. As shown in Figures 4 and 5, this additional information can prevent OOD samples from contaminating the adaptation process and protect the model from long-horizon collapse. However, performing this inspection at every step introduces latency, memory, and energy costs. This is directly connected to the two-level formulation introduced in Equation (3) where the meta level must determine not only which knowledge should guide adaptation, but also when acquiring this knowledge is worth the computational cost. Metacognitive monitoring is therefore a resource allocation problem: the system must decide when automatic evidence is sufficient and when deliberate inspection should be recruited.

Allocating the monitoring budget. We consider a cost-stratified architecture in which inexpensive fluency cues provide a continuously available first-pass estimate of reliability, while structured internal representations are inspected only when the available evidence indicates a sufficient level of ambiguity or novelty. We decompose the total cost of metacognitive monitoring C_{meta} into the cost of evaluating whether to inspect internal states C_{alloc}, and, conditionally, the costs of signal generation C_{sig}, semantic interpretation C_{int}, and control execution C_{ctrl}

$$C_{\text{meta}} = C_{\text{alloc}} + \mathbb{I}_{\text{meta}}(C_{\text{sig}} + C_{\text{int}} + C_{\text{ctrl}}), \qquad (16)$$

where $\mathbb{I}_{\text{meta}} \in \{0, 1\}$ indicates whether this channel is recruited. Following rational metareasoning principles (Lieder et al. 2018), this allocation decision can be expressed through the

Value of Computation (VOC) (Callaway et al. 2018) paradigm

$$\mathrm{VOC}(c,b) = \mathbb{E}_{B' \sim T_{\mathrm{meta}}(b,c,\cdot)}[V(B')] - V(b) - C(c), \quad (17)$$

where b is the current meta-level belief state, c is a candidate monitoring computation with execution cost $C(c) = C_{\mathrm{sig}} + C_{\mathrm{int}} + C_{\mathrm{ctrl}}$, and $V(b) = \max_{a \in \mathcal{A}} \mathbb{E}[\mathcal{R}(a \mid b)]$ denotes the expected utility of the optimal supervisory action $a \in \mathcal{A}$. The action space $\mathcal{A}$ includes decisions such as continuing adaptation, restricting or freezing updates, abstaining, deferring the decision or allocating additional computation. In TTA, $V(b)$ balances immediate accuracy on covariate-shifted ID samples against the longer-term risk of model drift caused by adapting on open-set contamination Q_t. Deliberate cues are recruited ($\mathbb{I}_{\text{meta}} = 1$) when its expected supervisory benefit exceeds its execution cost. Biological cognition follows a similar division of labor, estimating confidence from the evidence used for the initial decision and engaging separate circuits only when the task is ambiguous (Kiani and Shadlen 2009; Fleming and Dolan 2012; Fleming and Daw 2017).

Neuromorphic computing as an event-driven substrate. The allocation principle defined by Equations (16) and (17) finds a natural implementation in neuromorphic systems, where event-driven computation couples representational activity and temporal surprise to resource allocation while reducing monitoring costs through three core mechanisms (Mead 1990; Indiveri and Liu 2015; Frenkel, Bol, and Indiveri 2023). (i) Thresholded neuronal dynamics restrict communication to sparse events. The allocation gate can remain continuously active without executing the full monitoring pipeline at every time step, allowing C_{alloc} to scale with changes in input activity or model state (Merolla et al. 2014; Davies et al. 2018). This provides a physical implementation of continuously available automatic monitoring with selectively recruited deliberate inspection. (ii) Predictive coding and residual architectures naturally produce mismatch signals during inference through error units or residual spike emissions (Rao and Ballard 1999; Wacongne, Changeux, and Dehaene 2012). These signals can support C_{sig} without a separate monitoring network to recompute diagnostic evidence. Although they are generated by the adapting object level, their interpretation can remain independent: the meta level can compare them against a fixed concept-level reference stored in the pre-deployment state ϕ. Maintaining this separation is necessary because a monitor that adapts together with the object level may gradually lose the stable reference required to detect drift. (iii) Sparse population activity provides a natural substrate for representing and integrating monitoring evidence. Rather than compressing uncertainty into a single confidence score, population spike counts can encode evidence for multiple representational conditions, while parallel temporal accumulators track covariate shift, semantic novelty, and predictive instability. Because these conditions are non-exclusive, their evidence can accumulate simultaneously until it affects the supervisory decision. Once sufficient evidence has been collected, threshold dynamics can trigger actions as described in Table 2 (Wang 2002; Furman and Wang 2008).

The resulting benefit is not limited to energy efficiency. Event-driven computation provides a mechanism for translating changes in internal representations into temporally accumulated evidence and, eventually, into control over the adaptation process. However, this advantage is conditional, since under sustained novelty or persistent distribution change, the deliberate channel may remain continuously active, eliminating the savings obtained through sparse recruitment. Regime-change detection, bounded evidence retention, and explicit monitoring budgets are necessary to prevent the meta level itself from becoming an unbounded computational process.

