RELIABLE POSTERIOR SUPPORT FOR LONG-TAILED ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training under long-tailed distributions suffers from severe class-wise robustness degradation, especially on tail and worst-performing classes. Existing methods improve long-tailed robustness through rebalancing, distillation, geometric regularization, or posterior-aware supervision, but they largely overlook whether the posterior signal is supported by a reliable local neighborhood. We argue that robust supervision should depend not only on posterior confidence at a training sample, but also on the reliability of the posterior support explored by adversarial perturbations. This issue is most severe for tail classes, whose support is estimated from scarce data and is therefore noisy, high-variance, and easily distorted. We propose a reliable posterior support framework, **FEPT** (**F**ree **E**nergy-guided **P**osterior **T**ransport), which constructs free-energy posterior support to down-weight posterior mass in off-support, low-density, or unstable regions, and transports reliable low-free-energy support from related head/medium classes to tail classes. Our analysis formalizes why local margin stability and variance-reducing posterior transport are useful for reliable robust supervision, while extensive experiments show consistent improvements in tail and worst-class performance, balanced robustness, and posterior-support stability.

1 INTRODUCTION

Adversarial training is one of the most effective approaches for improving the robustness of deep neural networks against norm-bounded perturbations (Madry et al., 2018; Zhang et al., 2019; Wang et al., 2020). Its performance under long-tailed data, however, remains far less satisfactory. When a few head classes dominate the training set and many tail classes are represented by only a small number of samples, robust accuracy often becomes highly uneven across classes. A model may retain reasonable average robustness while its tail or worst-class robustness collapses (Wu et al., 2021; Yue et al., 2024; Cho et al., 2025). This class-wise failure is a central obstacle for deploying robust models in realistic imbalanced recognition settings.

Recent advances address this via margin rebalancing (Wu et al., 2021), loss reweighting, staged optimization, self-distillation, or adaptive perturbations (Li et al., 2023; Yue et al., 2024; Cho et al., 2025; Wang et al., 2025; Zhang et al., 2026a). Regarding optimization supervision, Zhang et al. (2026b) employ posterior-driven targets to alleviate posterior mismatch. From a geometric perspective, manifold-constrained training (Xian et al., 2026) regularizes adversarial examples near feature manifolds to explicitly prevent off-manifold geometric drift.

Despite these advances, a less examined issue is *where the posterior supervision used in adversarial training is actually supported*. A method may rebalance class contributions, improve posterior supervision at observed samples, or constrain adversarial features in representation space, yet the posterior signal can still be unreliable in the local region explored by adversarial perturbations. This occurs when the surrounding support is sparse, low-density, or highly sensitive, and is most severe for tail classes whose support must be estimated from very few samples. Such unreliable support can make adversarial training overfit to poorly supported tail neighborhoods, producing unstable tail boundaries and amplifying worst-class vulnerability. We refer to this failure mode as *posterior support unreliability*, as illustrated in left part of Figure 1.

To address posterior support unreliability, we draw inspiration from the free-energy principle in statistical physics and propose **FEPT** (**F**ree **E**nergy-guided **P**osterior **T**ransport), a reliable posterior

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

support framework for long-tailed adversarial robustness. The central idea of FEPT is to make adversarial training rely on posterior targets which should be trusted only when its mass is backed by a semantically close, sufficiently dense, and locally stable neighborhood. As shown in right part of Figure 1, FEPT realizes this idea through two coupled modules built on a Riemannian-inspired local semantic metric.

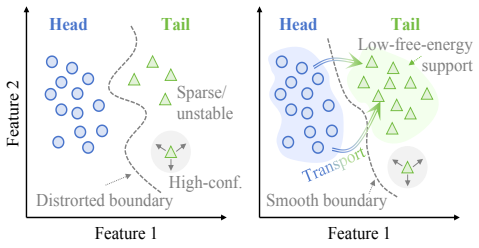


Figure 1: Motivation of FEPT. Left: tail posterior support can be sparse and unstable. Right: FEPT improves tail supervision via free-energy support and posterior transport.

First, **free-energy posterior support** turns posterior confidence into a support-aware reliability signal. We use a free-energy-inspired score to separate spurious point-wise confidence from statistically supported posterior mass: in statistical mechanics, stability is governed by free energy rather than energy alone, incorporating entropy-induced support in addition to energetic preference (Pathria & Beale, 2011). Analogously, FEPT combines posterior confidence with semantic proximity, local density, and posterior stability, down-weighting high-confidence mass in off-support, density-poor, or posterior-unstable regions. Second, since tail classes often lack enough samples to estimate reliable support on their own, **head-to-tail posterior transport** treats semantically related head/medium classes as low-variance support donors and transfers their low-free-energy support toward tail classes through a prototype-aligned map. The transfer strength is controlled by class similarity and tail sample scarcity, allowing FEPT to supplement tail support without assuming that head and tail classes share the same distribution.

We also provide theoretical analysis that motivates why reliable robust supervision requires stable posterior support in the local adversarial neighborhood and why head-to-tail transport can reduce tail support estimation variance, while extensive experiments show that FEPT improves tail and worst-class performance, balanced robustness, and posterior-support stability. Our contributions can be summarized as follows:

- We identify *posterior support reliability* as a key factor in long-tailed adversarial robustness, beyond sample reweighting and point-wise posterior correction.
- We propose FEPT, which combines free-energy posterior support with head-to-tail posterior transport to filter unreliable posterior mass and supplement tail support.
- We theoretically justify that robust supervision benefits from local posterior-support stability, while posterior transport reduces tail-support estimation variance with controllable bias under source alignment.
- Extensive experiments verify that FEPT improves over rebalancing, distillation, geometric, and posterior-aware baselines, especially on tail and worst-class robustness.

2 Preliminaries

2.1 Notation and Adversarial Setting

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a long-tailed training set with $y_i \in \{1, \dots, C\}$ and class counts $\{n_y\}_{y=1}^C$. The imbalance ratio is $\mathrm{IR} = \max_y n_y / \min_y n_y$. We write the model as $f_\theta(x) = h_\theta(\phi_\theta(x))$, where $\phi_\theta : \mathbb{R}^d \to \mathbb{R}^m$ maps an input to a representation $u = \phi_\theta(x)$, and h_θ outputs class logits. The predictive distribution is $p_\theta(\cdot \mid x) = \mathrm{softmax}(f_\theta(x))$. We consider the standard ℓ_∞ threat model $\mathcal{B}_\epsilon(x) = \{x' : \|x' - x\|_\infty \le \epsilon\}$, and let ℓ denote the cross-entropy loss. Standard adversarial training solves

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{x' \in \mathcal{B}_\epsilon(x)} \ell(f_\theta(x'), y) \right]. \tag{1}$$

2.2 Posterior Surrogate

Following the common posterior-surrogate practice in posterior-driven adversarial training (Zhang et al., 2026b), FEPT starts from a stopped-gradient posterior surrogate $q_\psi(y \mid u)$ in representation

space. We instantiate it by an exponential-moving-average (EMA) teacher with temperature scaling and class-prior correction:

$$q_\psi(y \mid u) = \mathrm{softmax}_y \left(\frac{z_{\mathrm{EMA},y}(x) - \lambda_\pi \log \hat{\pi}_y}{T} \right), \tag{2}$$

where $u = \phi_\theta(x)$, $z_{\mathrm{EMA},y}(x)$ is the EMA teacher logit, $\hat{\pi}_y = n_y/n$ is the empirical class prior, T is the temperature, and λ_π controls the strength of prior correction. The EMA teacher provides a stable soft posterior, while prior correction mitigates head-class bias under long-tailed training. Unless otherwise specified, this EMA teacher is the default posterior surrogate used by FEPT.

2.3 Free energy and posterior support

A high-confidence posterior target may still be unreliable when its surrounding region is sparse, low-density, or locally unstable. A free-energy-inspired perspective offers a useful way to organize this distinction. In statistical mechanics, the Helmholtz free energy is defined as

$$F = U - TS, \tag{3}$$

where U is energy, S is entropy, and T is temperature. This principle states that a stable state is not selected by energy alone, but by energy together with entropy-induced statistical support (Pathria & Beale, 2011). In our setting, posterior confidence acts as a point-wise preference, while reliable posterior support further requires class semantic support, local density, and posterior stability. This motivates the class-conditioned free-energy-inspired posterior support score introduced in Section 3.

3 Method

FEPT constructs support-aware posterior supervision in two stages. It first converts the EMA posterior surrogate into free-energy posterior support, which filters unreliable posterior mass in off-support, low-density, or unstable regions. It then transports low-free-energy support from semantically related head/medium classes to tail classes. The resulting transported posterior target is used for adversarial training in Section 3.4.

3.1 Local semantic geometry

FEPT evaluates posterior support in representation space rather than in the input space. To make this evaluation sensitive to the local class structure, we construct a lightweight Riemannian-inspired semantic metric for each class. For class y, we maintain a feature memory bank $\mathcal{B}_y = \{\phi_\theta(x_i) : y_i = y\}$. Given a representation $u = \phi_\theta(x)$, we retrieve its $\min(k, n_y)$ nearest neighbors in $\mathcal{B}_y$ and estimate a local tangent subspace via PCA, where this dynamically bounded neighborhood size ensures mathematical validity even for extreme tail classes whose total available historical samples n_y are strictly fewer than the baseline neighbor hyperparameter k. Let $P_T^y(u)$ denote the projection onto this tangent subspace and let $P_N^y(u) = I - P_T^y(u)$ denote the normal projection. We define the **class-conditional local metric**

$$G_y(u) = P_T^y(u) + \eta_y P_N^y(u),$$

where $\eta_y > 1$ makes normal displacement away from the estimated class support more costly than tangent variation along it. For tail classes, whose support is sparser and less reliably estimated, we use a stronger normal penalty $\eta_y = \eta_0 (n_{\max}/n_y)^\alpha$ with $\alpha \geq 0$. This metric is Riemannian-inspired in the local sense: it provides a position-dependent ruler that distinguishes semantic variation along the class support from off-support drift.

For nearby representations u and v, we define the **local quadratic distance**

$$\Delta_y(u, v) = (u - v)^\top G_y(u)(u - v).$$

Compared with an isotropic Euclidean distance, Δ_y better reflects the asymmetry between tangent movement and normal deviation. FEPT uses this distance to compute semantic support distance, local density, and posterior stability in the free-energy construction below.

3.2 FREE-ENERGY POSTERIOR SUPPORT

Given the local semantic geometry above, FEPT assesses whether the posterior mass around a representation u provides reliable supervision for class y. This reliability is characterized by the following three class-conditional factors.

Semantic support distance. The first factor measures how well u is supported by the observed class-y features. We define the **semantic support distance** as

$$d_y(u) = \min_{v \in \mathcal{B}_y} \Delta_y(u, v),$$

namely the distance from u to the observed support of class y under the local semantic metric. A large $d_y(u)$ indicates weak class-specific support, which is particularly problematic for tail classes because unsupported regions can produce unstable robust boundaries.

Local density score. The second factor measures whether the neighborhood around u is sufficiently populated by class-y samples. We define the **local density score** as

$$\rho_y(u) = \frac{1}{|\mathcal{N}_y(u)|} \sum_{v \in \mathcal{N}_y(u)} \exp\left(-\frac{\Delta_y(u, v)}{2\sigma_y^2}\right),$$

where $\mathcal{N}_y(u)$ denotes the k nearest neighbors of u in $\mathcal{B}_y$. This score is not intended as a generative likelihood; it serves as a local support indicator under the class-conditioned geometry.

Posterior sensitivity. The third factor measures whether the posterior remains stable under small semantic movements. We define the **posterior sensitivity** as

$$s_y(u) = \|\nabla_u \log q_\psi(y \mid u)\|^2_{G_y(u)^{-1}}.$$

A large $s_y(u)$ indicates that a small movement in the local semantic metric can cause a large posterior change, making the corresponding posterior target unreliable for robust supervision.

These factors yield the **class-conditional free-energy support score**

$$E_y(u) = d_y(u) - \tau \log(\rho_y(u) + \varepsilon) + \gamma s_y(u), \tag{4}$$

where τ and γ control the density and stability contributions. This mirrors physical free energy in a support-aware sense: energy penalizes unfavorable states, whereas entropy favors regions with many compatible microstates. In our setting, large support distance and posterior sensitivity are unfavorable, so $d_y(u)$ and $s_y(u)$ act as penalties, while high local density means that many class-y representations support the region, so $-\log(\rho_y(u) + \varepsilon)$ acts as a support reward. Thus, low free energy marks posterior mass that is confident, semantically supported, statistically populated, and locally stable.

Finally, FEPT converts posterior confidence into **free-energy weighted posterior support** over a localized candidate label space. For a training sample with true label y, we define the localized candidate class set $\mathcal{A}_y(u)$ as:

$$\mathcal{A}_y(u) = \{y\} \cup \mathrm{Top}_{M_{\mathrm{cand}}}(q_\psi(\cdot \mid u)) \cup \mathcal{H}_y \cup \left(\cup_{t \in \mathrm{Top}_{M_{\mathrm{cand}}}(q_\psi(\cdot|u)) \cap \mathcal{T}} \mathcal{H}_t\right), \tag{5}$$

where $\mathcal{T}$ denotes the global set of all tail classes, and $\mathcal{H}_y$, $\mathcal{H}_t$ represent the source-class donor sets for the true label and competing tail candidates, respectively. For any evaluated class $k \in \mathcal{A}_y(u)$, its free-energy weighted posterior support is normalized as:

$$\pi_k(u) = \frac{q_\psi(k \mid u) \exp[-E_k(u)]}{\sum_{j \in \mathcal{A}_y(u)} q_\psi(j \mid u) \exp[-E_j(u)]}. \tag{6}$$

For any class outside this unified candidate space, we strictly set $\pi_k(u) = 0$. This comprehensive candidate definition ensures that when evaluating the transported support for any competing tail class via Equation 7, its corresponding head/medium donor classes are guaranteed to reside within the localized normalization space, preventing the source support from being erroneously zeroed out while successfully preserving the $O(B(M_{\mathrm{cand}} + K_{\mathrm{src}} \cdot M_{\mathrm{cand}})K_{\mathrm{nn}}d)$ complexity.

