Enriching Knowledge Distillation with Cross-Modal Teacher Fusion

Supplementary Material

Overview

Section 7 presents further experiments; Section 8 shows additional visualizations; and Section 9 provides further theoretical support.

7. Further Experiments

Few-shot Learning Experiment. RichKD introduces additional knowledge diversity by enriching the conventional teacher's outputs with CLIP-driven multi-prompt fusion. This cross-modal enrichment produces more informative and varied supervisory signals, allowing the student to observe a broader spectrum of semantic cues even when the training data is limited. To evaluate this property, we conduct few-shot experiments using the ResNet-32×4 → ResNet-8×4×4 teacher–student setup under 25%, 50%, and 75% of the CIFAR-100 training set. In each scenario, the student is trained only on the reduced subset, while the teacher and CLIP remain fixed. As shown in Figure 6, RichKD consistently improves performance across all data fractions. These results indicate that RichKD enables the student to extract richer and more diverse knowledge from fewer samples, demonstrating strong effectiveness in low-data regimes.

Category-based Evaluation. As shown in Figure 1, fusing the teacher with CLIP's outputs leads to more confident correct predictions and reduces supervision from incorrect predictions. Table 9 presents the accuracy comparison between RichKD (L) and standard KD under different teacher prediction categories: correct and certain, correct and uncertain, and incorrect. As can be seen, distillation from the fused teacher improves the student's performance on both certain correct samples and incorrect predictions. Furthermore, fusing with CLIP results in more confident correct predictions, with RichKD improving the student's performance on uncertain correct samples by +5.5% compared to vanilla KD. Teacher and student architecters are ResNet-32×4 and ResNet-8×4, respectively.

Besides, Table 10 reports how often fusion with CLIP fixes or harms predictions. After fusing with CLIP logits, **1126** misclassified samples are corrected (**46%** of all teacher errors) . When the target class is not even in the teacher's top-5, fusion still corrects **219** cases (**40%**). Conversely, among teacher-correct cases, only **40** samples (**0.5%**) become wrong after fusion. Finally, we observe **no** case where the teacher is correct and the target class is outside fused teacher's top-5.

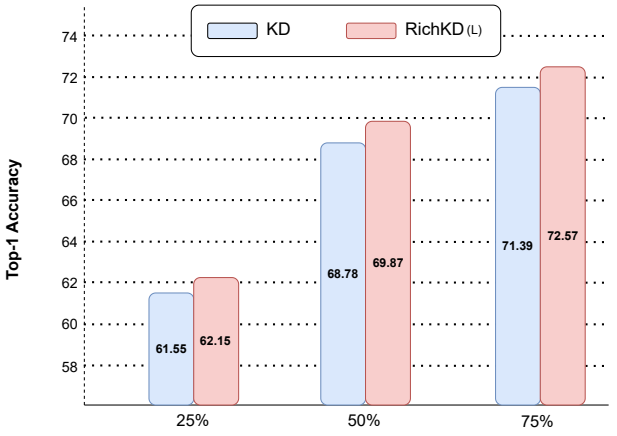

Figure 6. Few-shot CIFAR-100 distillation under limited data. Top-1 accuracy of the ResNet-32×4 → ResNet-8×4×4 student when trained on 25%, 50%, and 75% of the CIFAR-100 training set.

Table 9. Student performance (%) by teacher-based categories on CIFAR-100.

Category	KD		RichKD (L)	
	Top-1	**Top-5**	**Top-1**	**Top-5**
Teacher correct & **certain**	89.08	98.91	90.56	99.15
Teacher correct & **uncertain**	49.85	87.80	55.51	89.14
Teacher **wrong**	25.61	71.72	27.94	74.10

Table 10. Outcome changes with fusion. "Base" is the number of samples that qualify for the denominator of each condition. Abbrev.: T=Teacher, F=Fusion.

Condition	**#Base**	**Count (%)**
T wrong → F correct	2441	1126 (46%)
(Target ∉ Top-5 of T) → F correct	544	219 (40%)
T correct → F wrong	7559	40 (0.5%)
T correct → (Target ∉ Top-5 of F)	7559	0 (0%)

Experiments with Transformer Architecture. Distillation to ViTs using our method is straightforward, as our approach is model-agnostic. Experiments were conducted on CIFAR-100 following the same settings as [23, 24], with results summarized in Table 11. Notably, our method consistently and significantly improves KD performance across all ViTs. It outperforms vanilla KD, AutoKD [24], and LSKD, particularly when the student is hierarchical and relatively large. These results demonstrate the effectiveness of our approach in alleviating the data-hungry nature of ViTs.