Architectural Synthesis and Open Challenges. Existing neuromorphic systems already instantiate individual components of this architecture, including latency-adaptive classifiers with reliability guarantees (Chen, Park, and Simeone 2025), dynamic early-exit networks (Li et al. 2023a; Li, Jones, and Furber 2023), on-chip epistemic uncertainty estimation (Skatchkovsky, Jang, and Simeone 2022; Rajendran, Simeone, and Al-Hashimi 2026), spike-based OOD detection (Martinez-Seras et al. 2023), and always-on wake-up mechanisms (Chundi et al. 2021; Ballet et al. 2026). These systems show that monitoring, evidence accumulation, and conditional computation can be implemented efficiently. However, these capabilities are generally introduced as isolated reliability or efficiency mechanisms rather than as parts of a coordinated supervisory controller over an adapting object level. The remaining challenge is therefore architectural rather than merely computational. A complete metacognitive TTA system must connect the monitoring signal K_t to the control variables identified by our taxonomy. In particular, it should: (i) preserve an adaptation-independent reference $\mathbf{W}_s \in \phi$ against which object-level changes can be evaluated; (ii) use granular monitoring evidence to condition which samples and parameter subspaces are adapted, through $\mathcal{I}_t$ and $\Theta_{\mathrm{ad}}(K_t)$; and (iii) control the full action space of the model, including parameter restoration through (A_t) and abstention through $g(\cdot) \in \mathcal{Y} \cup \{\perp\}$. Neuromorphic principles offer a promising substrate for this closed loop because they make monitoring demand-dependent and allow evidence to accumulate over time without requiring dense computation at every step. Their role is not to replace the metacognitive framework, but to provide one possible implementation of bounded self-regulation.

Conclusion

In this work, we formalized Test-Time Adaptation through a two-level metacognitive framework that decouples task execution from supervisory monitoring and control. By analyzing existing protocols across their monitoring knowledge, adaptation loci, and trigger policies, we demonstrated that scalar fluency cues conflate task correctness with domain competence. We showed that inspection via concept-based interpretability provides an invariant competence gate to protect against contaminated TTA streams. Finally, by framing meta-level monitoring as a resource-allocation problem, we highlighted how event-driven neuromorphic principles enable bounded, self-regulated perception. Closing the loop between introspection and regulation provides a clear roadmap toward robust, self-aware perception systems.

Acknowledgements

This work was supported in part by NSF under Grants CNS-2312875 and OAC-2530896, DARPA under Cooperative Agreement D25AC00374-00, ONR under Award N00014-23-1-2221, and AFOSR under Award FA9550-23-1-0261.

References

Ackerman, R.; and Thompson, V. A. 2017. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, 21(8): 607–617.

Alain, G.; and Bengio, Y. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Ballet, T.; Romdhane, A.; Lovison-Franco, B.; Dupuis, T.; Renaudineau, A.; Alencar, F. P.; Iftakher, M. A.; Turck, C.; Harabi, K.-E.; Vianello, E.; Portal, J.-M.; Benoit, P.; Novo, D.; and Querlioz, D. 2026. Uncertainty-triggered wake-up enables energy-efficient, error-resilient edge AI with memristor front ends. arXiv:2605.29533.

Bhusal, D.; Clifford, M.; Rampazzi, S.; and Rastogi, N. 2025. FACE: Faithful Automatic Concept Extraction. In Belgrave, D.; Zhang, C.; Montoya, L. N.; Lin, H.; Pascanu, R.; Koniusz, P.; Ghassemi, M.; Chen, N.; Ruíz, I. V. M.; and Loaiza-Bonilla, A., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Bitterwolf, J.; Müller, M.; and Hein, M. 2023. In or Out? Fixing ImageNet Out-of-Distribution Detection Evaluation. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 2471–2506. PMLR.

Boudiaf, M.; Mueller, R.; Ayed, I. B.; and Bertinetto, L. 2022. Parameter-free Online Test-time Adaptation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8334–8343.

Callaway, F.; Gul, S.; Krueger, P. M.; Griffiths, T. L.; and Lieder, F. 2018. Learning to Select Computations. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, 776–785.

Chen, J.; Park, S.; and Simeone, O. 2025. Knowing When to Stop: Delay-Adaptive Spiking Neural Network Classifiers with Reliability Guarantees. *IEEE Journal on Selected Topics in Signal Processing*, 19(1): 88–102.

Chundi, P. K.; Wang, D.; Kim, S. J.; Yang, M.; Cerqueira, J. P.; Kang, J.; Jung, S.; Kim, S.; and Seok, M. 2021. Always-On Sub-Microwatt Spiking Neural Network Based on Spike-Driven Clock- and Power-Gating for an Ultra-Low-Power Intelligent Device. *Frontiers in Neuroscience*, 15: 684113.