3.3 HEAD-TO-TAIL POSTERIOR TRANSPORT

Free energy identifies reliable support, but tail classes may still have too little such support to estimate from their own samples. FEPT therefore borrows low-free-energy support from semantically related head/medium classes. The goal is to reduce the variance of tail support estimation using nearby data-rich classes.

Let $c_y = |\mathcal{B}_y|^{-1} \sum_{u \in \mathcal{B}_y} u$ be the prototype of class y. For a tail class t, we select its K_{src} nearest head/medium classes as $\mathcal{H}_t$ according to prototype distance. We use the prototype-aligned map $\mathcal{T}_{h \to t}(u) = u - c_h + c_t$ to move source support from the neighborhood of the source prototype to that of the target prototype. When evaluating source support at a target representation, we apply the inverse direction $\mathcal{T}_{t \to h}$ to query the source support in the source coordinate system.

For a tail class t, the **transported posterior support** is

$$\bar{\pi}_t(u) = (1 - \alpha_t)\pi_t(u) + \alpha_t \sum_{h \in \mathcal{H}_t} \omega_{ht}\, \pi_h(\mathcal{T}_{t \to h}(u)), \tag{7}$$

where $\alpha_t = \alpha_0(1 - n_t/n_{\max})$ is the tail-adaptive mixing coefficient, and

$$\omega_{ht} = \frac{\exp(-\|c_h - c_t\|_2^2/\sigma_c^2)}{\sum_{h' \in \mathcal{H}_t} \exp(-\|c_{h'} - c_t\|_2^2/\sigma_c^2)}$$

is the prototype-based transport coefficient from source class h to tail class t. Classes with fewer samples borrow more support, while data-rich classes rely primarily on their own support. For non-tail classes, we set $\bar{\pi}_y(u) = \pi_y(u)$.

This transport step reflects a bias–variance trade-off. Related head/medium classes provide lower-variance support estimates, while the transport coefficients and the tail-adaptive mixing coefficient control the bias introduced by source classes. FEPT therefore supplements tail support without assuming that head and tail classes share the same distribution.

3.4 FEPT ADVERSARIAL TRAINING

For a sample (x, y), the inner maximization seeks a high-loss perturbation while discouraging high-free-energy drift:

$$\delta^\star = \arg \max_{\|\delta\|_\infty \le \epsilon} \left[\ell(f_\theta(x + \delta), y) - \lambda_E E_y(\phi_\theta(x + \delta))\right]. \tag{8}$$

The penalty favors adversarial examples that are hard yet semantically supported. During each PGD step, retrieval and truncation objects, including kNN sets, PCA bases, candidate classes, donor sets, and memory states, are fixed; conditioned on them, the objective is differentiable almost everywhere. We further verify the absence of gradient masking with AutoAttack and adaptive attacks in Section 5.3.

The outer minimization uses a transported support-aware target. Rather than matching the pointwise surrogate q_ψ, FEPT converts it into free-energy support, transports reliable tail support, and normalizes

$$\bar{q}(k \mid u) = \frac{\bar{\pi}_k(u) + \varepsilon}{\sum_{j \in \mathcal{A}_y(u)} (\bar{\pi}_j(u) + \varepsilon)}, \qquad \bar{q}(k \mid u) = 0 \text{ for } k \notin \mathcal{A}_y(u).$$

The final objective is

$$\min_\theta \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\ell(f_\theta(x + \delta^\star), y) + \beta \mathrm{KL}\left(\mathrm{sg}[\bar{q}(\cdot \mid \phi_\theta(x))] \,\|\, p_\theta(\cdot \mid x + \delta^\star)\right)\right], \tag{9}$$

where $\mathrm{sg}[\cdot]$ stops gradients through the target. The first term keeps standard adversarial discrimination, and the KL term aligns predictions with tail support filtered by free energy and supplemented by transport. The full procedure is summarized in Appendix B.

4 THEORETICAL ANALYSIS

We analyze two properties that directly match the design of FEPT: local stability of posterior support and variance-reducing posterior transport. Full proofs are provided in Appendix C.

4.1 LOCAL STABILITY OF POSTERIOR SUPPORT

For class y, define the posterior margin $m_q^y(u) = q_\psi(y \mid u) - \max_{k \neq y} q_\psi(k \mid u)$. FEPT measures local perturbations by the class-conditional metric distance $d_{G_y}(u, v) = \sqrt{\Delta_y(u, v)}$, where G_y is the local semantic metric defined in Section 3.1. To model the effect of long-tailed sampling, we use a class-dependent Lipschitz constant $L_y = L_0 \left(n_{\max}/n_y\right)^\chi$, where $L_0 > 0$ is the head-class baseline stability constant and $\chi > 0$ controls how strongly posterior volatility increases as class support becomes scarce.

Theorem 1 (Posterior support requires local stability)**.** *Assume that m_q^y is L_y-Lipschitz in $\mathcal{B}_{G_y}(u, r)$ with respect to d_{G_y}, and that $\|q_\psi(\cdot \mid v) - p(\cdot \mid v)\|_\infty \leq \epsilon_p$ for all $v \in \mathcal{B}_{G_y}(u, r)$. Then class y remains posterior-preferred throughout $\mathcal{B}_{G_y}(u, r)$ if $m_q^y(u) > L_y r + 2\epsilon_p$. Conversely, if $m_q^y(u) \leq L_y r$, point-wise posterior accuracy at u alone cannot rule out a posterior decision change inside the local ball.*

The condition links posterior stability directly to class imbalance. The class-dependent scaling $L_y = L_0(n_{\max}/n_y)^\chi$ is an explicit modeling assumption for the empirically observed fact that posterior estimates become more volatile when class support is scarce. Under this assumption, the local stability threshold is $L_y r$, so tail classes with fewer samples require a larger posterior margin to keep their preference unchanged within the same adversarial neighborhood. Hence, correcting the posterior value at u is not sufficient when the surrounding support is sparse or sensitive. The free-energy term in Equation 4 addresses this issue by suppressing posterior mass in low-density or posterior-unstable regions before it is used for robust supervision.

4.2 BIAS–VARIANCE TRADE-OFF OF POSTERIOR TRANSPORT

We next formalize the role of head-to-tail transport. Let $\pi_t^\star$ be the ideal low-free-energy support of tail class t, and let $\hat{\pi}_t$ be its tail-only estimate with error scale σ_t^2/n_t. For each source class $h \in \mathcal{H}_t$, let $T_{h \to t}\hat{\pi}_h$ be the transported source estimate with error scale σ_h^2/n_h and transport bias b_{ht} relative to $\pi_t^\star$. FEPT materializes this formulation by constructing the empirical transported estimator $\hat{\pi}_t^{FEPT}$ over the representation space learned by the shared encoder.

Theorem 2 (Bias–variance trade-off with correlated estimators)**.** *Let $\hat{\pi}_t^{FEPT} = (1 - \alpha_t)\hat{\pi}_t + \alpha_t \sum_{h \in \mathcal{H}_t} \omega_{ht} T_{h \to t}\hat{\pi}_h$ serve as the empirical counterpart of the continuous transported posterior support $\bar{\pi}_t(u)$ defined in Equation 7, evaluated via spatial sampling. Assume that the estimation noises ξ have zero mean and that their cross-class correlations induced by the shared backbone are bounded by $\rho \in [0, 1]$, i.e., $\left|\mathbb{E}\langle \xi_a, \xi_b \rangle\right| \leq \rho \frac{\sigma_a \sigma_b}{\sqrt{n_a n_b}}$ for $a \neq b$. Then*

$$\mathbb{E}\|\hat{\pi}_t^{FEPT} - \pi_t^\star\|^2 \leq (1 - \alpha_t)^2 \frac{\sigma_t^2}{n_t} + \alpha_t^2 \sum_{h \in \mathcal{H}_t} \omega_{ht}^2 \frac{\sigma_h^2}{n_h} + \mathcal{C}_t(\rho) + \alpha_t^2 \left(\sum_{h \in \mathcal{H}_t} \omega_{ht} b_{ht} \right)^2, \tag{10}$$

where $\mathcal{C}_t(\rho) = 2\rho\alpha_t(1 - \alpha_t) \sum_{h \in \mathcal{H}_t} \omega_{ht} \frac{\sigma_t \sigma_h}{\sqrt{n_t n_h}} + \rho\alpha_t^2 \sum_{h \neq h'} \omega_{ht}\omega_{h't} \frac{\sigma_h \sigma_{h'}}{\sqrt{n_h n_{h'}}}$.

The first term is the high-variance tail-only estimate, the second term is the transported source variance, $\mathcal{C}_t(\rho)$ captures correlation induced by the shared encoder, and the last term is transport bias. Thus, transport is beneficial when source classes are data-rich and well aligned, while excessive or poorly aligned transport increases the bias and correlation terms. FEPT controls this trade-off through ω_{ht} and α_t. The theorem explains when such a transported estimator is expected to help, namely when donor classes are data-rich and sufficiently aligned. We test this alignment requirement empirically in Appendix E.3.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTING

Datasets and architectures. Following the common setting in recent work, we use CIFAR10-LT with imbalance ratio $\mathrm{IR} = 50$, and CIFAR100-LT and TinyImageNet-LT with $\mathrm{IR} = 10$ in main results. The class counts follow the standard exponential long-tailed construction, and the data-generation rule is given in Appendix D.1. Additional IR sweeps are also reported in Appendix E.5. WRN-28-10 is used as the primary backbone.

Table 1: Main results with WRN-28-10 under PGD-20 attack. The best result is bolded and the second best is underlined.

