NEURIPS 2026 · CONFERENCE TALK

BayesAT: Bayes-Guided Progressive Distillation for Semi-Supervised Adversarial Training

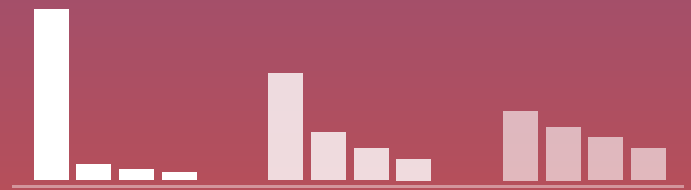
Yimo Guo Lilin Zhang Penglin Li Jinhui Hao Xianggen Liu*

Sichuan University · *Corresponding author

Code: github.com/Sherlockssa/BayesAT

SUPERVISION, RE-TEMPERED

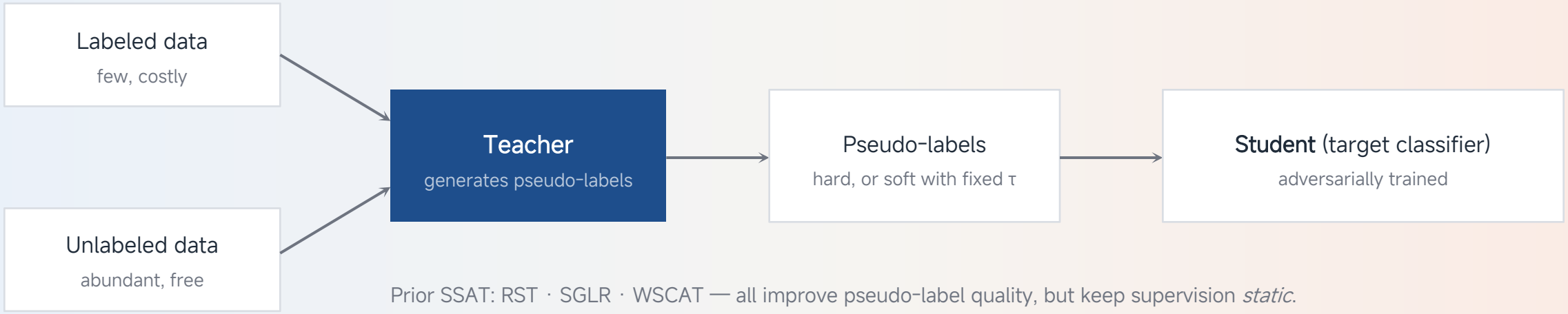
BayesAT



sharp → smooth temperature τ: low → high

Semi-supervised adversarial training, in one slide

Adversarial training (AT) is the standard defense against adversarial perturbations — but it is **label-hungry**. SSAT feeds it **unlabeled data** instead.



Prior SSAT: RST · SGLR · WSCAT — all improve pseudo-label quality, but keep supervision *static*.

THE PROBLEM

Supervision difficulty is **constant** throughout training, while the student keeps evolving: early on it needs **sharp, confident** signals to fix decision boundaries; later it needs **smoother** signals to control variance and see inter-class structure.

A fixed supply of supervision mismatches a moving demand — across the whole training trajectory.

