KNOWLEDGE DISTILLATION AS CONDITIONAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation (KD) typically treats the teacher representation as a deterministic target, overlooking the inherent uncertainty of knowledge transfer. In this paper, we reformulate KD as a conditional generative problem and propose *Conditional Generation for Knowledge Distillation* (*CondKD*), which models the conditional distribution of teacher representations given the student representation. However, directly applying conditional generation to KD faces two challenges: the curse of high-dimensional optimization and the lack of semantic supervision from labels. To address these issues, we introduce a *Split Tokenization (SplitTok)* strategy, achieving stable and effective unsupervised KD. Additionally, we develop the *Distribution Contraction* technique to integrate label supervision into the conditional reconstruction process. Our theoretical analysis shows that *Distribution Contraction* provides a stable gradient surrogate for multi-task learning, enabling efficient supervised training without explicit classification loss on multi-step sampling image representations. With such a simplified and unified objective, our training eliminates heavy hyperparameter tuning and effectively scales to strong training schedules. Experiments demonstrate that *CondKD* improves over the KL baseline by 16.29% under unsupervised KD on CC3M. With label supervision, our method achieves 82.36% top-1 accuracy with ResNet-50 on ImageNet in 600 epochs without extra data, establishing a new state-of-the-art.

1 INTRODUCTION

For natural language tasks, both inputs and outputs reside in the same domain, *i.e.*, language sequences, enabling the unification of diverse tasks within a single generative model optimized via next-token prediction. ChatGPT and GPT4V (Ouyang et al., 2022; Achiam et al., 2023) exemplify this approach with data scaling law and are often regarded as an early prototype of artificial general intelligence (AGI), showcasing the effectiveness of generative learning in natural language. Motivated by this success, researchers have begun extending generative modeling to vision and multimodal domains (Liu et al., 2023; Li et al., 2024; Tian et al., 2024; Zhou et al., 2024; Fan et al., 2025; Wu et al., 2024; Yang et al., 2025), with the long-term goal of building AGI systems.

Two prominent classes of generative models have gained popularity in the vision domain: autoregressive models (Li et al., 2024; Tian et al., 2024) and diffusion models (Ho et al., 2020; Yang et al., 2025). Autoregressive models adopt the next-token prediction paradigm to sequentially generate image content, whereas diffusion models transform images into Gaussian noise through a forward diffusion process and learn to recover them via a reverse denoising process. In this paper, we recast knowledge distillation (KD) (Hinton et al., 2015), typically formulated as a discriminative task minimizing the KL divergence between categorical output distributions of the teacher and student, as a conditional generative problem modeled with the diffusion mechanism.

Our Motivation: Uncertainty in KD. Knowledge Distillation (KD) (Hinton et al., 2015) has been widely adopted for knowledge transfer and model compression. Existing approaches typically formulate distillation as *consistency between prediction and target*: the student is trained to match a deterministic teacher target, either by minimizing the KL divergence between teacher and student output distributions (Hinton et al., 2015; Zhao et al., 2022; Cui et al., 2024a; Lv et al., 2024) or the Mean Squared Error (MSE) between intermediate-layer features (Chen et al., 2021; Romero et al., 2015; Park et al., 2019; Tian et al., 2020; Heo et al., 2019). While effective, these objectives primar-

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.

ily measure the discrepancy between paired teacher and student representations, implicitly treating the teacher representation as a deterministic target for each input.

However, the relationship between student and teacher representations can be inherently uncertain. For a given sample, the student's representation may provide a highly predictive description of the teacher's knowledge, while for another sample the correspondence may be less predictable. Moreover, when the teacher and student differ substantially in capacity or architecture, a given student representation may be compatible with multiple plausible teacher representations. Thus, rather than viewing knowledge transfer as a deterministic mapping, it is more natural to characterize it through a conditional distribution of teacher knowledge given the student's representation. Existing KD objectives quantify the discrepancy to the observed teacher target, but do not explicitly model this conditional uncertainty.

Conditional Generation for KD (CondKD). These observations motivate us to formulate knowledge distillation as *conditional generation* rather than deterministic target matching. Drawing inspiration from the success of generative learning (Li et al., 2024; Ho et al., 2020; Song et al., 2020b), we model the conditional distribution of teacher representations given the student's representation. As illustrated in Figure 1, instead of directly forcing student representation $\mathcal{F}_s(x)$ to match the teacher representation $\mathcal{F}_t(x)$, we train a diffusion model to generate $\mathcal{F}_t(x)$ conditioned on $\mathcal{F}_s(x)$. The resulting conditional likelihood provides a distributional characterization of teacher knowledge that can be inferred from the student representation, allowing the model to capture variations in the predictability of teacher knowledge across samples. In this way, *CondKD* explicitly models the *conditional uncertainty* and *one-to-many nature* of the student-to-teacher mapping, going beyond deterministic target matching.

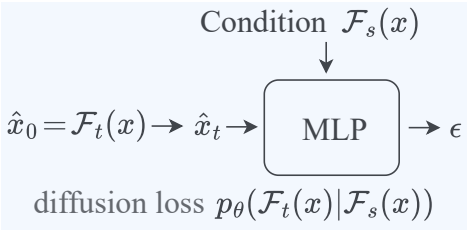


Figure 1: Conditional generation for KD.

Challenges of CondKD and Solutions. MAR (Li et al., 2024) deploys a diffusion loss for autoregressive image generation. Specifically, images are tokenized in continuous token sequences via VAE (Kingma et al., 2013) and then fed into autoregressive models. However, image tokens from VAE only have a dimension of 16, while image representations in classification often have a large dimension, reaching 2048. We empirically observe *the high-dimensional optimization disaster: the training can't converge or even crashes*. Moreover, diffusion models are optimized by a variational lower bound (VLB) to reconstruct inputs, which *lacks semantic constraints with labels and thus hinders model performance*.

To tackle the high-dimensional optimization challenge, we propose the *Split Tokenization (SplitTok)*: split the image representation $\mathcal{F}_t(x)$ into token sequences with positional indices. Conditioned on $\mathcal{F}_s(x)$, models are trained to reconstruct these tokens individually. Such a *SplitTok* operation effectively stabilizes the training of *CondKD*, achieving unsupervised KD. To enable label supervision of *CondKD*, we develop a *Distribution Contraction* technique. We theoretically prove that *CondKD* with *Distribution Contraction* serves as a more stable surrogate to multi-task learning (combining reconstruction and classification loss), eliminating explicit classification loss and multi-step sampling and thus leading to efficient and effective supervised KD. To avoid time-consuming sampling of diffusion models at inference, the linear probing scheme is deployed for practicality.

Experimental Results. To validate the effectiveness of our method, we conduct experiments on balanced data including CIFAR-100 (Krizhevsky & Hinton, 2009) and ImageNet (Russakovsky et al., 2015), imbalanced data like long-tailed ImageNet (Liu et al., 2019), and unlabeled data CC3M (Changpinyo et al., 2021). With the *SplitTok*, our *CondKD* model significantly surpasses the KL baseline by **16.29%** for the unsupervised KD. Moreover, the simplified and unified objective enables our method to scale effectively under strong training schedules (unlike previous methods). With the *Distribution Contraction* technique for label supervision, our approach not only significantly outperforms existing KD methods, but also achieves a **new state-of-the-art** performance for ResNet-50 on ImageNet with 600 epochs of training, without requiring additional data.

