EA-KD: Entropy-based Adaptive Knowledge Distillation
Supplementary Material

A. Additional Results and Comparison

A.1. Comparison with TTM and WTTM

TTM [48] removes student temperature in KD, revealing an inherent Rényi entropy regularizer that could improve generalization. WTTM further up-weights uncertain samples using the power sum of the teacher's output, akin to our w_{base}[5]. Notably, TTM's Rényi entropy is an *inherent effect* of its structural modification, while EA-KD *actively* uses Shannon entropy to prioritize valuable samples. Moreover, unlike WTTM's static teacher-based weighting, EA-KD dynamically adapts to the student's learning (via $H^{\mathcal{S}}$), yielding stronger enhancement to TTM (Tab. A1). Finally, while their methods combine with feature-based KDs by adding and balancing their losses, EA-KD offers direct integration to both logit- and feature-based KDs through reweighting.

Table A1. **TTM, WTTM, and EA-TTM on CIFAR-100 over 5 runs.** Unlike WTTM, EA-TTM consistently improves TTM, highlighting the superiority of dynamic over static weighting.

Teacher	Student	TTM	WTTM	EA-TTM (Ours)
ResNet32×4	ResNet8×4	76.17±0.28	76.06±0.27	**76.25**±0.04
WRN-40-2	WRN-40-1	74.32±0.31	74.58±0.26	**74.61**±0.07
ResNet32×4	SN-V2	76.57±0.26	76.55±0.08	**76.65**±0.16

A.2. MS-COCO

Tab. A2 reports additional results for MS-COCO, where EA-DKD is compared against DKD [47] using a ResNet-101 to ResNet-18 distillation setting. The results show that EA-DKD consistently improves upon DKD across all evaluation metrics.

Table A2. **More Results on MS-COCO.**

R101→R18	AP	AP_{50}	AP_{75}
DKD [39]	35.05	56.60	37.54
EA-DKD	**35.16**+0.11	**56.75** +0.15	**37.82** +0.28

A.3. LLM Distillation

Tab. A3 extends the comparison of EA-KD with Skewed KLD (SKLD) and SRKLD from the recent DistiLLM [15] method, under the same off-policy setting without pre-training corpus. EA-KD outperforms these stronger methods on most datasets, highlighting its overall effectiveness.

[5] In fact, they also noted teacher entropy as a potential weighting factor, but left its systematic exploration for future work.

Table A3. **More LLM Distillation Results.**

Method	Dolly	SInst	Vicuna	S-NI	UnNI	Avg.
SKLD (DistiLLM)	24.03	**10.66**	14.70	17.99	21.18	17.71
SRKLD (DistiLLM)	24.48	10.35	14.88	16.53	19.68	17.19
EA-KD	**24.95**	10.59	**16.41**	**18.27**	**21.46**	**18.34**

B. Additional Analysis and Visualizations

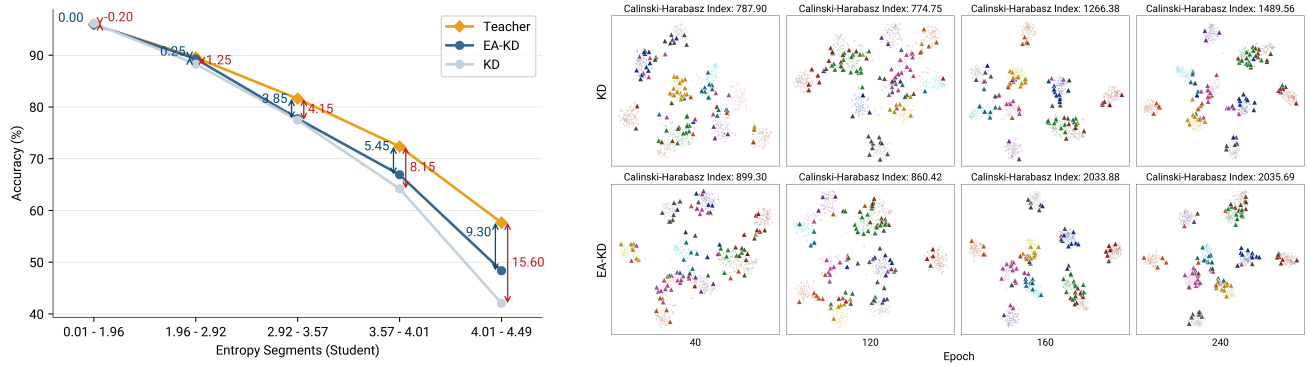
This section provides additional analysis to complement the main paper. It includes a detailed examination of high student entropy samples, loss landscape comparisons for KD and EA-KD, and t-SNE visualizations for various KD frameworks and their EA-enhanced variants.