Dataset	Method	Nat. all	Nat. tail	PGD all	PGD tail	AA all	AA tail	Worst-AA	Gap
CIFAR10-LT	AT	$58.31_{\pm 0.42}$	$48.64_{\pm 0.58}$	$27.35_{\pm 0.18}$	$13.76_{\pm 0.20}$	$26.18_{\pm 0.17}$	$12.46_{\pm 0.21}$	$6.74_{\pm 0.19}$	$22.31_{\pm 0.37}$
	AWP	$59.72_{\pm 0.25}$	$50.21_{\pm 0.31}$	$28.55_{\pm 0.32}$	$14.96_{\pm 0.36}$	$27.53_{\pm 0.30}$	$13.84_{\pm 0.33}$	$7.86_{\pm 0.26}$	$21.74_{\pm 0.44}$
	RoBal	$72.81_{\pm 0.18}$	$67.22_{\pm 0.22}$	$32.36_{\pm 0.44}$	$22.31_{\pm 0.38}$	$28.97_{\pm 0.35}$	$18.62_{\pm 0.34}$	$12.35_{\pm 0.28}$	$18.14_{\pm 0.41}$
	REAT	$67.85_{\pm 0.23}$	$60.94_{\pm 0.27}$	$28.76_{\pm 0.39}$	$17.25_{\pm 0.42}$	$27.20_{\pm 0.42}$	$15.63_{\pm 0.39}$	$9.85_{\pm 0.33}$	$20.09_{\pm 0.47}$
	AT-BSL	$77.18_{\pm 0.37}$	$72.55_{\pm 0.33}$	$38.08_{\pm 0.41}$	$28.74_{\pm 0.63}$	$34.68_{\pm 0.32}$	$25.03_{\pm 0.48}$	$18.94_{\pm 0.37}$	$15.86_{\pm 0.45}$
	TAET	$65.46_{\pm 1.85}$	$60.58_{\pm 1.69}$	$31.75_{\pm 0.49}$	$24.74_{\pm 0.55}$	$27.94_{\pm 0.73}$	$20.86_{\pm 0.41}$	$14.12_{\pm 0.39}$	$13.47_{\pm 0.46}$
	RobustLT	$77.83_{\pm 0.45}$	$73.94_{\pm 0.51}$	$42.24_{\pm 0.34}$	$36.13_{\pm 0.62}$	$37.62_{\pm 0.27}$	$30.70_{\pm 0.36}$	$23.16_{\pm 0.32}$	$10.42_{\pm 0.34}$
	MCAT	$76.94_{\pm 0.31}$	$72.81_{\pm 0.35}$	$41.63_{\pm 0.29}$	$35.74_{\pm 0.47}$	$36.91_{\pm 0.25}$	$30.26_{\pm 0.34}$	$22.62_{\pm 0.29}$	$9.88_{\pm 0.31}$
	PAT	$76.59_{\pm 0.34}$	$72.64_{\pm 0.44}$	$36.70_{\pm 0.34}$	$28.93_{\pm 0.68}$	$34.92_{\pm 0.26}$	$25.86_{\pm 0.36}$	$20.64_{\pm 0.40}$	$14.58_{\pm 0.43}$
	FEPT-BSL	$78.08_{\pm 0.30}$	$74.35_{\pm 0.33}$	$43.71_{\pm 0.26}$	$37.92_{\pm 0.42}$	$39.86_{\pm 0.24}$	$34.12_{\pm 0.31}$	$25.52_{\pm 0.28}$	$8.60_{\pm 0.30}$
	FEPT-PAT	$\mathbf{78.42}_{\pm 0.29}$	$\mathbf{74.70}_{\pm 0.31}$	$43.58_{\pm 0.27}$	$38.22_{\pm 0.40}$	$\mathbf{40.25}_{\pm 0.23}$	$34.42_{\pm 0.30}$	$25.78_{\pm 0.27}$	$8.74_{\pm 0.28}$
	FEPT-EMA	$78.26_{\pm 0.28}$	$74.62_{\pm 0.30}$	$\mathbf{43.82}_{\pm 0.28}$	$\mathbf{38.46}_{\pm 0.39}$	$40.18_{\pm 0.22}$	$\mathbf{34.85}_{\pm 0.29}$	$\mathbf{26.18}_{\pm 0.25}$	$\mathbf{8.37}_{\pm 0.27}$
CIFAR100-LT	AT	$44.40_{\pm 0.21}$	$40.02_{\pm 0.25}$	$15.36_{\pm 0.14}$	$13.35_{\pm 0.19}$	$15.21_{\pm 0.13}$	$13.20_{\pm 0.22}$	$7.48_{\pm 0.18}$	$4.82_{\pm 0.21}$
	AWP	$45.53_{\pm 0.36}$	$40.81_{\pm 0.34}$	$16.39_{\pm 0.08}$	$14.31_{\pm 0.24}$	$16.25_{\pm 0.09}$	$14.18_{\pm 0.23}$	$8.36_{\pm 0.19}$	$5.02_{\pm 0.25}$
	RoBal	$49.44_{\pm 0.14}$	$46.89_{\pm 0.13}$	$16.62_{\pm 0.07}$	$15.43_{\pm 0.10}$	$16.43_{\pm 0.08}$	$15.24_{\pm 0.12}$	$9.71_{\pm 0.10}$	$2.96_{\pm 0.13}$
	REAT	$47.31_{\pm 0.18}$	$45.52_{\pm 0.09}$	$15.48_{\pm 0.14}$	$14.55_{\pm 0.18}$	$15.32_{\pm 0.14}$	$14.34_{\pm 0.18}$	$9.12_{\pm 0.15}$	$2.75_{\pm 0.18}$
	AT-BSL	$55.08_{\pm 0.10}$	$52.83_{\pm 0.12}$	$19.88_{\pm 0.16}$	$18.86_{\pm 0.20}$	$19.62_{\pm 0.17}$	$18.51_{\pm 0.20}$	$12.42_{\pm 0.18}$	$2.43_{\pm 0.17}$
	TAET	$46.05_{\pm 0.19}$	$41.53_{\pm 0.13}$	$14.91_{\pm 0.15}$	$13.21_{\pm 0.22}$	$14.74_{\pm 0.15}$	$13.02_{\pm 0.21}$	$8.31_{\pm 0.17}$	$3.85_{\pm 0.20}$
	RobustLT	$55.72_{\pm 0.20}$	$53.61_{\pm 0.18}$	$23.91_{\pm 0.19}$	$23.48_{\pm 0.32}$	$20.03_{\pm 0.21}$	$19.16_{\pm 0.27}$	$14.92_{\pm 0.22}$	$1.95_{\pm 0.17}$
	MCAT	$56.47_{\pm 0.18}$	$54.68_{\pm 0.17}$	$24.96_{\pm 0.17}$	$24.42_{\pm 0.25}$	$21.36_{\pm 0.18}$	$20.78_{\pm 0.23}$	$16.11_{\pm 0.20}$	$1.42_{\pm 0.14}$
	PAT	$51.25_{\pm 0.33}$	$49.07_{\pm 0.38}$	$23.39_{\pm 0.33}$	$22.46_{\pm 0.22}$	$20.86_{\pm 0.19}$	$19.72_{\pm 0.24}$	$6.74_{\pm 0.20}$	$1.87_{\pm 0.15}$
	FEPT-BSL	$\mathbf{57.18}_{\pm 0.19}$	$55.10_{\pm 0.18}$	$\mathbf{25.82}_{\pm 0.18}$	$25.16_{\pm 0.24}$	$22.06_{\pm 0.17}$	$21.52_{\pm 0.22}$	$16.55_{\pm 0.19}$	$1.27_{\pm 0.13}$
	FEPT-PAT	$56.88_{\pm 0.18}$	$\mathbf{55.34}_{\pm 0.17}$	$25.54_{\pm 0.17}$	$25.25_{\pm 0.23}$	$22.18_{\pm 0.17}$	$21.64_{\pm 0.21}$	$16.68_{\pm 0.19}$	$1.22_{\pm 0.13}$
	FEPT-EMA	$57.05_{\pm 0.17}$	$55.26_{\pm 0.16}$	$25.68_{\pm 0.16}$	$\mathbf{25.44}_{\pm 0.22}$	$\mathbf{22.24}_{\pm 0.16}$	$\mathbf{21.88}_{\pm 0.20}$	$\mathbf{16.90}_{\pm 0.18}$	$\mathbf{1.15}_{\pm 0.12}$
TinyImageNet-LT	AT	$44.31_{\pm 0.38}$	$38.20_{\pm 0.26}$	$19.63_{\pm 0.41}$	$14.12_{\pm 0.55}$	$17.42_{\pm 0.31}$	$12.38_{\pm 0.25}$	$7.20_{\pm 0.21}$	$7.84_{\pm 0.34}$
	AWP	$40.82_{\pm 1.10}$	$34.32_{\pm 1.22}$	$20.62_{\pm 0.91}$	$15.05_{\pm 1.04}$	$18.65_{\pm 0.82}$	$13.46_{\pm 0.79}$	$7.83_{\pm 0.60}$	$7.36_{\pm 0.86}$
	RoBal	$49.94_{\pm 0.33}$	$47.98_{\pm 0.55}$	$21.63_{\pm 0.27}$	$18.66_{\pm 0.35}$	$18.58_{\pm 0.36}$	$15.45_{\pm 0.31}$	$10.52_{\pm 0.28}$	$4.72_{\pm 0.29}$
	REAT	$44.15_{\pm 0.78}$	$41.44_{\pm 0.31}$	$21.56_{\pm 0.72}$	$19.46_{\pm 0.94}$	$17.72_{\pm 0.37}$	$15.24_{\pm 0.58}$	$10.11_{\pm 0.41}$	$3.89_{\pm 0.36}$
	AT-BSL	$40.18_{\pm 0.88}$	$37.06_{\pm 1.03}$	$25.42_{\pm 0.38}$	$23.08_{\pm 0.41}$	$21.52_{\pm 0.55}$	$18.96_{\pm 0.63}$	$13.86_{\pm 0.47}$	$4.14_{\pm 0.43}$
	TAET	$31.85_{\pm 1.02}$	$28.70_{\pm 1.12}$	$17.42_{\pm 0.39}$	$15.36_{\pm 0.66}$	$12.96_{\pm 0.70}$	$11.08_{\pm 0.86}$	$6.40_{\pm 0.57}$	$3.55_{\pm 0.51}$
	RobustLT	$45.18_{\pm 1.20}$	$41.62_{\pm 1.30}$	$27.12_{\pm 0.31}$	$24.96_{\pm 0.22}$	$22.48_{\pm 0.21}$	$20.12_{\pm 0.12}$	$14.72_{\pm 0.14}$	$3.74_{\pm 0.15}$
	MCAT	$50.34_{\pm 0.42}$	$48.10_{\pm 0.40}$	$27.90_{\pm 0.27}$	$25.82_{\pm 0.23}$	$23.50_{\pm 0.22}$	$21.76_{\pm 0.19}$	$16.20_{\pm 0.17}$	$2.92_{\pm 0.18}$
	PAT	$49.28_{\pm 0.39}$	$46.75_{\pm 0.35}$	$27.36_{\pm 0.27}$	$25.44_{\pm 0.22}$	$22.87_{\pm 0.23}$	$20.92_{\pm 0.18}$	$15.38_{\pm 0.18}$	$3.36_{\pm 0.19}$
	FEPT-BSL	$50.86_{\pm 0.38}$	$48.70_{\pm 0.35}$	$\mathbf{28.62}_{\pm 0.26}$	$26.38_{\pm 0.23}$	$\mathbf{24.30}_{\pm 0.22}$	$22.36_{\pm 0.18}$	$16.68_{\pm 0.17}$	$2.64_{\pm 0.17}$
	FEPT-PAT	$\mathbf{51.18}_{\pm 0.37}$	$\mathbf{49.06}_{\pm 0.34}$	$28.40_{\pm 0.26}$	$26.50_{\pm 0.22}$	$24.12_{\pm 0.21}$	$22.58_{\pm 0.18}$	$16.88_{\pm 0.16}$	$2.53_{\pm 0.17}$
	FEPT-EMA	$51.04_{\pm 0.36}$	$48.96_{\pm 0.33}$	$28.54_{\pm 0.25}$	$\mathbf{26.72}_{\pm 0.21}$	$24.23_{\pm 0.20}$	$\mathbf{22.85}_{\pm 0.17}$	$\mathbf{17.12}_{\pm 0.16}$	$\mathbf{2.38}_{\pm 0.16}$

Training and attacks. All methods use the ℓ_∞ threat model with $\epsilon = 8/255$. Unless otherwise prescribed, training uses PGD-10 adversaries. Evaluation uses PGD-20 with step size $2/255$ and AutoAttack; PGD-100, CW-20, and adaptive attacks are reported in Section 5.3. Further details and sweeps are in Appendices D.2–E.6.

Baselines. We compare with standard adversarial training methods AT (Madry et al., 2018) and AWP (Wu et al., 2020); long-tailed robust methods RoBal (Wu et al., 2021), REAT (Li et al., 2023), AT-BSL (Yue et al., 2024), TAET (Wang et al., 2025), MCAT (Xian et al., 2026), and plug-and-play RobustLT (Zhang et al., 2026a) attached to AT-BSL; and PAT (Zhang et al., 2026b), the closest posterior-supervision baseline. Unless specified otherwise, FEPT uses the EMA posterior surrogate in Equation 2, and we denote this default configuration as FEPT or FEPT-EMA in the result tables. We additionally consider two variants, FEPT-PAT and FEPT-BSL, which replace the EMA surrogate by a PAT surrogate and a BSL-trained prior-corrected surrogate, respectively.

Metrics. We report natural, PGD-20, and AutoAttack accuracy on all and tail classes, denoted Nat., PGD, and AA, together with worst-class AutoAttack robustness (Worst-AA) and the head–tail gap $\mathrm{Gap}(M) = \mathrm{AA} * \mathrm{head}(M) - \mathrm{AA} * \mathrm{tail}(M)$. Tail classes are the 80% least frequent classes. Mechanism diagnostics include mean posterior sensitivity $\bar{s}$, mean free energy $\bar{E}$, robust posterior support volume $\widehat{V}^{\mathrm{RPS}}$, and empirical transport error $\widehat{\mathcal{E}}_{\mathrm{tr}} = |\hat{\pi}_t^{FEPT} - \hat{\pi}_t^{\star}|_2^2$, where $\hat{\pi}_t^{\star}$ is estimated by a class-balanced reference model for analysis only. All results are averaged over five runs and reported as mean with standard deviation.

5.2 Main Results

As shown in Table 1, FEPT-EMA consistently achieves the top AA tail and Worst-AA performance across all datasets. The two source-replacement variants, FEPT-PAT and FEPT-BSL, remain stronger than PAT and other competing methods, suggesting that the observed gain comes from the support-reliability transformation rather than from a specific posterior surrogate alone. This advantage indicates that correcting point-wise posterior mismatch at observed samples (as done by

Table 2: Robustness under stronger and adaptive attacks with WRN-28-10. Each entry reports all-class accuracy / tail-class accuracy.

Dataset	Method	PGD-100	CW-20	AA	Adaptive-E	Adaptive-s
CIFAR10-LT	AT-BSL	37.62/28.20$_{\pm 0.42}$	37.18/27.64$_{\pm 0.45}$	34.68/25.03$_{\pm 0.38}$	33.71/24.26$_{\pm 0.41}$	33.58/24.07$_{\pm 0.42}$
	RobustLT	41.73/35.51$_{\pm 0.36}$	40.25/33.76$_{\pm 0.39}$	37.62/30.70$_{\pm 0.32}$	36.58/29.82$_{\pm 0.35}$	36.34/29.55$_{\pm 0.36}$
	MCAT	41.18/35.26$_{\pm 0.32}$	39.56/33.17$_{\pm 0.33}$	36.91/30.26$_{\pm 0.29}$	35.96/29.42$_{\pm 0.30}$	35.74/29.20$_{\pm 0.31}$
	PAT	36.18/28.40$_{\pm 0.31}$	35.72/27.85$_{\pm 0.32}$	34.92/25.86$_{\pm 0.28}$	34.18/25.20$_{\pm 0.30}$	34.03/25.06$_{\pm 0.30}$
	FEPT-BSL	42.98/37.30$_{\pm 0.29}$	42.05/36.08$_{\pm 0.30}$	39.86/34.12$_{\pm 0.24}$	38.92/33.20$_{\pm 0.27}$	38.70/32.96$_{\pm 0.27}$
	FEPT-PAT	43.04/37.64$_{\pm 0.28}$	42.18/36.45$_{\pm 0.29}$	**40.25**/34.42$_{\pm 0.23}$	39.18/33.52$_{\pm 0.26}$	38.98/33.30$_{\pm 0.26}$
	FEPT-EMA	**43.12/37.91**$_{\pm 0.27}$	**42.37/36.72**$_{\pm 0.28}$	40.18/**34.85**$_{\pm 0.25}$	**39.34/33.91**$_{\pm 0.26}$	**39.12/33.70**$_{\pm 0.27}$
CIFAR100-LT	AT-BSL	19.54/18.40$_{\pm 0.19}$	20.18/19.04$_{\pm 0.20}$	19.62/18.51$_{\pm 0.19}$	18.93/17.86$_{\pm 0.20}$	18.81/17.73$_{\pm 0.20}$
	RobustLT	23.42/23.00$_{\pm 0.23}$	22.78/22.03$_{\pm 0.24}$	20.03/19.16$_{\pm 0.24}$	19.42/18.63$_{\pm 0.25}$	19.35/18.55$_{\pm 0.25}$
	MCAT	24.51/24.03$_{\pm 0.20}$	23.56/22.92$_{\pm 0.21}$	21.36/20.78$_{\pm 0.21}$	20.76/20.18$_{\pm 0.22}$	20.68/20.07$_{\pm 0.22}$
	PAT	23.89/23.32$_{\pm 0.21}$	23.02/22.18$_{\pm 0.22}$	20.86/19.72$_{\pm 0.22}$	20.02/18.94$_{\pm 0.23}$	19.92/18.84$_{\pm 0.23}$
	FEPT-BSL	25.14/24.72$_{\pm 0.19}$	24.52/23.98$_{\pm 0.20}$	22.06/21.52$_{\pm 0.17}$	21.48/20.96$_{\pm 0.20}$	21.36/20.84$_{\pm 0.20}$
	FEPT-PAT	25.10/24.88$_{\pm 0.19}$	24.50/24.08$_{\pm 0.20}$	22.18/21.64$_{\pm 0.17}$	21.60/21.10$_{\pm 0.20}$	21.48/20.98$_{\pm 0.20}$
	FEPT-EMA	**25.21/24.97**$_{\pm 0.18}$	**24.64/24.18**$_{\pm 0.19}$	**22.24/21.88**$_{\pm 0.18}$	**21.71/21.33**$_{\pm 0.19}$	**21.60/21.20**$_{\pm 0.19}$
TinyImageNet-LT	AT-BSL	24.81/22.41$_{\pm 0.43}$	23.05/20.39$_{\pm 0.48}$	21.52/18.96$_{\pm 0.52}$	20.83/18.25$_{\pm 0.55}$	20.70/18.08$_{\pm 0.56}$
	RobustLT	26.63/24.41$_{\pm 0.35}$	24.02/21.47$_{\pm 0.39}$	22.48/20.12$_{\pm 0.34}$	21.74/19.42$_{\pm 0.37}$	21.59/19.25$_{\pm 0.38}$
	MCAT	27.18/24.96$_{\pm 0.31}$	24.73/22.16$_{\pm 0.35}$	23.50/21.76$_{\pm 0.31}$	22.32/19.98$_{\pm 0.33}$	22.18/19.83$_{\pm 0.34}$
	PAT	26.92/24.71$_{\pm 0.32}$	24.46/21.92$_{\pm 0.36}$	22.87/20.92$_{\pm 0.32}$	21.96/19.61$_{\pm 0.34}$	21.83/19.48$_{\pm 0.35}$
	FEPT-BSL	28.02/25.70$_{\pm 0.29}$	25.40/22.88$_{\pm 0.31}$	**24.30**/22.36$_{\pm 0.22}$	23.48/21.30$_{\pm 0.29}$	23.28/21.18$_{\pm 0.30}$
	FEPT-PAT	27.96/25.92$_{\pm 0.29}$	25.20/23.12$_{\pm 0.31}$	24.12/22.58$_{\pm 0.21}$	23.32/21.42$_{\pm 0.29}$	23.10/21.24$_{\pm 0.30}$
	FEPT-EMA	**28.34/26.21**$_{\pm 0.28}$	**25.92/23.58**$_{\pm 0.30}$	24.23/**22.85**$_{\pm 0.20}$	**23.71/21.52**$_{\pm 0.29}$	**23.56/21.36**$_{\pm 0.30}$

Table 3: Ablation study on CIFAR10-LT and CIFAR100-LT with WRN-28-10. Mechanism metrics are averaged over tail classes.