Observation: teachers plateau — static supervision caps the frontier

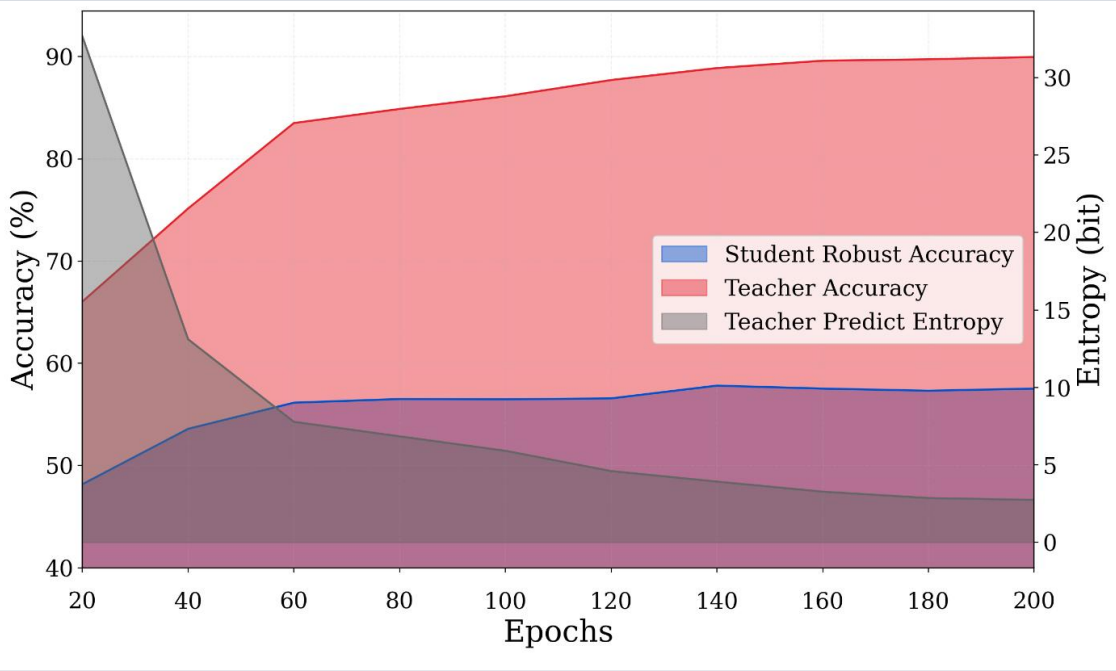


Fig. 1a · Teacher during training. Accuracy saturates while prediction entropy keeps dropping: the plateaued teacher becomes **over-confident**, and supervision variance takes over the generalization bound.

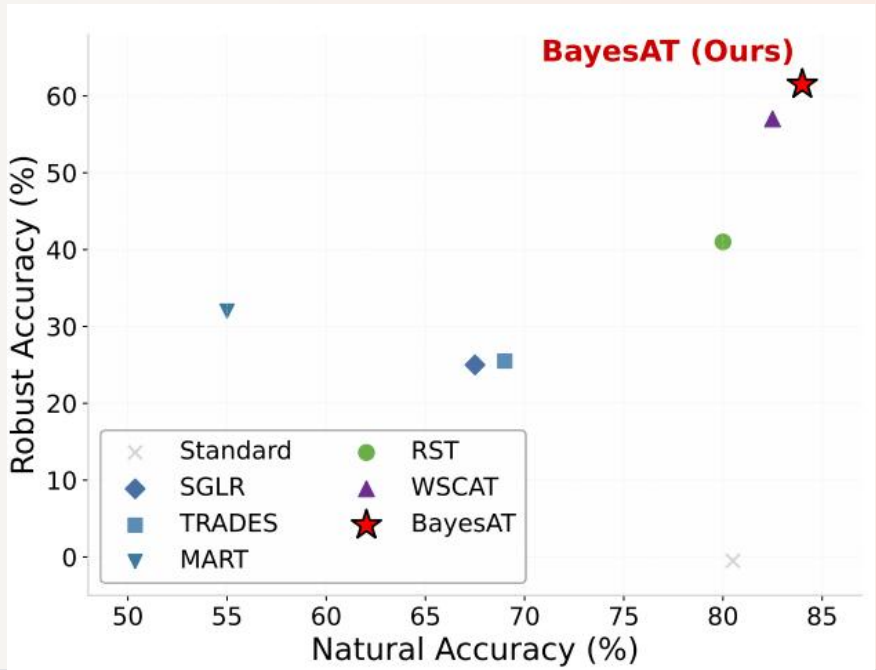


Fig. 1b · CIFAR-10 frontier. Existing SSAT methods cluster in a constrained natural–robust region — consistent with a ceiling imposed by static supervision.

