Deterministic Continuous Replacement: Fast and Stable Module Replacement in Pretrained Transformers

Rowan Bradbury
Bradbury Group
rowan@bradburygroup.org

Aniket Srinivasan Ashok
Bradbury Group
University of Waterloo
aniket@bradburygroup.org

Sai Ram Kasanagottu
Bradbury Group
SUNY Stony Brook
sairam@bradburygroup.org

Gunmay Jhingran
Bradbury Group
Delhi Technological University
gunmay@bradburygroup.org

Shuai Meng
Bradbury Group
UC Berkeley
shuai@bradburygroup.org

Abstract

Replacing modules in pretrained models—especially swapping quadratic self-attention for efficient attention alternatives—poses a hard optimization problem: cold-start reinitialization destabilizes frozen backbones. We isolate this core stability challenge in a controlled study. **Deterministic Continuous Replacement (DCR)** blends teacher and student outputs with a deterministic, annealed weight $\alpha(t)$. Theoretically, DCR eliminates gate-induced gradient variance inherent to stochastic replacement (Sec. 3.2). Empirically, DCR attains faster convergence and stronger alignment than stochastic gating and distillation baselines on controlled attention replacement, establishing a foundation for heterogeneous operator swaps.

1 Introduction

As training costs rise, model adaptation has become critical. Two trends converge: compression pipelines replace blocks with smaller surrogates [Han et al., 2015, Sanh et al., 2019], and efficient attention variants [Wang et al., 2020, Choromanski et al., 2020, Beltagy et al., 2020] promise $O(n)$ or $O(n \log n)$ complexity. However, replacing modules with cold-start operators inside frozen backbones destabilizes optimization: downstream blocks receive out-of-distribution features, leading to optimization instability, ineffective gradient updates, and slow recovery. Existing approaches face fundamental tradeoffs. Knowledge distillation methods [Hinton et al., 2015, Romero et al., 2014] require expensive teacher forward passes and enforce rigid feature matching, while stochastic replacement strategies like BERT-of-Theseus [Xu et al., 2020] introduce gradient variance and uneven recovery. We isolate the *core replacement stability problem*: integrating a randomly initialized module into a frozen backbone—the optimization challenge common to all replacement scenarios. By studying this in a controlled setting (replacing attention with re-initialized attention), we eliminate

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ScaleOPT: GPU-Accelerated and Scalable Optimization.

representational mismatch as a confound, allowing us to attribute differences solely to stability mechanisms and rigorously measure gradient dynamics.

Our contributions.

- **DCR method** (Sec. 3): deterministic blending that eliminates gate-induced gradient variance and naturally enables near-zero-cost feature alignment since both teacher and student outputs are computed for the blend.
- **Variance reduction theory** (Sec. 3.2): formal analysis isolating and eliminating the gate-induced variance term inherent to stochastic replacement (Props. 1–2), with bounds on curvature bias and loss-path smoothness.
- **Controlled validation** (Sec. 4): faster convergence and stronger alignment than stochastic gating and distillation baselines in a controlled self-replacement setting that isolates stability from representational mismatch.

While experiments focus on controlled self-replacement on smaller models (CIFAR-100, ViT-Small) to enable rigorous ablation, the method and theory are explicitly constructed for heterogeneous operator swaps (e.g., Linformer, Performer, sparse/Fourier attention), which is the immediate follow-on work. We present this study in workshop format because isolating and formalizing the replacement-stability gap is a prerequisite for scalable deployment in production settings, where understanding failure modes and convergence guarantees is critical. DCR's efficiency advantage over distillation is amplified in compute-saturated regimes—large language model or diffusion transformer replacement—where GPU utilization is high and the full teacher model forward pass required by distillation directly increases wall-clock cost relative to DCR's branch-local teacher evaluation.

2 Related Work

Knowledge Distillation and Model Compression. Knowledge distillation [Hinton et al., 2015] and feature-based variants [Romero et al., 2014, Touvron et al., 2021] require a full separate teacher model forward pass per training step and impose rigid alignment constraints. In contrast, DCR computes teacher outputs only at the replaced modules, avoiding full-model duplication. Compression methods [Han et al., 2015, Sanh et al., 2019, Jiao et al., 2019] assume parameter compatibility, which breaks down for heterogeneous operators.