Table 11. The Top-1 Acc. (%) of KD methods on CIFAR-100 for distillation to transformers. Teacher is ResNet56 and Hie. indicates if model is hierarchy structure.

Student	Hie.	Size	w/o distillation	AT [58]	LG [23]	AutoKD [24]	LSKD [45]	KD	RichKD (L)
DeiT-Ti	✗	5M	65.08	73.51	78.15	78.58	78.55	73.25	**79.01**$_{+5.76}$
T2T-ViT7	✗	4M	69.37	74.01	78.35	78.62	78.43	74.15	**78.79**$_{+4.64}$
PiT-Ti	✓	5M	73.58	76.03	78.48	78.51	78.76	75.47	**79.20**$_{+3.73}$
PVT-Ti	✓	13M	69.22	74.66	77.07	77.48	78.43	73.60	**78.84**$_{+5.24}$

Table 12. Computation comparison: training time per epoch and peak memory usage during batch training.

Method	**Time (s)**	**Memory (G)**
KD	7	2.278
RichKD (L)	41	3.048
CRD	26	3.226
RichKD (F)	45	4.136
RichKD (L+F)	46	4.138

Table 13. Comparison with KD method on class imbalance CIFAR-100. Reported accuracy (%).

Method	**ResNet32×4 ResNet8×4**	**WRN-40-2 WRN-40-1**	**VGG13 VGG8**
KD	47.68/79.85	47.53/79.64	46.49/77.01
RichKD (L)	49.18/79.90	48.1/80.83	47.89/78.56

Training Time Analysis. Fusing the logits and features of CLIP with a conventional teacher adds time and memory overhead to the distillation pipeline during training. Table 12 shows the time and GPU memory usage of RichKD compared to the KD and CRD methods. As shown, the overhead is considerable. However, this burden occurs only during training; there is no additional cost during inference.

To mitigate this overhead, we employed a caching approach in our experiments. First, we use all the samples in the dataset to compute and save the features and logits of each batch into tensors. Then, during each training epoch, we simply load these precomputed tensors. Using this method, the time and memory overhead becomes almost the same as the baseline, and the additional burden is practically negligible except for the one-time inference of CLIP used to generate the cached representations.

Class Imbalance Experiment. We follow the long-tailed data generation protocol commonly adopted in prior work on imbalanced learning [1]. Let the original CIFAR-100 training set contain $C = 100$ classes with $n_{\max} = 500$ samples per class. Given an imbalance factor $\rho \in (0, 1]$, the number of training samples assigned to class c is

$$n_c = n_{\max}\, \rho^{\frac{c}{C-1}}, \qquad c = 0, \ldots, C-1.$$

This exponential decay schedule produces a strictly decreasing class-frequency distribution, where early classes constitute the head and later classes the tail. For each class, we randomly permute all available indices and retain the first n_c samples without replacement. All label information remains unchanged; only the number of samples per class is reduced. The test split is kept intact and remains class-balanced. Table 14 reports the class-wise sample counts for the CIFAR-100 dataset under the long-tailed configuration with an imbalance ratio of $\rho = 0.01$.

Table 13 presents the comparison with standard KD on the imbalanced CIFAR-100 dataset. It can be observed that, across different teacher/student architectures, the proposed method consistently outperforms KD under the class-imbalance setting. Our cross-modal fusion mitigates teacher bias on head classes by injecting CLIP's text-aligned, class-balanced priors, improving performance on tail categories.

8. Further Visualizations

Class Activation Maps. Figure 7 presents a visual comparison using Gradient Class Activation Maps (Grad-CAM) [43]. Images were randomly selected from the CIFAR-100 dataset for this experiment. The student trained with RichKD demonstrates better attention to the target objects compared to both the baseline and standard KD students. In this setup, both the teacher and student are trained on CIFAR-100, with ResNet-32x4 as the teacher and ResNet-8x4 as the student.

Additional Output Samples. As shown in Figure 1, the logits of CLIP and the teacher exhibit different distributions for the input samples. We visualize additional samples in Figure 8. In the top example, the teacher predicts the wrong class; the target class is ranked second but with very low confidence. By contrast, CLIP predicts the correct class confidently, and fusing the two yields the correct prediction with high confidence. In the bottom example, the target class is not even in the teacher's top-5, yet after fusing with

Table 14. Number of samples per class for Imbalanced CIFAR-100.

