

ABKD: Pursuing a Proper Allocation of the Probability Mass in Knowledge Distillation via α - β -Divergence

Guanghui Wang, Zhiyong Yang, Zitai Wang,
Shi Wang, Qianqian Xu, Qingming Huang



Wechat



Paper

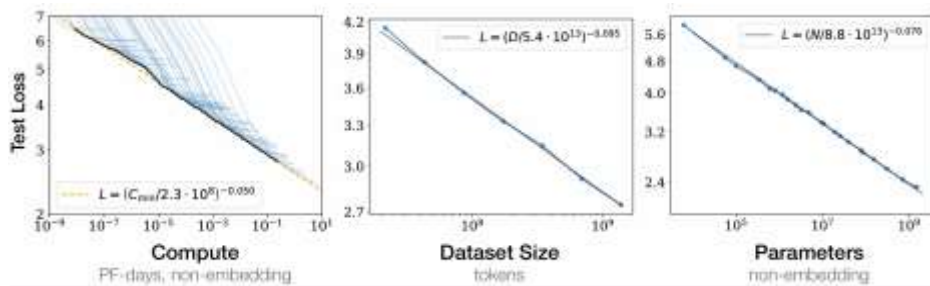


Code

The era of foundation models

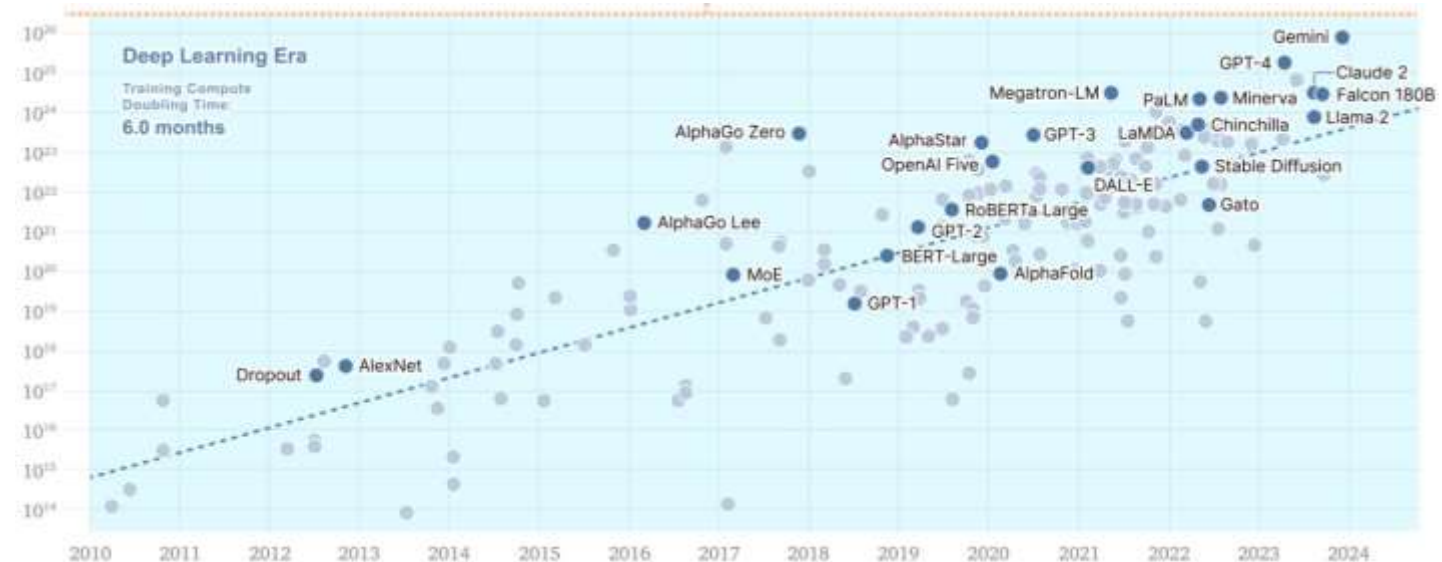


- ◆ The scaling law has facilitated the rise of foundation models with ever-growing capabilities and sizes.



The test loss scales as a power-law with **model size**, **dataset size**, and the **amount of training computation**

Scaling laws



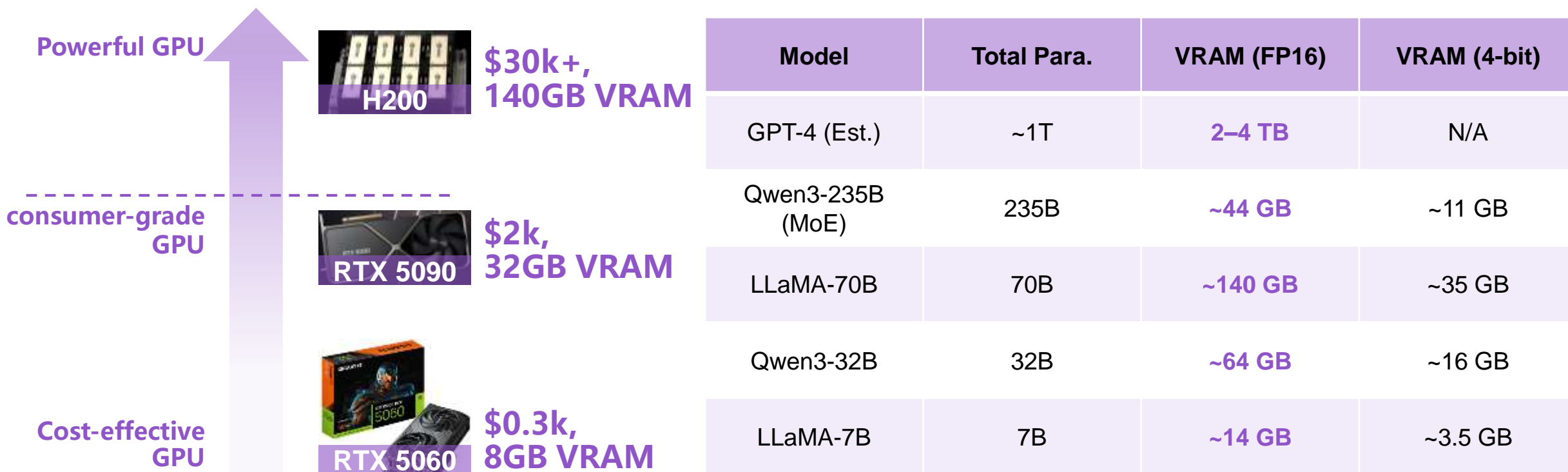
Model size scales exponentially

[1] Jared Kaplan et al., Scaling Laws for Neural Language Models. Arxiv 2020

The challenge of scaling laws



◆ As model sizes grow, compute capacity and cost-efficiency can no longer keep pace, making lightweight and on-device deployment essential