Variant	CIFAR10-LT				CIFAR100-LT			
	AA all	AA tail	$\bar{s}\downarrow$	$\widehat{\mathcal{E}}_{\mathrm{tr}}\downarrow$	AA all	AA tail	$\bar{s}\downarrow$	$\widehat{\mathcal{E}}_{\mathrm{tr}}\downarrow$
Full FEPT	**40.18**$_{\pm 0.22}$	**34.85**$_{\pm 0.29}$	**0.34**$_{\pm 0.12}$	**0.18**$_{\pm 0.10}$	**22.24**$_{\pm 0.16}$	**21.88**$_{\pm 0.20}$	**0.38**$_{\pm 0.11}$	**0.21**$_{\pm 0.10}$
w/o free-energy inner term	38.47$_{\pm 0.37}$	32.41$_{\pm 0.46}$	0.51$_{\pm 0.15}$	0.22$_{\pm 0.12}$	21.12$_{\pm 0.23}$	20.38$_{\pm 0.31}$	0.56$_{\pm 0.14}$	0.25$_{\pm 0.13}$
w/o support reweighting	37.92$_{\pm 0.29}$	31.70$_{\pm 0.38}$	0.58$_{\pm 0.17}$	0.24$_{\pm 0.15}$	20.81$_{\pm 0.18}$	20.06$_{\pm 0.34}$	0.61$_{\pm 0.18}$	0.27$_{\pm 0.16}$
w/o head-to-tail transport	37.66$_{\pm 0.41}$	30.82$_{\pm 0.54}$	0.39$_{\pm 0.13}$	0.37$_{\pm 0.19}$	20.72$_{\pm 0.25}$	19.54$_{\pm 0.37}$	0.43$_{\pm 0.12}$	0.42$_{\pm 0.18}$
fixed α_t	38.56$_{\pm 0.24}$	32.58$_{\pm 0.35}$	0.37$_{\pm 0.11}$	0.28$_{\pm 0.14}$	21.28$_{\pm 0.21}$	20.45$_{\pm 0.26}$	0.41$_{\pm 0.16}$	0.31$_{\pm 0.12}$
uniform source weights	38.73$_{\pm 0.32}$	32.83$_{\pm 0.43}$	0.36$_{\pm 0.16}$	0.30$_{\pm 0.13}$	21.34$_{\pm 0.24}$	20.63$_{\pm 0.30}$	0.40$_{\pm 0.10}$	0.33$_{\pm 0.17}$
w/o posterior KL	39.02$_{\pm 0.27}$	33.28$_{\pm 0.32}$	0.41$_{\pm 0.18}$	0.24$_{\pm 0.11}$	21.47$_{\pm 0.19}$	20.78$_{\pm 0.28}$	0.45$_{\pm 0.15}$	0.27$_{\pm 0.14}$

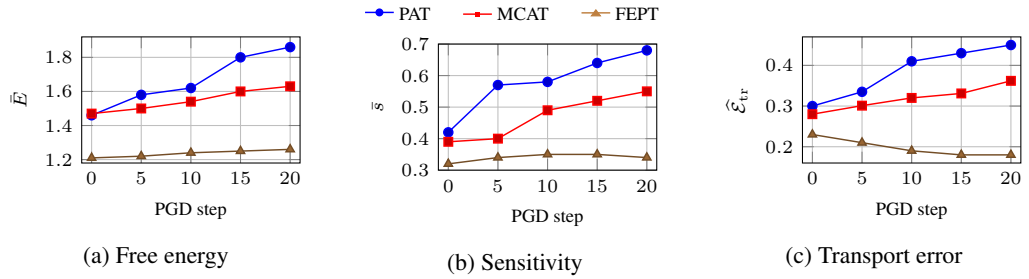
PAT) is not sufficient when local neighborhoods are sparse or unstable, supporting the need for free-energy support filtering and variance-reducing posterior transport. Consistent advantages are further corroborated on ResNet-18 architectures (Appendix E.1) and across varying imbalance ratios (Appendix E.5).

5.3 Stronger and adaptive attacks

We also evaluate FEPT under stronger attacks, including PGD-100, CW-20, AutoAttack, and two adaptive attacks. Adaptive-E maximizes $\ell + \mu E_y$, and Adaptive-s maximizes $\ell + \mu s_y$, with $\mu \in \{0.1, 0.2, 0.5, 1.0\}$; the strongest result is reported. As shown in Table 2, the AA entries are consistent with the main results in Table 1, and all FEPT variants remain stronger than PAT under stronger and adaptive evaluations. The default FEPT generally gives the strongest tail robustness, while FEPT-PAT and FEPT-BSL remain competitive across all-class metrics. This mixed ranking across posterior sources supports that the improvement comes from the support-reliability transformation rather than from a single privileged posterior surrogate. The adaptive variants reduce accuracy slightly compared with AutoAttack in most cases, but the ranking remains stable, suggesting that the gains are not caused by gradient masking or by an easier inner maximization.

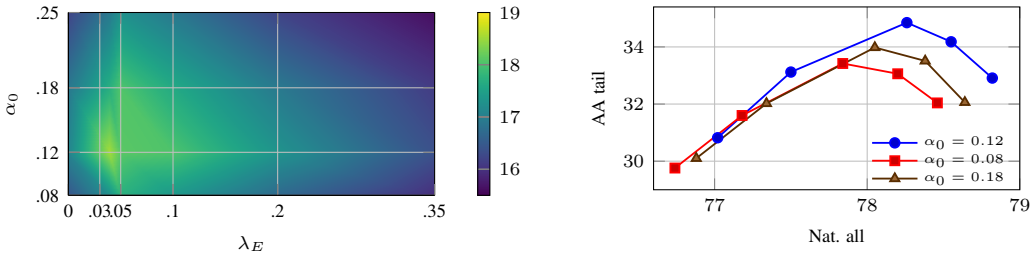
5.4 Ablation study

Table 3 isolates the role of each component. Removing the free-energy inner term lowers CIFAR10-LT AA tail from 34.85% to 32.41% and increases posterior sensitivity, supporting its role as a geometric regularizer rather than a weaker inner attack, consistent with Theorem 1. Removing head-to-tail transport hurts tail robustness more strongly, matching Theorem 2's bias–variance interpretation. Fixed α_t or uniform source weights are also sub-optimal, showing the importance of tail-adaptive



(a) Free energy (b) Sensitivity (c) Transport error

Figure 2: Posterior-support diagnostics on CIFAR10-LT.



(a) AA tail heat map (b) Natural–robust trade-off

Figure 3: Hyperparameter behavior on CIFAR10-LT.

mixing and prototype-aligned source weighting. The additional ablations test simpler alternatives: posterior matching without reliability filtering, tail-only support estimation, and the free-energy inner objective without posterior KL. TinyImageNet-LT results (Appendix E.2) and source-alignment diagnostics (Appendix E.3) show consistent trends.

5.5 Mechanism verification

Figure 2 reports posterior-support diagnostics on CIFAR10-LT. Along adversarial trajectories, FEPT maintains lower free energy $\bar{E}$, lower posterior sensitivity $\bar{s}$, and lower empirical transport error $\widehat{\mathcal{E}}_{\mathrm{tr}}$ than PAT and MCAT. These results support the mechanism predicted by Theorem 1 and Theorem 2: posterior supervision is useful only when local posterior preferences remain stable, and transport is beneficial only when it reduces tail-estimation variance without incurring excessive alignment bias. Appendix E.4 provides the corresponding diagnostics on CIFAR100-LT and TinyImageNet-LT, where a similar pattern holds.

5.6 Hyperparameter sensitivity

Figure 3 varies α_0 and λ_E on CIFAR10-LT with WRN-28-10. Tail robustness improves when transport is enabled and the free-energy penalty is nonzero, but it decreases when either term becomes too large. This behavior is expected: small α_0 leaves tail support under-estimated, whereas large α_0 over-borrows from related but non-identical classes; similarly, a very large λ_E may over-restrict the inner maximization. The results on CIFAR100-LT and TinyImageNet-LT are provided in Appendix E.6, showing a similar intermediate-optimum pattern.

6 Conclusion

We introduced FEPT, a free-energy guided posterior support framework for long-tailed adversarial robustness. The central claim is that tail-class robust failure is not only caused by posterior mismatch, but also by posterior support unreliability: tail classes occupy sparse, high-energy, and posterior-unstable regions in representation space. FEPT addresses this failure by combining free-energy support filtering, prototype-aligned posterior transport, and support-aware posterior supervision. The resulting formulation provides a direct route to improving tail and worst-class robustness beyond point-wise posterior matching.

REPRODUCIBILITY STATEMENT

We provide the full training procedure, optimization settings, attack configurations, model architectures, long-tailed data construction, and FEPT-specific hyperparameters in the main text and Appendix D. All reported results are averaged over five independent runs, and the evaluation protocol includes PGD-20, PGD-100, CW-20, AutoAttack, and adaptive attacks. We also specify the construction of class-wise memory banks, local semantic metrics, candidate class sets, source-class selection, and posterior transport. Upon publication, we will release the training and evaluation code, configuration files, random seeds, data-split generation scripts, and scripts for reproducing the reported tables and figures.

ETHICS STATEMENT

This work studies adversarial robustness under long-tailed data distributions using standard image classification benchmarks, including CIFAR10, CIFAR100, and TinyImageNet. The experiments do not involve human subjects, personally identifiable information, or private data. The adversarial attacks considered in this work are standard research tools for evaluating model robustness and are used solely to analyze and improve defensive learning methods. As with adversarial robustness research more broadly, the techniques studied here may have dual-use implications, but the proposed method is designed to improve the reliability and robustness of machine learning systems rather than to facilitate malicious attacks.

AI USE STATEMENT

Generative AI tools assisted with language editing, code drafting, and consistency checks. The authors remain responsible for all claims, citations, experimental results, and released artifacts, and will verify every AI-assisted component before submission.

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, 2001.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Yang Cao, Yanbo Chen, and Weiwei Liu. Prevalence of simplex compression in adversarial deep neural networks. In *Proceedings of the National Academy of Sciences*, 2025.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Seungju Cho, Hongsin Lee, and Changick Kim. Long-tailed adversarial training with self-distillation. In *International Conference on Learning Representations*, 2025.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 2019.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

Guanlin Li, Guowen Xu, and Tianwei Zhang. Alleviating the effect of data imbalance on adversarial training. *arXiv preprint arXiv:2307.10205*, 2023.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019.

Rafael Muller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 2019.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

R. K. Pathria and Paul D. Beale. *Statistical Mechanics*. Academic Press, 3 edition, 2011.

Gabriel Peyre and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, 2020.

Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, 2020.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, 2018.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6976–6987, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pp. 1195–1204, 2017.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Cedric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

Yu-Hang Wang, Junkang Guo, Aolei Liu, Kaihao Wang, Zaitong Wu, Zhenyu Liu, Wenfei Yin, and Jian Liu. Taet: Two-stage adversarial equalization training on long-tailed distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, 2020.

Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Guanmeng Xian, Ning Yang, and Philip S. Yu. Manifold-constrained adversarial training for long-tailed robustness via geometric alignment. *arXiv preprint arXiv:2605.02183*, 2026.

Xinli Yue, Ningping Mou, Qian Wang, and Lingchen Zhao. Revisiting adversarial training under long-tailed distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Lilin Zhang, Yimo Guo, Yue Li, Jiancheng Shi, and Xianggen Liu. Taming the long tail: Rebalancing adversarial training via adaptive perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Lilin Zhang, Yue Li, Jiancheng Shi, Jiancheng Lv, and Xianggen Liu. Posterior mismatch matters: Adversarial training for long-tailed robustness. In *Proceedings of the International Conference on Machine Learning*, 2026b.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

A Related Work

Adversarial examples, adversarial training, and robust evaluation. Early work showed that imperceptible input perturbations can reliably change deep-network predictions (Szegedy et al., 2014) and connected this vulnerability to the locally linear behavior of high-dimensional models (Goodfellow et al., 2015). Robust optimization and PGD adversarial training then became a standard empirical defense (Madry et al., 2018). Subsequent objectives refine this basic formulation: TRADES decomposes natural and boundary errors (Zhang et al., 2019), MART emphasizes the role of misclassified examples (Wang et al., 2020), and AWP regularizes robust training by adversarially perturbing model weights (Wu et al., 2020). Other studies clarify why robust learning is statistically and optimization-wise difficult, including the increased sample complexity of robust generalization (Schmidt et al., 2018), the accuracy–robustness tension (Tsipras et al., 2019), and robust overfitting (Rice et al., 2020). Reliable robustness claims also require careful attack evaluation: C&W attacks provided an early strong optimization-based benchmark (Carlini & Wagner, 2017), gradient masking was systematically exposed by Athalye et al. (2018), and AutoAttack provides a parameter-free ensemble for robust evaluation (Croce & Hein, 2020). FEPT is orthogonal to this line: it does not propose a new attack or threat model, but improves the reliability of posterior supervision used inside adversarial training under long-tailed data.

