STRUCTURED TEMPERATURE SCALING FOR GROUP KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation (KD) commonly uses temperature scaling to soften the teacher distribution and transfer dark knowledge. In this work, we revisit temperature from a structured prediction perspective. To this end, we derive an exact multi-group decomposition of vanilla KD into inter-group distillation (inter-KD) and probability-mass-weighted intra-group distillation (intra-KD). Inter-KD preserves the probability allocation across groups, while intra-KD transfers fine-grained conditional structures within each group. This decomposition reveals that temperature improves KD by adjusting the contribution of different probability regions, but conventional global temperature simultaneously alters inter-group allocation and flattens fine-grained intra-group structures. Based on this analysis, we propose Group Knowledge Distillation (GKD), which introduces structured temperature scaling by shifting temperature smoothing from the global category distribution to the group level. Specifically, GKD adaptively determines prediction groups based on prediction perplexity and applies temperature scaling at the group level, thereby rebalancing supervision across groups while preserving fine-grained intra-group structural knowledge. Without introducing additional grouping hyperparameters or complex structural designs, GKD only modifies the weighting strategy in the reconstructed KD formulation and achieves effective knowledge transfer. Experiments on CIFAR-100 and ImageNet demonstrate state-of-the-art Top-1 performance, providing a new perspective on temperature regulation in knowledge distillation.

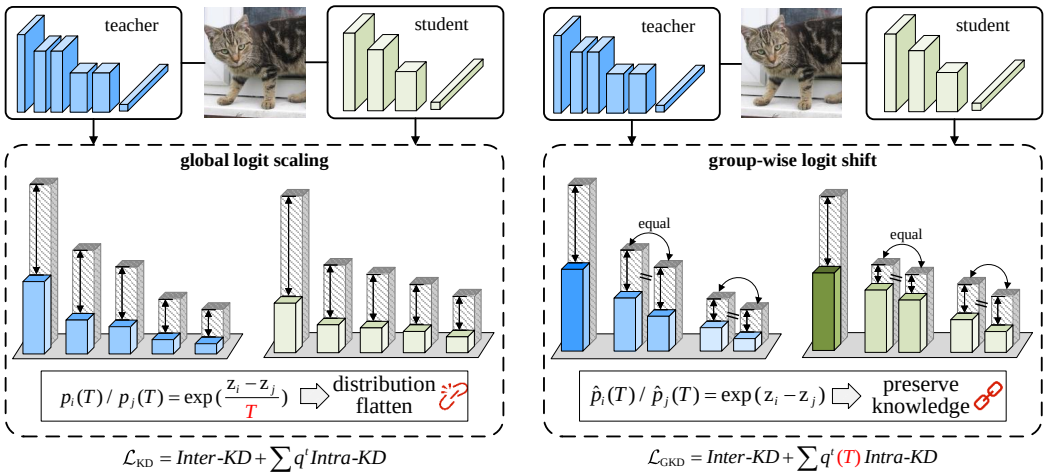
1 INTRODUCTION

Recent advances in deep learning have been largely driven by increasingly large neural networks, which bring substantial computational and storage costs. This makes it challenging to deploy state-of-the-art models on resource-constrained platforms such as mobile devices, edge systems, and latency-sensitive applications. Knowledge distillation (KD) addresses this issue by transferring knowledge from a high-capacity teacher model to a compact student model, enabling the student to achieve competitive performance with lower complexity. Owing to its effectiveness and flexibility, KD has been widely applied across diverse tasks and model architectures Liu et al. (2025); Lan & Tian (2025); Hao et al. (2025).

The seminal KD framework Hinton et al. (2015) minimizes the KL divergence between teacher and student predictive distributions, where temperature scaling softens predictions to expose dark knowledge. DKD Zhao et al. (2022) further reveals that the dominant target probability suppresses non-target knowledge and decouples the two components for independent regulation. More recently, LDRLD Xu et al. (2025) explores fine-grained local relations by recursively recombining ranked logits into pairwise distributions. These advances suggest that effective distillation benefits from prediction decomposition, which exposes finer-grained dark knowledge beyond global distribution matching. However, temperature is still applied as a global operation that simultaneously affects different levels of predictive structures.

To this end, we derive an equivalent multi-group decomposition of the KL divergence, reformulating vanilla KD into intra-group distillation (intra-KD) and inter-group distillation (inter-KD). Intra-KD transfers the conditional distribution within each group and is weighted by the teacher-assigned group probability mass, while inter-KD preserves the probability-mass allocation across

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.



(a) Vanilla Knowledge Distillation (KD) (b) Group Knowledge Distillation (GKD)

Figure 1: Illustration of vanilla KD and our GKD. Vanilla KD globally scales all logits by a shared temperature, uniformly reducing class-wise logit gaps and thereby flattening the output distribution. In contrast, GKD partitions classes to decouple inter-group probability mass from fine-grained intra-group relations, and applies temperature only across groups. This is equivalent to a group-wise logit shift, where all logits within the same group share the same shift, thus preserving their relative relations and intra-group knowledge.

groups. This decomposition reveals that vanilla KD emphasizes probability-mass dominant regions, while fine-grained structural relations contained in other probability regions receive insufficient supervision. Temperature scaling alleviates this imbalance by increasing the relative contributions of low-probability groups, but simultaneously smooths intra-group probability structures and alters inter-group importance. In particular, at high temperatures, group probability mass increasingly approaches a category-count-based allocation rather than reflecting the teacher's original confidence, an effect that cannot be recovered by the conventional T^2 gradient compensation.

Based on this analysis, we propose Group Knowledge Distillation (GKD), which realizes structured temperature scaling by moving temperature smoothing from the global category distribution to the group level. Specifically, GKD adaptively determines prediction groups based on prediction perplexity and groups categories with comparable probabilities. As illustrated in Fig. 1, GKD applies temperature only to rebalance group-level contributions rather than directly smoothing all category probabilities, thereby strengthening low-probability dark knowledge while preserving fine-grained intra-group structures. Meanwhile, inter-KD retains the original inter-group probability allocation as a regularizing constraint, complementing the local supervision of intra-KD with global distributional information. The intra-KD and inter-KD terms are independently weighted to accommodate different distillation settings.

Our contributions are summarized as follows:

- We derive the first equivalent KL decomposition under general multi-group probability partitioning, separating global probability-mass allocation from fine-grained intra-group knowledge transfer.
- Based on this decomposition, we reveal that conventional temperature scaling simultaneously alters inter-group probability mass and intra-group structures. We introduce structured temperature scaling in GKD, where temperature operates at the group level to rebalance contributions while preserving intra-group structural knowledge.
- We further introduce prediction perplexity as an adaptive grouping strategy to implement structured temperature scaling. Extensive experiments across architectures and tasks demonstrate consistent improvements and stronger preservation of local prediction structures.

2 Related Work

The seminal KD framework Hinton et al. (2015) transfers knowledge from a teacher to a compact student by matching temperature-softened predictions, where probabilities encode informative "dark knowledge". This paradigm has inspired attention transfer, knowledge flow, self-distillation, factor transfer, and information-theoretic distillation Zagoruyko & Komodakis (2017); Yim et al. (2017); Furlanello et al. (2018); Kim et al. (2018); Ahn et al. (2019). Subsequent methods further exploit intermediate representations Romero et al. (2015); Heo et al. (2019), relational structures Park et al. (2019), contrastive representations Tian et al. (2020), and cross-stage feature fusion Chen et al. (2021). While effective, these feature-based approaches often require additional feature alignment or architectural design.

Recent work has revisited the potential of logit-based distillation. DKD Zhao et al. (2022) decouples vanilla KD into target- and non-target-class components, motivating further decoupled formulations. NKD Yang et al. (2023) normalizes non-target logits, while SDD Wei et al. (2024) decomposes global logits into local predictions for finer-grained knowledge transfer. MLKD Jin et al. (2023) extends logit supervision across multiple levels with prediction augmentation. Logit Standardization Sun et al. (2024) shows that matching teacher and student logit scales is unnecessary and uses Z-score normalization to emphasize intrinsic logit relations. More recently, RLD Sun et al. (2025) refines unreliable teacher predictions while preserving class correlations, while LDRLD Xu et al. (2025) recursively constructs dense pairwise relations among ranked logits and adaptively emphasizes critical category pairs. These methods highlight the importance of refining local logit knowledge beyond global distribution matching. However, LDRLD still applies temperature directly to each local distribution, such that smoothing and relational supervision remain coupled. In contrast, GKD is derived from an equivalent multi-group KL decomposition and restricts temperature to regulating group-level knowledge weights while preserving the original intra-group conditional structures. This provides a complementary perspective on logit distillation by explicitly separating probability-mass reallocation from fine-grained structural transfer.