In summary, our contributions are as follows:

Table 1: Comparison with previous distillation methods.

Method	Gen. or Dis.	Multi-task	Uncertainty Modeling	Scaling
Logits-based	Discriminative	Yes	No	Hard
Feature-based	Discriminative	Yes	No	Hard
CondKD	**Generative**	**No (Unified Gen. objective)**	**Yes**	**Easy**

- We formulate KD as a conditional generation problem and propose the Conditional Generation for KD (*CondKD*), a unified generative objective that scales well on strong training settings and enables unsupervised KD.
- To address the high-dimensional optimization challenge, we propose a *SplitTok* strategy. To enable label supervision of *CondKD*, the *Distribution Contraction* technique is developed.
- We theoretically prove that *CondKD* with *Distribution Contraction* acts as a more stable gradient-level surrogate for multi-task learning, resulting in efficient and effective training.
- We empirically show the advantage of our method on balanced, imbalanced, and unlabeled data. Specifically, our method scales well to strong training schemes and achieves the state-of-the-art performance on ImageNet for both unsupervised and supervised KD.

2 Related Work

Generative Learning. In NLP, generative modeling based on next-token prediction (NTP) forms the foundation of the GPT series of language models (Radford et al., 2018; 2019; Brown et al., 2020; Ouyang et al., 2022; Achiam et al., 2023). This approach enables unsupervised learning from large-scale corpora and has driven significant advances in zero-shot and few-shot generalization, culminating in powerful systems such as ChatGPT and GPT4V (Ouyang et al., 2022; Achiam et al., 2023) that exhibit strong performance across diverse natural language tasks without task-specific fine-tuning. Subsequently, autoregressive (AR) models leveraging NTP have also gained popularity in vision (Li et al., 2024; Tian et al., 2024) and multi-modal (Liu et al., 2023; Zhu et al., 2023) domains, fostering the development of unified generalist models capable of both understanding and generation (Zhou et al., 2024; Fan et al., 2025; Yang et al., 2025) on multi-modal data.

Besides NTP and autoregressive (AR) models, diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2020b) have emerged as a powerful class of generative methods, demonstrating impressive sample quality and robustness (Rombach et al., 2022; Yang et al., 2025; Zhou et al., 2024). However, they often require multi-step iterative sampling during inference, which can be computationally expensive and time-consuming. Recently, flow matching approaches (Geng et al., 2025; Lipman et al., 2022; Gat et al., 2024) have been proposed to address these limitations by providing efficient and scalable generative modeling with fewer sampling steps while maintaining high fidelity.

Knowledge Distillation (KD). Knowledge distillation (KD) (Hinton et al., 2015) transfers knowledge from a teacher to a student using the teacher's soft predictions as supervision. Following its success, extensive studies have developed logit-based (Furlanello et al., 2018; Zhang et al., 2018; Cho & Hariharan, 2019; Huang et al., 2022; Zhao et al., 2022; Hao et al., 2023; Cui et al., 2024a; 2025; Sun et al., 2024; Beyer et al., 2022; Jin et al., 2023; Chen et al., 2022) and feature-based (Romero et al., 2015; Park et al., 2019; Tian et al., 2020; Heo et al., 2019; Chen et al., 2021; Huang et al., 2023; Yang et al., 2022; Wei et al., 2024) distillation methods. However, these approaches typically treat teacher supervision as deterministic and do not explicitly characterize the uncertainty of knowledge transfer. Because of the multi-task nature, they can also be sensitive to loss weighting and require careful hyperparameter tuning across teacher–student configurations (Hao et al., 2023). In contrast, *CondKD* formulates distillation as conditional generation, explicitly modeling the conditional distribution of teacher representations given the student representation. This formulation captures the uncertainty of knowledge transfer and provides a unified generative objective. See Table 1 for a comparison with existing methods.

3 CONDITIONAL GENERATION FOR KNOWLEDGE DISTILLATION (CONDKD)

3.1 PROBLEM FORMULATION

A typical classification model consists of a feature extractor $\mathcal{F}(\cdot)$ and a classifier $\mathcal{C}(\cdot)$. Given an input image x, it produces a feature representation $\mathcal{F}(x)$ and prediction $\mathcal{C}(\mathcal{F}(x))$. Given a pretrained teacher $(\mathcal{F}_t, \mathcal{C}_t)$, knowledge distillation aims to transfer its knowledge to a student $(\mathcal{F}_s, \mathcal{C}_s)$, which may differ in architecture or capacity.

Conventional KD treats $\mathcal{F}_t(x)$ as a deterministic target given $\mathcal{F}_s(x)$. However, the relationship between teacher and student representations can be inherently uncertain. In particular, due to differences in model capacity and architecture, a given student representation may be compatible with multiple plausible teacher representations. Thus, knowledge transfer need not follow a deterministic mapping. Motivated by this observation, we formulate KD as a conditional distribution,

$$p_\theta \left(\mathcal{F}_t(x) \mid \mathcal{F}_s(x) \right), \tag{1}$$

which explicitly captures the uncertainty of teacher knowledge given the student's representation and provides a natural basis for modeling knowledge transfer beyond deterministic target matching.

3.2 CONDKD WITH UNLABELED DATA

To model the conditional distribution in equation 1, we formulate knowledge distillation as conditional generation. Given an image $x \in X$, $\hat{x}_0 \sim q(\mathcal{F}_t(x))$, taking $\mathcal{F}_s(x)$ as condition, we learn to reconstruct the image representation of teacher model, *i.e.*, $\hat{x}_0$, with the following training objective,

$$\mathbb{E}_{x,m,\epsilon} \left[||\epsilon - \epsilon_\theta(\hat{x}_m, m, \mathcal{F}_s(x))||^2 \right], \tag{2}$$

where $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$, $m \in [0, M]$ is the sampled time step (M is the maximum), $\hat{x}_m = \sqrt{\bar{\alpha}_m}\hat{x}_0 + \sqrt{1-\bar{\alpha}_m}\epsilon$ is the noisy input at time step m, in particular, $\hat{x}_0 = \mathcal{F}_t(x)$, $\bar{\alpha}_m = \Pi_{i=1}^{m}\alpha_i$, and α is defined with a variance schedule (Ho et al., 2020; Nichol & Dhariwal, 2021).

At inference, with an input image x and a sampled $\hat{x}_M^{'} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$, the image representation $\hat{x}_0^{'}$ could be generated through iterative update from $m = M$ to $m = 0$:

$$\hat{x}_{m-1}^{'} = \frac{1}{\sqrt{\alpha_m}} \left(\hat{x}_m^{'} - \frac{1-\alpha_m}{\sqrt{1-\bar{\alpha}_m}} \epsilon_\theta(\hat{x}_m^{'}, m, \mathcal{F}_s(x)) \right) + \sigma_m \epsilon, \tag{3}$$

where $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$, σ_m could be learned or pre-defined. Note that, the reverse diffusion process could be respaced (Li et al., 2024; Song et al., 2020a) for efficient sampling. Then, the final prediction could be derived by inputting the image representation $\hat{x}_0^{'}$ into teacher model classifier, *i.e.*, $\arg\max \mathcal{C}_t(\hat{x}_0^{'})$.