Chung, Y.; Park, H.; Rim, P.; Zhang, X.; He, J.; Zeng, Z.; Cicek, S.; Hong, B.; Duncan, J. S.; and Wong, A. 2025. ETA: Energy-Based Test-Time Adaptation for Depth Completion. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, 1–12. IEEE.

Cox, M. T. 2005. Metacognition in Computation: A Selected Research Review. *Artificial Intelligence*, 169(2): 104–141.

Davies, M.; Srinivasa, N.; Lin, T.-H.; Chinya, G.; Cao, Y.; Choday, S. H.; Dimou, G.; Joshi, P.; Imam, N.; Jain, S.; Liao, Y.; Lin, C.-K.; Lines, A.; Liu, R.; Mathaikutty, D.; McCoy, S.; Paul, A.; Tse, J.; Venkataramanan, G.; Weng, Y.-H.; Wild, A.; Yang, Y.; and Wang, H. 2018. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro*, 38(1): 82–99.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Döbler, M.; Marsden, R. A.; and Yang, B. 2023. Robust Mean Teacher for Continual and Gradual Test-Time Adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7704–7714.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fel, T.; Picard, A. M.; Béthune, L.; Boissin, T.; Vigouroux, D.; Colin, J.; Cadène, R.; and Serre, T. 2023. CRAFT: Concept Recursive Activation FacTorization for Explainability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2711–2721. IEEE.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10): 906–911.

Fleming, S. M.; and Daw, N. D. 2017. Self-Evaluation of Decision-Making: A General Bayesian Framework for Metacognitive Computation. *Psychological Review*, 124(1): 91–114.

Fleming, S. M.; and Dolan, R. J. 2012. The Neural Basis of Metacognitive Ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594): 1338–1349.

Fleming, S. M.; and Lau, H. C. 2014. How to measure metacognition. *Frontiers in Human Neuroscience*, 8: 443.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

Frenkel, C.; Bol, D.; and Indiveri, G. 2023. Bottom-Up and Top-Down Approaches for the Design of Neuromorphic Processing Systems: Tradeoffs and Synergies Between Natural and Artificial Intelligence. *Proceedings of the IEEE*, 111(6): 623–652.

Furman, M.; and Wang, X.-J. 2008. Similarity Effect and Optimal Control of Multiple-Choice Decision Making. *Neuron*, 60(6): 1153–1168.

Galvin, S. J.; Podd, J. V.; Drga, V.; and Whitmore, J. 2003. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, 10(4): 843–876.

Gao, Z.; Zhang, X.-Y.; and Liu, C.-L. 2024. Unified Entropy Optimization for Open-Set Test-Time Adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23975–23984.

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards Automatic Concept-based Explanations. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Gong, T.; Jeong, J.; Kim, T.; Kim, Y.; Shin, J.; and Lee, S.-J. 2022. NOTE: Robust Continual Test-Time Adaptation Against Temporal Correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gong, T.; Kim, Y.; Lee, T.; Chottananurak, S.; and Lee, S.-J. 2023. SoTTA: Robust Test-Time Adaptation on Noisy Data Streams. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised Learning by Entropy Minimization. In *Advances in Neural Information Processing Systems 17 Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada*, 529–536.

Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1321–1330. JMLR.org.

Han, J.; Na, J.; and Hwang, W. 2025. Ranked Entropy Minimization for Continual Test-Time Adaptation. In *Forty-second International Conference on Machine Learning*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020. AugMix: A Simple Method to Improve Robustness and Uncertainty under Data Shift. In *International Conference on Learning Representations*.

Hoang, T.-H.; Vo, D. M.; and Do, M. N. 2024. Persistent Test-time Adaptation in Recurring Testing Scenarios. In *Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS)*.

Hong, J.; Lyu, L.; Zhou, J.; and Spranger, M. 2023. MECTA: Memory-Economic Continual Test-Time Model Adaptation. In *The Eleventh International Conference on Learning Representations*.

Howard, J. 2019. Imagenette: A smaller subset of 10 easily classified classes from Imagenet.

Indiveri, G.; and Liu, S.-C. 2015. Memory and Information Processing in Neuromorphic Systems. *Proceedings of the IEEE*, 103(8): 1379–1397.

Karmanov, A.; Guan, D.; Lu, S.; El Saddik, A.; and Xing, E. 2024. Efficient Test-Time Adaptation of Vision-Language Models. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kiani, R.; and Shadlen, M. N. 2009. Representation of Confidence Associated with a Decision and Uncertainty by Neurons in the Parietal Cortex. *Science*, 324(5928): 759–764.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C. J.; Wexler, J.; Viégas, F. B.; and Sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, 2673–2682. PMLR.