B.1. Analysis of High Student Entropy Samples

Complementary to the teacher entropy analysis in Sec. 1, this section emphasizes the critical role of high student entropy samples in the adaptive reweighting process of EA-KD. As shown in Fig. B1a, these samples correlate with larger teacher-student accuracy gaps similar to the teacher entropy segments (Fig. 2a). Additionally, Fig. B1b presents the students' t-SNE visualizations across training epochs for KD and EA-KD, with high $H^{\mathcal{S}}$ and high w_{EA} samples highlighted, respectively. Notably, these samples also cluster near decision boundaries similar to the teacher's t-SNE (Fig. 2b). Unlike the teacher, however, the student's top-entropy samples shift dynamically over training. For instance, in KD, the yellow class at epoch 40 is located near the center of most clusters and contains numerous high-entropy samples. By epoch 120, high-entropy samples are more prevalent in the green and purple classes as they shift closer to the center. With this dynamic integrated into w_{EA}, EA-KD continuously adapts to the student's evolving learning needs throughout training, complementing the static nature of w_{base} and fostering a tailored knowledge transfer.

B.2. Loss Landscape Analysis for KD and EA-KD

We further analyze the robustness of KD and EA-KD by examining their loss landscapes across training epochs. As shown in Fig. B2, EA-KD achieves a consistently smoother loss surface compared to KD (*e.g.* σ^2 of 217.68 vs. 305.21 at epoch 240) and larger Area@1.6 values. Notably, KD's Area@1.6 metric remains 0 at epochs 40 and 120, indicating the absence of low-loss regions. In contrast, EA-KD achieves values of 0.01 and 0.05, respectively, underscoring its improved learning efficiency by enabling the student to assimilate the key knowledge. These results highlight EA-KD's ability to ensure a more efficient optimization, facilitating enhanced generalizability compared to standard KD.



(a) Accuracy vs. Student Entropy Segments.

(b) t-SNE of Students Across Training Epochs.

Figure B1. **High Student Entropy Analysis in KD and EA-KD.** (a) Higher student entropy samples correlate with larger accuracy gaps in KD. EA-KD demonstrates closer alignment with the teacher, particularly in high-entropy regions. (b) High H^S samples consistently cluster near decision boundaries, reflecting the student's real-time learning needs. EA-KD adapts to this dynamic, achieving enhanced class separability over epochs.