3 Methodology

3.1 Multi-Group Probability Reconstruction of KD

Knowledge Distillation. We first revisit vanilla knowledge distillation (KD) in a C-class classification task. Let $\mathcal{C} = \{1, 2, \ldots, C\}$ denote the category set and $\mathbf{z} = [z_1, z_2, \ldots, z_C]$ the output logits. Given a temperature T, the softened category probability and the vanilla KD objective are respectively defined as

$$p_i = \frac{\exp(z_i/T)}{\sum_{j=1}^{C} \exp(z_j/T)}, \qquad \mathcal{L}_{\mathrm{KD}} = D_{\mathrm{KL}}(\mathbf{p}^t \| \mathbf{p}^s) = \sum_{i \in \mathcal{C}} p_i^t \log \frac{p_i^t}{p_i^s}. \tag{1}$$

The same softmax transformation is applied to the teacher and student logits $\mathbf{z}^t$ and $\mathbf{z}^s$, yielding $\mathbf{p}^t$ and $\mathbf{p}^s$. For notational simplicity, the explicit temperature argument is omitted when it is clear from the context.

Multi-Group Probability Reconstruction. We partition $\mathcal{C}$ into m non-empty and mutually exclusive groups $\{\mathcal{G}_r\}_{r=1}^m$, where $1 \le m \le C$. The group sizes are allowed to vary. For the r-th group, its probability mass is defined as $q_r = \sum_{i \in \mathcal{G}_r} p_i$. Applying this definition to the teacher and student predictions gives $\mathbf{q}^t = [q_1^t, q_2^t, \ldots, q_m^t]$ and $\mathbf{q}^s = [q_1^s, q_2^s, \ldots, q_m^s]$, with $\sum_{r=1}^m q_r^t = \sum_{r=1}^m q_r^s = 1$.

For a category $i \in \mathcal{G}_r$, we further define its normalized intra-group probability as $\hat{p}_{i,r} = p_i / q_r$. The same transformation is applied to the teacher and student predictions, yielding $\hat{\mathbf{p}}_r^t$ and $\hat{\mathbf{p}}_r^s$, respectively. Since $p_i = q_r \hat{p}_{i,r}$ for $i \in \mathcal{G}_r$, Eq. 1 can be reorganized as:

$$\begin{aligned} \mathcal{L}_{\mathrm{KD}} &= \sum_{r=1}^{m} \sum_{i \in \mathcal{G}_r} q_r^t \hat{p}_{i,r}^t \log \frac{q_r^t \hat{p}_{i,r}^t}{q_r^s \hat{p}_{i,r}^s} \\ &= \sum_{r=1}^{m} q_r^t \log \frac{q_r^t}{q_r^s} + \sum_{r=1}^{m} q_r^t \sum_{i \in \mathcal{G}_r} \hat{p}_{i,r}^t \log \frac{\hat{p}_{i,r}^t}{\hat{p}_{i,r}^s}. \end{aligned} \tag{2}$$

Thus, the overall KD loss can be equivalently written as:

$$\mathcal{L}_{\mathrm{KD}} = \underbrace{D_{\mathrm{KL}}\left(\mathbf{q}^t \| \mathbf{q}^s\right)}_{\text{inter-KD}} + \sum_{r=1}^{m} q_r^t \underbrace{D_{\mathrm{KL}}\left(\hat{\mathbf{p}}_r^t \| \hat{\mathbf{p}}_r^s\right)}_{\text{intra-KD}}. \tag{3}$$

This decomposition reveals two distinct levels of knowledge contained in vanilla KD. Inter-KD describes the probability allocation across groups, whereas intra-KD characterizes the fine-grained relative distribution among categories within each group. Notably, each intra-KD term is naturally weighted by the corresponding teacher group probability q_r^t, indicating that the contribution of intra-group knowledge is determined by the probability mass assigned to that group by the teacher.

The decomposition also covers several special cases. When $m = C$, each category forms an individual group, the normalized intra-group probability is one, and only inter-KD remains. When $m = 1$, the group probability mass is one, and the formulation reduces to intra-KD with unit weight. When the categories are partitioned into target and non-target groups, the formulation becomes the binary decomposition underlying DKD.

This reconstruction provides a new perspective for understanding logit-based knowledge distillation. Instead of viewing KD as a single global distribution-matching objective, it can be interpreted as simultaneously matching the probability structure between groups and the conditional probability structure within groups. More importantly, both q_r and $\hat{p}_{i,r}$ are derived from the same temperature-scaled logits. Therefore, a shared temperature inevitably acts on both inter-group probability allocation and intra-group probability structure at the same time. This coupling motivates us to further investigate the limitations of shared temperatures in vanilla KD.

3.2 Drawbacks of Shared Temperatures

Effect of Temperature Smoothing in KD. Based on the reconstructed KD formulation, we further examine how temperature affects inter-KD and intra-KD. For clarity, we consider a simple partition on CIFAR-100, where teacher predictions are sorted in descending order and uniformly divided into ten groups. According to Eq. 3, vanilla KD can be factorized into two complementary components. Intra-KD transfers the conditional probability structure within each group and is weighted by the teacher-assigned group mass q_r^t, while inter-KD preserves the probability-mass allocation across groups that is removed by intra-group normalization.

Since high-probability regions in the teacher prediction naturally carry larger group masses, they contribute more strongly to the distillation objective, whereas low-probability groups receive relatively weak intra-group supervision. Temperature smoothing reduces the probability contrast between high- and low-confidence regions by shrinking their effective logit gaps. Consequently, the relative masses of low-probability groups increase, redistributing the intra-KD weights toward these regions. This reveals an important role of temperature in KD: beyond softening predictions, it rebalances the contribution of different probability regions and exposes otherwise underemphasized dark knowledge.

Table 1: Effect of intra-KD probability-mass weights on CIFAR-100.

Method	Variant	Intra-KD weight	R32×4/R8×4 79.42/72.50	VGG13/VGG8 72.34/69.06	R32×4/SV1 79.42/70.50
KD	vanilla	$q_r^t(T)$	73.33	72.98	74.07
KD	w/o weight T	$q_r^t(1)$	71.79	72.42	73.26
KD	unit weight	1	**75.18**	**74.41**	**76.84**

The results in Table 1 support this analysis. Removing temperature from the intra-KD weights consistently reduces accuracy, showing that the original group masses underemphasize low-probability knowledge. In contrast, assigning unit weight to every group further improves performance by maximally strengthening all intra-group supervision. These results indicate that a major benefit of temperature smoothing comes from redistributing intra-KD weights toward low-probability groups, thereby exposing more dark knowledge. We regard the knowledge contained in these low-probability regions as fine-grained structural knowledge, which captures subtle relations among less

dominant categories. This group-wise perspective not only helps explain why temperature benefits KD, but also provides a new direction for identifying and exploiting dark knowledge more effectively.

Limitations of Shared Temperature. Although temperature smoothing increases the contribution of low-probability knowledge, it simultaneously alters the structural information carried by the teacher distribution. This limitation can be understood from both intra-KD and inter-KD.

For intra-KD, each group preserves a fine-grained conditional distribution among its categories. For two categories with $z_i > z_j$, temperature changes their probability ratio as

$$\frac{p_i(T)}{p_j(T)} = \exp\left(\frac{z_i - z_j}{T}\right). \tag{4}$$

As T increases, the ratio approaches one, directly flattening the relative differences within the group. Therefore, although temperature can increase the weight assigned to a low-probability group, it simultaneously weakens the fine-grained structure being transferred inside that group. Increasing the intra-KD weight alone cannot recover this structural information once it has been smoothed.

For inter-KD, temperature also changes the probability-mass allocation across groups. Let the r-th group contain $|\mathcal{G}_r|$ categories. Its teacher probability mass under temperature scaling is

$$q_r^t(T) = \frac{\sum_{i\in\mathcal{G}_r} \exp(z_i^t/T)}{\sum_{j=1}^{C} \exp(z_j^t/T)}. \tag{5}$$

When $T \to \infty$, $\exp(z_i^t/T) \to 1$, yielding $\lim_{T\to\infty} q_r^t(T) = |\mathcal{G}_r|/C$. Thus, high temperature progressively shifts the inter-group allocation away from the teacher confidence toward a category-count-based prior. For example, the high-confidence target prediction can rapidly approach $1/C$ under strong smoothing, substantially diluting the discriminative supervision originally encoded by the teacher. This distortion alters the relative importance of knowledge across groups and may therefore misguide inter-group knowledge transfer.

The results in Table 2 consistently support the above analyses. Most notably, removing temperature from the intra-group conditional distributions yields the largest improvements across all teacher–student pairs, demonstrating that preserving fine-grained intra-group structures is particularly important for effective knowledge transfer. Applying temperature only at the group level also consistently outperforms vanilla KD, indicating that its beneficial reweighting effect can be retained without uniformly smoothing all category predictions. This observation naturally motivates a distillation framework that uses temperature to rebalance group-level contributions while retaining the original structural knowledge within each group.