Once the teacher plateaus, bias is near its floor and variance dominates — a fixed temperature cannot follow this shift.

A Bayesian bias–variance view of SSAT

Reframe SSAT as **distillation toward a Bayes teacher** $p^*(x)$. The student's robust generalization gap decomposes into:

$$R(f) - \widehat{R}_t(f) \;\le\; \underbrace{\mathbb{E}_x[(p^*(x) - p_t(x;\tau))^\top \ell(\hat{x})]}_{Bias} \;+\; \underbrace{\sqrt{\frac{2\,\mathbb{V}[p_t^\top \ell]\log(1/\delta)}{N_u}} + \frac{\log(1/\delta)}{3N_u}}_{Variance}$$

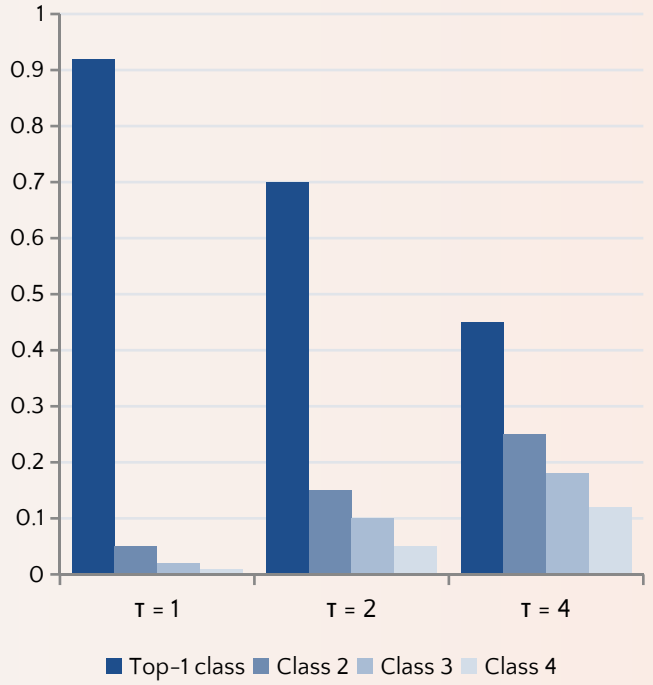
Thm. 3.1 — the Bayes teacher has zero bias and the lowest loss variance (Lem. 3.2)

Lem. 3.1 · Label smoothness. Raising τ spreads probability mass — soft labels get smoother monotonically.

Thm. 3.2 · Bias bound. Bias is small when the teacher stays close to one-hot labels → favors **low** τ.

Thm. 3.3 · Variance bound. Variance scales with $-p_t(x;\tau)-_2$, which τ shrinks → favors **high** τ.

Stage-dependent. The covariance term peaks mid-training (epoch ≈40), so no single τ is optimal throughout.



Temperature steers the teacher's softmax between sharp (low bias) and smooth (low variance).

τ is the knob of the bias–variance tradeoff — and the right setting drifts over training.