Class	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Count	500	488	476	465	453	442	431	420	409	399	389	379	369	359	350	341	332	323	315	306
Class	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**
Count	298	289	281	273	266	258	250	243	236	229	222	215	209	202	196	190	184	178	172	166
Class	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**
Count	161	155	150	145	140	135	130	125	120	116	111	107	102	98	94	90	86	82	79	75
Class	**60**	**61**	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**
Count	72	68	65	62	59	56	53	51	48	46	43	41	39	37	35	33	31	29	28	26
Class	**80**	**81**	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**	**91**	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**
Count	25	23	22	21	19	18	17	16	15	14	13	12	12	11	10	10	9	8	7	5

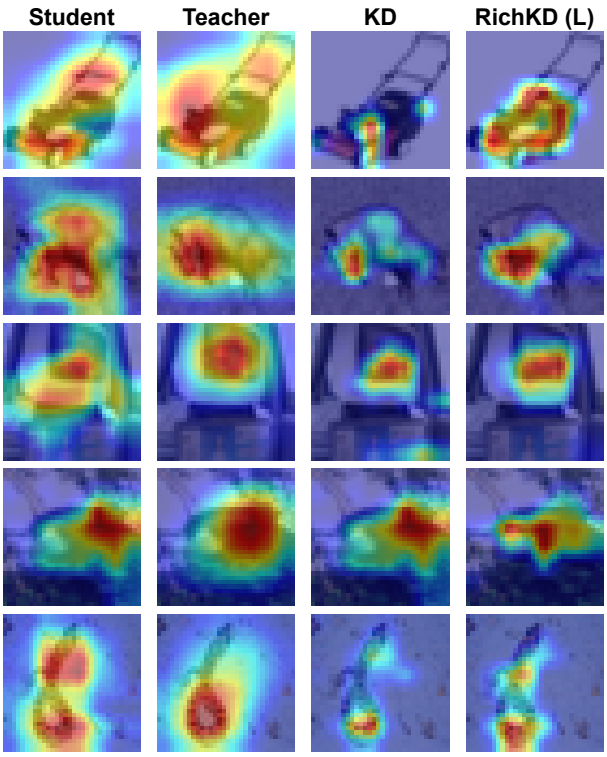

Figure 7. Visual comparison of Grad-CAMs.

CLIP's logits, the fused model correctly predicts the target class.

9. Further Theoretical Supports

Assumptions and Setup. We model the two teachers as providing scalar predictions for each input x. Let $f_t(x)$ and $f_c(x)$ denote the predictions (logits) from the task-specific teacher and CLIP teacher, respectively. Let $f^*(x)$ denote the true target (e.g., the true logit or an indicator for the correct class). Each teacher's prediction is assumed to be of the form

$$f_i(x) = f^*(x) + b_i + \varepsilon_i, \quad i \in \{t, c\},$$

where $b_i = \mathbb{E}[f_i(x) - f^*(x)]$ is the bias of teacher i, and

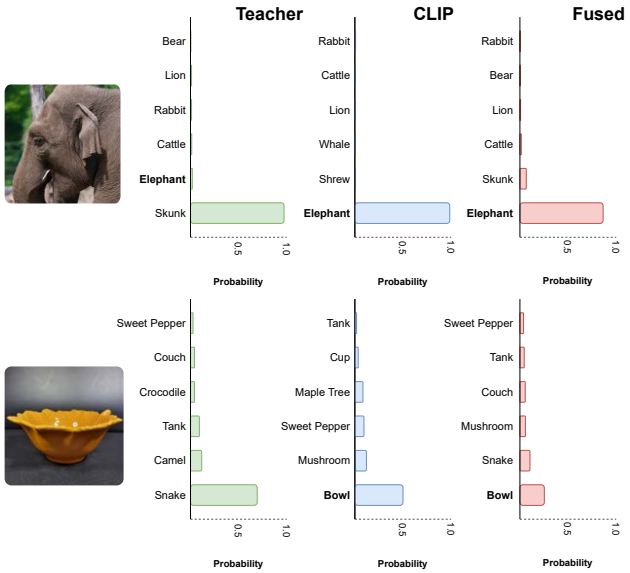

Figure 8. Impact of fusing CLIP logits with the teacher. The fusion frequently corrects teacher misclassifications.

ε_i is a zero-mean random error with variance $\mathbb{V}[\varepsilon_i] = \sigma_i^2$. We further assume that the teachers' errors are **partially independent**, such that the covariance $\mathrm{Cov}(\varepsilon_t, \varepsilon_c) = \rho\sigma_t\sigma_c$, where $\rho \in [-1, 1]$ is the correlation coefficient between the errors of the two teachers.