Stochastic Module Replacement. BERT-of-Theseus [Xu et al., 2020] randomly selects between teacher and student modules via a Bernoulli gate $z_\ell(t) \sim \text{Bernoulli}(p(t))$, enabling gradual knowledge transfer without explicit feature matching. However, this introduces gradient variance in the student-only gradient:

$$\nabla_{\theta_\ell} L = z_\ell \cdot J_{G_\ell}^\top \frac{\partial S_\ell}{\partial \theta_\ell},$$

producing high variance when $p(t)$ is mid-range. We replace this stochastic gate with a deterministic blend. For controlled ablation, we introduce **Theseus-Gumbel**, using soft Gumbel-Softmax gates $r_\ell(t) \in (0, 1)$ with temperature τ:

$$r_\ell(t) = \text{GumbelSoftmax}(p(t), \tau),$$

$$x_{\ell+1} = x_\ell + r_\ell(t) S_\ell(h_\ell) + (1 - r_\ell(t)) T_\ell(h_\ell).$$

This allows gradients to flow but retains gate-induced variance (GUM, GUM+DFG baselines in experiments).

Table 1 summarizes the key differences between DCR and prior module replacement approaches across three critical dimensions: gradient variance, computational overhead, and feature matching requirements.

Replacement Stability Gap. Prior methods assume parameter continuity; under cold-start reinitialization, downstream layers receive out-of-distribution features. **DCR targets this gap** with a deterministic, low-variance path stable under reinitialization.

Table 1: Comparison of module replacement methods. DCR achieves low gradient variance and minimal branch-local overhead **in our non-compute-saturated experimental regime**, addressing the stability bottleneck that limits both distillation and stochastic replacement.

Method	Gradient Variance	Extra Compute	Needs Feature Matching
Knowledge Distillation	Low	High (full teacher model forward)	Yes (rigid)
Theseus (Stochastic)	High (gate-induced)	Low (gate overhead only)	No
DCR (Ours)	Low (deterministic)	Low (branch-local teacher only at replaced layers)	No (optional via DFG)

3 Methodology

3.1 Problem Formulation

Given pretrained network F with L modules, we replace subset $\mathcal{I} \subseteq \{1, \ldots, L\}$. For $\ell \in \mathcal{I}$, let T_ℓ (frozen teacher) and $S_\ell(\cdot; \theta_\ell)$ (trainable student) share input/output shapes. Denote normalized input $h_\ell = \mathrm{LN}(x_\ell)$ and frozen tail G_ℓ. DCR blends on the residual branch:

$$x_{\ell+1}(t) \;=\; x_\ell(t) \;+\; \big[\alpha(t)\, T_\ell(h_\ell(t)) \;+\; (1-\alpha(t))\, S_\ell(h_\ell(t); \theta_\ell)\big], \tag{1}$$

with global gate $\alpha(t) \in [0, 1]$: $\alpha(0) = 1$ (teacher-only) $\to \alpha(T) = 0$ (student takeover).

3.2 Theoretical Properties (Stability)

Analysis scope. We analyze gate-induced variance and path geometry under frozen G_ℓ and scheduled gates, holding the student's input distribution fixed at each step. This provides local, conditional intuition—not full training-dynamics guarantees or global convergence proofs—validated empirically in Sec. 4. Standard assumptions: teacher runs in `eval()` mode, gates independent of minibatch, differentiable functions.

Lower gate-induced gradient variance. To our knowledge, this is the first formulation that analytically isolates and eliminates the gate-induced variance term central to stochastic replacement. Let

$$a \;:=\; J_{G_\ell}^\top \frac{\partial S_\ell}{\partial \theta_\ell} \;\in \mathbb{R}^{\dim(\theta_\ell)}.$$