High-dimensional Optimization Disaster. Following MAR (Li et al., 2024), we implement the diffusion head using a 3-layer MLP. MAR (Li et al., 2024) showcases the effectiveness of continuous tokenizers for autoregressive image generation, where images are first tokenized into sequences of continuous tokens using a VAE (Kingma et al., 2013). These tokens are then fed into an autoregressive model that learns the per-token distribution through a diffusion loss.

However, the dimensionality of each token in VAE (Kingma et al., 2013) is limited to 16, whereas feature representations in image classification tasks typically have much higher dimensionality, reaching up to 2048. Our empirical study reveals a high-dimensional optimization issue, particularly when the feature dimension of the student model, $\mathrm{Dim}(\mathcal{F}_s(x))$, is much lower than that of the teacher model, $\mathrm{Dim}(\mathcal{F}_t(x))$, often leading to training instability or failure. Refer to Section 4.4.

Split Tokenization (SplitTok). To address the challenges of high-dimensional optimization, we propose decomposing the feature representation $\hat{x}_0$ into a sequence of non-overlapped lower-dimensional tokens. Specifically, we define the *SplitTok* operation as:

$$\begin{aligned} \textit{SplitTok}(\mathcal{F}_t(x)) \quad &= [(\hat{x}_0^1, 1, \mathcal{F}_s(x)), (\hat{x}_0^2, 2, \mathcal{F}_s(x)), \\ &\quad \ldots, (\hat{x}_0^n, n, \mathcal{F}_s(x))], \end{aligned} \tag{4}$$

where each tuple consists of a token $\hat{x}_0^i$, its position index i, and the conditioning context from the student model $\mathcal{F}_s(x)$. Based on this structure, we reformulate the training objective in equation 2 into a token-wise form:

$$\mathbb{E}_{y,\mathrm{id},c,m,\epsilon} \left[||\epsilon - \epsilon_\theta(\hat{y}_m, m, \mathrm{id}, c)||^2 \right], \tag{5}$$

where $(y, \text{id}, c) \sim q\left(\textit{SplitTok}(\mathcal{F}_t(x))[\text{id}]\right)$. This token-based formulation allows the model to operate in lower-dimensional subspaces, thereby mitigating instability during optimization in high-dimensional feature spaces.

3.3 CondKD with Label Supervision

Conditioned on the student's feature representation, *CondKD* reconstructs the teacher's feature tokens, enabling unsupervised knowledge distillation. However, the reconstruction objective alone fails to exploit label supervision during training. To address this problem, we introduce a *Distribution Contraction* mechanism that enables *CondKD* to effectively incorporate label information into the optimization process.

Multi-task Learning. To incorporate label supervision, a straightforward baseline is multi-task learning, which combines the reconstruction objective with a standard cross-entropy loss:

$$\begin{aligned} &\min_{\theta_s} \mathcal{L}_{\text{CE}} = -y \log \mathcal{C}_s(\hat{x}_0^{'}), \\ &\text{s.t.} \min_{\theta} \mathbb{E}_{y,\text{id},c,m,\epsilon} \left[|\epsilon - \epsilon_\theta(\hat{y}_m, m, \text{id}, c)|^2 \right], \end{aligned} \tag{6}$$

where $\mathcal{C}_s$ denotes the classifier on top of the reconstructed representation $\hat{x}_0^{'}$, $\theta = (\theta_s, \theta_{diff})$ are parameters of student model and diffusion head respectively. As shown in Eq. equation 6, the cross-entropy loss encourages the generated feature $\hat{x}_0^{'}$ to be correctly classified, while the reconstruction loss regularizes the student feature space to align with that of the teacher. These two objectives can be optimized either alternately or simultaneously.

However, in practice, involving $\hat{x}_0^{'}$ directly in the training of the diffusion model is computationally inefficient. Since $\hat{x}_0^{'}$ must be sampled through a multi-step reverse diffusion process at each iteration, using it as an intermediate target for supervision substantially increases training time and resource consumption. Moreover, gradients cannot be efficiently propagated through the sampling chain, limiting the effectiveness of end-to-end optimization.

CondKD with Label Supervision. Instead of relying on conventional multi-task learning, we incorporate label supervision through the proposed *Distribution Contraction* technique, formally defined in Definition 1. Furthermore, Theorem 1 establishes that *CondKD*, trained with the *Distribution Contraction* technique, serves as an efficient and effective surrogate for multi-task learning.

Definition 1 (Distribution Contraction) Let $\hat{x}_0 \sim q(\mathcal{F}_t(x))$ be the feature representation produced by a well-trained teacher model $\mathcal{F}_t$ on input $x \in X$, where $\mathcal{C}_t(\hat{x}_0) \in \mathbb{R}^C$ is the logits vector over C categories. To incorporate label supervision y during diffusion model training, we enhance the semantic consistency of $\hat{x}_0$ by contracting it toward the class center c_y:

$$\tilde{x}_0 = \lambda \hat{x}_0 + (1 - \lambda) c_y, \tag{7}$$

where $\lambda \in [0, 1]$ controls the degree of contraction, c_y denotes the centroid of features for class y. Note that, we use the frozen weights of $\mathcal{C}_t$ as c_y in Eq. 7.

Theorem 1 (CondKD as a Teacher-Guided Surrogate of Multi-task Gradients) *Assume a fixed teacher model composed of a feature extractor $\mathcal{F}_t(\cdot)$ and a linear classifier $\mathcal{C}_t(z) = Wz$, where $W \in \mathbb{R}^{C \times d}$ and $c_y = W_y$.*

Under the single-step estimator of $\hat{x}_0^{'}$, the gradients of the CondKD objective:

$$\mathcal{L}_{CondKD} = \mathbb{E}_{x,m,\epsilon} \left[\|\epsilon - \epsilon_\theta(\tilde{x}_m, m, \mathcal{F}_s(x))\|^2 \right], \tag{8}$$

and the multi-task objective:

$$\mathcal{L}_{multi} = \gamma_0 \, \mathbb{E}_{x,m,\epsilon} \left[\|\epsilon - \epsilon_\theta(\hat{x}_m, m, \mathcal{F}_s(x))\|^2 \right] + \gamma_1 \, \mathbb{E}_x \left[\mathcal{L}_{\text{CE}} \left(\mathcal{C}_t(\hat{x}_0'), y \right) \right], \tag{9}$$

satisfy the relation:

$$\nabla_{\hat{x}_0'} \mathcal{L}_{CondKD} = \nabla_{\hat{x}_0'} \mathcal{L}_{multi} + \gamma_1 \big(\hat{x}_0 - \hat{x}_0^{''} \big), \;\; \hat{x}_0^{''} = \sum_{i=1}^{C} p(i|\hat{x}_0^{'}) W_i. \tag{10}$$

Furthermore:

- ***(Local approximation)*** *If* $\|\hat{x}_0^{''} - \hat{x}_0\| \leq \delta$*, then*
$$\|\nabla \mathcal{L}_{CondKD} - \nabla \mathcal{L}_{multi}\| \leq \gamma_1 \delta.$$
- ***(High-confidence regime)*** *If* $p(y|\hat{x}_0^{'}) \to 1$*, then* $\hat{x}_0^{''} \to c_y$*, and the multi-task classification gradient vanishes, while CondKD introduces a non-vanishing contraction term* $(\hat{x}_0 - c_y)$*.*
- ***(Uncertain regime)*** *When* $\hat{x}_0^{''}$ *deviates from* $\hat{x}_0$*, CondKD replaces the prediction-dependent direction* $(\hat{x}_0^{''} - c_y)$ *with a teacher-anchored direction* $(\hat{x}_0 - c_y)$*.*

Therefore, CondKD can be interpreted as a teacher-guided surrogate of the multi-task gradient, which matches it locally and provides a more stable optimization direction when the classifier predictions are uncertain or saturated.