Table 2: Analysis of the limitations of shared temperature on CIFAR-100.

Method	Variant	Inter-KD Dist.	Intra-KD Dist.	R32×4/R8×4 79.42/72.50	VGG13/VGG8 72.34/69.06	R32×4/SV1 79.42/70.50
KD	vanilla	$\mathbf{q}^t(T)$	$\hat{\mathbf{p}}^t(T)$	73.33	72.98	74.45
KD	w/o intra T	$\mathbf{q}^t(T)$	$\hat{\mathbf{p}}^t(1)$	**75.65**	**74.03**	**76.63**
KD	group T	$\tilde{\mathbf{q}}^t(T)$	$\hat{\mathbf{p}}^t(T)$	74.60	73.40	75.38

3.3 Group Knowledge Distillation

Based on these findings, vanilla KD can be viewed as a combination of inter-group probability matching and intra-group conditional distribution matching. A shared temperature enhances dark knowledge by redistributing the probability-mass weights of intra-KD, but it also changes the intra-group conditional structure and weakens the original inter-group probability order. To address this issue, we propose structured temperature scaling through Group Knowledge Distillation (GKD), which applies temperature smoothing only to group-level probability masses while preserving the fine-grained relative structure within each group.

Perplexity-Guided Group Partition. A key issue in GKD is how to construct suitable groups for each sample. We use the prediction perplexity of the student to estimate the effective number of activated categories. Given a student probability distribution $\mathbf{p}^s$, the effective category number is defined as

$$n_{\text{eff}}(\mathbf{p}^s) = \exp\left(-\sum_{i\in\mathcal{C}} p_i^s \log p_i^s\right). \tag{6}$$

The value of n_{eff} estimates the effective number of categories supported by the current prediction (Appendix A.4). Sharp predictions yield smaller values, whereas flatter predictions involve more effective categories. For batch-wise parallel grouping, we apply a mild sublinear scaling and use $n_{\text{eff}}^{0.9}$ as the group-size estimate, which reduces abrupt variations in group sizes across samples and leads to more stable partitions within a mini-batch.

Specifically, categories are sorted by the student probabilities, with the target category assigned to an individual group. For the remaining categories, we iteratively compute $n_{\text{eff}}^{0.9}$ over the residual distribution, select the corresponding leading categories as one group, and repeat until all categories are assigned. The detailed procedure is provided in Algorithm 1. This strategy groups categories with comparable response strengths while separating different probability levels. More importantly, the number and sizes of groups are determined adaptively by the prediction itself, eliminating the need to specify the group number for different datasets or tasks.

GKD Objective. Given the group partition $\{\mathcal{G}_r\}_{r=1}^m$, the teacher and student group masses are denoted as $q_r^t = \sum_{i\in\mathcal{G}_r} p_i^t$ and $q_r^s = \sum_{i\in\mathcal{G}_r} p_i^s$. To obtain the inter-group probability allocation, we first construct the equivalent teacher group logit as

$$g_r^t = \log \sum_{i\in\mathcal{G}_r} \exp(z_i^t), \qquad \tilde{q}_r^t(T) = \frac{\exp(g_r^t/T)}{\sum_{k=1}^m \exp(g_k^t/T)}. \tag{7}$$

Thus, GKD applies temperature only to the equivalent group logits, which is equivalent to a group-wise logit shift that preserves the intra-group structure (Appendix A.2). The intra-group conditional probabilities are kept unchanged and defined as $\hat{p}_{i,r}^t = p_i^t/q_r^t$ and $\hat{p}_{i,r}^s = p_i^s/q_r^s$ for $i \in \mathcal{G}_r$. The GKD loss is formulated as

$$\begin{aligned}\mathcal{L}_{\text{GKD}} &= \alpha D_{\text{KL}}\left(\mathbf{q}^t \| \mathbf{q}^s\right) + \beta \sum_{r=1}^m \tilde{q}_r^t(T) D_{\text{KL}}\left(\hat{\mathbf{p}}_r^t \| \hat{\mathbf{p}}_r^s\right) \\ &= \alpha\mathcal{L}_{\text{inter}} + \beta \sum_r \tilde{q}_r^t(T)\mathcal{L}_{\text{intra},r}.\end{aligned} \tag{8}$$

where α and β control inter-KD and intra-KD, respectively, and the overall objective is $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{GKD}}$. Compared with the reconstructed KD objective, GKD replaces the original intra-KD weight q_r^t with its temperature-smoothed group mass $\tilde{q}_r^t(T)$, while retaining the original inter-KD supervision. For $T > 1$, this smoothing reduces probability gaps among groups without changing their relative order, thereby increasing the contribution of low-probability groups and exposing more dark knowledge. Meanwhile, the intra-group conditional distributions remain unchanged, preserving fine-grained relations within each group, while inter-KD preserves the global probability allocation lost by intra-group normalization. The pseudo-code is provided in Algorithm 2.

4 Experiments

Datasets. We evaluate GKD on CIFAR-100 Krizhevsky (2009) and ImageNet-1K Deng et al. (2009). CIFAR-100 contains 50K training and 10K test images from 100 classes of 32×32, while ImageNet-1K contains 1.28M training and 50K validation images from 1,000 classes.

Baselines. We compare GKD with representative logit-based methods, including KD Hinton et al. (2015), DKD Zhao et al. (2022), MLKD Jin et al. (2023), RLD Sun et al. (2025) and LDRLD Xu et al. (2025), as well as feature-based methods including RKD Park et al. (2019), CRD Tian et al. (2020), OFD Heo et al. (2019), ReviewKD Chen et al. (2021), and CAT-KD Guo et al. (2023).

Table 3: Top-1 accuracy (%) on CIFAR-100 with heterogeneous architectures. GKD* denotes GKD with the prediction augmentation of MLKD.

Teacher	ResNet32×4	WRN-40-2	VGG13	ResNet50	ResNet32×4
	79.42	75.61	74.64	79.34	79.42
Student	ShuffleNet-V1	ShuffleNet-V1	MobileNet-V2	MobileNet-V2	ShuffleNet-V2
	70.50	70.50	64.60	64.60	71.82
RKD	72.28	72.21	64.52	64.43	73.21
CRD	75.11	76.05	69.73	69.11	75.65
OFD	75.98	75.85	69.48	69.04	76.82
ReviewKD	77.45	77.14	70.37	69.89	77.78
CAT-KD	78.26	77.35	69.13	71.36	78.41
KD	74.07	74.83	67.37	67.35	74.45
DKD	76.45	76.70	69.71	70.35	77.07
MLKD	77.18	77.44	70.57	71.04	78.44
RLD	–	–	69.97	70.76	77.56
LDRLD	76.46	77.09	70.11	70.74	77.33
GKD	77.29	77.31	70.39	71.15	78.24
GKD*	**78.73**	**77.81**	**70.92**	**72.68**	**79.18**

Implementation Details. We follow the training protocol of DKD Zhao et al. (2022). On CIFAR-100, models are optimized with SGD for 240 epochs. The initial learning rate is 0.01 for MobileNet/ShuffleNet and 0.05 for the other architectures, and is divided by 10 at epochs 150, 180, and 210. ImageNet experiments follow the DKD setting. All results are averaged over three runs. For fair comparison, GKD* denotes GKD equipped with the same prediction augmentation strategy as MLKD Jin et al. (2023). Training and hyperparameter details are provided in the appendix.

4.1 Main Results

Results on CIFAR-100. Tables 3 and 4 report results under heterogeneous and homogeneous teacher–student architectures, respectively. Without additional prediction augmentation, GKD already achieves leading or highly competitive performance across most settings, demonstrating the effectiveness of group-wise knowledge transfer itself. When equipped with the same prediction augmentation as MLKD, GKD* achieves the best accuracy across all evaluated CIFAR-100 settings. The gains are consistent across both heterogeneous and homogeneous architectures, and are particularly pronounced for heterogeneous teacher–student pairs with larger architectural gaps. Incorporating prediction augmentation further improves GKD by approximately 0.5–1.0 percentage points in most settings, indicating that GKD enables the student to exploit richer knowledge from stronger teachers more effectively.

Results on ImageNet. Table 5 further evaluates GKD on large-scale ImageNet. GKD consistently achieves state-of-the-art performance under both homogeneous (ResNet34/ResNet18) and heterogeneous (ResNet50/MobileNet-V1) teacher–student architectures, demonstrating its effectiveness on large-scale and challenging datasets.