Long-tailed visual recognition. Long-tailed recognition addresses the mismatch between highly imbalanced training data and more balanced evaluation requirements. Representative approaches include effective-number reweighting (Cui et al., 2019), label-distribution-aware margins and deferred reweighting (Cao et al., 2019), dynamic meta-embedding for open long-tailed recognition (Liu et al., 2019), decoupled representation and classifier learning (Kang et al., 2020), bilateral-branch/cumulative learning (Zhou et al., 2020), balanced softmax (Ren et al., 2020), logit adjustment (Menon et al., 2021), and multi-expert routing for bias–variance control (Wang et al., 2021). These methods mainly target natural accuracy and balanced recognition, whereas long-tailed adversarial training must additionally control how adversarial neighborhoods are shaped around minority classes. FEPT borrows the long-tailed principle that tail estimates are high-variance, but applies it to posterior-support reliability rather than only to class-frequency reweighting or classifier adjustment.

Long-tailed adversarial robustness. Long-tailed adversarial robustness combines class imbalance with robust optimization. RoBal first systematically studied adversarial robustness under long-tailed distributions and introduced margin-aware rebalancing (Wu et al., 2021). Later methods alleviate robust tail collapse through data-imbalance correction and reweighting (Li et al., 2023), balanced or staged adversarial training (Yue et al., 2024; Wang et al., 2025), self-distillation (Cho et al., 2025), and adaptive perturbation or plug-and-play robust rebalancing (Zhang et al., 2026a). The closest posterior-supervision baseline, PAT, argues that posterior mismatch is a central cause of long-tailed robust degradation (Zhang et al., 2026b). MCAT instead introduces an explicit geometric constraint to keep adversarial examples near class feature manifolds and aligns the classifier with ETF-inspired geometry (Xian et al., 2026). FEPT differs from these methods in its object of correction: rather than only reweighting classes, matching point-wise posteriors, or constraining feature drift, it asks whether the posterior mass used as supervision is locally supported, dense, and stable along the adversarial neighborhood.

Posterior targets, calibration, and consistency regularization. Soft supervision is widely used in modern training. Knowledge distillation transfers teacher information through soft targets (Hinton et al., 2015), label smoothing regularizes over-confident targets (Szegedy et al., 2016), and subsequent analysis shows that label smoothing improves calibration but can also remove class-similarity information useful for distillation (Muller et al., 2019). Calibration studies further show that modern neural networks can be poorly calibrated and that temperature scaling is an effective post-hoc correction (Guo et al., 2017). EMA teachers have also been used to stabilize consistency targets (Tarvainen & Valpola, 2017), while virtual adversarial training promotes local distributional smoothness of the conditional output distribution (Miyato et al., 2019). FEPT uses an EMA posterior surrogate and temperature/prior correction, but it does not directly trust the resulting posterior value. It first converts the posterior into a free-energy weighted support signal and then normalizes this signal over a localized candidate set, so the target is both posterior-aware and support-aware.

Manifold geometry, local support, and robust features. The local-geometry component of FEPT is related to the broader manifold-learning view that high-dimensional data often admit lower-dimensional local structure, as studied by Isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis & Saul, 2000), and Laplacian eigenmaps (Belkin & Niyogi, 2001). In adversarial robustness, manifold-based analyses distinguish between on-manifold and off-manifold adversarial examples (Stutz et al., 2019), and feature-based analyses argue that adversarial vulnerability can arise from predictive but brittle non-robust features (Ilyas et al., 2019). The Riemannian and information-geometric perspective is also connected to natural gradient and Fisher geometry (Amari, 1998), while neural collapse and simplex ETF geometry characterize late-stage class-feature structures (Papyan et al., 2020; Kothapalli, 2022); recent work further observes simplex compression in adversarially robust networks (Cao et al., 2025). FEPT does not assume access to the true data manifold. It uses a lightweight class-conditional local metric, estimated from feature memory banks, as a model-relative probe of semantic support and posterior stability.

Transport, source alignment, and support borrowing. Transport-based views provide a mathematical language for moving mass between distributions. Classical optimal transport formalizes distributional matching (Villani, 2009), entropic regularization makes OT computationally tractable via Sinkhorn distances (Cuturi, 2013), and computational OT has become a standard tool for scalable distribution comparison and transfer (Peyre & Cuturi, 2019). In long-tailed learning, several methods implicitly borrow information across classes through shared representations, classifier adjustment, meta-embedding, or expert routing (Liu et al., 2019; Kang et al., 2020; Wang et al., 2021). FEPT uses a simpler and more targeted form of transport: it does not solve a global OT problem, but performs prototype-aligned head/medium-to-tail posterior-support transfer. The transport coefficients and tail-adaptive mixing strength explicitly control the bias–variance trade-off introduced by borrowing source support for scarce tail classes.

B Training Algorithm

Algorithm 1 summarizes the comprehensive training pipeline of FEPT. The posterior surrogate q_ψ is instantiated by an EMA teacher and treated as a stopped-gradient signal during optimization. We utilize K_{nn} for local semantic metric estimation and K_{src} for dynamically selecting the low-variance donor classes in head-to-tail transport.

C Proofs

C.1 Proof of Theorem 1

By assumption, m_q^y is L_y-Lipschitz in $\mathcal{B}_{G_y}(u,r)$ with respect to d_{G_y}:

$$|m_q^y(u) - m_q^y(v)| \le L_y d_{G_y}(u,v), \qquad \forall v \in \mathcal{B}_{G_y}(u,r).$$

Therefore,

$$m_q^y(v) \ge m_q^y(u) - L_y d_{G_y}(u,v) \ge m_q^y(u) - L_y r.$$

Let $m_p^y(v) = p(y \mid v) - \max_{k\neq y} p(k \mid v)$ be the true posterior margin. Since $\|q_\psi(\cdot \mid v) - p(\cdot \mid v)\|_\infty \le \epsilon_p$, the true-class posterior may decrease by at most ϵ_p, while the largest competing posterior may increase by at most ϵ_p. Hence

$$m_p^y(v) \ge m_q^y(v) - 2\epsilon_p \ge m_q^y(u) - L_y r - 2\epsilon_p.$$

Thus, if $m_q^y(u) > L_y r + 2\epsilon_p$, then $m_p^y(v) > 0$ for all $v \in \mathcal{B}_{G_y}(u,r)$, so class y remains posterior-preferred. Conversely, if $m_q^y(u) \le L_y r$, the Lipschitz lower bound can become non-positive inside the ball, so point-wise posterior accuracy at u alone does not guarantee local preference preservation. Since $L_y = L_0(n_{\max}/n_y)^\chi$, this stability threshold is larger for tail classes.

C.2 Proof of Theorem 2

Let $\tilde{\pi}_{h\to t} = T_{h\to t}\hat{\pi}_h$ be the transported source estimate. Decompose the tail and transported source estimators as $\hat{\pi}_t = \pi_t^\star + \xi_t, \tilde{\pi}_{h\to t} = \pi_t^\star + b_h + \xi_h$, where $\mathbb{E}[\xi_t] = 0$, $\mathbb{E}[\xi_h] = 0$, $\mathbb{E}\|\xi_t\|^2 \le \sigma_t^2/n_t$,

Algorithm 1 FEPT training

Require: Dataset $\mathcal{D}$, classifier f_θ, EMA posterior surrogate q_ψ, perturbation budget ϵ, PGD steps T, PGD step size η, free-energy weight λ_E, posterior target weight β, candidate set size M_{cand}, local neighbor number k, source-class number K_{src}, transport strength α_0.

for each training iteration **do**
  Sample a mini-batch $\{(x_i, y_i)\}_{i=1}^{B}$ and compute representations $u_i = \phi_\theta(x_i)$.
  Update class-wise feature memory banks $\{\mathcal{B}_a\}_{a=1}^{C}$.
  Estimate local semantic metrics $G_a(\cdot)$ using $\min(k, n_a)$ nearest neighbors in each class memory bank $\mathcal{B}_a$.
  Compute class prototypes $c_a = |\mathcal{B}_a|^{-1} \sum_{u \in \mathcal{B}_a} u$ for all classes $a \in \{1, \dots, C\}$.
  for each sample i in the mini-batch **do**
    Construct the localized candidate class set $\mathcal{A}_{y_i}(u_i) = \{y_i\} \cup \mathrm{Top}_{M_{\text{cand}}}(q_\psi(\cdot \mid u_i)) \cup \mathcal{H}_{y_i} \cup \left(\cup_{t \in \mathrm{Top}_{M_{\text{cand}}}(q_\psi(\cdot|u_i)) \cap \mathcal{T}} \mathcal{H}_t\right)$.
    for each class $a \in \mathcal{A}_{y_i}(u_i)$ **do**
      Compute the class-conditional free energy $E_a(u_i)$ via Equation 4.
    end for
    Compute free-energy posterior support $\pi_a(u_i)$ via Equation 6 for all $a \in \mathcal{A}_{y_i}(u_i)$.
    for each tail class $t \in \mathcal{A}_{y_i}(u_i)$ **do**
      Evaluate source support strengths $\pi_h(\mathcal{T}_{t \to h}(u_i))$ for all $h \in \mathcal{H}_t$ via Equation 6.
      Construct the transported posterior support $\bar{\pi}_t(u_i)$ via Equation 7.
    end for
    For each non-tail class $a \in \mathcal{A}_{y_i}(u_i)$, set $\bar{\pi}_a(u_i) = \pi_a(u_i)$.
    Normalize $\{\bar{\pi}_a(u_i) : a \in \mathcal{A}_{y_i}(u_i)\}$ into the localized soft target $\bar{q}(\cdot \mid u_i)$, setting $\bar{q}(k \mid u_i) = 0$ for $k \notin \mathcal{A}_{y_i}(u_i)$.
  end for
  Generate adversarial examples by solving $\max_{\|\delta\|_\infty \le \epsilon} [\ell(f_\theta(x_i + \delta), y_i) - \lambda_E E_{y_i}(\phi_\theta(x_i + \delta))]$ with T steps of PGD.
  Update θ by minimizing $\frac{1}{B} \sum_{i=1}^{B} \left[\ell(f_\theta(x_i^{\text{adv}}), y_i) + \beta \,\mathrm{KL}\left(\mathrm{sg}[\bar{q}(\cdot \mid u_i)] \,\|\, p_\theta(\cdot \mid x_i^{\text{adv}})\right)\right]$.
  Update the EMA teacher that instantiates q_ψ.
end for
return Robust classifier f_θ.

$\mathbb{E}\|\xi_h\|^2 \le \sigma_h^2/n_h$, and $\|b_h\| \le b_{ht}$. Then $\hat{\pi}_t^{FEPT} - \pi_t^\star = (1-\alpha_t)\xi_t + \alpha_t \sum_{h \in \mathcal{H}_t} \omega_{ht}\xi_h + \alpha_t \sum_{h \in \mathcal{H}_t} \omega_{ht} b_h$. Taking squared norm and expectation, the cross terms between zero-mean noise and deterministic bias vanish. We obtain

$$\mathbb{E}\|\hat{\pi}_t^{FEPT} - \pi_t^\star\|^2 = (1-\alpha_t)^2 \mathbb{E}\|\xi_t\|^2 + \alpha_t^2 \mathbb{E}\left\|\sum_{h \in \mathcal{H}_t} \omega_{ht}\xi_h\right\|^2 + 2\alpha_t(1-\alpha_t) \sum_{h \in \mathcal{H}_t} \omega_{ht} \mathbb{E}\langle \xi_t, \xi_h \rangle + \alpha_t^2 \left\|\sum_{h \in \mathcal{H}_t} \omega_{ht} b_h\right\|^2.$$

The transported source variance expands as $\mathbb{E}\left\|\sum_{h \in \mathcal{H}_t} \omega_{ht}\xi_h\right\|^2 = \sum_{h \in \mathcal{H}_t} \omega_{ht}^2 \mathbb{E}\|\xi_h\|^2 + \sum_{h \neq h'} \omega_{ht}\omega_{h't} \mathbb{E}\langle \xi_h, \xi_{h'} \rangle$. Using the variance bounds and the correlation assumption, $|\mathbb{E}\langle \xi_a, \xi_b \rangle| \le \rho \frac{\sigma_a \sigma_b}{\sqrt{n_a n_b}}, a \neq b$, we get

$$\mathbb{E}\|\hat{\pi}_t^{FEPT} - \pi_t^\star\|^2 \le (1-\alpha_t)^2 \frac{\sigma_t^2}{n_t} + \alpha_t^2 \sum_{h \in \mathcal{H}_t} \omega_{ht}^2 \frac{\sigma_h^2}{n_h} + 2\rho\alpha_t(1-\alpha_t) \sum_{h \in \mathcal{H}_t} \omega_{ht} \frac{\sigma_t \sigma_h}{\sqrt{n_t n_h}} + \rho\alpha_t^2 \sum_{h \neq h'} \omega_{ht}\omega_{h't} \frac{\sigma_h \sigma_{h'}}{\sqrt{n_h n_{h'}}} + \alpha_t^2 \left(\sum_{h \in \mathcal{H}_t} \omega_{ht} b_{ht}\right)^2.$$

D Additional Experimental Details

D.1 Long-tailed data generation

For a dataset with C classes and imbalance ratio IR, classes are sorted from head to tail and the number of training samples in class c is set by the exponential rule

$$N_c = N_{\max} \cdot \mathrm{IR}^{-\frac{c-1}{C-1}}, \qquad c = 1, \ldots, C.$$

CIFAR10-LT and CIFAR100-LT are generated from the original CIFAR10 and CIFAR100 training sets. TinyImageNet-LT applies the same exponential construction to its training split. Test sets remain unchanged. Unless otherwise specified, head, medium, and tail groups are defined by sorting classes according to training-set frequency and splitting them into three equal groups. Tail accuracy in the main tables follows the common long-tailed adversarial training protocol and averages over the 80% of classes with the fewest training samples. The same splits are used across all methods.