Proposition 1 (Variance decomposition for Theseus). If a hard gate $z \sim \mathrm{Bernoulli}(p)$ selects student vs. teacher (independent of the data), then

$$\nabla_{\theta_\ell} L \;=\; z\, a, \qquad \mathbb{E}[\nabla_{\theta_\ell} L] = p\, \mathbb{E}[a],$$
$$\mathrm{Var}[\nabla_{\theta_\ell} L] \;=\; p\, \mathrm{Var}[a] \;+\; p(1-p)\, \|\mathbb{E}[a]\|^2 \;\le\; p\, \mathrm{Var}[a] + p(1-p)\, \mathbb{E}\|a\|^2.$$

This decomposition reveals the gate-induced variance component that DCR eliminates. *(Proof in Appendix A.2.1.)*

Proposition 2 (Deterministic gate removes gate-induced variance). Let $Y_\alpha := (1-\alpha)\, T_\ell + \alpha\, S_\ell$ and define $a(y) := J_{G_\ell}(y)^\top \frac{\partial S_\ell}{\partial \theta_\ell}$, so that the DCR gradient is

$$\nabla_{\theta_\ell} L_{\mathrm{DCR}} \;=\; \alpha\, a(Y_\alpha),$$

whereas under Theseus (hard gate $z \sim \mathrm{Bernoulli}(p)$) the student gradient is $\nabla_{\theta_\ell} L_{\mathrm{Th}} = z\, a(S_\ell)$. Let X denote the minibatch. Then the *gate-induced* component of gradient variance,

$$\mathbb{E}\big[\mathrm{Var}\big(\nabla_{\theta_\ell} L \,\big|\, X\big)\big]\,,$$

is zero for DCR and equals $p(1-p)\, \mathbb{E}\|a(S_\ell; X)\|^2$ for Theseus. Hence,

$$\mathbb{E}\big[\mathrm{Var}\big(\nabla_{\theta_\ell} L_{\mathrm{Th}} \,\big|\, X\big)\big] \;-\; \mathbb{E}\big[\mathrm{Var}\big(\nabla_{\theta_\ell} L_{\mathrm{DCR}} \,\big|\, X\big)\big] \;=\; p(1-p)\, \mathbb{E}\|a(S_\ell; X)\|^2 \;\ge\; 0.$$

This is the core theoretical justification for DCR: strictly lower gradient variance than stochastic gating. *(Proof in Appendix A.2.2.)*

Remark (Soft gates / Theseus-Gumbel). Let $r \in (0, 1)$ be a random soft gate with $\mathbb{E}[r] = p$, $\mathrm{Var}(r) > 0$ (e.g., Gumbel-Softmax, temperature τ). Then

$$\mathrm{Var}[ra] = \mathbb{E}[r^2]\mathrm{Var}[a] + \mathrm{Var}(r)\, \|\mathbb{E}[a]\|^2 = p^2\, \mathrm{Var}[a] \;+\; \underbrace{\mathrm{Var}(r)\, \mathbb{E}\|a\|^2}_{\text{extra, gate-induced}},$$

so any stochastic gate incurs an additional nonnegative term that DCR does not.

Curvature bias. Stochastic mixing through nonlinearities introduces a curvature-dependent bias: the expected output after a nonlinearity differs from the nonlinearity applied to the expected blend. DCR's deterministic path avoids this entirely (Proposition 3, Appendix A.2.3).

Summary. Props. 1–2 show DCR removes gate-induced variance. Proposition 3 (Appendix A.2.3) shows DCR avoids curvature bias from stochastic mixing through nonlinearities. Proposition 4 (Appendix A.2.4) bounds the loss path via Lipschitz continuity. Under the stated assumptions (independent gate scheduling, fixed frozen tail, local smoothness), deterministic blending yields a better-conditioned optimization path. These results are local and conditional, not full training-dynamics guarantees. Empirical validation in Sec. 4. Algorithm 1 (Appendix) details the full procedure.