Most real-world GPUs fall far short of the memory needed to serve large models

Knowledge Distillation (KD)



◆ Knowledge distillation has been a common practice to achieve cost-efficient model inference

2.4. Distillation: Empower Small Models with Reasoning Capability

To equip more efficient smaller models with reasoning capabilities like DeepSeek-R1, we directly fine-tuned open-source models like Qwen (Qwen, 2024b) and Llama (AI@Meta, 2024) using the 800k samples curated with DeepSeek-R1, as detailed in §2.3.3. Our findings indicate that this straightforward distillation method significantly enhances the reasoning abilities of smaller models. The base models we use here are Qwen2.5-Math-1.5B, Qwen2.5-Math-7B, Qwen2.5-14B, Qwen2.5-32B, Llama-3.1-8B, and Llama-3.3-70B-Instruct. We select Llama-3.3 because its reasoning capability is slightly better than that of Llama-3.1.

3.2. Distilled Model Evaluation

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

DeepSeek-R1 technical report

4 Post-training

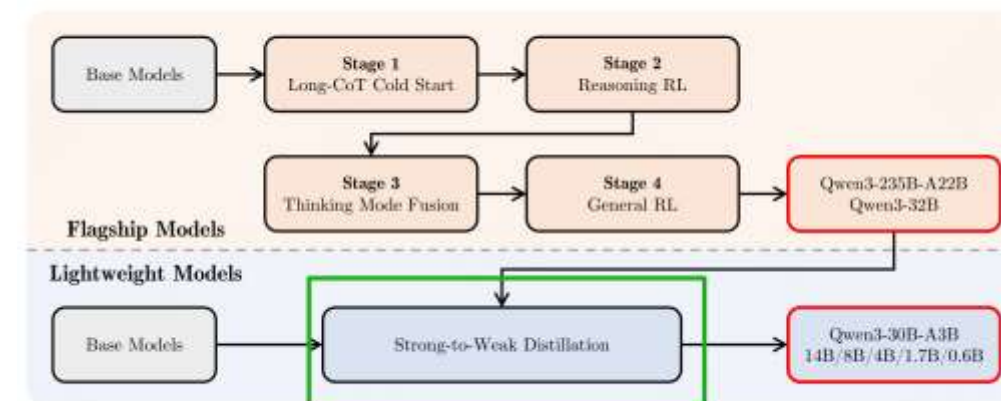


Figure 1: Post-training pipeline of the Qwen3 series models.

Table 21: Comparison of reinforcement learning and on-policy distillation on Qwen3-8B. Numbers in parentheses indicate pass@64 scores.

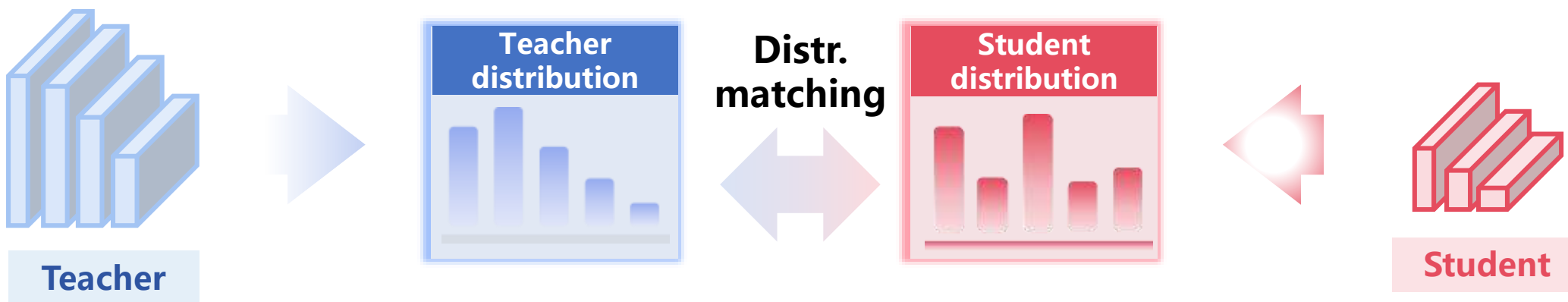
Method	AIME'24	AIME'25	MATH500	LiveCodeBench v5	MMLU -Redux	GPQA -Diamond	GPU Hours
Off-policy Distillation	55.0 (90.0)	42.8 (83.3)	92.4	42.0	86.4	55.6	-
+ Reinforcement Learning	67.6 (90.0)	55.5 (83.3)	94.8	52.9	86.9	61.3	17,920
+ On-policy Distillation	74.4 (93.3)	65.5 (86.7)	97.0	60.3	88.3	63.3	1,800

Qwen3 technical report

Knowledge Distillation | Distribution Matching



- ◆ Let the student distribution q_θ to **imitate** the teacher distribution p
- ◆ **Minimize** a pre-defined divergence measure between their output distributions



$$\min_{\theta} \underbrace{\ell_{\text{CE}}}_{\text{student's performance}} + \underbrace{\text{div}(p, q_{\theta})}_{\text{distribution divergence}}$$

student's performance distribution divergence

How to find a proper divergence ?


◆ Two basic divergence measures

- **Forward** Kullback-Leibler divergence (FKLD)

$$\mathbb{D}_{\text{KL}}(p||q_{\theta}) = \sum_k p(k) \log \frac{p(k)}{q_{\theta}(k)}$$

- **Reverse** Kullback-Leibler divergence (RKLD)

$$\mathbb{D}_{\text{KL}}(q_{\theta}||p) = \sum_k q_{\theta}(k) \log \frac{q_{\theta}(k)}{p(k)}$$


$$\mathbb{D}_{\text{KL}}(p||q_{\theta}) \neq \mathbb{D}_{\text{KL}}(q_{\theta}||p)$$



Asymmetric !