BayesAT: Bayes-guided progressive distillation

Phase 1 — train & freeze a reliable teacher

Phase 2 — adversarially train the student

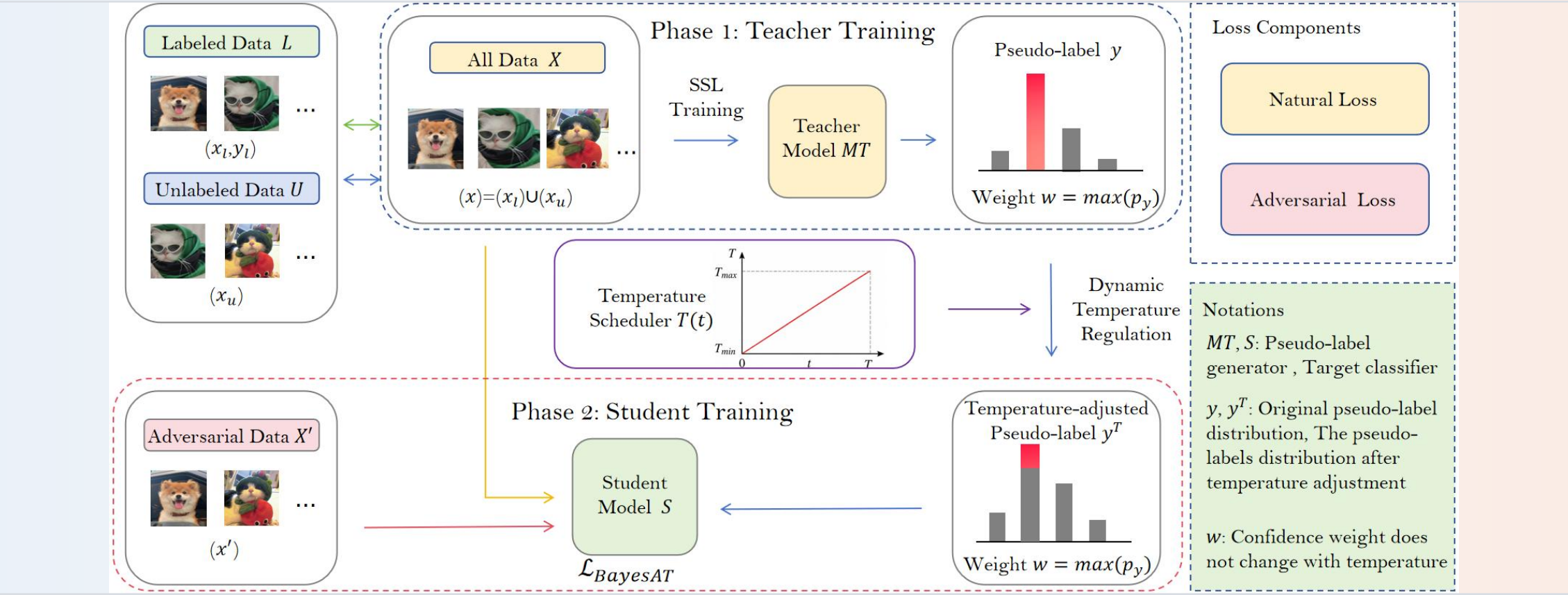


Fig. 3 · Framework. Phase 1 trains a Mean Teacher on labeled + unlabeled data, then freezes it. Phase 2 distills *temperature-adjusted* soft labels into the adversarially trained student, with *confidence-adaptive* per-sample weights.

Two complementary mechanisms

ACROSS STAGES · TEMPERATURE WARMING

$$\tau(k) = \tau_{min} + (\tau_{max} - \tau_{min}) \cdot \frac{k}{K}$$



Early, low τ: sharp, confident labels → reliable decision boundaries

Late, high τ: smooth labels → variance control, inter-class structure

An **easy-to-hard curriculum** over supervision difficulty, grounded in Thm. 3.3: later stages need larger τ to keep the variance term of the generalization bound in check.

Defaults: τ_{min} = 1.0, τ_{max} = 4.0 (CIFAR-10) / 5.0 (CIFAR-100) / 3.0 (ImageNet-200), warmed linearly over 100 epochs.

ACROSS SAMPLES · CONFIDENCE REWEIGHTING

$$w(x) = \max_{i \in [C]} p_{t,i}(x;\ \tau = 1)$$

Teacher confidence is a **proxy for per-sample supervision bias** (Thm. 3.2): samples whose pseudo-labels are reliable get larger weight, ambiguous ones are down-weighted.

Early in training the reliable subset dominates the objective; as the student matures, focus shifts toward harder, more ambiguous samples.

Both mechanisms plug into existing SSAT objectives — we instantiate BayesAT on the WSCAT backbone (clean KL + adversarial KL + weakly-supervised contrastive loss), with near-zero extra compute.

Temperature differentiates supervision across stages; confidence differentiates samples within a stage.