Koriat, A. 1997. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4): 349–370.

Koriat, A.; and Levy-Sadot, R. 1999. Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In *Dual-process theories in social psychology*, 483–502. New York, NY, US: The Guilford Press. ISBN 978-1-57230-421-5.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F.; Burges, C.; Bottou, L.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Lee, J.; Jung, D.; Lee, S.; Park, J.; Shin, J.; Hwang, U.; and Yoon, S. 2024. Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors. In *The Twelfth International Conference on Learning Representations*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Li, C.; Jones, E. G.; and Furber, S. 2023. Unleashing the Potential of Spiking Neural Networks with Dynamic Confidence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13350–13360.

Li, Y.; Geller, T.; Kim, Y.; and Panda, P. 2023a. SEENN: Towards Temporal Spiking Early-Exit Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Li, Y.; Xu, X.; Su, Y.; and Jia, K. 2023b. On the Robustness of Open-World Test-Time Training: Self-Training with Dynamic Prototype Expansion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11802–11812.

Liang, J.; He, R.; and Tan, T. 2024. A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts. *International Journal of Computer Vision*, 133(1): 31–64.

Lieder, F.; Shenhav, A.; Musslick, S.; and Griffiths, T. L. 2018. Rational Metareasoning and the Plasticity of Cognitive Control. *PLOS Computational Biology*, 14(4): e1006043.

Lim, H.; Kim, B.; Choo, J.; and Choi, S. 2023. TTN: A Domain-Shift Aware Batch Normalization in Test-Time Adaptation. In *The Eleventh International Conference on Learning Representations*.

Liu, Y.; Kothari, P.; van Delft, B. G.; Bellot-Gurlet, B.; Mordan, T.; and Alahi, A. 2021. TTT++: When Does Self-Supervised Test-Time Training Fail or Thrive? In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Maharana, S. K.; Zhang, B.; and Guo, Y. 2025. PALM: Pushing Adaptive Learning Rate Mechanisms for Continual Test-Time Adaptation. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, 19378–19386. AAAI Press.

Marsden, R. A.; Döbler, M.; and Yang, B. 2024. Universal Test-time Adaptation through Weight Ensembling, Diversity Weighting, and Prior Correction. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2543–2553.

Martinez-Seras, A.; Del Ser, J.; Lobo, J. L.; Garcia-Bringas, P.; and Kasabov, N. 2023. A Novel Out-of-Distribution Detection Approach for Spiking Neural Networks: Design, Fusion, Performance Evaluation and Explainability. *Information Fusion*, 100: 101943.

Mead, C. 1990. Neuromorphic Electronic Systems. *Proceedings of the IEEE*, 78(10): 1629–1636.

Merolla, P. A.; Arthur, J. V.; Alvarez-Icaza, R.; Cassidy, A. S.; Sawada, J.; Akopyan, F.; Jackson, B. L.; Imam, N.; Guo, C.; Nakamura, Y.; Brezzo, B.; Vo, I.; Esser, S. K.; Appuswamy, R.; Taba, B.; Amir, A.; Flickner, M. D.; Risk, W. P.; Manohar, R.; and Modha, D. S. 2014. A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface. *Science*, 345(6197): 668–673.

Nelson, T. O.; and Narens, L. 1990. Metamemory: A Theoretical Framework and New Findings. In *Psychology of Learning and Motivation*, volume 26, 125–173. Elsevier. ISBN 978-0-12-543326-6.

Nguyen, A. T.; Nguyen-Tang, T.; Lim, S.-N.; and Torr, P. H. 2023. Tipi: Test time adaptation with transformation invariance. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24162–24171. IEEE.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient Test-Time Model Adaptation without Forgetting. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 16888–16905. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-time Adaptation in Dynamic Wild World. In *The Eleventh International Conference on Learning Representations*.

Park, J.; Kim, J.; Kwon, H.; Yoon, I.; and Sohn, K. 2024. Layer-wise Auto-Weighting for Non-Stationary Test-Time Adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1414–1423.

Park, K.; Choe, Y. J.; and Veitch, V. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Press, O.; Schneider, S.; Kümmerer, M.; and Bethge, M. 2023. RDumb: A simple approach that questions our progress in continual test-time adaptation. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Rajendran, B.; Simeone, O.; and Al-Hashimi, B. 2026. Towards efficient and reliable artificial intelligence through neuromorphic principles. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 384(2315): 20240522.

Rao, R. P. N.; and Ballard, D. H. 1999. Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nature Neuroscience*, 2(1): 79–87.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do ImageNet Classifiers Generalize to ImageNet? In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5389–5400. PMLR.

Sahoo, S.; ElAraby, M.; Ngnawé, J.; Pequignot, Y. B.; Precioso, F.; and Gagné, C. 2025. A Layer Selection Approach to Test Time Adaptation. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, 20237–20245. AAAI Press.

Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; and Bethge, M. 2020. Improving robustness against common corruptions by covariate shift adaptation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Shea, N.; Boldt, A.; Bang, D.; Yeung, N.; Heyes, C.; and Frith, C. D. 2014. Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, 18(4): 186–193.

Skatchkovsky, N.; Jang, H.; and Simeone, O. 2022. Bayesian Continual Learning via Spiking Neural Networks. *Frontiers in Computational Neuroscience*, 16: 1037976.

Snoek, J.; Ovadia, Y.; Fertig, E.; Lakshminarayanan, B.; Nowozin, S.; Sculley, D.; Dillon, J. V.; Ren, J.; and Nado, Z. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 13969–13980.

Song, J.; Lee, J.; Kweon, I. S.; and Choi, S. 2023. EcoTTA: Memory-Efficient Continual Test-time Adaptation via Self-distilled Regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A. A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Thompson, V. A. 2009. Dual-process theories: A metacognitive perspective. In Evans, J. S. B. T.; and Frankish, K., eds., *In Two Minds: Dual Processes and Beyond*, 171–195. Oxford: Oxford University Press.

Thompson, V. A.; Prowse Turner, J. A.; and Pennycook, G. 2011. Intuition, reason, and metacognition. *Cognitive Psychology*, 63(3): 107–140.

Veenman, M. V. J.; Van Hout-Wolters, B. H. A. M.; and Afflerbach, P. 2006. Metacognition and Learning: Conceptual and Methodological Considerations. *Metacognition and Learning*, 1(1): 3–14.

Wacongne, C.; Changeux, J.-P.; and Dehaene, S. 2012. A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. *Journal of Neuroscience*, 32(11): 3665–3678.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual Test-Time Domain Adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7191–7201.

Wang, X.-J. 2002. Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron*, 36(5): 955–968.

Wang, Z.; Luo, Y.; Zheng, L.; Chen, Z.; Wang, S.; and Huang, Z. 2024. In Search of Lost Online Test-Time Adaptation: A Survey. *International Journal of Computer Vision (IJCV)*, 133(3): 1106–1139.

Wells, A.; and Matthews, G. 1996. Modelling Cognition in Emotional Disorder: The S-REF Model. *Behaviour Research and Therapy*, 34(11–12): 881–888.

Yuan, L.; Xie, B.; and Li, S. 2023. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15922–15932.

Zhang, M.; Levine, S.; and Finn, C. 2022. MEMO: test time robustness via adaptation and augmentation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Zhang, Q.; Bian, Y.; Kong, X.; Zhao, P.; and Zhang, C. 2025. COME: Test-time Adaption by Conservatively Minimizing Entropy. In *The Thirteenth International Conference on Learning Representations*.

Zhao, H.; Liu, Y.; Alahi, A.; and Lin, T. 2023. On Pitfalls of Test-Time Adaptation. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 42058–42080. PMLR.

Corruption	1	2	3	4	5
Noise (σ)	.08	.12	**.18**	.26	.38
Blur (r, σ)	3, .1	4, .5	**6**, **.5**	8, .5	10, .5
Fog (i, d)	1.5, 2	2, 2	**2.5**, **1.7**	2.5, 1.5	3, 1.4
Contrast (c)	.4	.3	**.2**	.1	.05

Table 4: Corruption parameters by severity level, parameters used for our experiments in **bold**.

Implementation Details

All image corruptions were obtained by leveraging the `imagecorruptions==1.1.2` python package, details for reproducing the corruptions are available in Table 4.

Additional Results

For better readability, we report our experimental results in tabular forms. Figure 4 is shown in Table 6, Figure 5 is shown in Table 5. Additionally, we report the fraction of admitted OOD samples by cue for the four covariate shift corruptions, at various levels of stream contamination π in Table 7, showing the superior filtering capabilities of concept-based gating compared to prediction entropy.

Together with Figure 3, we report some additional concept crops extracted from the datasets. First, Figure 6 shows concepts extracted on the OOD NINCO dataset (Bitterwolf, Müller, and Hein 2023). Other most related concept crops are shown across multiple combinations of ID/OOD and corruptions in Figures 7 to 10. We notice that OOD images tend to retrieve semantically unrelated crops that share similar structural qualities, such as texture or material.



Figure 6: (*Left*) a novel NINCO image under Gaussian noise, severity 3; (*right*) the top two most relevant concept crops extracted by FACE (Bhusal et al. 2025).