Proof. See Appendix A.1.

Remark. Theorem 1 shows that $\mathcal{L}_{\text{CondKD}}$ serves as a teacher-guided surrogate for the multi-task objective $\mathcal{L}_{\text{multi}}$ at the gradient level. Specifically, *CondKD* replaces the prediction-dependent classification gradient with a teacher-anchored direction, yielding a close approximation to the multi-task gradient when classifier predictions are reliable, and introducing a stable class-directed contraction when the cross-entropy gradient diminishes. This distinction becomes particularly important in the uncertain regime, where the prediction-derived direction $\hat{x}_0^{''}$ deviates from the teacher feature $\hat{x}_0$. Such deviation typically arises when the estimated feature $\hat{x}_0^{'}$ is ambiguous or inaccurate, causing the classifier to assign probability mass to multiple competing classes. Consequently, the multi-task gradient becomes highly dependent on the predicted class probabilities and may vary significantly under prediction uncertainty. In contrast, *CondKD* maintains a stable optimization direction through the teacher-anchored contraction term $(\hat{x}_0 - c_y)$. As a result, *CondKD* avoids explicit optimization of the classification loss and eliminates the need for multi-step sampling to obtain $\hat{x}_0^{'}$ during training, leading to improved computational efficiency while preserving effective label supervision.

Linear Probing Scheme for Speeding Up Inference. To avoid the inference-time latency introduced by iterative samplings of the diffusion head, we train a linear classifier on top of the frozen representations learned by *CondKD*. Interestingly, the performance of this lightweight linear classifier matches that of the diffusion head, demonstrating that *CondKD* has successfully learned good representations via knowledge transfer from the teachers.

4 Experiments

Section 4.1 presents the competitive performance of *CondKD* under the unsupervised KD setting. When label supervision is incorporated via *Distribution Contraction*, *CondKD* achieves strong results on both balanced and imbalanced datasets, as shown in Sections 4.2 and 4.3. Finally, we perform ablation studies in Section 4.4 to assess the impact of the proposed *Split Tokenization (SplitTok)* and *Distribution Contraction* techniques, and inference efficiency. More analysis of the hyperparameter sensitivity is included in Table 7 of the Appendix.

Experimental Settings. Following prior work (Li et al., 2024), we implement the diffusion head using a 3-layer MLP. For training, the maximum diffusion step is set to $M = 1000$. At inference, the diffusion head will be removed, and the learned linear classifier is used in the same manner as conventional KD. For *SplitTok*, the feature representation $\hat{x}_0$ is divided into non-overlapping tokens, each with a dimensionality of 64.

For the unsupervised KD setting, we evaluate *CondKD* on the target dataset, *i.e.*, ImageNet (Deng et al., 2009), and non-target dataset CC3M (Changpinyo et al., 2021). Under the supervised KD setting, we train various teacher-student configurations on balanced (including ImageNet (Deng et al., 2009) and CIFAR (Krizhevsky et al., 2009)) and imbalanced data.

4.1 CondKD in Unsupervised Setting

To evaluate the effectiveness of *CondKD* for unsupervised knowledge distillation (KD), we train models on both target data (ImageNet) and non-target data (CC3M (Changpinyo et al., 2021)), and assess their performance on the ImageNet validation set. The results are summarized in Table 3.

Table 2: **Top-1 accuracy (%) on the ImageNet with supervised CondKD**. Note that LSKD combines DKD without AutoAug for fair comparison. *CondKD* scales much better than DKD (logits-based) and ReviewKD (feature-based) methods with strong training schedules that are used in work (Hao et al., 2023).

	Teacher	Student	Discriminative Point-wise Distillation							Gen.D.D.
			SD-KD	CRD	ReviewKD	DKD*	LSKD	IKL-KD*	KD*	**CondKD**
ResNet-34, *ResNet-18*, *Regular recipe*, *100 epochs*										
Top-1	73.31	69.75	71.44	71.17	71.61	71.87	71.88	71.91	71.24	**72.40**
Top-5	91.42	89.07	90.05	90.13	90.51	90.45	90.58	90.52	90.23	**90.63**
ResNet-50, *MobileNet*, *Regular recipe*, *100 epochs*										
Top-1	76.16	68.87	72.24	71.37	72.56	72.55	72.85	73.19	71.44	**73.76**
Top-5	92.86	88.76	90.71	90.41	91.00	91.05	91.23	91.47	90.35	**91.56**
BEiTv2, *ResNet-50*, *Strong recipe*, *300 (A2) or 600 (A1) epochs*										
Top-1 (BEiT-L-A2)	88.01	79.80	-	79.48	79.11	80.77	-	80.98	80.89	**81.79**
Top-1 (BEiT-B-A2)	86.12	79.80	-	-	-	-	-	-	80.96	**81.96**
Top-1 (BEiT-L-A1)	88.01	80.38	-	-	-	81.83	-	-	81.68	**82.36**

Table 3: **Top-1 accuracy(%) on ImageNet with unsupervised CondKD.**

Method	Teacher	Student	Accuracy
w/o Label On Target Data, *i.e.*, ImageNet-1K			
KL	ResNet-50	MobileNet	71.40
CondKD	ResNet-50	MobileNet	**72.03**
w/o Label On Non-target Data, *i.e.*, CC3M			
KL	ResNet-50	MobileNet	51.60
CondKD	ResNet-34	ResNet-18	**66.90**
CondKD	ResNet-50	MobileNet	**67.89**

Table 4: **Top-1 accuracy(%) on ImageNet-LT with CondKD.** "*" represents the unsupervised setting.

Method	Teacher	Student	Accuracy
Baseline	-	ResNet-18	41.15
Baseline	-	ResNet-50	45.47
KD	ResNeXt-101	ResNet-18	44.32
KD	ResNeXt-101	ResNet-50	48.31
IKL-KD	ResNeXt-101	ResNet-18	45.21
IKL-KD	ResNeXt-101	ResNet-50	49.29
CondKD*	ResNext-101	ResNet-18	**45.54**
CondKD*	ResNeXt-101	ResNet-50	**49.31**

For teacher-student configurations such as (ResNet-34, ResNet-18) and (ResNet-50, MobileNet), we adopt pre-trained teacher models from PyTorch. Since these teachers have been trained on the target dataset, their predictions closely approximate the ground-truth, allowing conventional KL-based distillation without cross-entropy to perform competitively relative to *CondKD*. However, in practical unsupervised KD scenarios, custom training data, along with their annotations, can be both private and inaccessible. To simulate this setting, we train student models on non-target data, specifically CC3M (Changpinyo et al., 2021), where teacher models have never been exposed to the data. In this case, teacher predictions become less reliable, and naive KL-based distillation without cross-entropy for label supervision fails to produce satisfactory results. As illustrated in Table 3, with *CondKD*, our MobileNet achieves 67.89% top-1 accuracy, significantly outperforming the KL baseline by **16.29%**.