Which one should we choose?

How the Teacher Guides the Student



Will F/RKLD necessarily provide a more reliable student distribution?

◆ **Result:** F/RKLD automatically tunes the mass reallocation rate, but with different direction!

◆ **Tool:** keep track of the probability mass change in each gradient update step

$$\text{LogR}_t^{\mathcal{A}}(y) \triangleq \log \left(\frac{q_{t+1}^{\mathcal{A}}(y)}{q_t(y)} \right) \quad \text{the reallocation rate} \quad \frac{d \log (q_t(y))}{dt}$$
$$\sum_y q_t(y) = 1$$

reallocation speed provide information of mass competition among class channels

[2] Yi Ren, Danica J. Sutherland, Learning Dynamics of LLM Finetuning. ICLR 2025

[3] Fahim Tajwar, Anikait Singh, et al. Preference Fine-Tuning of LLMs Should Leverage Suboptimal, On-Policy Data. ICML 2024

How the Teacher Guides the Student | F/RKLD



- ◆ FKLD and RKLD as Two **Extreme Cases** (inspired by Tajwar et al. 2024)

Proposition 1. (Tajwar et al. 2024)

The updates induced by FKLD and RKLD within one gradient :

- ◆ (FKLD, **conservative**) $\text{LogR}_t^{\mathcal{F}}(y) \propto_y 1 \cdot p(y) - q_t(y)$

- ◆ (RKLD, **aggressive**) $\text{LogR}_t^{\mathcal{F}}(y) \propto_y q_t(y) \cdot (\log p(y) - \log q_t(y) + \mathbb{D}_{KL}(q_t, p))$

$$\text{rate} \propto_y \text{Confidence Concentration} \times \text{Hardness Concentration}$$

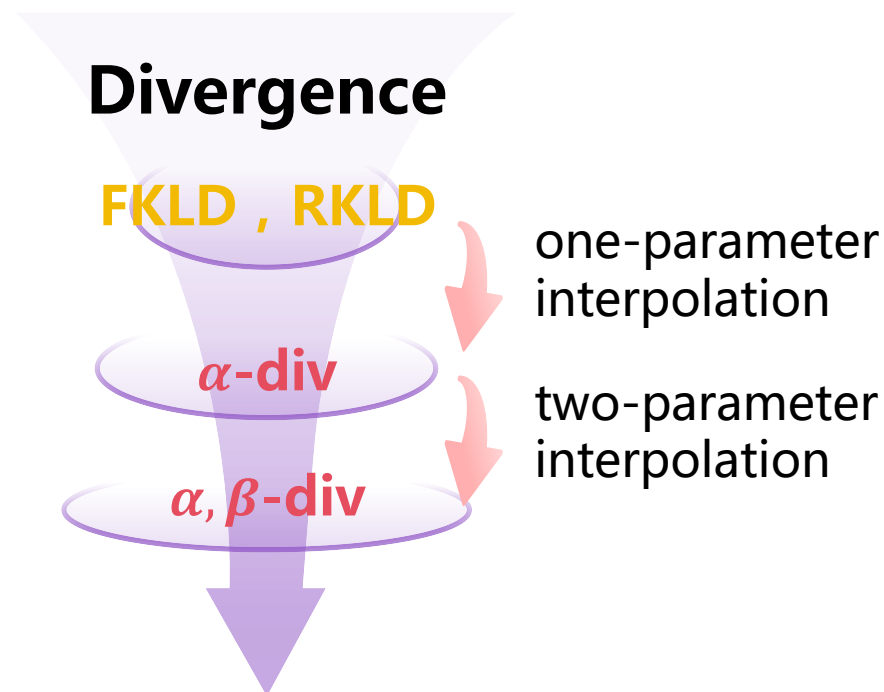
- FKLD causes the student to treat all mismatches equally
- RKLD emphasizes errors from the target class and suppress others

Since Teacher might be wrong,
student should learn with a proper rate!



Our Idea:

Find **interpolations**
between the dynamics of
FKLD and RKLD



α -Divergence: one-parameter interpolation



Introduce **one** trade-off parameter to interpolate between the dynamics of FKLD and RKLD

Definition 1 (α -divergence).

Consider $\alpha \in \mathbb{R} \setminus \{0, 1\}$, the α -divergence of two distributions is given by:

$$\mathbb{D}_\alpha(p \parallel q) \triangleq \frac{1}{\alpha(\alpha - 1)} \left[\sum_k p(k)^\alpha q(k)^{1-\alpha} - 1 \right]$$

- ◆ When $\alpha \rightarrow 1$, $\mathbb{D}_\alpha(p \parallel q_\theta)$ becomes the FKLD $\mathbb{D}_{KL}(p \parallel q_\theta)$
- ◆ When $\alpha \rightarrow 0$, $\mathbb{D}_\alpha(p \parallel q_\theta)$ becomes the RKLD $\mathbb{D}_{KL}(q_\theta \parallel p)$

α -Divergence: one-parameter interpolation



- ◆ (FKLD, conservative) $\text{LogR}_t^{\mathcal{F}}(y) \propto_y 1 \cdot p(y) - q_t(y)$
- ◆ (RKLD, aggressive) $\text{LogR}_t^{\mathcal{F}}(y) \propto_y q_t(y) \cdot (\log p(y) - \log q_t(y) + \mathbb{D}_{KL}(q_t, p))$



Proposition 2.

The updates induced by α -divergence within one gradient step are given by:

$$\text{LogR}_t^{\alpha}(y) \propto_y q_t(y)^{1-\alpha} \left(\frac{p(y)^{\alpha} - q_t(y)^{\alpha}}{\alpha} \right) + q_t(y) \sum_k q_t(k)^{1-\alpha} \left(\frac{q_t(k)^{\alpha} - p(k)^{\alpha}}{\alpha} \right)$$



sum-to-one -> increase one means reduce the other

α - β Divergence: two-parameter interpolation



Introduce **two** trade-off parameters separately

Definition 2 (α - β -divergence).