Ablation on α and β. We study the GKD weights α and β on ResNet32×4/ResNet8×4 while keeping all other settings fixed. As shown in Table 6, removing either component degrades performance, while the larger drop at $\beta = 0$ highlights the critical role of fine-grained intra-group structural knowledge in effective distillation. Since GKD does not use the conventional T^2 gradient compensation, relatively larger coefficients are generally required to maintain the distillation strength. The performance remains stable over a broad range, with α and β around 8 and 32 giving the best results; we therefore use $(\alpha, \beta) = (8, 32)$ as the default setting.

4.2 Extensions

Comparison of Grouping Strategies. We evaluate how different grouping strategies affect GKD while keeping the inter-KD and intra-KD objectives unchanged. Specifically, we compare random grouping, sequential grouping with fixed group numbers, and the proposed prediction-perplexity-

Table 4: Top-1 accuracy (%) on CIFAR-100 with homogeneous architectures. GKD* denotes GKD with the prediction augmentation of MLKD.

Teacher	ResNet56	ResNet110	ResNet32×4	WRN-40-2	WRN-40-2	VGG13
	72.34	74.31	79.42	75.61	75.61	74.64
Student	ResNet20	ResNet32	ResNet8×4	WRN-16-2	WRN-40-1	VGG8
	69.06	71.14	72.50	73.26	71.98	70.36
RKD	69.61	71.82	71.90	73.35	72.22	71.48
CRD	71.16	73.48	75.51	75.48	74.14	73.94
OFD	70.98	73.23	74.95	75.24	74.33	73.95
ReviewKD	71.89	73.89	75.63	76.12	75.09	74.84
CAT-KD	71.62	73.62	76.91	75.60	74.82	74.65
KD	70.66	73.08	73.33	74.92	73.54	72.98
DKD	71.97	74.11	76.32	76.24	74.81	74.68
MLKD	72.19	74.11	77.08	76.63	75.35	75.18
RLD	72.00	74.02	76.64	–	74.88	74.93
LDRLD	72.20	74.16	77.20	76.35	74.98	75.06
GKD	72.04	74.25	77.55	76.16	75.29	75.15
GKD*	**72.27**	**74.57**	**78.25**	**76.68**	**75.50**	**75.56**

Table 5: Top-1 and Top-5 accuracy (%) on ImageNet. Best and second-best results are highlighted in **bold** and underlined, respectively.

T/S	Metric	Baseline	Feature-based			Logit-based						
			CRD	ReviewKD	CATKD	KD	DKD	MLKD	RLD	LDRLD	GKD	GKD*
R34/R18	Top-1	73.31/69.75	71.17	71.61	71.26	71.03	71.70	71.90	71.91	71.88	72.06	**72.14**
	Top-5	91.42/89.07	90.13	90.51	90.45	90.05	90.41	90.55	90.59	90.58	90.58	**90.77**
R50/MV1	Top-1	76.16/68.87	71.37	72.56	72.24	70.50	72.05	73.01	72.75	73.12	72.86	**73.16**
	Top-5	92.87/88.76	90.41	91.00	91.13	89.80	91.05	91.42	91.18	91.43	91.51	**91.54**

Table 6: Sensitivity to α and β on CIFAR-100 using ResNet32×4/ResNet8×4.

Coefficient	0	1	4	8	16	Coefficient	0	1	4	16	32
α ($\beta = 16$)	75.85	76.27	77.06	77.12	76.19	β ($\alpha = 4$)	73.59	74.71	75.93	77.06	77.16
α ($\beta = 32$)	76.16	76.59	77.16	**77.55**	76.96	β ($\alpha = 8$)	73.98	74.22	75.65	77.12	**77.55**

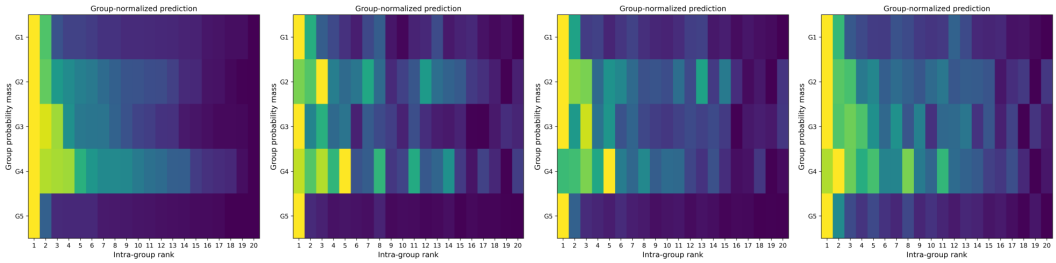
based grouping. For prediction-dependent strategies, both teacher-guided and student-guided partitions are considered. This comparison aims to examine whether the gain of GKD mainly comes from the grouped formulation itself or from the proposed adaptive partition.

As shown in Table 7, the optimal fixed group number varies across teacher–student pairs, whereas perplexity-based grouping consistently achieves the best performance without introducing an additional group-number hyperparameter. Interestingly, random grouping also yields competitive results. Although it does not preserve semantic or probability-rank structure, splitting the prediction into multiple groups still allows low-probability categories to receive more explicit intra-group supervision than in global KD. This supports our main motivation that the benefit of grouping does not necessarily rely on semantic clustering; rather, grouping provides multiple local spaces in which fine-grained knowledge can be aligned at different probability levels.

Student-guided grouping consistently outperforms teacher-guided grouping. Unlike the fixed teacher partition, the student partition evolves throughout training and therefore reflects the student's current uncertainty and learning state. Categories that the student incorrectly assigns high probabilities are more likely to be grouped with other high-response categories, where intra-KD provides stronger local supervision to correct these mistakes. In contrast, a teacher-guided partition remains fixed and may not align with the student's current errors. These results suggest that adaptive grouping is most effective when it follows the evolving student prediction, while prediction perplexity further removes the need to manually determine the number of groups.

Table 7: Comparison of different grouping strategies and grouping guidance on CIFAR-100. T-guided and S-guided denote partitions constructed from teacher and student predictions.

T/S	Random ($K = 10$)	Sequential ($K = 10$)		Sequential ($K = 5$)		Perplexity (N_{eff})	
		T-guided	S-guided	T-guided	S-guided	T-guided	S-guided
R32×4/R8×4	77.26	75.70	76.52	76.97	<u>77.35</u>	76.73	**77.55**
R32×4/SV2	77.20	75.73	<u>77.91</u>	76.47	77.40	75.81	**78.24**



(a) Teacher (b) KD (c) DKD (d) GKD

Figure 2: Visualization of intra-group prediction structures on CIFAR-100. The 100 classes are randomly split into five groups of 20 with the same grouping for all methods. Closer agreement with the teacher within each group reflects better preservation of local structural knowledge.

Local Structure Preservation. To further examine the structural information transferred by different distillation methods, we visualize teacher–student predictions under random class grouping in Fig. 2. Specifically, the classes are randomly partitioned into several equal-sized groups, independently of the grouping used during training. Within each group, teacher probabilities are sorted in descending order, and the same class ordering is applied to all student models. We then independently normalize each group to $[0, 1]$, removing the influence of absolute probability magnitude and highlighting the relative intra-group structure.

As shown in Fig. 2, KD and DKD exhibit more dispersed responses, with frequent shifts of dominant positions and secondary peaks. In contrast, GKD preserves the teacher-induced descending patterns and dominant positions more consistently across randomly constructed groups. Since the random partition is shared by all methods and is unrelated to GKD training, this comparison provides a fair probe of the learned prediction structure. The results support our motivation that, by reducing the influence of temperature on intra-group knowledge transfer, GKD better preserves fine-grained structural relations within local probability regions rather than simply enforcing stronger global probability matching.

5 Conclusion

In this work, we investigate how temperature should participate in knowledge distillation. Our central view is that the essential dark knowledge exposed by KD is reflected in the relative predictive structures encoded by the teacher, namely the conditional relations among categories after separating their overall probability mass. From this perspective, the key role of temperature is to rebalance the contribution of these structures to knowledge transfer, rather than to smooth the structures themselves. We propose an equivalent multi-group decomposition of vanilla KD and, from this perspective, reveal both the benefits and limitations of temperature scaling. We further introduce GKD, which organizes prediction knowledge using prediction perplexity and restricts temperature to group-level regulation while preserving intra-group structures and inter-group supervision. With a simple modification to vanilla KD, GKD achieves state-of-the-art performance across diverse experimental settings, demonstrating consistent effectiveness. Beyond GKD itself, the proposed decomposition provides a useful perspective for other learning paradigms based on distribution matching, such as incremental learning and related objectives involving KL divergence.

AI USE STATEMENT

The authors acknowledge the use of large language model (LLM) technology in the preparation of this manuscript. Specifically, an LLM was used to assist with language refinement, clarity improvement, and prose polishing in certain sections.

REFERENCES

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9155–9163, 2019. doi: 10.1109/CVPR.2019.00938.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10925–10934, 2022.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5006–5015, 2021. doi: 10.1109/CVPR46437.2021.00497.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616. PMLR, 10–15 Jul 2018.

Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11868–11877, 2023. doi: 10.1109/CVPR52729.2023.01142.

Jitai Hao, Qiang Huang, Hao Liu, Xinyan Xiao, Zhaochun Ren, and Jun Yu. A token is worth over 1,000 tokens: Efficient knowledge distillation through low-rank clone. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 53678–53710. Curran Associates, Inc., 2025. doi: 10.52202/085713-1791.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1921–1930, 2019. doi: 10.1109/ICCV.2019.00201.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24276–24285, 2023. doi: 10.1109/CVPR52729.2023.02325.

Jangho Kim, Seonguk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Qizhen Lan and Qing Tian. Acam-kd: Adaptive and cooperative attention masking for knowledge distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3957–3966, 2025. doi: 10.1109/ICCV51701.2025.00377.

Chenxin Li, Mingbao Lin, Zhiyuan Ding, Nie Lin, Yihong Zhuang, Yue Huang, Xinghao Ding, and Liujuan Cao. Knowledge condensation distillation. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2022. doi: 10.1007/978-3-031-20083-0_2.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1504–1512, 2023.

Hou-I Liu, Christine Wu, Jen-Hao Cheng, Wenhao Chai, Shian-Yun Wang, Gaowen Liu, Hugo Latapie, Jhih-Ciang Wu, Jenq-Neng Hwang, Hong-Han Shuai, and Wen-Huang Cheng. Monotakd: Teaching assistant knowledge distillation for monocular 3d object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22266–22275, 2025. doi: 10.1109/CVPR52734.2025.02074.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971, 2019. doi: 10.1109/CVPR.2019.00409.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740, 2024. doi: 10.1109/CVPR52733.2024.01489.

Wujie Sun, Defang Chen, Siwei Lyu, Genlang Chen, Chun Chen, and Can Wang. Knowledge distillation with refined logits. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1110–1119, 2025. doi: 10.1109/ICCV51701.2025.00111.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4928–4937, 2019.

Shicai Wei, Chunbo Luo, and Yang Luo. Scale decoupled distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15975–15983, 2024. doi: 10.1109/CVPR52733.2024.01512.

Guodong Xu, Ziwei Liu, and Chen Change Loy. Computation-efficient knowledge distillation via uncertainty-aware mixup. *Pattern Recognition*, 138:109338, 2023. doi: 10.1016/j.patcog.2023.109338.

Liuchi Xu, Kang Liu, Jinshuai Liu, Lu Wang, Lisheng Xu, and Jun Cheng. Local dense logit relations for enhanced knowledge distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4539–4549, 2025. doi: 10.1109/ICCV51701.2025.00432.

Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17139–17148, 2023. doi: 10.1109/ICCV51070.2023.01576.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11943–11952, 2022. doi: 10.1109/CVPR52688.2022.01165.

A APPENDIX

A.1 DETAILS ABOUT THE REFORMULATION IN SEC 3.1

We provide the detailed derivation of the multi-group reconstruction used in the main paper. Consider an arbitrary partition of the category set $\mathcal{C}$ into m non-empty and mutually exclusive groups $\{\mathcal{G}_r\}_{r=1}^m$, where $1 \le m \le C$. For each group, its probability mass is defined as $q_r = \sum_{i\in\mathcal{G}_r} p_i$. Applying this definition to the teacher and student predictions yields $\mathbf{q}^t = [q_1^t, \dots, q_m^t]$ and $\mathbf{q}^s = [q_1^s, \dots, q_m^s]$, both normalized over the m groups.

For each category $i \in \mathcal{G}_r$, we define the normalized intra-group probability as $\hat{p}_{i,r} = p_i/q_r$. Accordingly, $\sum_{i\in\mathcal{G}_r} \hat{p}_{i,r} = 1$, and the original category probability can be exactly reconstructed as $p_i = q_r \hat{p}_{i,r}$. The same transformation gives the teacher and student intra-group distributions $\hat{\mathbf{p}}_r^t$ and $\hat{\mathbf{p}}_r^s$.

Starting from vanilla KD, the summation over all categories can first be reorganized according to the group partition:

$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} &= \sum_{r=1}^{m} \sum_{i\in\mathcal{G}_r} p_i^t \log \frac{p_i^t}{p_i^s} \\ &= \sum_{r=1}^{m} \sum_{i\in\mathcal{G}_r} q_r^t \hat{p}_{i,r}^t \log \frac{q_r^t \hat{p}_{i,r}^t}{q_r^s \hat{p}_{i,r}^s}.\end{aligned} \tag{9}$$

Separating the group-level and intra-group terms gives

$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} = &\sum_{r=1}^{m} q_r^t \log \frac{q_r^t}{q_r^s} \left(\sum_{i\in\mathcal{G}_r} \hat{p}_{i,r}^t \right) \\ &+ \sum_{r=1}^{m} q_r^t \sum_{i\in\mathcal{G}_r} \hat{p}_{i,r}^t \log \frac{\hat{p}_{i,r}^t}{\hat{p}_{i,r}^s}.\end{aligned} \tag{10}$$

Since the intra-group probabilities are normalized, the first inner summation equals one. Therefore,

$$\mathcal{L}_{\mathrm{KD}} = \underbrace{D_{\mathrm{KL}}(\mathbf{q}^t \| \mathbf{q}^s)}_{\text{inter-KD}} + \sum_{r=1}^{m} q_r^t \underbrace{D_{\mathrm{KL}}(\hat{\mathbf{p}}_r^t \| \hat{\mathbf{p}}_r^s)}_{\text{intra-KD}}. \tag{11}$$

The first term matches the probability-mass allocation across groups, while the second transfers the conditional probability structure within each group. Importantly, q_r^t naturally serves as the weight of the corresponding intra-group knowledge, and these weights satisfy $\sum_{r=1}^m q_r^t = 1$. The decomposition is exact for any valid partition and therefore introduces no information loss.

A.2 EQUIVALENCE OF GROUP TEMPERATURE SMOOTHING AND GROUP LOGIT SHIFT

To further clarify how temperature operates at the group level, we show that applying temperature smoothing to the equivalent group logits is exactly equivalent to applying a uniform logit shift to all categories within each group. This equivalence provides a logit-level interpretation of our formulation: temperature changes the relative probability allocation across groups without modifying the original logit differences, and hence the structural knowledge, within each group. It also establishes the equivalence between the explicitly grouped representation and a corresponding full-class logit transformation. Since the derivation applies identically to the teacher and student, the corresponding superscripts are omitted for simplicity.

$$g_r = \log \sum_{i\in\mathcal{G}_r} \exp(z_i), \qquad q_r(T) = \frac{\exp(g_r/T)}{\sum_{k=1}^m \exp(g_k/T)}. \tag{12}$$

Temperature scaling changes the group logit from g_r to g_r/T. We therefore define the corresponding group-wise logit shift as

$$\Delta_r(T) = g_r - \frac{g_r}{T} = \left(1 - \frac{1}{T}\right) g_r, \qquad \tilde{z}_i = z_i - \Delta_r(T), \quad i \in \mathcal{G}_r. \tag{13}$$

Since the same shift $\Delta_r(T)$ is applied to all categories within $\mathcal{G}_r$, the equivalent logit of the shifted group becomes

$$\log \sum_{i \in \mathcal{G}_r} \exp(\tilde{z}_i) = \log \left[\exp(-\Delta_r(T)) \sum_{i \in \mathcal{G}_r} \exp(z_i) \right] = g_r - \Delta_r(T) = \frac{g_r}{T}. \tag{14}$$

Therefore, applying temperature smoothing to the equivalent group logits is exactly equivalent to subtracting the same group-specific shift from all category logits within each group.

Importantly, such a uniform shift does not alter the intra-group structure. For any $i, j \in \mathcal{G}_r$, $\tilde{z}_i - \tilde{z}_j = z_i - z_j$, and consequently

$$\frac{\exp(\tilde{z}_i)}{\sum_{j \in \mathcal{G}_r} \exp(\tilde{z}_j)} = \frac{\exp(z_i)}{\sum_{j \in \mathcal{G}_r} \exp(z_j)} = \hat{p}_{i,r}. \tag{15}$$

Hence, group-level temperature smoothing only changes the relative probability allocation among groups while preserving the original probability structure within each group.