4.2 CondKD with Label Supervision on Balanced Data

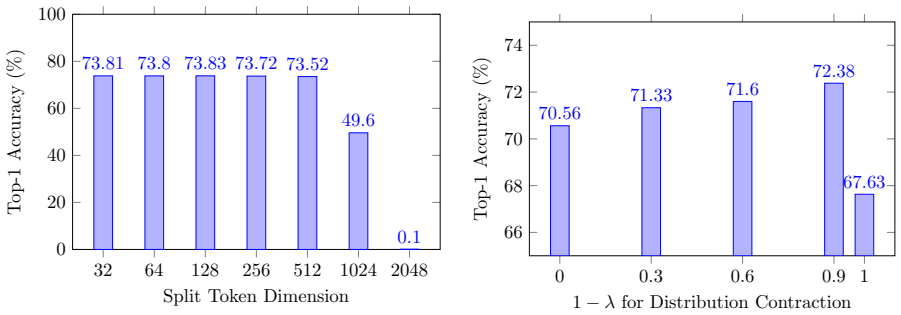
Experimental Results on ImageNet. On ImageNet, we evaluate a range of teacher-student configurations, covering diverse network architectures (CNNs with regular or depth-wise convolutions, and Transformers), training recipes (standard vs. strong augmentation), and model scales (e.g., ResNet-34, ResNet-50, BEiT-Large). *Despite the significant variation across configurations, we employ a consistent $\lambda = 0.1$ for the Distribution Contraction in Definition 1, highlighting the generalizability, robustness, and practical convenience of CondKD.*

The Classical Training Setting. Under the regular training recipe (including *RandomResizedCrop* and horizontal flip), we train models 100 epochs with a cosine learning rate schedule. For fair comparisons, we reproduce the results of KD, DKD, and IKL-KD with their open-sourced code and just replace the step learning rate schedule with the cosine learning rate schedule. Our ResNet-18 achieves a top-1 accuracy of 72.40%, outperforming KD, IKL-KD, and DKD by 1.16%, 0.49%, and 0.53%, respectively. Similarly, our MobileNet reaches 73.76% top-1 accuracy, surpassing KD, IKL-KD, and DKD by 2.32%, 0.57%, and 1.21%, respectively.

Scaling to Strong Training Settings. Under strong training recipes, prior work (Hao et al., 2023) shows that advanced KD methods, such as DKD and ReviewKD, perform only on par with or even

Table 5: **Top-1 accuracy (%) on the CIFAR-100.** Teachers and students are in the **same** architectures. Δ represents the improvements over the KD (Hinton et al., 2015). Results are the average over three trials.

Distillation Manner							
	Teacher	ResNet56 72.34	ResNet110 74.31	ResNet32×4 79.42	WRN-40-2 75.61	WRN-40-2 75.61	VGG13 74.64
	Student	ResNet20 69.06	ResNet32 71.14	ResNet8×4 72.50	WRN-16-2 73.26	WRN-40-1 71.98	VGG8 70.36
Discriminative Distillation	FitNet	69.21	71.06	73.50	73.58	72.24	71.02
	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
	DKD	71.97	74.11	76.32	76.24	74.81	74.68
	IKL-KD	71.44	74.26	76.59	76.45	74.98	**74.98**
	KD	70.66	73.08	73.33	74.92	73.54	72.98
Generative.	**CondKD**	**72.63**	**74.95**	**77.47**	**76.83**	**75.98**	74.24
	Δ	**+1.97**	**+1.87**	**+4.14**	**+1.91**	**+2.44**	**+1.26**



(a) *SplitTok* for token dimension (b) $1 - \lambda$ for *Distribution Contraction*

Figure 2: **Ablation studies on *SplitTok* and *Distribution Contraction*.** (a) Top-1 Accuracy with different token dimension for *SplitTok*. The teacher-student configuration of (ResNet-50, MobileNet) is used on ImageNet.; (b) Top-1 Accuracy under different $1 - \lambda$ values for *Distribution Contraction*. The teacher-student configuration of (ResNet-34, ResNet-18) is used on ImageNet. Note that $\lambda = 0.1$ is used for all of settings on ImageNet, demonstrating robustness of CondKD.

worse than vanilla KD, which can be caused by their complex optimization. The evaluated settings include:

- A1: *RandAug(7/0.5)*, *MixUp*: 0.2, *CutMix*: 1.0, *Label Smoothing*: 0.1, 600 epochs.
- A2: *RandAug(7/0.5)*, *MixUp*: 0.1, *CutMix*: 1.0, *Label Smoothing*: 0.0, 300 epochs.

In contrast, *CondKD*, a simple and unified objective, consistently outperforms these baselines by a clear margin under the same settings. In particular, with BEiTv2-Large as the teacher, our ResNet-50 achieves 82.36% top-1 accuracy under the A1 recipe, establishing a new state-of-the-art. These results demonstrate that *CondKD* scales effectively under strong training schedules.

Experimental Results on CIFAR. Following previous work (Cui et al., 2024a), we consider the distillation among the architectures having the same unit structures, like ResNets and VGGNets. On the other hand, we explore the distillation among architectures made up of different unit structures, like WideResNet and ShuffleNet, VGGNet, and ResNet. Specifically, we train models for 240 epochs with a learning rate that decays by 0.1 at the 150th, 180th, and 210th epoch.

Experimental results on CIFAR-100 are summarized in Table 5 and Table 8(Appendix). Table 5 lists the comparisons with previous methods under the setting that the architectures of the teacher and student have the same unit structures. As shown in Table 5, *CondKD* models can achieve much better or comparable performance in all considered settings. Specifically, we achieve the best performance in 5 out of 6 training settings. Table 8 lists the comparisons with previous methods under the setting that the architectures of the teacher and student have different unit structures. As shown in Table 8, we achieve the best performance in 4 out of 5 training configurations.

Table 6: **Ablation on efficiency**. Speed is measured on 1 GeForce RTX3090 GPU for inference and 4 RTX 6000 Ada GPUs for training with a batch of 512 images.

Teacher-student	Mode	R34-R18 Acc/Speed	BEiT-B-R50(A2) Acc/Speed	BEiT-L-R50(A2) Acc/Speed
KD	Inference	71.44/0.20s	80.96/0.67s	81.68/0.67s
	Training	0.12s/iter	0.48s/iter	1.05s/iter
CondKD	Inference	73.76/0.20s	81.96/0.67s	82.36/0.67s
	Training	0.17s/iter	0.54s/iter	1.13s/iter

4.3 CondKD with Label Supervision on Imbalanced Data

Real-world data often exhibits a long-tailed distribution, making long-tailed recognition a critical challenge for practical applications. Extensive research has been devoted to addressing this problem through both algorithmic and theoretical advances (Cui et al., 2019; Cao et al., 2019; Kang et al., 2019; Cui et al., 2022; Menon et al., 2020; Cui et al., 2021; 2023; 2024b). Following recent efforts (Cui et al., 2024a; 2025), we also evaluate the effectiveness of *CondKD* under data imbalance using the ImageNet-LT benchmark (Liu et al., 2019). We train ResNet models for 90 epochs using *RandomResizedCrop* and horizontal flipping as standard preprocessing.