Consider α and $\beta \in \mathbb{R}$, satisfying $\alpha + \beta \neq 0$, the $\alpha - \beta$ -divergence of two distributions is given by:

$$\mathbb{D}_{AB}^{(\alpha, \beta)}(p \parallel q) \triangleq -\frac{1}{\alpha\beta} \sum_k \left[p(k)^\alpha q(k)^\beta - \frac{\alpha}{\alpha + \beta} p(k)^{\alpha+\beta} - \frac{\beta}{\alpha + \beta} q(k)^{\alpha+\beta} \right].$$

- ◆ When $\alpha + \beta = 1$, $\mathbb{D}_{AB}^{(\alpha, \beta)}(p \parallel q)$ becomes the α -divergence.
- ◆ When $\alpha = 1$, $\mathbb{D}_{AB}^{(\alpha, \beta)}(p \parallel q)$ becomes the β -divergence.

ABKD: The Proposed Method



- ◆ (FKLD, conservative) $\text{LogR}_t^{\mathcal{F}}(y) \propto_y 1 \cdot p(y) - q_t(y)$
- ◆ (RKLD, aggressive) $\text{LogR}_t^{\mathcal{F}}(y) \propto_y q_t(y) \cdot (\log p(y) - \log q_t(y) + \mathbb{D}_{KL}(q_t, p))$



Proposition 2.

The updates induced by α - β -divergence within one gradient step are given by:

$$\text{LogR}_t^{(\alpha, \beta)}(y) = \eta q_t(y)^{\beta} \left(\frac{p(y)^{\alpha} - q_t(y)^{\alpha}}{\alpha} \right) + \eta q_t(y) \sum_k q_t(k)^{\beta} \left(\frac{q_t(k)^{\alpha} - p(k)^{\alpha}}{\alpha} \right)$$



α, β controls hardness and confidence concentration separately

◆ Hyperparameter tuning guidelines

Theorem 1 (Informal).

Let q_t denote the student distribution before a gradient update. The α - β -divergence induces the following trends in the probability mass update of the student q_t under different parameter settings:

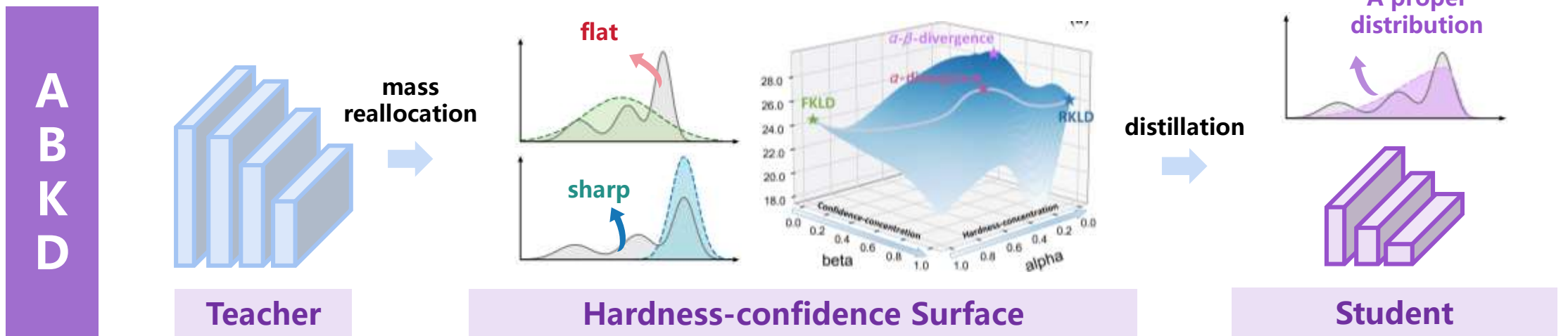
1. The α - β -divergence more aggressively reallocates probability mass across classes as α decreases.
2. The α - β -divergence becomes more (less) preferential in focusing the error on classes with higher student confidence as β increases (decreases).

- **Smaller α** help escape local optima and facilitates **faster convergence** when distributions are **far apart**.
- **Smaller β** encourage learning from non-target class in **higher-dimensional** output distribution.

ABKD: The Proposed Method



◆ The goal of ABKD is to find a **global minimum** of the following objective



O b j.

$$\min_{\theta} \ell_{\text{CE}} + \mathbb{D}_{\text{AB}}^{(\alpha, \beta)}(p \parallel q)$$

cross entropy loss

distillation loss

Attention of error w.r.t α

Attention of confidence w.r.t β

Further Discussion | Benefits of ABKD



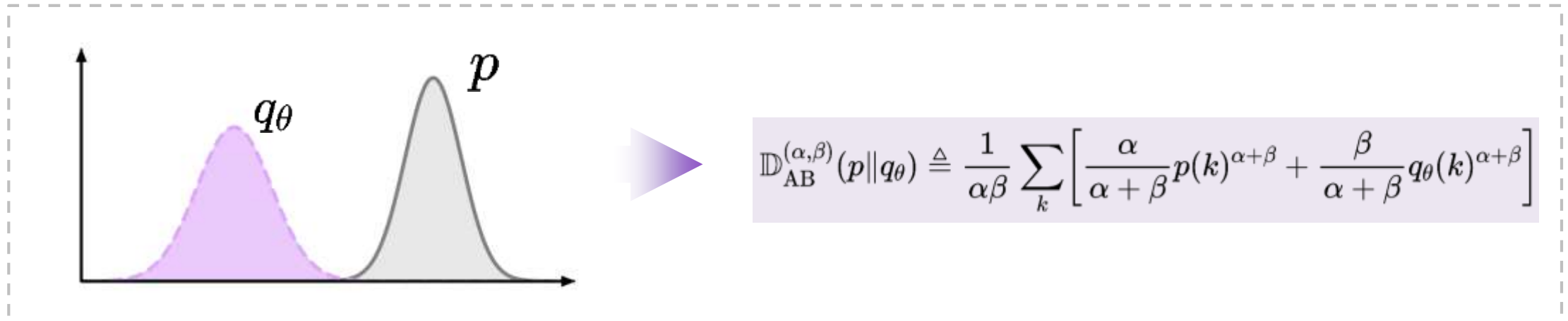
- ◆ ABKD is **insensitive** to extreme modes