Moreover, Eq. 13 provides the equivalent shifted logit for every category. Let $r(i)$ denote the group containing category i; the complete transformed logits can be written as $\tilde{z}_i = z_i - \Delta_{r(i)}(T)$. The resulting full-class probability distribution satisfies:

$$\tilde{p}_i = \frac{\exp(\tilde{z}_i)}{\sum_{j \in \mathcal{C}} \exp(\tilde{z}_j)} = q_{r(i)}(T) \hat{p}_{i,r(i)} \tag{16}$$

Thus, the explicitly grouped representation and the full-class representation based on the shifted logits are mathematically identical. In practice, the distillation objective can therefore be computed either by explicitly constructing the group probabilities and intra-group distributions or directly from the complete shifted logits. We adopt the grouped formulation because it provides a clearer interpretation of inter- and intra-group knowledge and is more convenient for implementation.

A.3 Two Equivalent Implementations of Group-Level Temperature Smoothing

In GKD, temperature smoothing is applied to group-level probability allocation rather than individual categories. It can be implemented in two equivalent ways. The first explicitly constructs equivalent group logits:

$$g_r = \log \sum_{i \in \mathcal{G}_r} \exp(z_i), \qquad q_r(T) = \frac{\exp(g_r/T)}{\sum_k \exp(g_k/T)}. \tag{17}$$

Alternatively, the same operation can be directly performed on the original group probability masses. Since

$$q_r(1) = \frac{\exp(g_r)}{\sum_k \exp(g_k)}, \tag{18}$$

we have

$$q_r(T) = \frac{q_r(1)^{1/T}}{\sum_k q_k(1)^{1/T}}. \tag{19}$$

Therefore, temperature scaling on equivalent group logits is exactly equivalent to applying a power transformation to the original group probability masses followed by renormalization.

In our Algorithm 2, we adopt the first formulation by explicitly computing equivalent group logits, which is consistent with conventional temperature scaling. In practice, the second formulation is more efficient since it directly transforms group probabilities and avoids repeated log-sum-exp operations over categories, requiring fewer category-level traversals.

A.4 Details about the Perplexity in Sec 3.3

In this part, we provide the theoretical justification for prediction perplexity from two perspectives: first, why N_{eff} can be interpreted as the effective number of categories carrying distillation information; and second, why selecting the leading top-k probabilities is a reasonable strategy for identifying these informative categories.

Effective category number. We provide the theoretical motivation for using prediction perplexity to estimate the number of informative categories. For a normalized probability distribution $\mathbf{q} = [q_1, \ldots, q_k]$, its entropy and effective category number are defined as

$$H(\mathbf{q}) = -\sum_{i=1}^{k} q_i \log q_i, \qquad N_{\text{eff}}(\mathbf{q}) = \exp(H(\mathbf{q})). \tag{20}$$

Let $\mathbf{u}_k = [1/k, \ldots, 1/k]$ denote the uniform distribution over k categories. The entropy can be equivalently written as

$$H(\mathbf{q}) = \log k - D_{\text{KL}}(\mathbf{q}\|\mathbf{u}_k), \qquad N_{\text{eff}}(\mathbf{q}) = k \exp[-D_{\text{KL}}(\mathbf{q}\|\mathbf{u}_k)] \leq k. \tag{21}$$

Hence, N_{eff} measures the size of an equiprobable distribution having the same entropy as $\mathbf{q}$. In particular, when the k probabilities are equal, $D_{\text{KL}}(\mathbf{q}\|\mathbf{u}_k) = 0$ and $N_{\text{eff}}(\mathbf{q}) = k$; when they are only approximately equal, $N_{\text{eff}}(\mathbf{q}) \approx k$. Therefore, prediction perplexity naturally quantifies how many categories effectively contribute comparable information.

Top-k category selection. We next explain why the effective categories are selected from the high-probability head. Let the complete prediction $\mathbf{p}$ be sorted as $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(C)}$, and define the probability mass of its top-k categories as $m_k = \sum_{i=1}^{k} p_{(i)}$. The normalized head and residual distributions are respectively defined by $q_i = p_{(i)}/m_k$ and $r_i = p_{(k+i)}/(1 - m_k)$. The entropy of the complete prediction admits the exact decomposition

$$H(\mathbf{p}) = h(m_k) + m_k H(\mathbf{q}) + (1 - m_k) H(\mathbf{r}), \tag{22}$$

where $h(m_k) = -m_k \log m_k - (1 - m_k) \log(1 - m_k)$ is the binary entropy between the selected head and the remaining categories. Exponentiating both sides gives

$$N_{\text{eff}}(\mathbf{p}) = e^{h(m_k)} N_{\text{eff}}(\mathbf{q})^{m_k} N_{\text{eff}}(\mathbf{r})^{1-m_k}. \tag{23}$$

Equation 23 shows that the influence of the residual distribution is weighted by its remaining probability mass $1 - m_k$. As the selected head captures most of the prediction mass, i.e., $m_k \to 1$, the contribution of the residual categories vanishes and $N_{\text{eff}}(\mathbf{p})$ approaches $N_{\text{eff}}(\mathbf{q})$. If the head further contains k categories with comparable probabilities, Eq. 21 gives $N_{\text{eff}}(\mathbf{q}) \approx k$. Therefore, prediction perplexity provides a natural estimate of the number of dominant categories carrying effective distillation information.

This also motivates selecting categories in descending order of probability. For a fixed k, the top-k categories retain the largest probability mass m_k, thereby suppressing the contribution of the residual distribution in Eq. 23. Meanwhile, selecting the leading probabilities in descending order naturally favors a subset whose members are both relatively large and close in magnitude, making the normalized head more balanced and allowing $N_{\text{eff}}(\mathbf{q})$ to better reflect the number of selected categories. Thus, sorting and selecting the high-probability head provides a simple way to identify the dominant informative categories.

In practice, diffuse tail probabilities can slightly inflate the perplexity estimate. We therefore apply a mild sublinear scaling, N_{eff}^{γ}, to obtain a more conservative and smoother estimate of the effective category number. This is particularly useful for batch-wise parallel grouping, as it reduces abrupt variations in group sizes across samples and yields more stable partitions within a mini-batch. We empirically set $\gamma = 0.9$; this choice has little influence on the final accuracy and is mainly adopted for stable and efficient batch-wise implementation rather than performance optimization.

A.5 Algorithm Implementation

For completeness, we summarize the grouping and optimization procedures of GKD in a PyTorch-like style. The algorithms are presented for a single sample for clarity, while the actual implementation performs the same operations in parallel over a mini-batch. The target category forms an individual group, while prediction perplexity is recursively applied to the remaining categories. Here, $H(\cdot)$ denotes Shannon entropy.

For mini-batch implementation, although the prediction distributions vary across samples, the resulting number of groups is generally stable for a fixed model. We therefore perform grouping

recursively in parallel and use the first sample that reaches the stopping condition to determine the group number for the current mini-batch. For samples whose grouping has not yet terminated, all remaining categories are assigned to the final group, as the probability differences in these tail regions are already relatively small. This strategy maintains efficient batch-wise parallelization while avoiding excessive partitioning of low-probability tail categories and the redundant noise it may introduce.

Algorithm 1 Prediction-Perplexity Grouping

Require: Student logits $\mathbf{z}^s$, target y
Ensure: Groups $\mathcal{G}$

$\mathbf{p} = \mathrm{softmax}(\mathbf{z}^s)$
$\mathrm{idx} = \mathrm{argsort}(\mathbf{p}, \mathrm{descending} = \mathrm{True})$
$\mathcal{G} = [\{y\}]$; $\mathcal{R} = \mathrm{idx}[\mathrm{idx} \neq y]$
while $|\mathcal{R}| > 0$ **do**
  $\mathbf{p}' = \mathbf{p}[\mathcal{R}]/\mathbf{p}[\mathcal{R}].\mathrm{sum}()$
  $k = \lceil \exp(H(\mathbf{p}'))^{0.9} \rceil$
  $\mathcal{G}.\mathrm{append}(\mathcal{R}[:k])$
  $\mathcal{R} = \mathcal{R}[k:]$
end while
return $\mathcal{G}$

Algorithm 2 Group Knowledge Distillation

Require: Student logits $\mathbf{z}^s$, teacher logits $\mathbf{z}^t$, groups $\mathcal{G} = \{\mathcal{G}_r\}_{r=1}^m$, group temperature T, hyper-parameters α, β
Ensure: $\mathcal{L}_{\mathrm{GKD}}$