As shown in Table 4 and Table 9 (Appendix A.5), we observe an interesting phenomenon: *CondKD without label supervision can even achieve slightly better performance than IKL-KD incorporating labels. However, there are few accuracy gains after applying the label supervision with Distribution Contraction, which is a different behaviour compared to balanced data.* This phenomenon gives us new insight into KD for imbalanced data: the necessity of labels for KD on imbalanced data. As this work focuses on generative learning of KD, we leave it as our future work.

4.4 Ablation Studies

Ablation of *SplitTok*. We validate the necessity of the proposed *SplitTok* under the (ResNet-50, MobileNet) teacher-student configuration on ImageNet. As shown in Figure 2a, the model maintains competitive performance when the token dimension is ≤ 256. However, accuracy drops sharply to 0.1% as the token dimension increases from 512 to 2048, highlighting the high-dimensional optimization challenge and the effectiveness of *SplitTok* in mitigating it.

Ablation on λ for *Distribution Contraction*. We validate the effectiveness of the proposed *Distribution Contraction* technique under the (ResNet-34, ResNet-18) teacher configuration on ImageNet. As illustrated in Figure 2b, *CondKD* achieves competitive performance in the unsupervised KD setting with $1 - \lambda = 0.0$. Deploying *Distribution Contraction* technique with a proper $1 - \lambda = 0.9$, our model achieves significant performance gains. Interestingly, we observe that the model achieves much worse accuracy when the sample features contract to class centers, which indicates the importance of the continuity of the sample feature space. Inspired by this phenomenon, we apply the unsupervised mixup for diffusion training.

Ablation on Inference Efficiency. Diffusion head involves iterative sampling during inference to generate image representations, which can introduce significant latency and undermine the intent of KD methods. Thus, we adopt the linear probing strategy to approximate the diffusion head with only a single linear classifier. As shown in Table 6, our method with a linear classifier (the same as vanilla KD) achieves significant improvements while keeping high efficiency.

5 Conclusion

In this paper, we propose the Conditional Generation for Knowledge Distillation (*CondKD*) algorithm, formulating the KD as a conditional generation problem. The straightforward *CondKD* pipeline suffers from the high-dimensional optimization disaster and the lack of label supervision. We propose the *Split Tokenization (SplitTok)* and *Distribution Contraction* techniques to address the above issues, respectively. With theoretical analysis, we prove that *CondKD* with *Distribution Contraction* is a more stable surrogate to the multi-task learning (combining the reconstruction loss and the cross-entropy loss), while eliminating the multi-step sampling during training and achieving efficient training optimization. Experimental results on popular benchmarks, including CC3M, ImageNet, and CIFAR, in unsupervised/supervised KD demonstrate the effectiveness and great potential of our method.

AI USE STATEMENT

In this work, We used generative AI tools to polish the writing of the manuscript. The authors reviewed the AI-assisted edits and take full responsibility for the final content of this work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11933–11942, 2022.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, pp. 4794–4802, 2019.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.

Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 3695–3706, 2022.

Jiequan Cui, Zhisheng Zhong, Zhuotao Tian, Shu Liu, Bei Yu, and Jiaya Jia. Generalized parametric contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled kullback-leibler divergence loss. *Advances in Neural Information Processing Systems*, 37:74461–74486, 2024a.

Jiequan Cui, Beier Zhu, Xin Wen, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Classes are not equal: An empirical study on image recognition fairness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23283–23292, 2024b.

Jiequan Cui, Beier Zhu, Qingshan Xu, Zhuotao Tian, Xiaojuan Qi, Bei Yu, Hanwang Zhang, and Richang Hong. Generalized kullback-leibler divergence loss. *arXiv preprint arXiv:2503.08038*, 2025.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pp. 1607–1616. PMLR, 2018.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Zhiwei Hao, Jianyuan Guo, Kai Han, Han Hu, Chang Xu, and Yunhe Wang. Vanillakd: Revisit the power of vanilla knowledge distillation from small scale to large scale. *arXiv preprint arXiv:2305.15781*, 2023.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, 2022.

Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. *Advances in Neural Information Processing Systems*, 36:65299–65316, 2023.

Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24276–24285, 2023.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546, 2019.

Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals kullback-leibler divergence for knowledge distillation. *Advances in Neural Information Processing Systems*, 37:65445–65475, 2024.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. ImageNet large scale visual recognition challenge. *IJCV*, 2015.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *The Fourteenth International Conference on Learning Representations*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, volume 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15731–15740, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020.

Shicai Wei, Chunbo Luo, and Yang Luo. Scaled decoupled distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15975–15983, 2024.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *European conference on computer vision*, pp. 53–69. Springer, 2022.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pp. 4320–4328, 2018.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A APPENDIX

A.1 PROOF OF THEOREM 1

We begin by restating the *CondKD* objective and the *multi-task* objective. The *CondKD* objective is the expected noise prediction loss over noisy data $\tilde{x}_m$:

$$\mathcal{L}_{\text{CondKD}} = \mathbb{E}_{x,m,\epsilon}\left[\|\epsilon - \epsilon_\theta(\tilde{x}_m, m, \mathcal{F}_s(x))\|^2\right], \tag{11}$$

where $\tilde{x}_m$ is given by:

$$\tilde{x}_m = \sqrt{\bar{\alpha}_m}\tilde{x}_0 + \sqrt{1-\bar{\alpha}_m}\epsilon, \quad \tilde{x}_0 = \lambda\hat{x}_0 + (1-\lambda)c_y \textbf{ (by Definition 1)}, \tag{12}$$

where $\hat{x}_0$ is the teacher feature and c_y=W_y for class y.

The multi-task objective consists of two parts:

- The reconstruction loss via noise prediction:

$$\mathcal{L}_{\text{noise}} = \mathbb{E}_{x,m,\epsilon}\left[\|\epsilon - \epsilon_\theta(\hat{x}_m, m, \mathcal{F}_s(x))\|^2\right],$$

- The classification loss via cross-entropy:

$$\mathcal{L}_{\text{CE}} = \mathbb{E}_x\left[\mathcal{L}_{\text{CE}}\left(\mathcal{C}_t(\hat{x}_0^{'}), y\right)\right],$$

where $\hat{x}_m$ is the noisy version of $\hat{x}_0$, $\mathcal{C}_t$ is the teacher classifier, and $\hat{x}_0^{'}$ is the generated feature.

Thus, the multi-task loss is:

$$\mathcal{L}_{\text{multi}} = \gamma_0\mathcal{L}_{\text{noise}} + \gamma_1\mathcal{L}_{\text{CE}},$$

where γ_0 and γ_1 are scaling constants.

We derive the gradients of the *CondKD* and multi-task objectives with respect to the single-step estimate $\hat{x}_0^{'}$, and establish their exact relationship.

Step 1: Gradient of $\mathcal{L}_{CondKD}$.