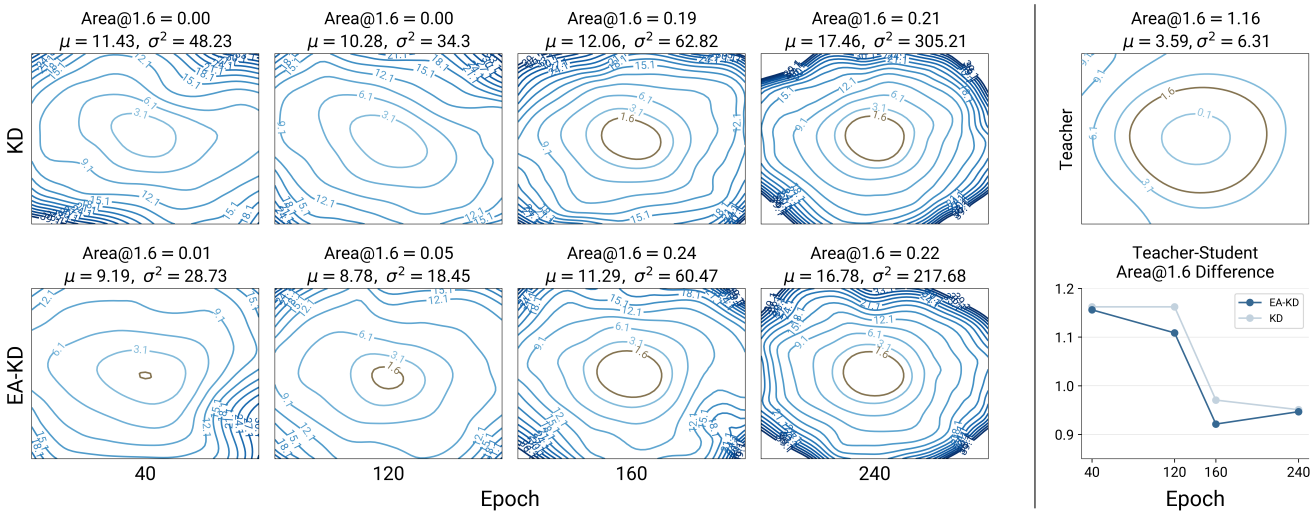


Figure B2. **Loss Surface and Differences in Area@1.6 for KD and EA-KD Students Across Epochs.** Contour plots illustrate the loss landscapes of KD (first row) and EA-KD (second row) across training epochs. The line plot (lower right) tracks the differences in Area@1.6 between the teacher and students over epochs. EA-KD exhibits a more stable and robust loss surface, with greater Area@1.6 earlier in training, signifying a more efficient learning process.

B.3. t-SNE of KD Frameworks and EA-methods

The t-SNE visualizations of students from various KD frameworks and their EA-method variants, alongside the teacher's representation, are shown in Fig. B3. EA-methods consistently exhibit more distinct and well-defined class clusters compared to their respective baselines, as evidenced by higher CH indices. Remarkably, the SOTA EA-MLD+LS achieves the highest CH index of 814.96, indicating its superior performance and closer alignment with the teacher. Furthermore, although EA-DKD ranks fourth in performance (Tab. 2), it demonstrates the second-best class separability, underscoring its enhanced robustness as discussed in Sec. 4.3. These findings highlight EA-KD's versatility in enhancing class separability and improving knowledge transfer across diverse KD frameworks.

C. Statistics for Table 2 and 4

Tab. C4 and Tab. C5 present the mean and standard deviation of EA-methods on CIFAR-100 and ImageNet across multiple runs, respectively. Since baseline papers typically report only mean values, formal statistical tests such as p-value computation are not applicable. Nonetheless, EA-methods consistently demonstrate improved performance with low variance, highlighting the stable and reliable gains.