When $p(k) > 0$ and $q_\theta(k) \approx 0$

When $q_\theta(k) > 0$ and $p(k) \approx 0$

$$\left(p(k)^\alpha q(k)^\beta - \frac{\alpha}{\alpha + \beta} p(k)^{\alpha + \beta} - \frac{\beta}{\alpha + \beta} q(k)^{\alpha + \beta} \right) \rightarrow \infty$$

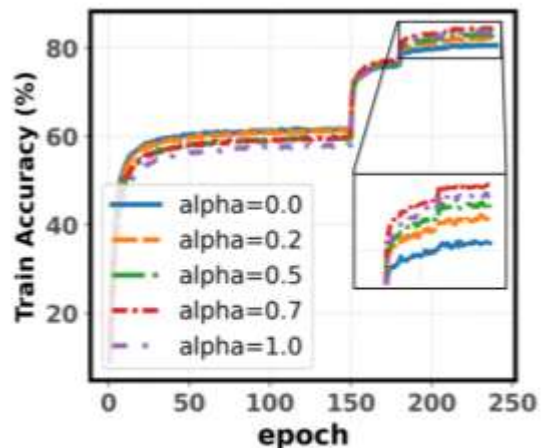
- ◆ ABKD **avoids the vanishing gradient** when the distributions are far apart



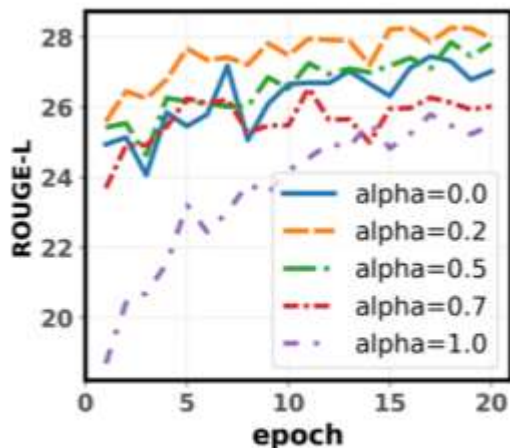
Experiments | How to Tune α, β



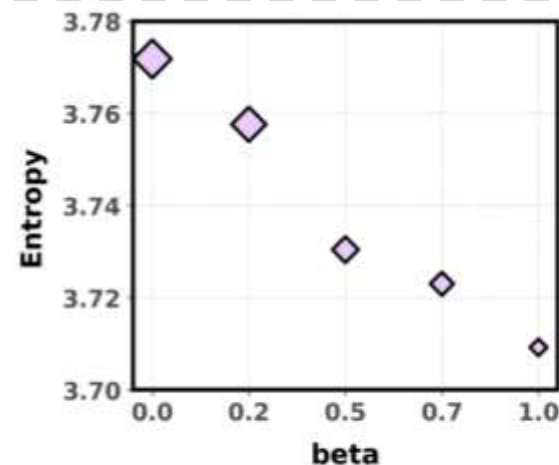
◆ Sensitivity Analysis



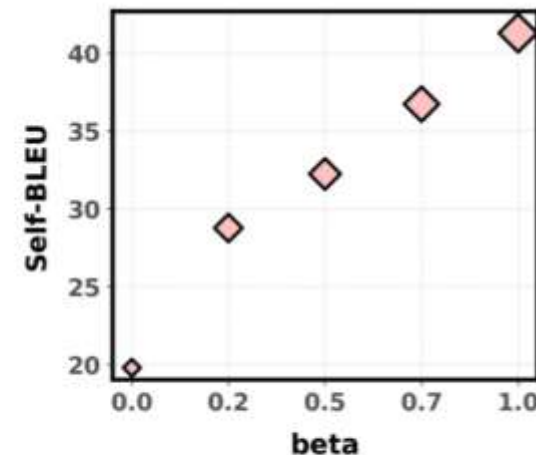
(a) alpha on CIFAR-100



(b) alpha on Dolly



(c) beta on CIFAR-100



(d) beta on Dolly

Smaller α aggressively reallocates probabilities, key for **high-dimensional, distant distributions**.

Smaller β focus more on matching the soft label information, leading to **smoother distributions**

◆ Image Classification

Table 6. Hyperparameters for different image datasets

Dataset	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101
α	0.5	0.8	0.8	0.6	0.9	0.5	0.6	0.8	1.0	0.6	0.8
β	0.5	0.2	0.4	0.4	0.1	0.5	0.5	0.2	0.2	0.5	0.2



Large α and Small β suffice for some tasks with **low-dimensional** output distributions (*e.g.*, simple image classification).

◆ Instruction-Following

Table 7. Hyperparameters for different instruction datasets

Dataset	Dolly Eval	Self-Instruct	Vicuna Eval	Super-Natural	Unnatural
α	0.2	0.2	0.2	0.2	0.2
β	0.7	0.7	0.7	0.7	0.7



Small α and Large β are crucial for some tasks with **high-dimensional** output distributions (*e.g.*, instruction-following).