Recall the forward process with distribution contraction:

$$\tilde{x}_0 = \lambda\hat{x}_0 + (1-\lambda)c_y, \quad \tilde{x}_m = \sqrt{\bar{\alpha}_m}\tilde{x}_0 + \sqrt{1-\bar{\alpha}_m}\epsilon. \tag{13}$$

Using the single-step estimator:

$$\hat{x}_0^{'} = \frac{1}{\sqrt{\bar{\alpha}_m}}\left(\tilde{x}_m - \sqrt{1-\bar{\alpha}_m}\epsilon_\theta\right), \tag{14}$$

we obtain:

$$\epsilon - \epsilon_\theta = \sqrt{\frac{\bar{\alpha}_m}{1-\bar{\alpha}_m}}\left(\hat{x}_0^{'} - \tilde{x}_0\right). \tag{15}$$

Substituting into $\mathcal{L}_{CondKD}$ yields:

$$\mathcal{L}_{CondKD} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta\|^2\right]. \tag{16}$$

Taking gradients with respect to $\hat{x}_0^{'}$:

$$\nabla_{\hat{x}_0^{'}}\mathcal{L}_{CondKD} = \frac{2\bar{\alpha}_m}{1-\bar{\alpha}_m}\left(\hat{x}_0^{'} - \tilde{x}_0\right). \tag{17}$$

Substituting $\tilde{x}_0$:

$$\nabla_{\hat{x}_0^{'}}\mathcal{L}_{CondKD} = \frac{2\bar{\alpha}_m}{1-\bar{\alpha}_m}\left(\hat{x}_0^{'} - \hat{x}_0\right) + \frac{2(1-\lambda)\bar{\alpha}_m}{1-\bar{\alpha}_m}\left(\hat{x}_0 - c_y\right). \tag{18}$$

—

Step 2: Gradient of $\mathcal{L}_{multi}$.

The noise prediction term gives:

$$\nabla_{\hat{x}_0'}\mathcal{L}_{noise} = \frac{2\bar{\alpha}_m}{1-\bar{\alpha}_m}\left(\hat{x}_0^{'} - \hat{x}_0\right). \tag{19}$$

For the cross-entropy term, let $\mathcal{C}_t(z) = Wz$. Then:

$$\nabla_{\hat{x}_0'}\mathcal{L}_{CE} = \sum_{i=1}^{C} p(i|\hat{x}_0^{'})W_i - W_y. \tag{20}$$

Define:

$$\hat{x}_0^{''} = \sum_{i=1}^{C} p(i|\hat{x}_0^{'})W_i, \quad c_y = W_y. \tag{21}$$

Thus:

$$\nabla_{\hat{x}_0'}\mathcal{L}_{multi} = \gamma_0\frac{2\bar{\alpha}_m}{1-\bar{\alpha}_m}(\hat{x}_0^{'} - \hat{x}_0) + \gamma_1(\hat{x}_0^{''} - c_y). \tag{22}$$

—

Step 3: Exact gradient relationship.

Let

$$\gamma_0 = 1, \quad \gamma_1 = \frac{2(1-\lambda)\bar{\alpha}_m}{1-\bar{\alpha}_m}.$$

Then subtracting the two gradients yields:

$$\nabla\mathcal{L}_{CondKD} - \nabla\mathcal{L}_{multi} = \gamma_1(\hat{x}_0 - c_y) - \gamma_1(\hat{x}_0^{''} - c_y) \tag{23}$$

$$= \gamma_1(\hat{x}_0 - \hat{x}_0^{''}). \tag{24}$$

This establishes:

$$\nabla\mathcal{L}_{CondKD} = \nabla\mathcal{L}_{multi} + \gamma_1(\hat{x}_0 - \hat{x}_0^{''}). \tag{25}$$

—

Step 4: Local approximation.

If $\|\hat{x}_0^{''} - \hat{x}_0\| \leq \delta$, then:

$$\|\nabla\mathcal{L}_{CondKD} - \nabla\mathcal{L}_{multi}\| = \gamma_1\|\hat{x}_0 - \hat{x}_0^{''}\| \leq \gamma_1\delta. \tag{26}$$

—

Step 5: High-confidence regime.

If $p(y|\hat{x}_0^{'}) \to 1$, then:

$$\hat{x}_0^{''} \to c_y.$$

Thus:

$$\nabla\mathcal{L}_{multi} \to \gamma_0\nabla\mathcal{L}_{noise}, \tag{27}$$

$$\nabla\mathcal{L}_{CondKD} \to \gamma_0\nabla\mathcal{L}_{noise} + \gamma_1(\hat{x}_0 - c_y). \tag{28}$$

—

Step 6: Uncertain regime.

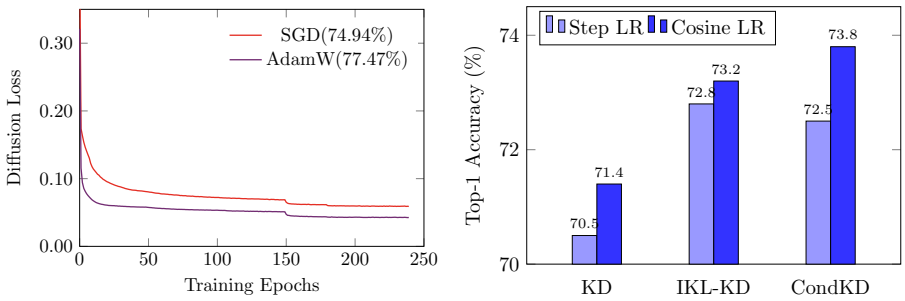
When $\hat{x}_0^{''}$ deviates from $\hat{x}_0$, the difference term $\gamma_1(\hat{x}_0 - \hat{x}_0^{''})$ captures the deviation between the prediction-dependent gradient and the teacher-anchored direction, completing the proof. □

A.2 Comparisons of Sensitivity to Hyperparameters with Previous Methods on ImageNet

Table 7 summarizes the comparisons between our method and advanced previous KD methods regarding hyperparameter sensitivity. With a single hyper-parameter $1 - \lambda = 0.9$, our method consistently achieves better performance, demonstrating the robustness and practicality of our method.

Table 7: **Sensitivity to hyperparameters.** Advanced KD methods often involve complex hyperparameter tuning. "*" represents that MLKD uses RandAug. Our *CondKD* method consistently works well across diverse teacher-student configurations with $\lambda = 0.1$ on ImageNet, demonstrating simplicity and robustness.

Method	Teacher-Student	Hyperparameter configuration
KD	ResNet-34 — ResNet-18 ResNet-50 — MobileNet	$w_{kl} = 0.5, w_{ce} = 0.5, T = 1.0$
DKD	ResNet-34 — ResNet-18 ResNet-50 — MobileNet	$w_{ce} = 1.0, w_{\alpha} = 1.0, w_{\beta} = 0.5, T = 1.0$ $w_{ce} = 1.0, w_{\alpha} = 1.0, w_{\beta} = 2.0, T = 1.0$
MLKD*	ResNet-34 — ResNet-18 ResNet-50 — MobileNet	$w_{ce} = 0.5, w_{kl} = 0.5, T = 1.0, 2.0, 3.0, 5.0, 6.0$, AutoAug $w_{ce} = 0.5, w_{kl} = 0.5, T = 1.0, 2.0, 3.0, 5.0, 6.0$, AutoAug
IKL-KD	ResNet-34 — ResNet-18 ResNet-50 — MobileNet	$w_{ce} = 1.0, w_{\alpha} = 1.0, w_{\beta} = 0.5, T = 1.0$ $w_{ce} = 1.0, w_{\alpha} = 4.0, w_{\beta} = 1.0, T = 1.0$
WKD	ResNet-34 — ResNet-18 ResNet-50 — MobileNet	$w_{ce} = 1.0, w_{l} = 10.0, w_{f} = 1, T = 2.0$, Hint-layer=4 $w_{ce} = 1.0, w_{l} = 25.0, w_{f} = 0.005, T = 2.0$, Hint-layer=-1
DiffKD	ResNet-34 — ResNet-18 ResNet-50 — MobileNet	w_{ce} for task, λ_1 for diffusion model, λ_2 for autoencoder, λ_3 for DiffKD w_{ce} for task, λ_1 for diffusion model, λ_2 for autoencoder, λ_3 for DiffKD
SDD-NKD	ResNet-34 — ResNet-18 ResNet-50 — MobileNet	$w_{ce} = 1.0, w_{\alpha} = 1.0, w_{\beta} = 0.5, T = 1.0$ $w_{ce} = 1.0, w_{\alpha} = 1.0, w_{\beta} = 2.0, T = 1.0$
CondKD	ResNet-34 — ResNet-18 ResNet-50 — MobileNet BEiT-L/B — ResNet-50	**λ=0.1 (Unified reconstruction objective)**