D.2 Optimization details

All methods are trained with the ℓ_∞ threat model under $\epsilon = 8/255$. The default training-time inner maximization uses 10-step PGD with step size $2/255$. Evaluation uses PGD-20, PGD-100, CW-20, AutoAttack, and the adaptive attacks described in Section 5. All methods use the same data augmentation: random crop, horizontal flip, and normalization.

For WRN-28-10 and ResNet-18, we use SGD with Nesterov momentum 0.9, weight decay 5×10^{-4}, batch size 128, and initial learning rate 0.1. Unless the original implementation of a baseline specifies otherwise, models are trained for 200 epochs with learning-rate drops at epochs 100 and 150. When a baseline has an official training recipe, we keep its schedule and adversarial training loss unchanged and only replace the method-specific components needed for fair comparison. All reported numbers are averaged over five independent runs and shown as $\mathrm{mean}_{\pm\mathrm{std}}$.

D.3 FEPT details

The memory bank stores normalized penultimate-layer features and is updated after each mini-batch. Class prototypes are updated with an exponential moving average. The local tangent projection is estimated from the K_{nn} nearest same-class memory features. We set the tangent dimension by retaining 90% of local variance, with a minimum of 4 and a maximum of 32 dimensions. The class-wise metric is refreshed every five epochs to reduce computational overhead. For each tail class, K_{src} source classes are selected from head and medium classes by prototype distance. Source weights are computed by the softmax kernel in Section 3.

Unless otherwise stated, we use the default hyperparameters in Table 4 for all main-table FEPT results. The values in the hyperparameter figures are sweep values and should not be read as additional defaults. Here K_{nn} controls local PCA, K_{src} controls the number of source classes used for transport, M_{cand} controls the number of high-posterior candidate classes used in candidate-restricted support evaluation, η_0 controls the base normal penalty, α_0 controls the maximum transport strength, λ_E controls the free-energy term in the inner maximization, γ weights posterior sensitivity, τ weights local density, and β weights the transported posterior KL term.

Table 4: Default hyperparameters for FEPT.

Dataset	K_{nn}	K_{src}	M_{cand}	η_0	α_0	λ_E	γ	τ	β
CIFAR10-LT	20	5	20	2.0	0.12	0.05	0.20	0.60	0.18
CIFAR100-LT	20	8	30	2.2	0.15	0.05	0.25	0.60	0.20
TinyImageNet-LT	30	8	30	2.5	0.18	0.05	0.25	0.70	0.20

D.4 Compute budget

Table 5 reports the per-seed compute budget for FEPT with WRN-28-10. The overhead mainly comes from memory-bank queries, local PCA updates, and source-support transport. Since the local metric is refreshed only periodically, the additional cost remains moderate relative to the adversarial training loop.

Table 5: Compute budget per seed for FEPT with WRN-28-10.

Dataset	GPU	Epochs	Wall-clock	Peak memory
CIFAR10-LT	A100 40GB	200	7.84 $h_{\pm 0.09}$	10.28 $GB_{\pm 0.12}$
CIFAR100-LT	A100 40GB	200	8.63 $h_{\pm 0.11}$	10.91 $GB_{\pm 0.14}$
TinyImageNet-LT	A100 40GB	200	22.46 $h_{\pm 0.21}$	19.34 $GB_{\pm 0.20}$

Table 6: Main results with ResNet-18 under PGD-20 attack. The best result is bolded and the second best is underlined.

Dataset	Method	Nat. all	Nat. tail	PGD all	PGD tail	AA all	AA tail	Worst-AA	Gap
CIFAR10-LT	AT	$55.94_{\pm 0.46}$	$45.86_{\pm 0.61}$	$25.91_{\pm 0.20}$	$12.84_{\pm 0.24}$	$24.76_{\pm 0.19}$	$11.62_{\pm 0.25}$	$6.21_{\pm 0.21}$	$22.84_{\pm 0.40}$
	AWP	$57.12_{\pm 0.28}$	$47.38_{\pm 0.34}$	$26.84_{\pm 0.30}$	$13.82_{\pm 0.39}$	$25.93_{\pm 0.29}$	$12.94_{\pm 0.35}$	$7.15_{\pm 0.27}$	$22.31_{\pm 0.43}$
	RoBal	$69.86_{\pm 0.21}$	$64.13_{\pm 0.25}$	$30.64_{\pm 0.46}$	$20.68_{\pm 0.40}$	$27.36_{\pm 0.37}$	$17.18_{\pm 0.36}$	$11.43_{\pm 0.29}$	$18.93_{\pm 0.42}$
	REAT	$64.48_{\pm 0.27}$	$57.63_{\pm 0.31}$	$27.12_{\pm 0.41}$	$16.08_{\pm 0.44}$	$25.68_{\pm 0.43}$	$14.51_{\pm 0.41}$	$9.12_{\pm 0.35}$	$20.72_{\pm 0.48}$
	AT-BSL	$72.86_{\pm 0.59}$	$68.49_{\pm 0.59}$	$38.38_{\pm 0.14}$	$31.50_{\pm 0.28}$	$34.03_{\pm 0.23}$	$27.88_{\pm 0.32}$	$18.02_{\pm 0.39}$	$15.10_{\pm 0.47}$
	TAET	$62.18_{\pm 1.71}$	$57.19_{\pm 1.56}$	$29.82_{\pm 0.53}$	$22.96_{\pm 0.58}$	$26.41_{\pm 0.76}$	$19.54_{\pm 0.45}$	$13.05_{\pm 0.42}$	$14.23_{\pm 0.49}$
	RobustLT	$73.39_{\pm 0.48}$	$69.08_{\pm 0.23}$	$38.86_{\pm 0.20}$	$32.49_{\pm 0.59}$	$34.72_{\pm 0.28}$	$29.14_{\pm 0.37}$	$21.41_{\pm 0.35}$	$11.38_{\pm 0.36}$
	MCAT	$72.91_{\pm 0.34}$	$68.72_{\pm 0.37}$	$39.21_{\pm 0.30}$	$33.58_{\pm 0.45}$	$34.88_{\pm 0.27}$	$28.61_{\pm 0.36}$	$20.88_{\pm 0.31}$	$10.76_{\pm 0.32}$
	PAT	$73.41_{\pm 0.55}$	$69.31_{\pm 0.58}$	$36.71_{\pm 0.18}$	$30.31_{\pm 0.11}$	$34.82_{\pm 0.26}$	$28.96_{\pm 0.33}$	$20.81_{\pm 0.44}$	$11.18_{\pm 0.35}$
	FEPT-BSL	$73.86_{\pm 0.32}$	$70.42_{\pm 0.35}$	$\mathbf{41.24}_{\pm 0.29}$	$35.92_{\pm 0.42}$	$37.28_{\pm 0.26}$	$32.10_{\pm 0.32}$	$24.12_{\pm 0.28}$	$9.62_{\pm 0.31}$
	FEPT-PAT	$\mathbf{74.22}_{\pm 0.31}$	$\mathbf{71.05}_{\pm 0.34}$	$40.92_{\pm 0.28}$	$36.12_{\pm 0.40}$	$\mathbf{37.66}_{\pm 0.25}$	$32.38_{\pm 0.31}$	$24.36_{\pm 0.27}$	$9.42_{\pm 0.30}$
	FEPT-EMA	$\underline{74.05_{\pm 0.30}}$	$\underline{70.82_{\pm 0.33}}$	$\underline{41.08_{\pm 0.27}}$	$\mathbf{36.41}_{\pm 0.39}$	$\underline{37.58_{\pm 0.24}}$	$\mathbf{32.74}_{\pm 0.30}$	$\mathbf{24.67}_{\pm 0.26}$	$\mathbf{9.18}_{\pm 0.28}$
CIFAR100-LT	AT	$41.82_{\pm 0.24}$	$37.46_{\pm 0.28}$	$14.28_{\pm 0.17}$	$12.36_{\pm 0.20}$	$14.10_{\pm 0.16}$	$12.18_{\pm 0.24}$	$6.86_{\pm 0.20}$	$5.16_{\pm 0.24}$
	AWP	$42.71_{\pm 0.39}$	$38.12_{\pm 0.35}$	$15.26_{\pm 0.12}$	$13.18_{\pm 0.26}$	$15.02_{\pm 0.10}$	$13.06_{\pm 0.25}$	$7.68_{\pm 0.21}$	$5.32_{\pm 0.26}$
	RoBal	$46.88_{\pm 0.18}$	$44.20_{\pm 0.16}$	$15.54_{\pm 0.09}$	$14.42_{\pm 0.12}$	$15.31_{\pm 0.11}$	$14.19_{\pm 0.14}$	$8.94_{\pm 0.13}$	$3.24_{\pm 0.16}$
	REAT	$44.56_{\pm 0.20}$	$42.58_{\pm 0.12}$	$14.42_{\pm 0.16}$	$13.53_{\pm 0.20}$	$14.22_{\pm 0.16}$	$13.38_{\pm 0.21}$	$8.43_{\pm 0.17}$	$3.06_{\pm 0.19}$
	AT-BSL	$52.64_{\pm 0.13}$	$50.42_{\pm 0.15}$	$18.82_{\pm 0.17}$	$17.61_{\pm 0.22}$	$18.22_{\pm 0.18}$	$17.12_{\pm 0.21}$	$11.61_{\pm 0.20}$	$2.76_{\pm 0.19}$
	TAET	$43.72_{\pm 0.20}$	$39.48_{\pm 0.15}$	$13.86_{\pm 0.18}$	$12.24_{\pm 0.23}$	$13.70_{\pm 0.17}$	$12.10_{\pm 0.22}$	$7.82_{\pm 0.18}$	$4.12_{\pm 0.22}$
	RobustLT	$53.11_{\pm 0.21}$	$51.36_{\pm 0.20}$	$22.28_{\pm 0.21}$	$21.84_{\pm 0.33}$	$18.88_{\pm 0.23}$	$18.04_{\pm 0.29}$	$13.85_{\pm 0.24}$	$2.18_{\pm 0.20}$
	MCAT	$53.62_{\pm 0.20}$	$51.94_{\pm 0.18}$	$23.18_{\pm 0.19}$	$22.82_{\pm 0.26}$	$19.96_{\pm 0.19}$	$19.32_{\pm 0.24}$	$15.02_{\pm 0.21}$	$1.58_{\pm 0.16}$
	PAT	$48.03_{\pm 0.19}$	$45.97_{\pm 0.12}$	$22.11_{\pm 0.19}$	$21.28_{\pm 0.20}$	$19.42_{\pm 0.21}$	$18.40_{\pm 0.25}$	$5.63_{\pm 0.12}$	$2.05_{\pm 0.17}$
	FEPT-BSL	$\mathbf{54.40}_{\pm 0.20}$	$52.45_{\pm 0.19}$	$\mathbf{24.36}_{\pm 0.19}$	$23.72_{\pm 0.25}$	$20.70_{\pm 0.19}$	$20.18_{\pm 0.23}$	$15.48_{\pm 0.20}$	$1.42_{\pm 0.16}$
	FEPT-PAT	$54.10_{\pm 0.19}$	$\mathbf{52.78}_{\pm 0.18}$	$24.10_{\pm 0.18}$	$23.82_{\pm 0.24}$	$\mathbf{20.96}_{\pm 0.18}$	$20.32_{\pm 0.22}$	$15.66_{\pm 0.20}$	$1.36_{\pm 0.15}$
	FEPT-EMA	$\underline{54.28_{\pm 0.18}}$	$\underline{52.64_{\pm 0.17}}$	$\underline{24.22_{\pm 0.17}}$	$\mathbf{23.98}_{\pm 0.23}$	$\underline{20.88_{\pm 0.18}}$	$\mathbf{20.54}_{\pm 0.21}$	$\mathbf{15.86}_{\pm 0.19}$	$\mathbf{1.29}_{\pm 0.14}$
TinyImageNet-LT	AT	$41.26_{\pm 0.42}$	$35.28_{\pm 0.31}$	$18.14_{\pm 0.45}$	$13.10_{\pm 0.59}$	$16.24_{\pm 0.35}$	$11.41_{\pm 0.30}$	$6.42_{\pm 0.24}$	$8.26_{\pm 0.38}$
	AWP	$37.84_{\pm 1.15}$	$31.62_{\pm 1.30}$	$18.96_{\pm 0.95}$	$13.92_{\pm 0.82}$	$17.42_{\pm 0.84}$	$12.54_{\pm 0.77}$	$6.98_{\pm 0.60}$	$7.84_{\pm 0.78}$
	RoBal	$47.18_{\pm 0.37}$	$45.29_{\pm 0.59}$	$20.38_{\pm 0.31}$	$17.35_{\pm 0.39}$	$17.42_{\pm 0.38}$	$14.42_{\pm 0.34}$	$9.64_{\pm 0.32}$	$5.08_{\pm 0.33}$
	REAT	$41.93_{\pm 0.82}$	$39.36_{\pm 0.35}$	$20.11_{\pm 0.76}$	$18.08_{\pm 0.97}$	$16.58_{\pm 0.42}$	$14.13_{\pm 0.63}$	$9.26_{\pm 0.45}$	$4.37_{\pm 0.39}$
	AT-BSL	$38.24_{\pm 0.92}$	$35.22_{\pm 1.05}$	$23.84_{\pm 0.41}$	$21.31_{\pm 0.45}$	$20.06_{\pm 0.56}$	$17.52_{\pm 0.64}$	$12.92_{\pm 0.50}$	$4.56_{\pm 0.46}$
	TAET	$29.46_{\pm 1.16}$	$26.45_{\pm 1.21}$	$16.28_{\pm 0.46}$	$14.28_{\pm 0.71}$	$12.04_{\pm 0.72}$	$10.18_{\pm 0.91}$	$5.78_{\pm 0.60}$	$3.82_{\pm 0.56}$
	RobustLT	$42.83_{\pm 1.16}$	$39.48_{\pm 1.26}$	$25.64_{\pm 0.35}$	$23.41_{\pm 0.28}$	$20.94_{\pm 0.26}$	$18.74_{\pm 0.18}$	$13.82_{\pm 0.18}$	$4.16_{\pm 0.19}$
	MCAT	$47.82_{\pm 0.46}$	$45.68_{\pm 0.48}$	$26.35_{\pm 0.31}$	$24.12_{\pm 0.27}$	$21.94_{\pm 0.24}$	$20.10_{\pm 0.21}$	$15.12_{\pm 0.20}$	$3.28_{\pm 0.20}$
	PAT	$41.16_{\pm 0.16}$	$38.75_{\pm 0.27}$	$17.03_{\pm 0.27}$	$16.06_{\pm 0.33}$	$14.92_{\pm 0.25}$	$13.84_{\pm 0.22}$	$5.96_{\pm 0.32}$	$3.48_{\pm 0.25}$
	FEPT-BSL	$48.42_{\pm 0.41}$	$46.32_{\pm 0.38}$	$\mathbf{27.58}_{\pm 0.30}$	$25.32_{\pm 0.26}$	$\mathbf{22.96}_{\pm 0.24}$	$20.98_{\pm 0.21}$	$15.74_{\pm 0.20}$	$3.12_{\pm 0.21}$
	FEPT-PAT	$\mathbf{48.86}_{\pm 0.40}$	$\mathbf{46.76}_{\pm 0.37}$	$27.26_{\pm 0.29}$	$25.48_{\pm 0.25}$	$22.68_{\pm 0.23}$	$21.12_{\pm 0.20}$	$15.92_{\pm 0.19}$	$3.02_{\pm 0.20}$
	FEPT-EMA	$\underline{48.66_{\pm 0.39}}$	$\underline{46.58_{\pm 0.35}}$	$\underline{27.42_{\pm 0.27}}$	$\mathbf{25.61}_{\pm 0.23}$	$\underline{22.82_{\pm 0.22}}$	$\mathbf{21.34}_{\pm 0.19}$	$\mathbf{16.08}_{\pm 0.17}$	$\mathbf{2.91}_{\pm 0.18}$