Fusion and Bias-Variance Decomposition. The fused teacher's prediction is a weighted average of the task-specific and CLIP teacher's predictions

$$f_f(x) = \alpha f_t(x) + (1 - \alpha) f_c(x),$$

where $\alpha \in [0, 1]$ is the fusion weight. The bias of the fused teacher relative to the true target is

$$b_f \equiv \mathbb{E}[f_f - f^*] = \alpha b_t + (1 - \alpha) b_c.$$

The variance of the fused prediction is

$$\sigma_f^2 \equiv \mathbb{V}[f_f] = \alpha^2\sigma_t^2 + (1 - \alpha)^2\sigma_c^2 + 2\alpha(1 - \alpha)\,\mathrm{Cov}(\varepsilon_t, \varepsilon_c).$$

Thus, the MSE of the fused teacher can be written as

$$\mathbb{E}[(f_f - f^*)^2] = (b_f)^2 + \sigma_f^2.$$

Substituting the expressions for b_f and σ_f^2, we get

$$\mathbb{E}\big[(f_f - f^*)^2\big] = (\alpha b_t + (1-\alpha) b_c)^2 + \alpha^2 \sigma_t^2 + (1-\alpha)^2 \sigma_c^2 + 2\alpha(1-\alpha)\rho\sigma_t\sigma_c.$$

This decomposition shows how the fusion weight α affects the **bias** and **variance** of the fused teacher's error. If the errors of the two teachers are uncorrelated (i.e., $\rho \approx 0$), the variance of the fused prediction becomes a simple weighted average of the two individual variances.

Optimal Fusion Weight. To minimize the expected error, we differentiate the last equation with respect to α and set the derivative equal to zero

$$\frac{d}{d\alpha}\mathbb{E}[(f_f - f^*)^2] = 0.$$

Solving this, we obtain the optimal fusion weight

$$\alpha^* = \frac{\sigma_c^2 - \rho\sigma_t\sigma_c}{\sigma_t^2 + \sigma_c^2 - 2\rho\sigma_t\sigma_c}.$$

For the special case where the teachers' errors are independent ($\rho = 0$), this simplifies to

$$\alpha^* = \frac{\sigma_c^2}{\sigma_t^2 + \sigma_c^2},$$

which suggests that the fusion weight should be proportional to the inverse of the teacher's variance. If one teacher is much more accurate (i.e., $\sigma_t^2 \ll \sigma_c^2$), the optimal weight will lean toward the more reliable teacher.

Fusion Benefits and Entropy Regularization. The regularization effect of the CLIP teacher comes from its softer (higher-entropy) predictions. In particular, the CLIP teacher's logits $f_c(x)$ are typically less confident and more spread out compared to the task-specific teacher's logits $f_t(x)$. This difference in certainty has a regularizing effect: it prevents the student from being overly confident about incorrect classes, which can help improve generalization.

Let $q_t(y|x)$ and $q_c(y|x)$ be the predicted probabilities (softmax outputs) from the task-specific teacher and CLIP teacher, respectively. The fused probability is

$$q_f(y|x) = \alpha q_t(y|x) + (1-\alpha) q_c(y|x).$$

The fused teacher provides a distribution that is a weighted average of the two individual teachers' predictions. The use of soft targets from the CLIP teacher (which are higher-entropy) helps smooth the student's learning and reduces overfitting, as it discourages the student from being overly confident about the class label, especially when the teachers disagree. This behavior is akin to label smoothing [47], a well-known technique in machine learning where soft targets (with non-zero probability spread across multiple classes) improve the robustness and generalization ability of the student. Empirically, it has been shown that using soft targets from a diverse teacher model (like CLIP) can help prevent overfitting and improve the robustness of the student model.

Diversity and Complementarity of Teachers. The covariance term $\mathrm{Cov}(\varepsilon_t, \varepsilon_c)$ plays a key role in determining how much variance reduction is achieved by the fusion. If the teachers are **highly correlated** (i.e., $\rho \approx 1$), the fused model will not benefit much from ensemble averaging, as the predictions of the two teachers are too similar. On the other hand, if the teachers have **low correlation** or are independent (i.e., $\rho \approx 0$), the variance of the fused teacher will be significantly lower than that of each individual teacher.

Conclusion. The theoretical analysis confirms that fusing a task-specific teacher with CLIP helps reduce variance and average out biases, leading to improved student generalization. By leveraging CLIP's softer, high-entropy predictions, RichKD regularizes the student's learning and prevents overfitting. The analysis also suggests that teacher diversity (low correlation and complementary biases) is crucial for the fusion to be beneficial. Under the right conditions, fused teaching provides a more informative, stable target for the student, resulting in better overall performance.