Experiments | Instruction-Following GPT-2 XL



◆ Task-agnostic instruction-following tasks

Method	Dolly Eval	Self-Instruct	Vicuna Eval	Super-Natural	Unnatural
GPT-2 XL (Teacher)	26.94 (0.23)	13.31 (0.63)	16.23 (0.62)	24.28 (0.43)	29.05 (0.14)
<i>GPT-2 XL (1.5B) → GPT-2 (0.1B)</i>					
SFT	23.14 (0.23)	10.22 (0.44)	15.15 (0.31)	17.41 (0.18)	19.76 (0.09)
KD (Hinton, 2015)	23.80 (0.37)	10.01 (0.75)	15.25 (0.65)	17.69 (0.26)	18.99 (0.05)
SeqKD (Kim & Rush, 2016)	24.28 (0.22)	11.24 (0.30)	14.94 (0.58)	20.66 (0.28)	23.59 (0.13)
MiniLLM (Gu et al., 2024a)	24.62 (0.33)	12.49 (0.56)	17.30 (0.41)	23.76 (0.38)	24.30 (0.14)
GKD (Agarwal et al., 2024)	24.49 (0.16)	11.41 (0.14)	16.01 (0.37)	18.25 (0.24)	21.41 (0.11)
DISTILLM (Ko et al., 2024)	25.32 (0.14)	11.65 (0.28)	16.76 (0.66)	23.52 (0.47)	25.79 (0.08)
Ours (ABKD)	25.65 (0.24)	13.47 (0.42)	16.06 (0.25)	26.47 (0.31)	29.32 (0.08)
<i>GPT-2 XL (1.5B) → GPT-2 Medium (0.3B)</i>					
SFT	25.30 (0.31)	12.56 (0.62)	16.36 (0.22)	23.32 (0.13)	23.42 (0.07)
KD (Hinton, 2015)	24.71 (0.17)	10.33 (0.54)	16.23 (0.50)	23.74 (0.32)	23.97 (0.12)
SeqKD (Kim & Rush, 2016)	25.93 (0.44)	12.98 (0.24)	16.68 (0.30)	21.95 (0.19)	25.23 (0.08)
MiniLLM (Gu et al., 2024a)	25.34 (0.25)	13.36 (0.62)	17.25 (0.46)	25.68 (0.41)	26.63 (0.12)
GKD (Agarwal et al., 2024)	24.75 (0.27)	12.76 (0.85)	16.54 (0.39)	24.94 (0.14)	26.42 (0.15)
DISTILLM (Ko et al., 2024)	26.21 (0.29)	13.53 (0.13)	16.96 (0.66)	25.78 (0.19)	28.51 (0.26)
Ours (ABKD)	26.08 (0.36)	13.86 (0.40)	16.63 (0.26)	27.25 (0.38)	29.69 (0.21)
<i>GPT-2 XL (1.5B) → GPT-2 Large (0.8B)</i>					
SFT	25.42 (0.32)	12.91 (0.46)	16.31 (0.51)	23.76 (0.28)	25.72 (0.07)
KD (Hinton, 2015)	26.02 (0.43)	12.34 (0.52)	16.26 (0.44)	25.11 (0.37)	26.44 (0.12)
SeqKD (Kim & Rush, 2016)	26.29 (0.47)	13.53 (0.34)	16.39 (0.36)	25.81 (0.40)	27.51 (0.10)
MiniLLM (Gu et al., 2024a)	26.12 (0.25)	13.79 (0.31)	17.35 (0.51)	26.12 (0.37)	28.53 (0.17)
GKD (Agarwal et al., 2024)	26.06 (0.34)	13.21 (0.45)	16.64 (0.45)	26.13 (0.41)	27.13 (0.21)
DISTILLM (Ko et al., 2024)	26.56 (0.36)	13.97 (0.36)	16.61 (0.45)	26.73 (0.36)	29.24 (0.23)
Ours (ABKD)	26.51 (0.22)	14.38 (0.43)	16.63 (0.42)	28.05 (0.21)	29.92 (0.14)

- Distillation can outperform SFT, but **relies heavily on well-chosen objectives** (KD on Unnatural).
- Our method surpasses KD and SFT across datasets **by only modifying the final distillation loss**.
- Outperforms or matches SGO-based methods (MINILLM, GKD, DISTILLM), especially on **Super-Natural and Unnatural**.

◆ Efficiency Comparison & Effects of SGO

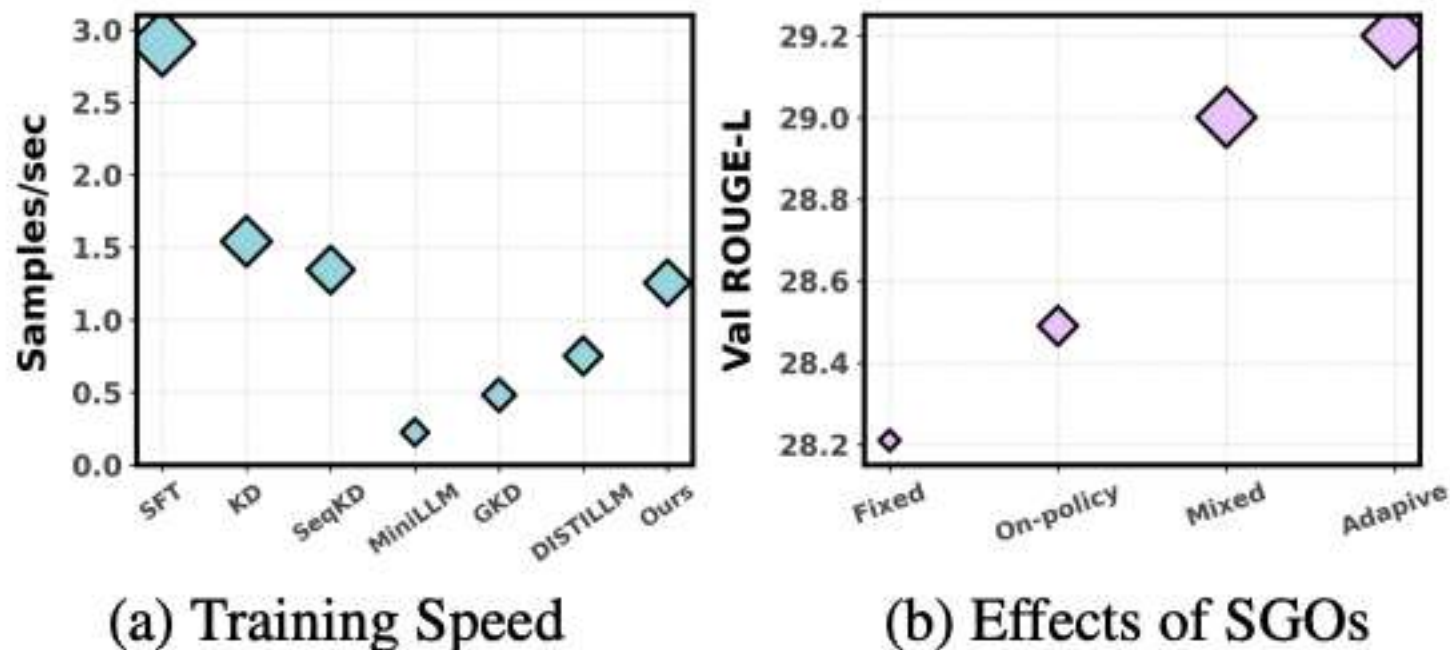


Figure 3. Comparison of training speeds and the effects of using SGOs. Please see Sec. I.1.2 for details of different SGOs strategies.