E Additional Results

E.1 Results with ResNet-18

Table 6 reports the corresponding main results with ResNet-18. Compared with WRN-28-10, all methods obtain lower absolute robustness, but the relative ranking is consistent: FEPT-EMA achieves the best tail and worst-class robustness across all three benchmarks, while FEPT-PAT and FEPT-BSL remain consistently competitive. This suggests that the gains are not tied to the capacity of WRN-28-10 or to a single posterior-surrogate instantiation.

Table 7 reports the same stronger-attack and adaptive evaluation with ResNet-18. The adaptive attacks reduce absolute robustness, but all FEPT variants remain stronger than PAT. The default FEPT gives the best tail robustness in most cases, while FEPT-BSL and FEPT-PAT are occasionally stronger on all-class metrics, matching the mixed posterior-source pattern observed with WRN-28-10.

E.2 TinyImageNet-LT Ablation

Table 8 reports the ablation results on TinyImageNet-LT. The conclusions are consistent with the CIFAR benchmarks: free-energy support mainly controls posterior sensitivity, while head-to-tail transport mainly reduces transport error and improves tail robustness.

Table 7: Robustness under stronger and adaptive attacks with ResNet-18. Each entry reports all-class accuracy / tail-class accuracy. Best and second-best all/tail numbers are bolded and underlined separately.

Dataset	Method	PGD-100	CW-20	AA	Adaptive-E	Adaptive-s
CIFAR10-LT	AT-BSL	38.05/31.18$_{\pm 0.31}$	37.44/30.53$_{\pm 0.34}$	34.03/27.88$_{\pm 0.23}$	33.10/27.05$_{\pm 0.27}$	32.94/26.82$_{\pm 0.28}$
	RobustLT	38.52/32.04$_{\pm 0.29}$	37.89/31.36$_{\pm 0.32}$	34.72/29.14$_{\pm 0.28}$	33.82/28.31$_{\pm 0.31}$	33.65/28.08$_{\pm 0.32}$
	MCAT	38.92/33.15$_{\pm 0.28}$	38.21/32.38$_{\pm 0.30}$	34.88/28.61$_{\pm 0.27}$	33.96/27.82$_{\pm 0.29}$	33.80/27.60$_{\pm 0.30}$
	PAT	38.62/32.72$_{\pm 0.27}$	37.80/31.90$_{\pm 0.29}$	34.82/28.96$_{\pm 0.26}$	33.54/27.92$_{\pm 0.28}$	33.36/27.70$_{\pm 0.29}$
	FEPT-BSL	**40.88**/35.40$_{\pm 0.28}$	**40.14**/34.58$_{\pm 0.29}$	37.28/32.10$_{\pm 0.26}$	36.42/31.22$_{\pm 0.27}$	36.24/31.02$_{\pm 0.28}$
	FEPT-PAT	40.60/35.70$_{\pm 0.27}$	39.90/34.86$_{\pm 0.28}$	**37.66**/32.38$_{\pm 0.25}$	36.60/31.54$_{\pm 0.26}$	36.38/31.32$_{\pm 0.27}$
	FEPT-EMA	40.74/**35.96**$_{\pm 0.25}$	40.03/**35.12**$_{\pm 0.26}$	37.58/**32.74**$_{\pm 0.24}$	**36.73/31.86**$_{\pm 0.25}$	**36.51/31.62**$_{\pm 0.26}$
CIFAR100-LT	AT-BSL	18.38/17.24$_{\pm 0.19}$	18.85/17.68$_{\pm 0.20}$	18.22/17.12$_{\pm 0.18}$	17.58/16.46$_{\pm 0.19}$	17.44/16.31$_{\pm 0.19}$
	RobustLT	22.07/21.60$_{\pm 0.22}$	21.48/20.76$_{\pm 0.23}$	18.88/18.04$_{\pm 0.23}$	18.31/17.52$_{\pm 0.24}$	18.22/17.43$_{\pm 0.24}$
	MCAT	22.91/22.46$_{\pm 0.20}$	22.14/21.52$_{\pm 0.21}$	19.96/19.32$_{\pm 0.19}$	19.38/18.78$_{\pm 0.20}$	19.27/18.66$_{\pm 0.20}$
	PAT	22.47/21.88$_{\pm 0.21}$	21.73/20.96$_{\pm 0.22}$	19.42/18.40$_{\pm 0.21}$	18.73/17.76$_{\pm 0.22}$	18.62/17.64$_{\pm 0.22}$
	FEPT-BSL	**24.02**/23.30$_{\pm 0.19}$	**23.36**/22.66$_{\pm 0.20}$	20.70/20.18$_{\pm 0.19}$	20.18/19.62$_{\pm 0.20}$	20.08/19.50$_{\pm 0.20}$
	FEPT-PAT	23.74/23.42$_{\pm 0.19}$	23.10/22.74$_{\pm 0.20}$	**20.96**/20.32$_{\pm 0.18}$	20.28/19.82$_{\pm 0.20}$	20.18/19.72$_{\pm 0.20}$
	FEPT-EMA	23.86/**23.51**$_{\pm 0.18}$	23.28/**22.86**$_{\pm 0.19}$	20.88/**20.54**$_{\pm 0.18}$	**20.36/20.02**$_{\pm 0.19}$	**20.24/19.88**$_{\pm 0.19}$
TinyImageNet-LT	AT-BSL	23.28/20.92$_{\pm 0.47}$	21.62/19.03$_{\pm 0.52}$	20.06/17.52$_{\pm 0.56}$	19.42/16.88$_{\pm 0.58}$	19.28/16.72$_{\pm 0.59}$
	RobustLT	25.22/23.06$_{\pm 0.32}$	22.64/20.12$_{\pm 0.35}$	20.94/18.74$_{\pm 0.26}$	20.26/18.08$_{\pm 0.29}$	20.12/17.91$_{\pm 0.30}$
	MCAT	25.94/23.76$_{\pm 0.29}$	23.48/20.96$_{\pm 0.32}$	21.94/20.10$_{\pm 0.24}$	21.18/19.36$_{\pm 0.27}$	21.02/19.20$_{\pm 0.28}$
	PAT	16.82/15.84$_{\pm 0.30}$	16.08/14.86$_{\pm 0.33}$	14.92/13.84$_{\pm 0.25}$	14.28/13.08$_{\pm 0.28}$	14.12/12.92$_{\pm 0.29}$
	FEPT-BSL	**27.24**/24.96$_{\pm 0.29}$	**24.92**/22.22$_{\pm 0.31}$	**22.96**/20.98$_{\pm 0.24}$	**22.30**/20.42$_{\pm 0.26}$	**22.12**/20.24$_{\pm 0.27}$
	FEPT-PAT	27.02/25.08$_{\pm 0.28}$	24.66/22.36$_{\pm 0.30}$	22.68/21.12$_{\pm 0.23}$	22.02/20.58$_{\pm 0.25}$	21.88/20.40$_{\pm 0.26}$
	FEPT-EMA	27.08/**25.22**$_{\pm 0.26}$	24.78/**22.58**$_{\pm 0.28}$	22.82/**21.34**$_{\pm 0.22}$	22.18/**20.72**$_{\pm 0.24}$	22.02/**20.54**$_{\pm 0.25}$

Table 8: Ablation study on TinyImageNet-LT with WRN-28-10. Mechanism metrics are averaged over tail classes.

Variant	AA all	AA tail	$\bar{s} \downarrow$	$\widehat{\mathcal{E}}_{\mathrm{tr}} \downarrow$
Full FEPT	**24.23**$_{\pm 0.24}$	**22.85**$_{\pm 0.19}$	**0.41**$_{\pm 0.11}$	**0.23**$_{\pm 0.13}$
w/o free-energy inner term	22.86$_{\pm 0.36}$	21.17$_{\pm 0.43}$	0.60$_{\pm 0.15}$	0.27$_{\pm 0.12}$
w/o support reweighting	22.31$_{\pm 0.29}$	20.72$_{\pm 0.51}$	0.66$_{\pm 0.18}$	0.29$_{\pm 0.16}$
w/o head-to-tail transport	21.84$_{\pm 0.44}$	20.01$_{\pm 0.38}$	0.47$_{\pm 0.13}$	0.45$_{\pm 0.19}$
fixed α_t	22.93$_{\pm 0.27}$	21.26$_{\pm 0.34}$	0.45$_{\pm 0.12}$	0.34$_{\pm 0.15}$
uniform source weights	23.04$_{\pm 0.32}$	21.43$_{\pm 0.46}$	0.44$_{\pm 0.16}$	0.36$_{\pm 0.18}$
w/o posterior KL	23.36$_{\pm 0.25}$	21.72$_{\pm 0.31}$	0.49$_{\pm 0.19}$	0.30$_{\pm 0.14}$

E.3 Source-class alignment diagnostics

Table 9 examines whether the source classes used for transport provide well-aligned support for tail classes. Randomly selected head/medium sources lead to a larger prototype distance and a higher transport error. Selecting nearest source classes reduces the prototype distance, while using prototype-based weights further lowers the transport error and improves tail robustness. These results support the use of the transport coefficients ω_{ht} in Equation 7, rather than uniform aggregation over candidate source classes.

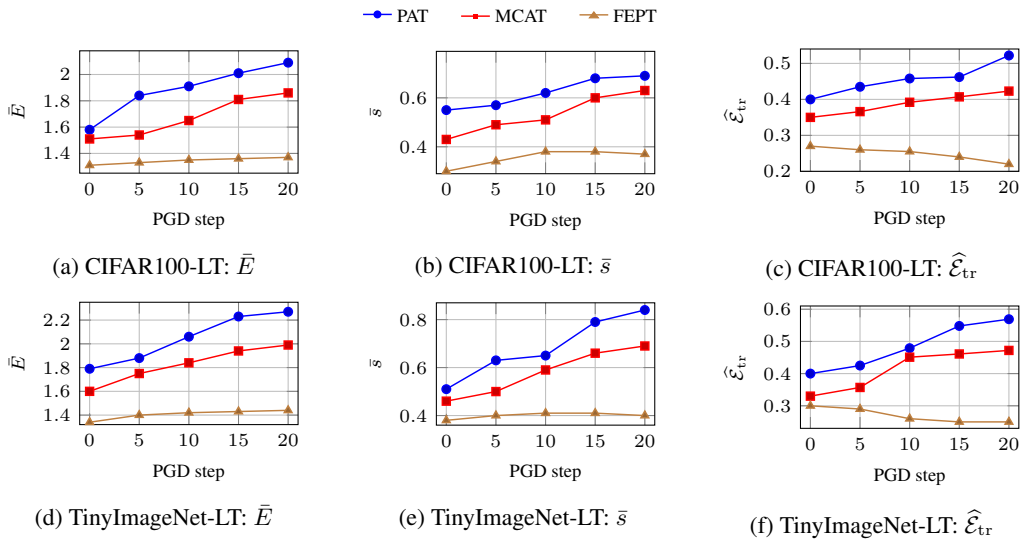
Table 9: Source-class alignment diagnostics on CIFAR100-LT with WRN-28-10.

Source selection	AA all	AA tail	Prototype distance $\downarrow$	Transport error $\downarrow$
Random head/medium sources	20.90$_{\pm 0.19}$	19.64$_{\pm 0.24}$	1.84$_{\pm 0.05}$	0.39$_{\pm 0.02}$
Nearest sources, uniform weights	21.34$_{\pm 0.17}$	20.63$_{\pm 0.21}$	1.12$_{\pm 0.03}$	0.33$_{\pm 0.02}$
Nearest sources, prototype weights	**22.24**$_{\pm 0.16}$	**21.88**$_{\pm 0.20}$	**1.12**$_{\pm 0.03}$	**0.21**$_{\pm 0.01}$

E.4 Supplementary posterior-support diagnostics

This appendix provides additional mechanism evidence complementing Figure 2. We first report class-group diagnostics on CIFAR10-LT, and then provide supplementary trajectory plots on CIFAR100-LT and TinyImageNet-LT. We also visualize the robust posterior support volume and the empirical bias–variance trade-off induced by transport strength.