3.3 Deep Feature Guidance (DFG)

While DCR ensures a smooth replacement, the student can benefit from direct interface alignment. Since DCR already evaluates both $T_\ell(h_\ell)$ and $S_\ell(h_\ell)$ for the blend at replaced layers, the auxiliary loss adds near-zero marginal cost—no additional forward passes are required. This contrasts with standard knowledge distillation, which requires a full teacher model pass. We add an auxiliary loss on the residual outputs at the replaced sites:

$$\mathcal{L}_{\mathrm{DFG}} \;=\; \sum_{\ell \in \mathcal{I}} \big\| S_\ell(h_\ell) - T_\ell(h_\ell) \big\|_2^2, \qquad h_\ell = \mathrm{LN}(x_\ell). \tag{2}$$

Let $\hat{y} = F_t(x)$ denote the model output under the current global gate $\alpha(t)$. The overall objective is

$$\mathcal{L}_{\mathrm{total}} \;=\; \mathcal{L}_{\mathrm{task}}(\hat{y},\, y^\star) \;+\; \lambda\, \mathcal{L}_{\mathrm{DFG}}, \qquad \lambda \geq 0 \tag{3}$$

Since DCR already computes both $T_\ell(h_\ell)$ and $S_\ell(h_\ell)$ for blending, DFG adds negligible cost ($\lambda \geq 0$ controls strength). In our non-compute-saturated regime, DCR's overhead scales with the number of replaced modules $|\mathcal{I}|$, whereas knowledge distillation incurs a full teacher forward regardless of $|\mathcal{I}|$. We anneal λ following the same aggr20 schedule as α; full schedule in Appendix A.1. For Theseus-Gumbel+DFG, we evaluate the teacher branch locally even when the gate selects the student.

4 Controlled Evaluation of Replacement Stability

4.1 Experimental Setup

Datasets and Models. We evaluate DCR on ImageNet-pretrained ViT Small models [Dosovitskiy et al., 2020], serving as the teacher backbone finetuned on CIFAR100 [Krizhevsky et al., 2009]. Student modules replace attention blocks and are randomly re-initialized (Kaiming initialization [He et al., 2015]). All DCR blending is applied post-softmax and prior to residual addition.

Replacement Schedules. DCR (aggr20): α transitions $1.0 \rightarrow 0.0$ over first 20% of training. Theseus variants use inverse probability p. DFG anneals with aggr20 schedule.

4.2 Results

Alignment Dynamics. DCR and DCR+DFG achieve consistently higher interface cosine similarity than stochastic baselines (Figure 1), with largest gains in mid and late blocks. Crucially, deterministic blending ensures downstream blocks receive in-distribution features from the start, enabling later layers to learn earlier without wasted gradients—avoiding the plateauing seen in GUM and BERN where gate-induced starvation delays deep-layer convergence.

Accuracy Recovery and DFG Effect. DCR variants reach target accuracy sooner in both epoch and wall-clock views (Figure 2), despite similar final accuracies (≈78–80%). Adding DFG accelerates takeover without full teacher passes, with strongest gains in deeper blocks—confirming near-zero-cost feature guidance compounds with deterministic blending.

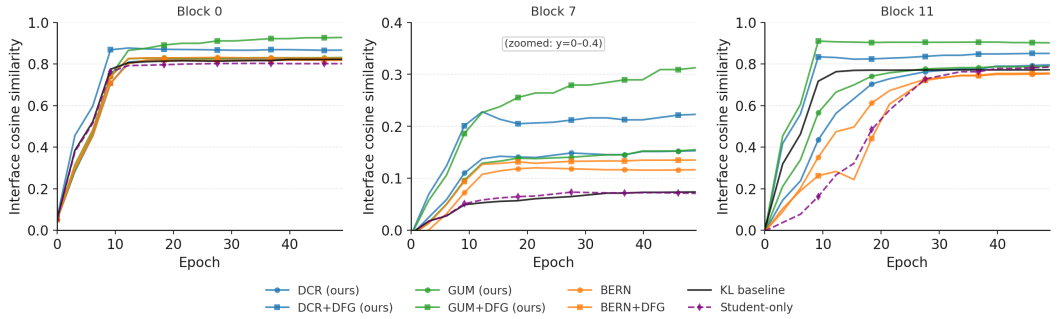

Figure 1: Interface cosine similarity (cosine similarity of residual outputs) between teacher and student outputs at different layers (Block 0, 7, 11) across training epochs.