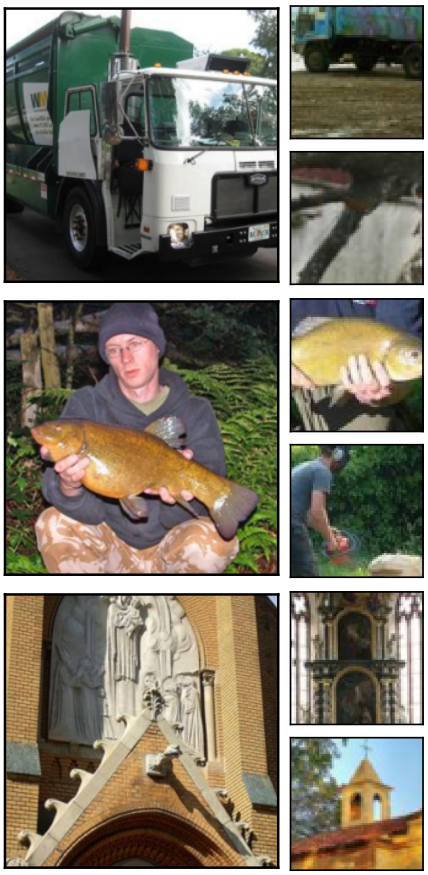


Figure 7: ID, clean.

Step	No gate	Entropy	Residual (ours)	OWTTT	Perfect filter
200	$0.671_{\pm.009}$	$0.730_{\pm.030}$	$0.755_{\pm.012}$	$\mathbf{0.779}_{\pm.010}$	$0.823_{\pm.010}$
400	$0.659_{\pm.019}$	$0.690_{\pm.039}$	$0.774_{\pm.017}$	$\mathbf{0.794}_{\pm.021}$	$0.851_{\pm.021}$
600	$0.602_{\pm.029}$	$0.608_{\pm.043}$	$0.765_{\pm.011}$	$\mathbf{0.795}_{\pm.024}$	$0.857_{\pm.027}$
800	$0.527_{\pm.041}$	$0.530_{\pm.056}$	$0.755_{\pm.011}$	$\mathbf{0.788}_{\pm.029}$	$0.860_{\pm.029}$
1000	$0.474_{\pm.037}$	$0.488_{\pm.068}$	$0.751_{\pm.013}$	$\mathbf{0.789}_{\pm.029}$	$0.860_{\pm.032}$
1200	$0.434_{\pm.025}$	$0.471_{\pm.056}$	$0.749_{\pm.011}$	$\mathbf{0.791}_{\pm.030}$	$0.860_{\pm.032}$
1400	$0.413_{\pm.030}$	$0.461_{\pm.059}$	$0.745_{\pm.013}$	$\mathbf{0.790}_{\pm.029}$	$0.861_{\pm.031}$
1600	$0.395_{\pm.031}$	$0.439_{\pm.054}$	$0.740_{\pm.012}$	$\mathbf{0.789}_{\pm.030}$	$0.862_{\pm.034}$
1800	$0.376_{\pm.039}$	$0.433_{\pm.058}$	$0.741_{\pm.011}$	$\mathbf{0.790}_{\pm.028}$	$0.861_{\pm.031}$
2000	$0.369_{\pm.037}$	$0.412_{\pm.066}$	$0.739_{\pm.018}$	$\mathbf{0.789}_{\pm.025}$	$0.861_{\pm.032}$

Table 5: Clean ID accuracy against adaptation step on Gaussian noise at $\pi = 0.6$ (10 seeds, mean$_{\pm SD}$).