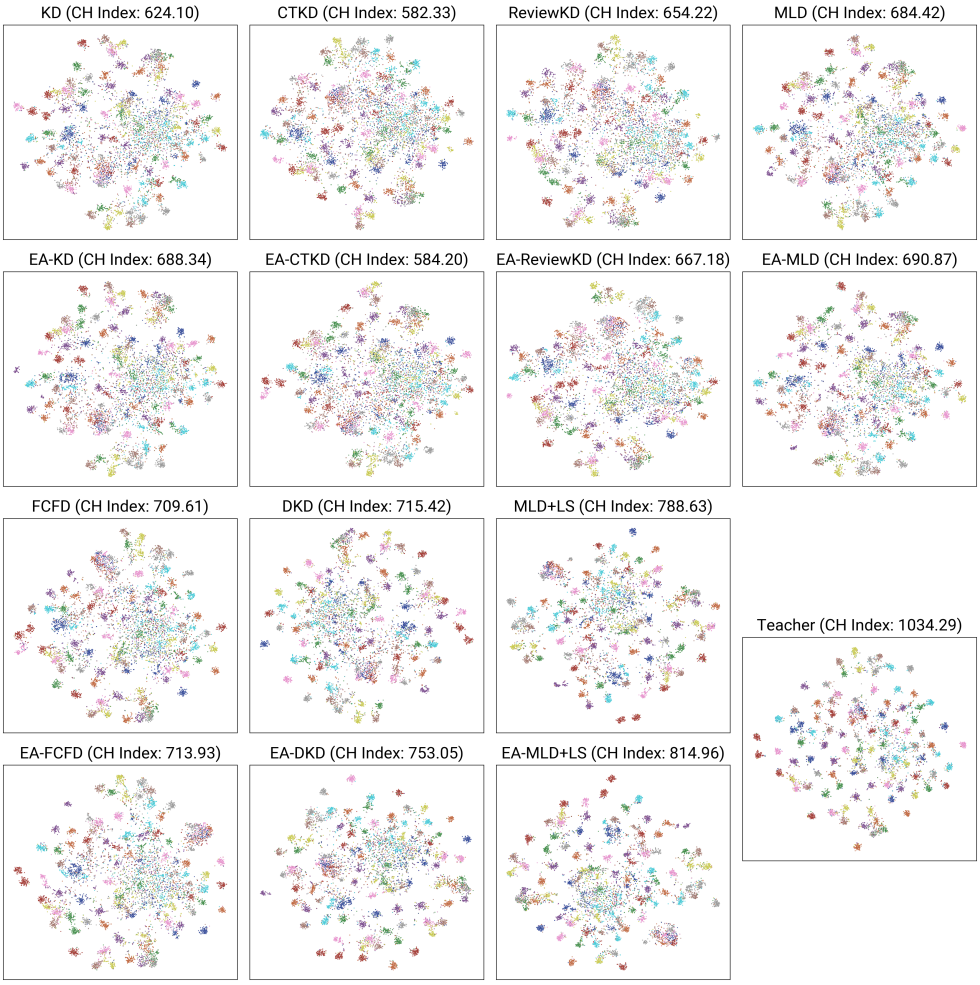


Figure B3. **t-SNE Visualizations of EA-methods vs. Baselines.** The students from various KD frameworks (first and third row) and their EA-enhanced counterparts (second and fourth row), along with the teacher (lower right), are shown. EA-methods consistently achieve higher CH indices, indicating better class separability.

Table C4. **Results on CIFAR-100.** Mean accuracy (%) and standard deviation across five runs are reported.

Type								
	Teacher	ResNet32×4 79.42	WRN-28-4 78.60	WRN-40-2 75.61	VGG13 74.64	VGG13 74.64	ResNet50 79.34	ResNet32×4 79.42
	Student	ResNet8×4 72.50	WRN-16-2 73.26	WRN-40-1 71.98	VGG8 70.36	MN-V2 64.60	MN-V2 64.60	SN-V2 71.82
Logit	EA-KD	75.46±0.15	75.79±0.14	74.38±0.10	74.08±0.10	69.17±0.08	69.67±0.26	75.91±0.25
	EA-CTKD	75.18±0.24	75.72±0.16	74.03±0.05	73.79±0.10	69.19±0.26	69.38±0.36	76.02±0.15
	EA-DKD	76.80±0.05	76.74±0.09	74.98±0.15	75.07±0.14	70.39±0.14	70.98±0.13	77.72±0.07
	EA-MLD	77.65±0.05	77.47±0.12	75.77±0.21	75.28±0.22	70.72±0.15	71.43±0.18	78.85±0.05
	EA-MLD+LS	**78.38**±0.10	**77.60**±0.07	**75.78**±0.10	**75.38**±0.15	70.67±0.25	71.36±0.21	**79.13**±0.23
Feature	EA-ReviewKD	76.10±0.13	76.95±0.16	75.43±0.13	74.56±0.11	70.55±0.09	69.80±0.18	78.22±0.15
	EA-FCFD	77.50±0.08	77.15±0.15	75.30±0.03	75.36±0.06	**71.02**±0.26	**71.97**±0.29	78.75±0.32

Table C5. **Results on ImageNet.** Mean accuracy (%) and standard deviation across three runs are reported for EA-methods.

Teacher	Student	KD [10]	EA-KD	KD+LS [30]	DKD [39]	EA-DKD	DKD+LS [30]	EA-DKD+LS	PAD [38]
73.31	69.75	71.03	**71.79**±0.02	71.42	71.70	71.96±0.08	71.88	**71.99**±0.01	71.71