$\mathbf{p}^s = F.\mathrm{softmax}(\mathbf{z}^s)$; $\mathbf{p}^t = F.\mathrm{softmax}(\mathbf{z}^t)$
$\mathbf{q}^s = [\,\mathbf{p}^s[\mathcal{G}_r].\mathrm{sum}()$ for $\mathcal{G}_r$ in $\mathcal{G}\,]$
$\mathbf{q}^t = [\,\mathbf{p}^t[\mathcal{G}_r].\mathrm{sum}()$ for $\mathcal{G}_r$ in $\mathcal{G}\,]$
$\mathbf{g}^t = [\,\mathrm{logsumexp}(\mathbf{z}^t[\mathcal{G}_r])$ for $\mathcal{G}_r$ in $\mathcal{G}\,]$
$\tilde{\mathbf{q}}^t(T) = F.\mathrm{softmax}(\mathbf{g}^t/T)$
$\mathcal{L}_{\mathrm{inter}} = F.\mathrm{kl_div}(\log \mathbf{q}^s, \mathbf{q}^t)$
$\mathcal{L}_{\mathrm{intra}} = 0$
for $r = 1$ **to** m **do**
  $\hat{\mathbf{p}}_r^s = \mathbf{p}^s[\mathcal{G}_r]/\mathbf{p}^s[\mathcal{G}_r].\mathrm{sum}()$
  $\hat{\mathbf{p}}_r^t = \mathbf{p}^t[\mathcal{G}_r]/\mathbf{p}^t[\mathcal{G}_r].\mathrm{sum}()$
  $\mathcal{L}_{\mathrm{intra}} \mathrel{+}= \tilde{q}_r^t(T)\, F.\mathrm{kl_div}(\log \hat{\mathbf{p}}_r^s, \hat{\mathbf{p}}_r^t)$
end for
$\mathcal{L}_{\mathrm{GKD}} = \alpha\, \mathcal{L}_{\mathrm{inter}} + \beta\, \mathcal{L}_{\mathrm{intra}}$
return $\mathcal{L}_{\mathrm{GKD}}$

To facilitate understanding of the grouping procedure, Fig. 3 provides a simple example of prediction-perplexity grouping. The target category is first isolated as an individual group. At each subsequent iteration, the remaining probabilities are conditionally renormalized, and the effective category number is recomputed to determine the next group size. This procedure is repeated until all categories are assigned, resulting in three groups in this example. Their probability masses are 0.70, 0.285, and 0.015, which are further smoothed across groups with $T = 4$. In this way, structured temperature scaling operates only on group-level weighting, while the relative probability structure within each group is preserved.

A.6 Implementation: Experiments in Sec 3.2

Basic setup. All ablation experiments follow the CIFAR-100 training protocol of DKD Zhao et al. (2022). We train each student for 240 epochs using SGD. The initial learning rate is 0.01 for ShuffleNet and 0.05 for the other architectures, and is decayed by a factor of 0.1 at epochs 150, 180, and 210. Unless otherwise specified, the distillation temperature is fixed to $T = 4$. Following the standard KD configuration, the cross-entropy and distillation losses are weighted by 0.1 and 0.9, respectively.

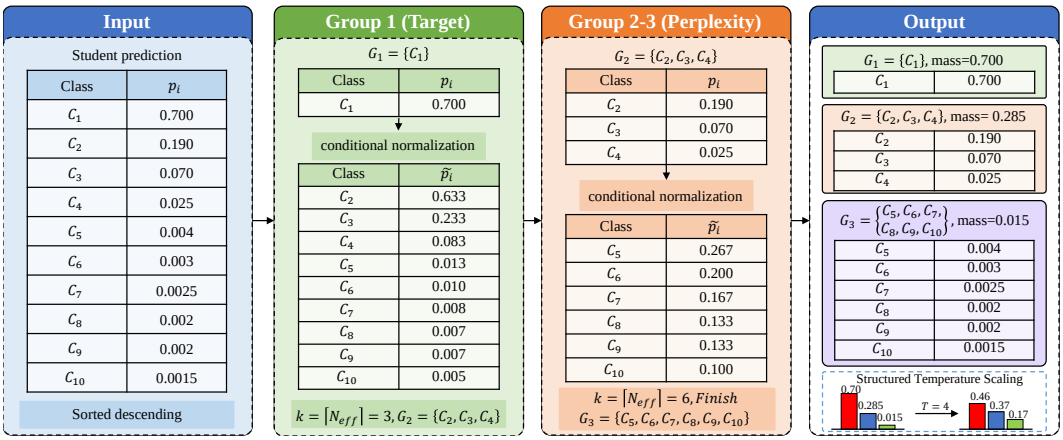


Figure 3: Toy example of prediction-perplexity grouping. By iteratively estimating the effective category number over the conditionally normalized residual distribution, the predictions are partitioned into three groups with clear inter-group differences and relatively consistent intra-group probability levels.

Grouping setup. To isolate the effect of temperature from the proposed perplexity-based grouping, we reconstruct KD using a fixed sequential partition with $m = 10$ groups. The target category forms an individual group, while the remaining categories are ordered by prediction probability and sequentially assigned to the other groups. The same partitioning strategy is used for all variants below.

Intra-KD weight ablation. Table 1 examines whether temperature mainly benefits KD by reallocating the weights of intra-group knowledge. The vanilla reconstruction uses the temperature-smoothed teacher group masses $q_r^t(T)$. In *w/o weight* T, only the intra-KD weights are replaced by $q_r^t(1)$, removing temperature smoothing from the group weights while keeping the remaining loss unchanged. In *unit weight*, every intra-group term is assigned a weight of 1 without normalization, giving equal emphasis to the dark knowledge contained in all groups.

Shared-temperature ablation. Table 2 separately examines the effects of temperature on intra- and inter-group knowledge. For *w/o intra* T, the conditional distributions of both teacher and student within each group are computed with $T = 1$, while the original temperature-smoothed group weights and inter-KD supervision are retained. This removes intra-group smoothing while preserving temperature-induced reweighting. For *group* T, we replace $q_r^t(T)$, where temperature is applied before grouping, with a distribution obtained by applying temperature directly to the grouped representation. Specifically, we first compute the equivalent group logits from the original teacher logits and then apply $T = 4$ at the group level to obtain $\tilde{\mathbf{q}}^t(T)$. The intra-KD term is kept unchanged. This setting preserves the original relative structure among groups before smoothing and isolates the effect of temperature on inter-group probability allocation.

A.7 Implementation: Experiments in Sec 4

Training protocol. We follow the training protocol of DKD Zhao et al. (2022). **On CIFAR-100**, all students are trained for 240 epochs with a batch size of 64 using SGD with momentum 0.9 and weight decay 5×10^{-4}. The initial learning rate is 0.01 for MobileNet and ShuffleNet and 0.05 for the other architectures, and is divided by 10 at epochs 150, 180, and 210. **On ImageNet**, models are trained for 100 epochs with a batch size of 512, an initial learning rate of 0.2, momentum 0.9, and weight decay 10^{-4}; the learning rate is divided by 10 at epochs 30, 60, and 90.

GKD configuration. Unless otherwise specified, we set $(\alpha, \beta) = (8, 32)$ on CIFAR-100 and $(1, 2)$ on ImageNet. These coefficients can be adjusted according to the teacher–student gap, with detailed tuning guidance provided in Appendix A.8. The category partition is determined adaptively by prediction perplexity, which jointly determines the number and sizes of groups; therefore, GKD introduces no predefined group-number hyperparameter.

Table 8: Representative hyperparameter settings for GKD on CIFAR-100. $\Delta_{\mathrm{T/S}}$ denotes the Top-1 accuracy gap between teacher and student.

T/S	Teacher	Student	$\Delta_{\mathrm{T/S}}$	α	β	Test Acc.	Train Acc.
ResNet32×4 ResNet8×4	79.42	72.50	6.92	4	32	77.16	94.01
				8	4	75.65	98.88
				8	32	**77.55**	96.29
ResNet56 ResNet20	72.34	69.06	3.28	4	16	**72.25**	82.38
				8	32	71.57	80.36
ResNet110 ResNet32	74.31	71.14	3.17	4	16	**74.27**	87.67
				8	32	73.57	87.50
WRN-40-2 WRN-16-2	75.61	73.26	2.35	4	16	**76.57**	91.60
				8	32	76.29	91.26

Prediction augmentation. To separate the gain of GKD itself from that of stronger input augmentation, we report both GKD and GKD*. GKD* adopts the weak/strong augmentation protocol used in the official MLKD implementation Jin et al. (2023). The weak view uses standard random cropping and horizontal flipping, while the strong view additionally applies RandAugment followed by Cutout. Previous studies have shown that augmentation strategies such as Mixup and related perturbation-based approaches can further improve knowledge distillation by increasing training diversity and enhancing the robustness of transferred knowledge Beyer et al. (2022); Xu et al. (2023); Li et al. (2022). For GKD*, both the CE and GKD losses are computed on the weak and strong views, and the corresponding losses are averaged across the two views to maintain a comparable optimization scale with the non-augmented setting. Except for this augmentation branch, GKD and GKD* use identical training and hyperparameter settings. Unless explicitly specified as GKD*, all experiments in this paper use GKD without additional augmentation.