◆ Task-agnostic instruction-following

Table 10. ROUGE-L scores (\uparrow) on five task-agnostic instruction-following datasets when distilling OpenLLaMA2-7B into OpenLLaMA2-3B. Experiments are conducted on eight RTX 3090 24GB GPUs. * indicates that SGOs are used.

Method	Dolly	Self-Instruct	Vicuna	Super-Natural	Unnatural
SFT	24.54 (0.51)	16.80 (0.64)	16.15 (0.15)	29.29 (0.13)	27.43 (0.21)
FKLD	25.23 (0.44)	18.90 (1.20)	16.67 (0.35)	31.68 (0.22)	29.36 (0.13)
RKLD	27.74 (0.45)	20.61 (0.80)	18.83 (0.40)	35.31 (0.24)	33.86 (0.16)
Jensen's KL	26.28 (0.43)	18.84 (0.66)	17.81 (0.38)	30.92 (0.12)	29.79 (0.17)
BDKD	26.78 (0.53)	18.94 (0.68)	17.81 (0.52)	32.15 (0.34)	30.89 (0.24)
AKL	26.38 (0.41)	17.69 (0.46)	16.72 (0.48)	33.02 (0.16)	31.29 (0.08)
DISTILLM*	28.24 (0.48)	21.00 (0.72)	19.12 (0.53)	37.06 (0.35)	35.05 (0.13)
AlphaNet	28.11 (0.29)	21.30 (0.63)	18.70 (0.23)	37.86 (0.44)	35.40 (0.17)
Ours (ABKD)	30.25 (0.37)	22.39 (0.62)	20.83 (0.42)	38.51 (0.32)	38.66 (0.10)

ABKD **refines only the final distillation loss** yet outperforms baselines in OpenLLaMA2-7B \rightarrow 3B, with ROUGE-L gains of **0.65–3.26%**.

◆ Mathematical reasoning task

Table 12. The distillation results of Qwen2.5-Math on English mathematical benchmarks. Models are evaluated with chain-of-thought prompting.

MODEL \ BENCHMARK						
	GSM8K	MATH	GaoKao 2023 En	Olympiad Bench	College Math	Avg.
Qwen2.5-Math-7B-Instruct (Teacher)	95.5	82.8	66.8	38.5	37.7	64.3
Qwen2.5-1.5B-Instruct (Student)	73.3	54.9	45.9	18.9	30.3	44.7
SeqKD	75.8	57.3	47.3	17.7	31.3	45.9
KD	75.9	58.1	45.5	21.1	31.3	46.3
ABKD	77.4	58.6	48.5	20.4	32.0	47.4

ABKD applies a **simple FKLD calibration**, boosting pass@1 by **1.1%** on average in Qwen2.5-Math-7B \rightarrow 1.5B, with **strong gains** on **GSM8K and GAOKAO 2023 En**.

◆ Effects of Loss Functions

Table 10. ROUGE-L scores (\uparrow) of different loss functions on five task-agnostic instruction-following datasets when distilling GPT-2 XL (1.5B) into GPT-2 (0.1B). We report the average and standard deviation of ROUGE-L scores across five random seeds [10, 20, 30, 40, 50].

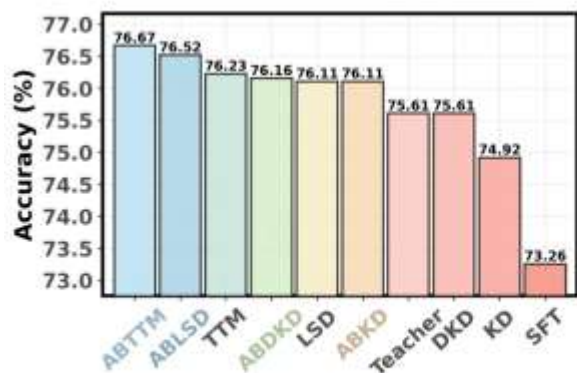
Loss Function	Dolly	Self-Instruct	Vicuna	Super-Natural	Unnatural
FKLD	23.80 (0.37)	10.01 (0.75)	15.25 (0.65)	17.69 (0.26)	18.99 (0.05)
RKLD	24.77 (0.37)	12.02 (0.48)	15.06 (0.28)	23.27 (0.29)	26.01 (0.11)
WSD	23.33 (0.52)	10.52 (0.47)	14.83 (0.61)	19.67 (0.13)	21.21 (0.21)
BDKD	23.94 (0.24)	11.83 (0.39)	15.21 (0.23)	19.56 (0.23)	21.66 (0.23)
Jensen-Shannon divergence	23.79 (0.24)	11.52 (0.18)	15.35 (0.80)	21.36 (0.17)	21.97 (0.10)
AKL	23.83 (0.59)	10.87 (0.42)	15.63 (0.66)	20.07 (0.32)	21.97 (0.13)
AlphaNet	25.13 (0.27)	12.46 (0.46)	15.64 (0.40)	25.27 (0.20)	27.56 (0.15)
SKL	25.01 (0.23)	12.47 (0.29)	<u>15.98</u> (0.84)	25.56 (0.31)	27.51 (0.07)
SRKL	<u>25.75</u> (0.39)	11.58 (0.49)	15.56 (0.17)	<u>26.13</u> (0.25)	27.37 (0.18)
α -divergence	25.15 (0.41)	<u>12.92</u> (0.22)	15.60 (0.27)	24.83 (0.21)	<u>27.81</u> (0.10)
β -divergence	24.12 (0.38)	11.18 (0.27)	14.95 (0.33)	20.98 (0.23)	23.15 (0.14)
α - β -divergence (Ours)	25.65 (0.24)	13.47 (0.42)	16.06 (0.25)	26.47 (0.31)	29.32 (0.08)