Main results on CIFAR-10 / 100 and ImageNet-200

Dataset	Metric	RST	WSCAT	BayesAT (Ours)
CIFAR-10	Natural	79.95	80.10	**81.36**
	PGD	40.86	56.47	**60.95**
	AA	38.22	50.94	**52.58**
	Mean	49.56	59.58	**61.94**
CIFAR-100	PGD	8.44	25.28	**30.41**
	Mean	17.92	30.94	**34.57**
ImageNet-200	PGD	0.25	9.24	14.42
	Mean	4.85	13.04	16.32

Natural / robust accuracy (%), mean over three seeds; ϵ = 8/255 (CIFAR), 4/255 (ImageNet-200). Full tables incl. FGSM, CW and more baselines (SGLR, TRADES, MART, MT) in the paper.

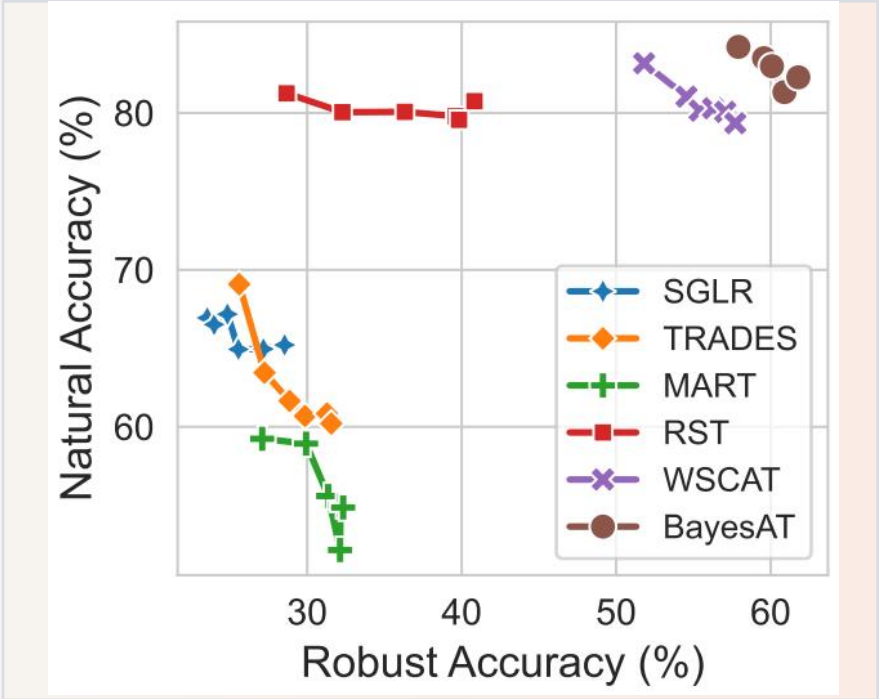


Fig. 4 · Natural vs. PGD on CIFAR-10. BayesAT pushes the frontier toward the top-right.

+3.8 PGD on CIFAR-10, +5.1 on CIFAR-100 over WSCAT — with natural accuracy preserved.