A.8 Guidance for Tuning α and β

GKD decomposes distillation into inter-KD and intra-KD, where inter-KD preserves global probability allocation across groups and intra-KD transfers fine-grained structure within each group. Their effects on optimization are also different: stronger inter-KD generally reinforces global prediction alignment and improves training accuracy, whereas excessive intra-KD may suppress training accuracy by imposing stronger constraints on fine-grained teacher structure. Therefore, an appropriate balance between α and β is important for effective knowledge transfer.

The preferred loss scale is also related to the teacher–student gap, particularly when their recognition abilities are close. In this case, the student already captures much of the discriminative information available from the teacher, making detailed intra-group matching less informative while introducing additional constraints that may compete with the label-based CE objective. Reducing the overall GKD strength therefore allows greater flexibility to fit the ground-truth labels. In contrast, when the teacher is substantially stronger, its richer prediction structure provides more transferable knowledge, making stronger distillation beneficial.

As shown in Table 8, $(\alpha, \beta) = (8, 32)$ performs best for R32×4/R8×4, which has a relatively large accuracy gap. For pairs with smaller gaps, including R56/R20, R110/R32, and WRN-40-2/WRN-16-2, reducing the weights to $(4, 16)$ consistently yields better results. These observations suggest using stronger GKD supervision for larger teacher–student gaps and milder weighting when their recognition abilities are similar.

A.9 More Analysis

Object Detection Results on MS-COCO2017. We further evaluate GKD on the MS-COCO2017 object detection benchmark using Faster R-CNN with FPN. Following the detection implementation of DKD Zhao et al. (2022), distillation is applied to the classification logits of the RoI head, while the bounding-box regression branch remains unchanged. We retain the original DKD training pipeline and hyperparameters, with $\alpha = 1$ and $\beta = 1$.

Table 9: Object detection results on MS-COCO2017. The teacher and student detectors are based on Faster R-CNN with FPN.

Method	R-101 → R-18			R-101 → R-50		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Teacher	42.04	62.48	45.88	42.04	62.48	45.88
Student	33.26	53.61	35.26	37.93	58.84	41.05
FitNet Romero et al. (2015)	34.13	54.16	36.71	38.76	59.62	41.80
FGFI Wang et al. (2019)	35.44	55.51	38.17	39.44	60.27	43.04
ReviewKD Chen et al. (2021)	36.75	56.72	34.00	40.36	60.97	44.08
KD Hinton et al. (2015)	33.97	54.66	36.62	38.35	59.41	41.71
CTKD Li et al. (2023)	34.56	55.43	36.91	–	–	–
DKD Zhao et al. (2022)	34.88	56.16	37.08	39.01	60.41	42.33
LDRLD Xu et al. (2025)	35.12	**56.79**	37.57	**39.31**	61.06	42.56
GKD (Ours)	**35.20**	56.62	**37.73**	39.28	**61.20**	**42.75**

Table 10: Training efficiency comparison on CIFAR-100 using a ResNet32×4 teacher and ResNet8×4 student on a GeForce RTX 4090. The training time denotes the average time per epoch, and the memory usage is measured during training.

Method	Time (s/epoch) ↓	Memory (MB) ↓	Top-1 Acc. (%) ↑
CRD	16.84	3211	75.51
OFD	8.65	2337	74.95
ReviewKD	13.76	3407	75.63
CAT-KD	6.53	2131	76.91
KD	6.42	2131	73.33
DKD	6.48	2131	76.32
RLD	6.65	2131	76.64
MLKD	14.70	2377	77.08
LDRLD	9.20	2131	77.20
GKD	7.45	2131	**77.55**

As shown in Table 9, GKD consistently outperforms DKD and remains competitive with the recent LDRLD method. Notably, GKD achieves the best AP_{75} in both settings, indicating stronger localization performance and effective generalization of GKD to object detection.

Training Efficiency Analysis. Table 10 compares the training efficiency of different distillation methods on CIFAR-100. Since all methods share the same student architecture during inference, the comparison focuses on the additional training overhead introduced by different distillation strategies. GKD maintains the same GPU memory consumption as other logit-based distillation methods while introducing only a minor increase in training time. Meanwhile, it achieves higher accuracy than existing approaches, demonstrating a favorable trade-off between effectiveness and training efficiency.

Semantic Structure of Prediction-Induced Groups. As shown in Fig. 4, the high-probability groups exhibit clear semantic consistency with the target. For example, the *stone wall* sample is associated with outdoor and structural categories, while the *macaque* sample is mainly grouped with other primates and visually related mammals. In contrast, low-probability tail groups contain substantially weaker correlations, such as *pizza* or *airliner*. The probability-mass changes also illustrate how group-level smoothing redistributes supervision from dominant groups toward lower-probability knowledge. Notably, our grouping criterion is derived purely from prediction probabilities rather than semantic labels, but the resulting groups still exhibit meaningful clustering characteristics. This observation further supports prediction-guided grouping as a practical way to organize different levels of distillation knowledge.

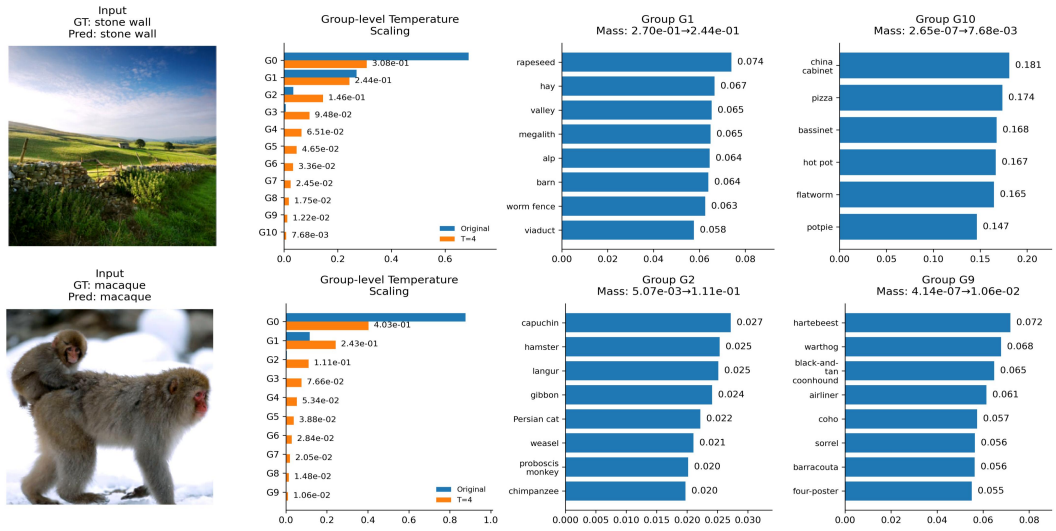
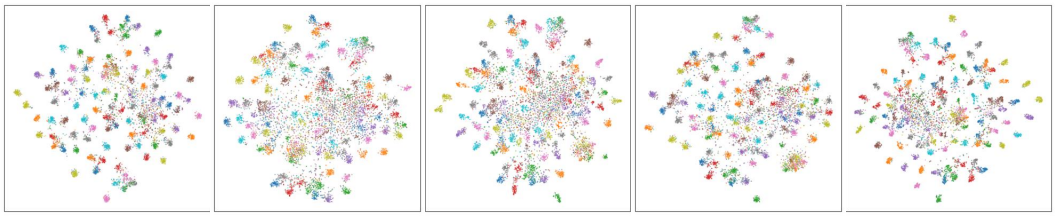


Figure 4: Visualization of prediction-induced groups on ImageNet. For each example, we show the group probability masses before and after temperature smoothing, together with representative Top-8 predictions from the head and tail groups.



(a) Teacher R32×4 (b) KD R8×4 (c) GKD R8×4 (d) KD SV1 (e) GKD SV1

Figure 5: The t-SNE visualization of feature representations on CIFAR-100. ResNet32×4 serves as the teacher, with ResNet8×4 and ShuffleNet-V1 as students.

Feature Visualization. To qualitatively compare the representations learned by different distillation methods, we visualize the deep features of CIFAR-100 test samples using t-SNE. As shown in Fig. 5, we compare KD and GKD under two teacher–student settings, where ResNet32×4 serves as the teacher and ResNet8×4 or ShuffleNet-V1 serves as the student. Compared with KD, GKD produces more compact and separable feature clusters, indicating that the proposed GKD helps the student learn more discriminative representations.

A.10 Limitations and Future Work

Several aspects remain worth further investigation. First, although α and β control complementary inter- and intra-group supervision, their optimal relationship is not yet theoretically determined; we currently provide practical tuning guidance in Appendix A.8. Developing an adaptive or theoretically grounded balancing strategy would further reduce manual tuning. Second, the improvement of GKD is relatively limited when the teacher and student have similar capacities. We conjecture that, in such cases, emphasizing teacher predictions may provide less additional knowledge and can compete with the label-based CE supervision. Future work will investigate teacher–student-gap-aware weighting and adaptive coordination between distillation and label supervision.