(a) *AdamW vs. SGD* (b) *Step vs. Cosine LR*

Figure 3: **Ablation studies on optimizer and learning rate schedule.** (a) Comparison between AdamW and SGD optimizer for CondKD with teacher-student configuration of (ResNet-32x4, ResNet-8x4) on CIFAR; (b) Comparison between Step and Cosine learning rate schedule for CondKD with teacher-student configuration of (ResNet50, MobileNet) on ImageNet.

A.3 More Ablation Studies

AdamW vs. SGD Optimizer for CondKD. We investigate the impact of different optimizers on training *CondKD*. While SGD is commonly used for CNNs and AdamW/Adam are standard choices for Transformers, we adopt AdamW for *CondKD* following previous work MAR (Liu et al., 2023) and DDPM (Ho et al., 2020). As shown in Figure 3a, our empirical results demonstrate that AdamW leads to more stable and effective optimization for *CondKD*.

Cosine vs. Step Learning Rate Schedule for CondKD. We study the impact of learning rate schedules on *CondKD*, focusing on step decay and cosine annealing strategies. For fair comparison,

we adopt the step schedule on CIFAR and reproduce the results of previous work with a cosine learning rate in their open-sourced code on ImageNet. Our empirical results show that the cosine schedule is critical for *CondKD*, particularly on large-scale datasets such as ImageNet. As illustrated in Figure 3b, cosine learning rates significantly accelerate training convergence and improve overall performance.

A.4 More Results on CIFAR-100

Table 8: **Top-1 accuracy (%) on the CIFAR-100 validation.** Teachers and students are in **different** architectures. Δ represents the improvements over the KD (Hinton et al., 2015) baseline. All results are the average over three trials.

Distillation Manner	Teacher	ResNet32×4	WRN-40-2	VGG13	WRN-40-2	ResNet32×4
		79.42	75.61	74.64	75.61	79.42
	Student	ShuffleNet-V1	ShuffleNet-V1	ResNet20	ResNet8x4	ShuffleNet-V2
		70.50	70.50	69.06	72.50	71.82
Discriminative Point-wise Distillation	FitNet	73.59	73.73	69.27	74.74	73.54
	RKD	72.28	72.21	69.83	72.43	73.21
	CRD	75.11	76.05	71.16	76.64	75.65
	OFD	75.98	75.85	-	74.50	76.82
	ReviewKD	77.45	77.14	-	75.48	77.78
	DKD	76.45	76.70	70.57	75.23	77.07
	IKL-KD	76.64	**77.19**	70.88	76.12	77.16
	KD	74.07	74.83	70.20	74.01	74.45
Gen.D.D.	**CondKD**	**77.58**	76.73	**72.17**	**77.55**	**78.13**
	Δ	**+3.51**	**+1.90**	**+1.97**	**+3.54**	**+3.68**

A.5 Detailed Performance on Few-, Medium-, and Many-shot

Table 9: **Top-1 accuracy(%) on the ImageNet-LT validation with CondKD.** "*" represents the unsupervised setting.

Method	Teacher	Student	All	Few	Medium	Many
Baseline	-	ResNet-18	63.16	33.47	5.88	41.15
Baseline	-	ResNet-50	67.25	38.56	8.21	45.47
KD	ResNeXt-101	ResNet-18	64.60	37.88	9.53	44.32
KD	ResNeXt-101	ResNet-50	68.83	42.31	11.37	48.31
IKL-KD	ResNeXt-101	ResNet-18	66.60	38.53	8.19	45.21
IKL-KD	ResNeXt-101	ResNet-50	70.06	43.47	10.99	49.29
CondKD*	ResNeXt-101	ResNet-18	66.71	39.02	8.66	45.54
CondKD*	ResNeXt-101	ResNet-50	70.12	43.52	10.84	49.31

A.6 Application to Self-supervised Distillation

Recent studies have shown that self-supervised representations can significantly improve image generation models through representation alignment (e.g., REG (Wu et al., 2026), REPA (Yu et al., 2024), and iREPA (Singh et al., 2025)). As a preliminary step toward applying *CondKD* to representation distillation for image generation, we first evaluate its effectiveness on self-supervised representation distillation using DINO, where DINO-Base serves as the teacher and DINO-Small as the student.

Following iREPA, we measure Local Descriptor Similarity (LDS), which quantifies the preservation of spatial structures and has been shown to correlate positively with image generation quality. The results are summarized in Table 10. Compared with MSE-based feature regression, *CondKD* substantially improves the LDS score from 0.34 to 0.41, indicating that it better preserves local spatial structures during representation distillation. These results suggest that *CondKD* is a promising representation distillation framework for downstream image generation models.

Table 10: Comparison of self-supervised representation distillation on ImageNet. LDS is computed on the 50,000-image ImageNet validation set.

Method	Training Epochs	LDS(↑)
DINO-small	800 × 2 (2 global crops)	0.44
MSE	100	0.34
CondKD	100	0.41

A.7 Comparison with DiffKD

We would like to clarify that *CondKD* and DiffKD have different objectives. DiffKD introduces diffusion models as a feature enhancement mechanism to reconstruct teacher features. It trains the unconditional diffusion model with teacher features. After training, the student feature is input to the diffusion model for enhancement. In contrast, *CondKD* formulates KD as conditional generative modeling by explicitly learning the distribution of teacher representations conditioned on student features. Therefore, *CondKD* focuses on modeling the conditional distribution $p_\theta(\mathcal{F}_t|\mathcal{F}_s)$, rather than improving student features through auxiliary feature generation enhancement.

Additionally, *CondKD* has a unified objective, while DiffKD incorporates multiple objectives, like cross-entropy, reconstruction loss for the autoencoder, diffusion loss for student feature enhancement, and KD loss. Please refer to Table 7 in the Appendix.

Comparisons with DiffKD are conducted on CIFAR-100. The experimental results are summarized in Table 11. *CondKD* achieves better performance with a simpler and unified training objective.

Table 11: Comparison with DiffKD on CIFAR-100.

Teacher-student configuration	DiffKD	CondKD
Wrn40-2-Wrn40-1	74.09	75.98
ResNet56-ResNet20	71.92	72.63
ResNet32x4-ResNet8x4	76.72	77.47