Ablations: direction of the schedule matters

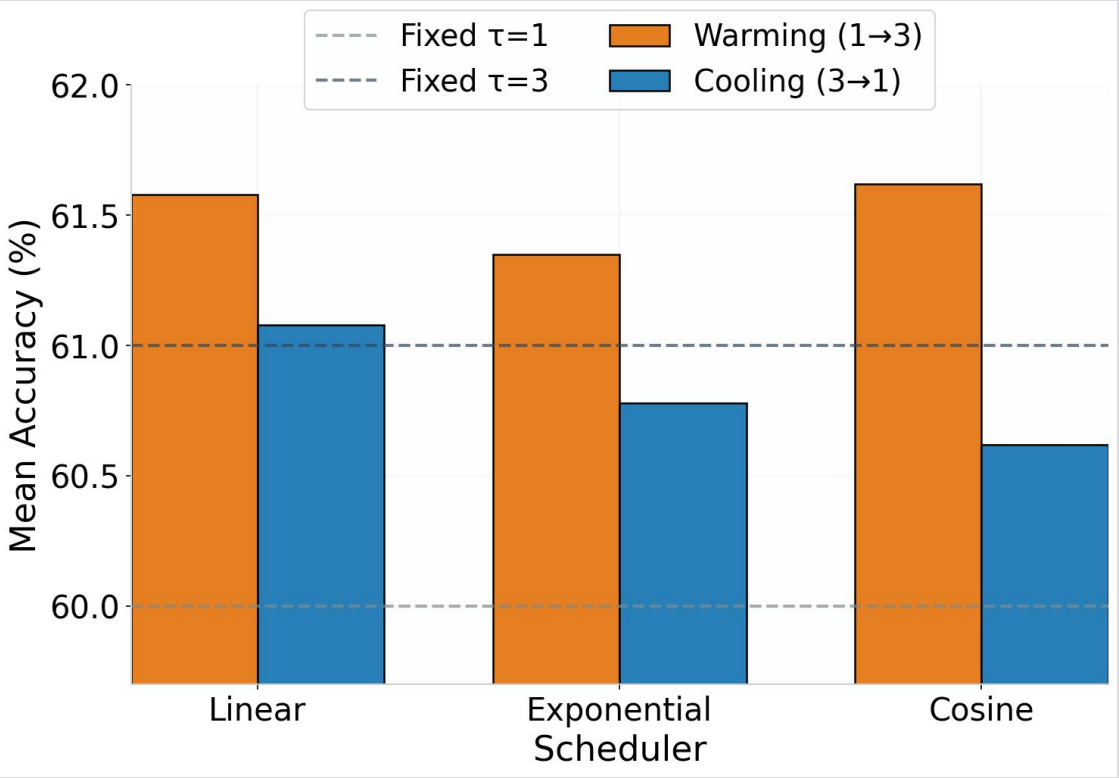


Fig. 5b · Mean accuracy per schedule (CIFAR-10). Across linear / exponential / cosine variants, **warming (1→3)** consistently beats both cooling (3→1) and fixed-τ baselines (dashed lines). A cosine oscillation between τ = 1 and 4 is also weaker (61.34 vs. 64.22).

COMPONENT ABLATION · CIFAR-10

Configuration	Mean (%)
Base model (WSCAT)	60.04
+ Soft labels	60.75
+ Confidence weighting	60.73
+ Temperature warming	**61.88**

Temperature warming gives the largest single gain (PGD 57.06 → 60.89), and it composes with confidence weighting — the two act on complementary axes.

Teacher-quality sensitivity is controlled: a Mean Teacher nearly matches ground-truth supervision (mean 64.27 vs. 64.34), while a labeled-only teacher drops to 61.72.

Supervision, re-tempered.

static labels → temperature-warmed distillation → a student that keeps improving



cold · τ = 1 · sharp labels · low bias

hot · τ = 4 · smooth labels · low variance

Takeaways

1 **Static supervision is the bottleneck.** Reframing SSAT as distillation toward a Bayes teacher exposes a stage-dependent bias–variance tradeoff governed by temperature — no fixed τ stays optimal as the student evolves (Thm. 3.1–3.3).

2 **BayesAT = low→high temperature warming + confidence-adaptive reweighting.** Two plug-in mechanisms that balance bias and variance across training stages and across samples — no explicit difficulty metric needed.

3 **Consistent gains on CIFAR-10, CIFAR-100 and ImageNet-200.** +3.8 to +5.2 PGD points over WSCAT at preserved natural accuracy — dynamic, theory-driven supervision breaks the static-supervision frontier.

Limitations & future work. Two-phase pipeline needs a pre-trained teacher; τ bounds are tuned per dataset; scope is ℓ_∞ image classification. Next: adaptive warming driven by online bias–variance estimates, and end-to-end teacher–student training.

Thank you — questions welcome!

Code: github.com/Sherlockssa/BayesAT

Sichuan University · NeurIPS 2026