The α - β -divergence, as a **unified and theoretically supported** optimization goal, **performs better than previous baselines.**

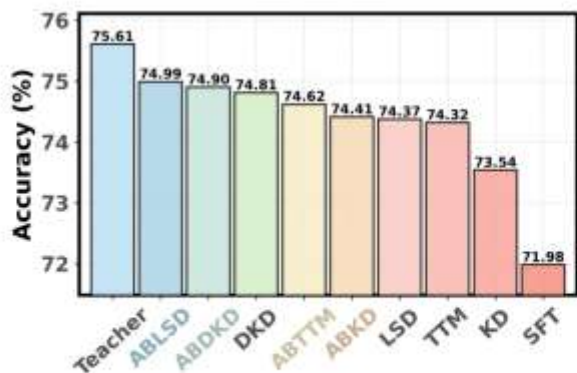
Experiments



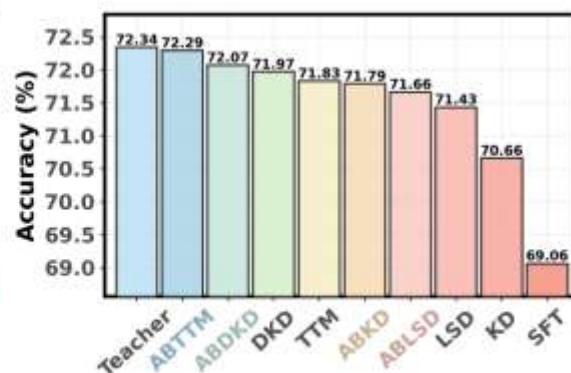
Image Classification



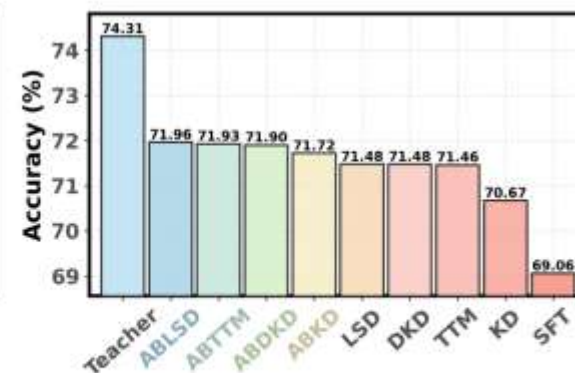
(a) WRN-40-2 → WRN-16-2



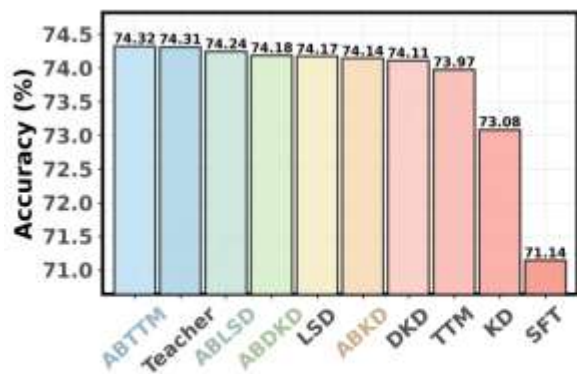
(b) WRN-40-2 → WRN-40-1



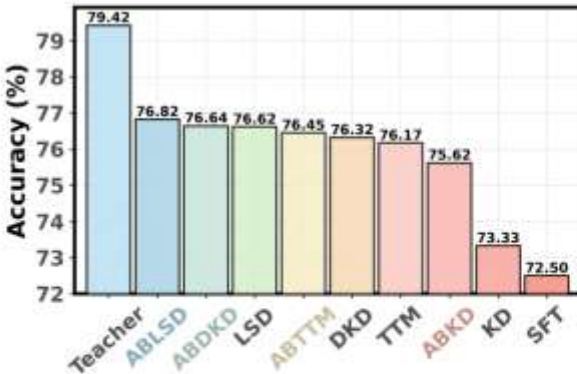
(c) resnet56 → resnet20



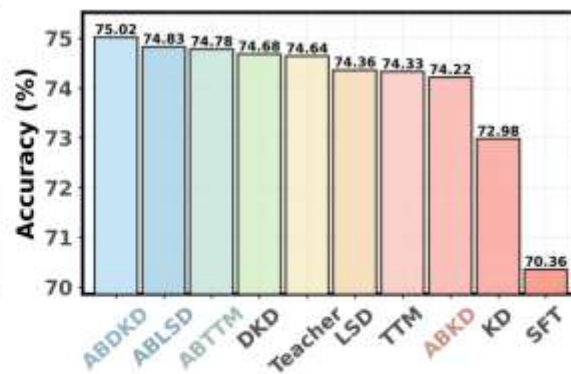
(d) resnet110 → resnet20



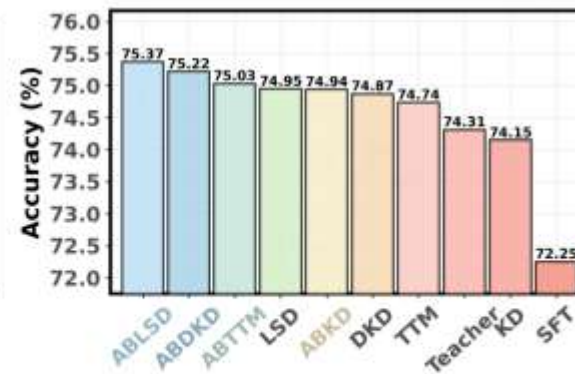
(e) resnet110 → resnet32



(f) resnet32x4 → resnet8x4



(g) vgg13 → vgg8



(h) resnet110 → resnet44

Conclusions



- ◆ **Theoretically:** identify the limitations of FKLD and RKLD via hardness and confidence concentration effects and show that α - β -divergence flexibly balances the two.
- ◆ **Methodologically:** introduce ABKD, a unified and extensible framework covering FKLD, RKLD, and other unexplored divergences.
- ◆ **Empirically:** extensive experiments on NLP and vision tasks demonstrate ABKD' s effectiveness across diverse teacher-student settings.



Wechat



Paper



Code



Email: guanghui6691@gmail.com