Protocol	$\pi=0$	$\pi=0.2$	$\pi=0.4$	$\pi=0.6$	$\pi=0.8$
Gaussian noise					
No gate	$0.859_{\pm.006}$	$0.820_{\pm.005}$	$0.755_{\pm.007}$	$0.670_{\pm.008}$	$0.521_{\pm.014}$
Entropy	$\mathbf{0.876}_{\pm.005}$	$\mathbf{0.849}_{\pm.007}$	$0.809_{\pm.010}$	$0.731_{\pm.026}$	$0.569_{\pm.021}$
Residual (ours)	$0.813_{\pm.006}$	$0.833_{\pm.005}$	$0.801_{\pm.009}$	$0.759_{\pm.011}$	$0.683_{\pm.012}$
OWTTT	$0.811_{\pm.006}$	$0.837_{\pm.008}$	$\mathbf{0.812}_{\pm.007}$	$\mathbf{0.782}_{\pm.011}$	$\mathbf{0.724}_{\pm.032}$
Perfect filter	$0.859_{\pm.006}$	$0.854_{\pm.007}$	$0.843_{\pm.009}$	$0.824_{\pm.011}$	$0.796_{\pm.018}$
Unadapted			0.390		
Defocus blur					
No gate	$0.853_{\pm.004}$	$0.802_{\pm.009}$	$0.716_{\pm.009}$	$0.598_{\pm.010}$	$0.423_{\pm.017}$
Entropy	$\mathbf{0.869}_{\pm.006}$	$\mathbf{0.838}_{\pm.009}$	$0.781_{\pm.017}$	$0.662_{\pm.031}$	$0.444_{\pm.028}$
Residual (ours)	$0.814_{\pm.005}$	$0.831_{\pm.007}$	$0.803_{\pm.007}$	$0.754_{\pm.011}$	$0.608_{\pm.018}$
OWTTT	$0.809_{\pm.006}$	$0.836_{\pm.011}$	$\mathbf{0.813}_{\pm.007}$	$\mathbf{0.776}_{\pm.014}$	$\mathbf{0.695}_{\pm.021}$
Perfect filter	$0.853_{\pm.004}$	$0.851_{\pm.006}$	$0.834_{\pm.010}$	$0.810_{\pm.014}$	$0.760_{\pm.016}$
Unadapted			0.422		
Fog					
No gate	$0.892_{\pm.003}$	$0.855_{\pm.003}$	$0.806_{\pm.003}$	$0.744_{\pm.004}$	$0.646_{\pm.007}$
Entropy	$\mathbf{0.903}_{\pm.004}$	$0.876_{\pm.004}$	$0.838_{\pm.005}$	$0.786_{\pm.007}$	$0.694_{\pm.012}$
Residual (ours)	$0.858_{\pm.008}$	$0.865_{\pm.004}$	$0.848_{\pm.005}$	$0.821_{\pm.006}$	$0.772_{\pm.009}$
OWTTT	$0.872_{\pm.005}$	$\mathbf{0.881}_{\pm.006}$	$\mathbf{0.866}_{\pm.006}$	$\mathbf{0.839}_{\pm.007}$	$\mathbf{0.789}_{\pm.008}$
Perfect filter	$0.892_{\pm.003}$	$0.888_{\pm.005}$	$0.880_{\pm.005}$	$0.865_{\pm.007}$	$0.842_{\pm.011}$
Unadapted			0.549		
Contrast					
No gate	$0.902_{\pm.005}$	$0.854_{\pm.009}$	$0.799_{\pm.007}$	$0.729_{\pm.007}$	$0.616_{\pm.011}$
Entropy	$\mathbf{0.911}_{\pm.005}$	$0.880_{\pm.007}$	$0.832_{\pm.007}$	$0.762_{\pm.011}$	$0.638_{\pm.022}$
Residual (ours)	$0.866_{\pm.009}$	$0.889_{\pm.004}$	$0.879_{\pm.006}$	$0.849_{\pm.008}$	$0.795_{\pm.016}$
OWTTT	$0.879_{\pm.004}$	$\mathbf{0.899}_{\pm.004}$	$\mathbf{0.888}_{\pm.006}$	$\mathbf{0.866}_{\pm.008}$	$\mathbf{0.823}_{\pm.010}$
Perfect filter	$0.902_{\pm.005}$	$0.900_{\pm.005}$	$0.894_{\pm.008}$	$0.881_{\pm.008}$	$0.860_{\pm.011}$
Unadapted			0.572		

Table 6: Clean ID accuracy after 200 adaptation steps, per corruption and contamination rate (15 seeds, mean$_{\pm SD}$).

Corruption	$\pi = 0.2$			$\pi = 0.4$			$\pi = 0.6$			$\pi = 0.8$		
	Entropy	Residual	Ratio	Entropy	Residual	Ratio	Entropy	Residual	Ratio	Entropy	Residual	Ratio
Gaussian Noise	$0.381_{\pm 0.021}$	$0.157_{\pm 0.012}$	$2.4_{\pm 0.2}\times$	$0.377_{\pm 0.018}$	$0.165_{\pm 0.008}$	$2.3_{\pm 0.2}\times$	$0.425_{\pm 0.034}$	$0.152_{\pm 0.011}$	$2.8_{\pm 0.3}\times$	$0.513_{\pm 0.021}$	$0.228_{\pm 0.016}$	$2.3_{\pm 0.2}\times$
Defocus Blur	$0.368_{\pm 0.021}$	$0.166_{\pm 0.025}$	$2.3_{\pm 0.3}\times$	$0.388_{\pm 0.022}$	$0.206_{\pm 0.019}$	$1.9_{\pm 0.2}\times$	$0.450_{\pm 0.028}$	$0.198_{\pm 0.020}$	$2.3_{\pm 0.3}\times$	$0.550_{\pm 0.021}$	$0.347_{\pm 0.029}$	$1.6_{\pm 0.1}\times$
Fog	$0.473_{\pm 0.020}$	$0.058_{\pm 0.006}$	$8.2_{\pm 1.0}\times$	$0.456_{\pm 0.017}$	$0.061_{\pm 0.005}$	$7.5_{\pm 0.7}\times$	$0.479_{\pm 0.016}$	$0.047_{\pm 0.003}$	$10.2_{\pm 0.8}\times$	$0.537_{\pm 0.018}$	$0.036_{\pm 0.006}$	$15.2_{\pm 2.6}\times$
Contrast	$0.512_{\pm 0.024}$	$0.069_{\pm 0.009}$	$7.5_{\pm 1.1}\times$	$0.495_{\pm 0.019}$	$0.064_{\pm 0.007}$	$7.8_{\pm 1.0}\times$	$0.516_{\pm 0.016}$	$0.043_{\pm 0.004}$	$12.0_{\pm 1.1}\times$	$0.570_{\pm 0.024}$	$0.027_{\pm 0.005}$	$22.1_{\pm 4.5}\times$