Class-group diagnostics. Table 10 reports posterior-support quantities across head, medium, and tail classes. The main gain of FEPT appears on tail classes: it lowers free energy and posterior sensitivity while increasing robust posterior support volume. This agrees with the intended role of



(a) CIFAR100-LT: $\bar{E}$ (b) CIFAR100-LT: $\bar{s}$ (c) CIFAR100-LT: $\widehat{\mathcal{E}}_{\mathrm{tr}}$

(d) TinyImageNet-LT: $\bar{E}$ (e) TinyImageNet-LT: $\bar{s}$ (f) TinyImageNet-LT: $\widehat{\mathcal{E}}_{\mathrm{tr}}$

Figure 4: Supplementary posterior-support diagnostics on CIFAR100-LT and TinyImageNet-LT. FEPT consistently maintains lower free energy, lower posterior sensitivity, and lower empirical transport error along adversarial trajectories.

posterior transport, which should be most useful when the empirical tail support estimate is unreliable.

Table 10: Class-group diagnostics on CIFAR10-LT.

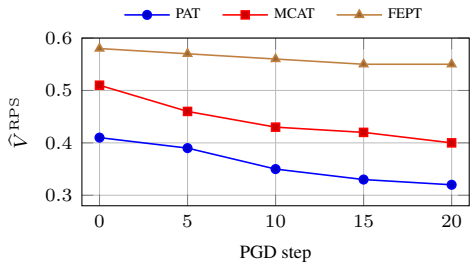
Group	Method	$\bar{E}\downarrow$	$\bar{s}\downarrow$	$\widehat{V}^{\mathrm{RPS}}\uparrow$	AA
Head	PAT	$1.20_{\pm 0.14}$	$0.31_{\pm 0.12}$	$0.66_{\pm 0.13}$	$41.79_{\pm 0.31}$
	MCAT	$1.16_{\pm 0.12}$	$0.30_{\pm 0.13}$	$0.68_{\pm 0.17}$	$40.14_{\pm 0.36}$
	FEPT	$\mathbf{1.10}_{\pm 0.10}$	$\mathbf{0.28}_{\pm 0.10}$	$\mathbf{0.71}_{\pm 0.12}$	$\mathbf{43.22}_{\pm 0.22}$
Medium	PAT	$1.48_{\pm 0.16}$	$0.42_{\pm 0.14}$	$0.52_{\pm 0.12}$	$34.62_{\pm 0.33}$
	MCAT	$1.40_{\pm 0.13}$	$0.39_{\pm 0.12}$	$0.55_{\pm 0.15}$	$34.94_{\pm 0.38}$
	FEPT	$\mathbf{1.23}_{\pm 0.11}$	$\mathbf{0.33}_{\pm 0.10}$	$\mathbf{0.62}_{\pm 0.12}$	$\mathbf{38.08}_{\pm 0.24}$
Tail	PAT	$1.88_{\pm 0.18}$	$0.68_{\pm 0.15}$	$0.30_{\pm 0.13}$	$31.58_{\pm 0.42}$
	MCAT	$1.64_{\pm 0.16}$	$0.56_{\pm 0.14}$	$0.39_{\pm 0.10}$	$30.26_{\pm 0.47}$
	FEPT	$\mathbf{1.26}_{\pm 0.12}$	$\mathbf{0.34}_{\pm 0.11}$	$\mathbf{0.55}_{\pm 0.14}$	$\mathbf{34.85}_{\pm 0.29}$

Supplementary mechanism curves on other datasets. Figure 4 reports the same posterior-support diagnostics on CIFAR100-LT and TinyImageNet-LT. The trends are consistent with the CIFAR10-LT results in Figure 2: FEPT keeps posterior preferences more stable along adversarial trajectories and produces smaller transport error for tail classes.

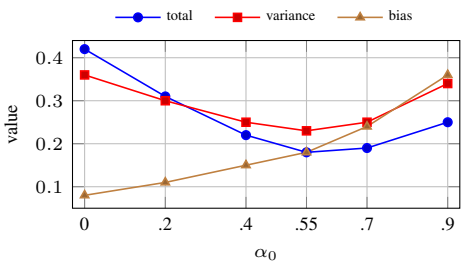
Supplementary mechanism plots. Figure 5 provides two complementary mechanism diagnostics. Figure 5a shows that FEPT preserves a larger robust posterior support volume along adversarial trajectories, indicating that the transported posterior target remains better supported in the local adversarial neighborhood. Figure 5b visualizes the empirical bias–variance trade-off induced by transport strength: moderate transport gives the lowest total error, whereas excessive transport increases the bias proxy.

E.5 Imbalance-ratio sweep

Tables 11 and 12 show that the gains of FEPT persist under both moderate and severe imbalance. These sweeps complement the main-table settings and directly test the support-scarcity premise. The relative improvement is larger when the imbalance becomes more severe, which is expected because transport is most useful when the tail-only support estimate has high variance.



(a) Robust posterior support volume

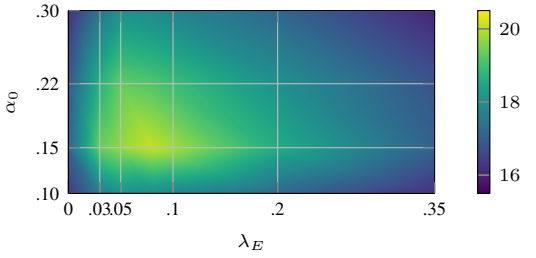


(b) Transport bias–variance trade-off

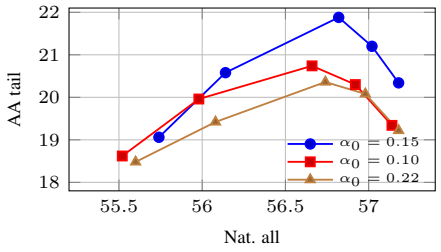
Figure 5: Supplementary mechanism plots on CIFAR10-LT. (a) FEPT preserves a larger robust posterior support volume along adversarial trajectories. (b) The empirical transport error is minimized at an intermediate transport strength, reflecting the bias–variance trade-off of posterior transport.

Table 11: CIFAR10-LT with different imbalance ratios using WRN-28-10.

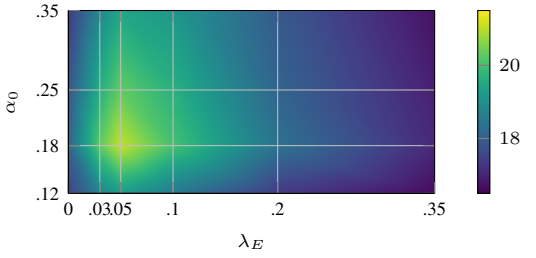
Method	IR = 10				IR = 100			
	Nat. all	Nat. tail	AA all	AA tail	Nat. all	Nat. tail	AA all	AA tail
AT-BSL	$83.55_{\pm 0.19}$	$80.70_{\pm 0.30}$	$43.82_{\pm 0.21}$	$37.24_{\pm 0.32}$	$75.21_{\pm 0.48}$	$70.80_{\pm 0.73}$	$32.85_{\pm 0.24}$	$24.60_{\pm 0.36}$
RobustLT	$83.74_{\pm 0.08}$	$81.36_{\pm 0.12}$	$46.10_{\pm 0.30}$	$41.76_{\pm 0.55}$	$73.72_{\pm 0.64}$	$69.08_{\pm 0.92}$	$36.92_{\pm 0.13}$	$30.36_{\pm 0.47}$
MCAT	$83.92_{\pm 0.10}$	$81.70_{\pm 0.14}$	$46.48_{\pm 0.24}$	$42.32_{\pm 0.41}$	$74.34_{\pm 0.51}$	$69.85_{\pm 0.72}$	$37.26_{\pm 0.15}$	$31.04_{\pm 0.39}$
PAT	$83.80_{\pm 0.11}$	$81.44_{\pm 0.15}$	$46.35_{\pm 0.23}$	$42.05_{\pm 0.40}$	$74.12_{\pm 0.53}$	$69.52_{\pm 0.74}$	$36.88_{\pm 0.16}$	$31.18_{\pm 0.38}$
FEPT	$\mathbf{84.36}_{\pm 0.09}$	$\mathbf{82.20}_{\pm 0.13}$	$\mathbf{48.02}_{\pm 0.20}$	$\mathbf{44.48}_{\pm 0.35}$	$\mathbf{75.18}_{\pm 0.45}$	$\mathbf{71.02}_{\pm 0.66}$	$\mathbf{39.40}_{\pm 0.14}$	$\mathbf{34.26}_{\pm 0.34}$



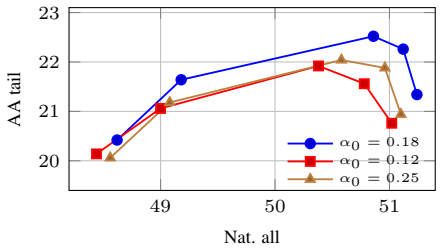
(a) CIFAR100-LT: AA tail heat map



(b) CIFAR100-LT: natural–robust trade-off



(c) TinyImageNet-LT: AA tail heat map



(d) TinyImageNet-LT: natural–robust trade-off

Figure 6: Hyperparameter sensitivity on CIFAR100-LT and TinyImageNet-LT. The best tail robustness appears at intermediate transport strength and moderate free-energy regularization, consistent with the CIFAR10-LT trend in Figure 3, while the detailed response surfaces vary across datasets.

E.6 Hyperparameter Sensitivity

Figure 6 details the hyperparameter sensitivity of α_0 and λ_E on CIFAR100-LT and TinyImageNet-LT, both exhibiting clear intermediate optima where performance degrades if transport α_0 overborrows from distinct classes or free-energy weight λ_E excessively filters out informative adversarial examples. Additionally, the candidate set size M_{cand} governs the cost-quality trade-off. With the expanded candidate set in Equation 5, the worst-case class-conditioned support-query complexity is truncated from $O(BCK_{\mathrm{nn}}d)$ to $O(B\min(C, 1 + M_{\mathrm{cand}} + K_{\mathrm{src}}(1 + M_{\mathrm{cand}}))K_{\mathrm{nn}}d)$, where B, C,

Table 12: CIFAR100-LT with different imbalance ratios using WRN-28-10.

Method	IR = 5				IR = 50			
	Nat. all	Nat. tail	AA all	AA tail	Nat. all	Nat. tail	AA all	AA tail
AT-BSL	$58.60_{\pm 0.27}$	$57.14_{\pm 0.23}$	$24.22_{\pm 0.20}$	$23.78_{\pm 0.25}$	$46.45_{\pm 0.11}$	$42.42_{\pm 0.18}$	$17.40_{\pm 0.12}$	$15.44_{\pm 0.10}$
RobustLT	$59.24_{\pm 0.04}$	$57.98_{\pm 0.06}$	$24.52_{\pm 0.29}$	$24.48_{\pm 0.34}$	$46.90_{\pm 0.08}$	$42.80_{\pm 0.06}$	$17.92_{\pm 0.32}$	$16.48_{\pm 0.28}$
MCAT	$59.82_{\pm 0.05}$	$58.66_{\pm 0.07}$	$25.18_{\pm 0.24}$	$25.02_{\pm 0.28}$	$47.62_{\pm 0.08}$	$43.72_{\pm 0.08}$	$18.76_{\pm 0.29}$	$17.48_{\pm 0.26}$
PAT	$59.50_{\pm 0.05}$	$58.30_{\pm 0.08}$	$24.86_{\pm 0.25}$	$24.58_{\pm 0.30}$	$47.24_{\pm 0.09}$	$43.20_{\pm 0.09}$	$18.32_{\pm 0.30}$	$16.98_{\pm 0.27}$
FEPT	$\mathbf{60.28}_{\pm 0.04}$	$\mathbf{59.18}_{\pm 0.06}$	$\mathbf{25.92}_{\pm 0.23}$	$\mathbf{25.86}_{\pm 0.25}$	$\mathbf{48.30}_{\pm 0.07}$	$\mathbf{44.42}_{\pm 0.07}$	$\mathbf{19.84}_{\pm 0.26}$	$\mathbf{18.86}_{\pm 0.24}$

Table 13: Sensitivity to candidate set size M_{cand} on CIFAR100-LT with WRN-28-10. The base query factor is $\min(C, 1 + M_{\text{cand}} + K_{\text{src}})$ with the CIFAR100-LT default $K_{\text{src}} = 8$; the expanded candidate set in Equation 5 has worst-case factor $\min(C, 1 + M_{\text{cand}} + K_{\text{src}}(1 + M_{\text{cand}}))$, but duplicated donor classes are cached in implementation.

M_{cand}	5	10	20	50	All classes
Base query factor	14	19	29	59	100
Worst-case expanded factor	54	99	100	100	100
AA all	$21.42_{\pm 0.18}$	$21.86_{\pm 0.17}$	$\mathbf{22.24}_{\pm 0.16}$	$22.31_{\pm 0.16}$	$22.38_{\pm 0.15}$
AA tail	$20.74_{\pm 0.22}$	$21.35_{\pm 0.21}$	$\mathbf{21.88}_{\pm 0.20}$	$21.96_{\pm 0.19}$	$22.03_{\pm 0.19}$
Wall-clock / seed	7.96 $\text{h}_{\pm 0.10}$	8.21 $\text{h}_{\pm 0.10}$	8.63 $\text{h}_{\pm 0.11}$	10.42 $\text{h}_{\pm 0.13}$	15.86 $\text{h}_{\pm 0.18}$
Peak memory	10.32 $\text{GB}_{\pm 0.13}$	10.58 $\text{GB}_{\pm 0.13}$	10.91 $\text{GB}_{\pm 0.14}$	12.36 $\text{GB}_{\pm 0.16}$	16.84 $\text{GB}_{\pm 0.22}$

and d denote the batch size, class count, and feature dimension, respectively. In practice, repeated donor classes are cached and many top posterior candidates are non-tail, so the measured overhead is milder than this worst-case bound. As shown in Table 13, scaling $M_{\text{cand}} > 20$ yields diminishing robustness returns while severely inflating wall-clock overhead, justifying our default choice of $M_{\text{cand}} = 20$.