Weak-to-Strong Knowledge Distillation Accelerates Visual Learning

Baiang Li [1] Wenhao Chai [1] Felix Heide [1]

[1]Princeton University

Problem & Motivation

- Training large vision models requires **long schedules** and high compute cost. Modern foundation models (DINOv2, EVA-02, SigLIP) are built iteratively from prior checkpoints.
- Knowledge distillation (KD) typically transfers from a **strong** teacher to a **weak** student for **model compression**.
- Weaker models from prior runs or public checkpoints are freely available but **never used for training acceleration**.

Can a weaker teacher speed up training of a stronger student?

Method: 3-Step Plug-and-Play Recipe

Step 1 — Freeze a weak teacher

Use any existing weaker model (prior checkpoint, smaller architecture, public model). The teacher is frozen — no retraining or fine-tuning needed.

Step 2 — Add early-stage distillation

Augment the base loss with a KD term using a **warmup–hold–decay** weight schedule. The rest of the training pipeline stays **completely unchanged**.

Step 3 — Stop after surpass

Once the student exceeds teacher-level performance for $k{=}2$ consecutive validations, turn off KD **permanently**. Training continues as baseline.

$$\mathcal{L}(u) = \mathcal{L}_{\text{base}} + \gamma\,\lambda(u)\,\mathcal{L}_{\text{distill}}$$

$\lambda(u)$: warmup–hold–decay, set to 0 after surpass. When $\lambda{=}0$, training is **exactly** the baseline.

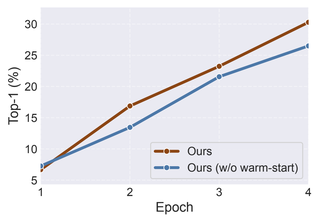
Task-agnostic design

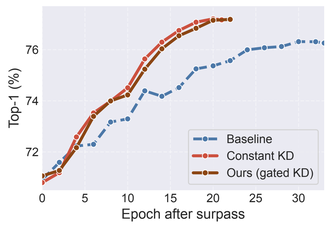
The **same recipe** applies across all tasks:

- **Classification** — KL divergence on softened posteriors (temperature $6 \to 1$)
- **Object Detection** — focal-logit distillation with confidence masking
- **Diffusion Generation** — MSE on noise predictions with timestep masking

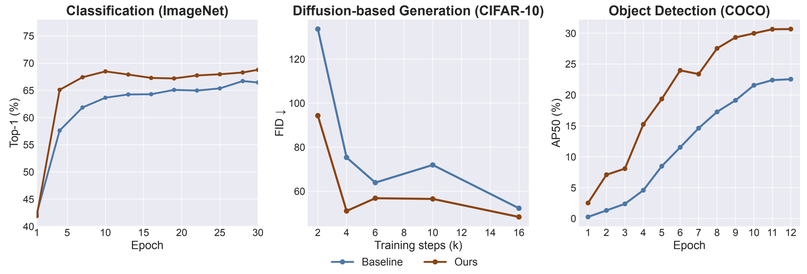
Shared: **frozen weak teacher**, **early-stage schedule**, **stop-after-surpass**.

Ablation: Warm-Start & Stop-After-Surpass





Training Acceleration Across Visual Tasks

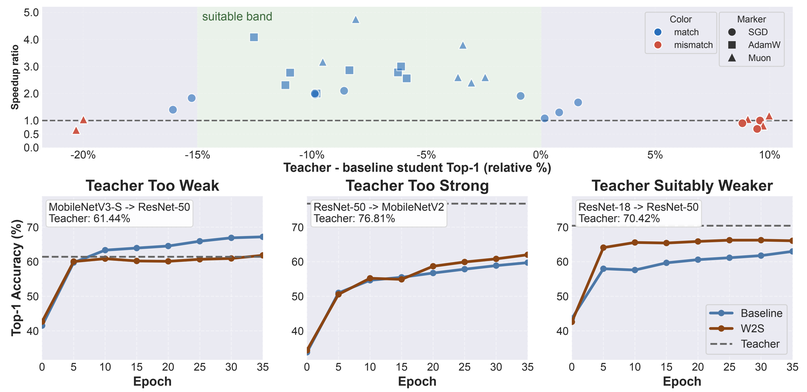


■**Baseline** ■**Ours (weak-to-strong KD)**

Our method reaches target quality earlier in **all three visual learning tasks**:

- **4.75**× fewer epochs — ImageNet classification (R18→R50, Muon, $\tau{=}65$)
- **2.67**× fewer steps — CIFAR-10 diffusion generation ($\tau{=}$FID 60)
- **1.67**× fewer epochs — COCO object detection ($\tau{=}$AP50 20%)

The Teacher Operating Band



Acceleration depends on the teacher–student accuracy gap:

- **Too weak** (>15% gap): uninformative — **speedup** $\leq 1\times$
- **Suitably weaker** (5–15%): informative + learnable — **strongest acceleration**
- **Too strong** (teacher $\geq$ student): too sharp early — **near-parity**

Consistent across **classification, detection, and generation**.

Classification Results

Dataset	Teacher→Student	Opt.	Speedup	Top-1
ImageNet	R18 → R50	Muon	**4.75**×	77.11
ImageNet	R18 → R50	AdamW	**2.86**×	77.72
ImageNet	R18 → R50	SGD	**1.62**×	76.81
ImageNet	MNv2 → R50	Muon	**3.17**×	77.09
ImageNet	MNv2 → R50	AdamW	**2.00**×	77.56
CIFAR-100	DN40 → DN100	SGD	**1.83**×	82.51
CIFAR-10	MNv2 → R50	AdamW	**1.60**×	94.00

Final accuracy **preserved or improved** in all 11 tested settings.

Detection, Generation & Wall-Clock

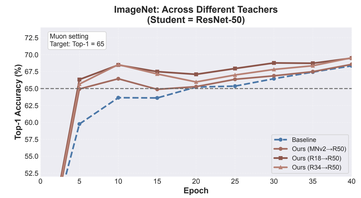
Task	Teacher→Student	τ	Speedup	Best
Detection	RetR34 → RetR50	AP50 20	**1.67**×	30.67
Detection	FRCNN-R18 → R50	AP50 20	**1.33**×	36.72
Generation	nc64 → nc128	FID 60	**2.67**×	47.22
Generation	nc64 → nc160	FID 60	**1.50**×	47.67

Wall-clock validation (ImageNet Muon, R18→R50)

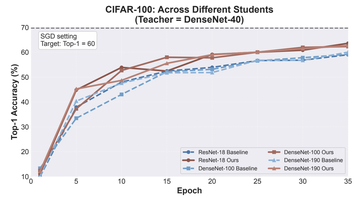
Setting	Time-to-τ	Speedup
Baseline (student-only)	6.38 h	1.00×
Online Teacher	1.43 h	**4.45**×
Cached Teacher	1.34 h	**4.75**×

Epoch speedup → **real wall-clock savings**. Teacher amortizes over ≥3 runs.

Teacher/Student Generalization



Fix student, **swap teachers**.



Fix teacher, **swap students**.

Key Takeaways

1. Up to **4.75**× fewer epochs (ImageNet), **2.67**× (diffusion), **1.67**× (detection). Accuracy preserved.
2. Plug-and-play: freeze weak teacher, add early KD, stop after surpass.
3. 5–15% weaker teachers give the best acceleration.

baiang.li@princeton.edu