Table 7: Fraction of OOD samples admitted for adaptation for the entropy and concept-residual gates. Mean $\pm$ SD over 15 seeds.

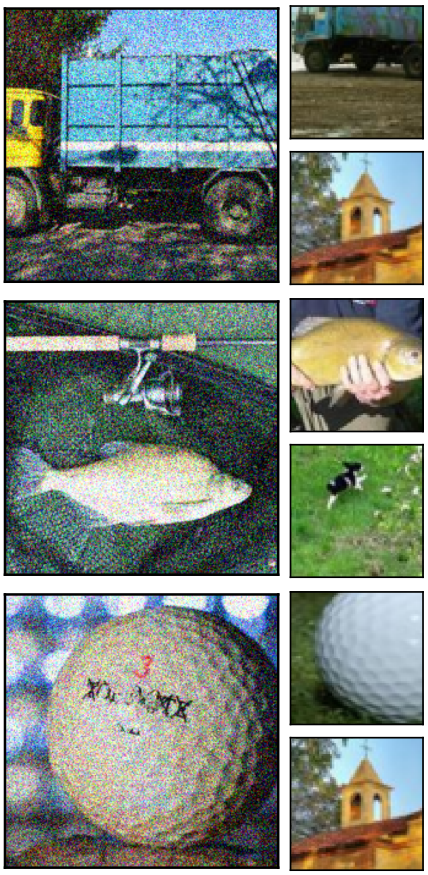


Figure 8: ID, Gaussian noise, severity 3.

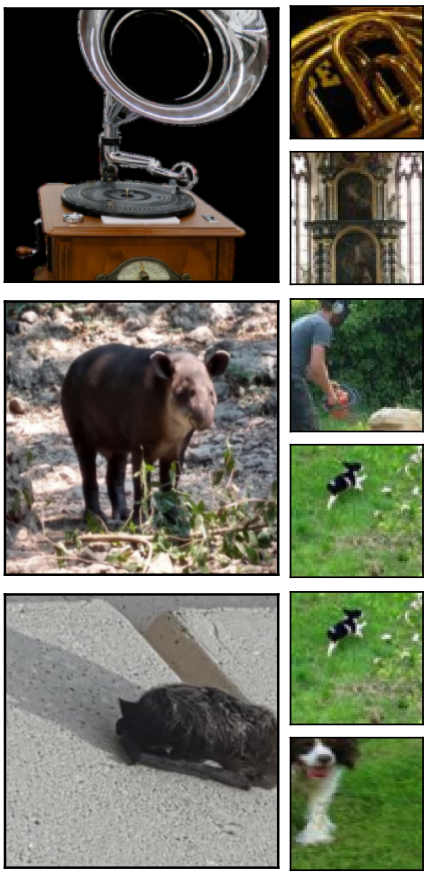


Figure 9: OOD, clean.

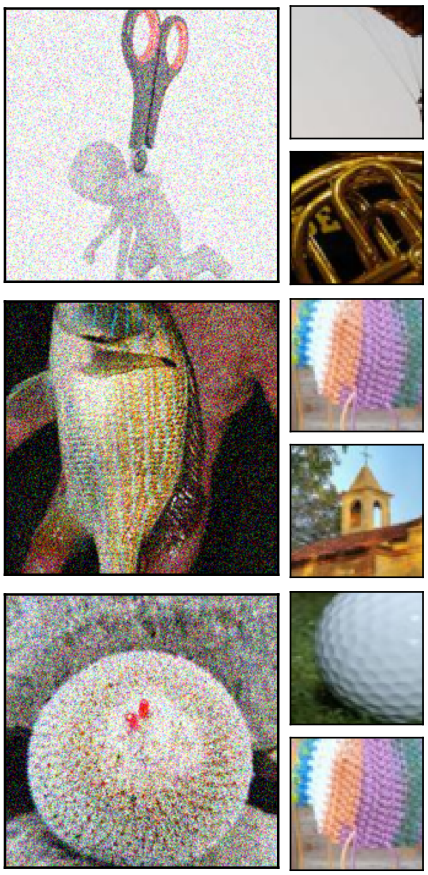


Figure 10: OOD, Gaussian noise, severity 3.