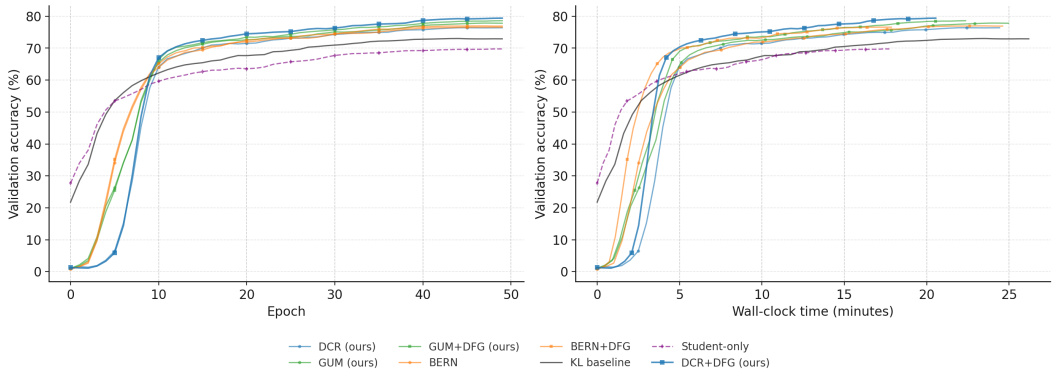

Figure 2: Validation accuracy during module replacement on CIFAR-100 (ViT-Small/16). Left: epochs. Right: wall-clock time.

5 Research Positioning and Scope

Our controlled design prioritizes internal validity: single model family (ViT-Small), dataset (CIFAR100), single seed, I/O-bound regime, pre-norm residual Transformers, and self-replacement (attention $\rightarrow$ re-initialized attention) to isolate stability from representational mismatch. Intentional scoping choices include: (i) global gate $\alpha(t)$ rather than per-layer or progressive schedules, (ii) comparison to Theseus variants and student-only baselines without function-preserving initialization (Net2Net) or stronger alignment methods (CKA, Gram matching, learned adapters), (iii) no exhaustive hyperparameter tuning. Architectures with batch normalization, extensive simultaneous replacements across many layers, or other normalization schemes may exhibit different stability dynamics. These constraints enable causal attribution of variance reduction effects—rarely possible when varying architecture, operators, and compute simultaneously—and establish methodological clarity as a prerequisite for scaling to heterogeneous operators. Results should be interpreted as feasibility evidence rather than definitive benchmarking. Extensions to compute-saturated regimes, heterogeneous operators, and diverse architectures are natural next steps.

6 Conclusion

We introduced **Deterministic Continuous Replacement (DCR)**, which eliminates gate-induced gradient variance in cold-start module replacement. In controlled experiments, DCR+DFG outperforms stochastic gating and distillation baselines, establishing a foundation for heterogeneous operator swaps. For frozen-backbone replacement under our assumptions, DCR provides a stable, efficient alternative to stochastic methods. Next steps include: heterogeneous operators (efficient attention variants), larger models, compute-saturated regimes, and per-layer adaptive α schedules conditioned on interface similarity for deep architectures.

Acknowledgements

This work was conducted by the Bradbury Group, an independent non-profit AI research lab. Beyond the listed authors, we thank the broader Bradbury Group research community for internal reviews, discussion of early prototypes, and support with maintaining the shared training and evaluation infrastructure, with special thanks to Elea Zhong and Joseph Haynes for their detailed manuscript review and feedback. This project received no external funding and was enabled entirely through the lab's internal resources.

References

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

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

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arxiv 2014. *arXiv preprint arXiv:1412.6550*, 2014.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. Bert-of-theseus: Compressing bert by progressive module replacing. *arXiv preprint arXiv:2002.02925*, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015. URL `http://arxiv.org/abs/1502.01852`.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL `http://arxiv.org/abs/1711.05101`.

A Technical Appendices and Supplementary Material

A.1 Detailed Experimental Setup

Training Procedure. Experiments are conducted on NVIDIA A100 GPUs with mixed-precision (BF16) on CIFAR100 dataset. Training follows two stages: (i) **Head warmup**: only the classification head is trained for 6 epochs with 2 epoch warmup and lr 1×10^{-3}, cosine annealing to 1×10^{-6} full weights unfrozen and further finetune at 1×10^{-4} for 6 further epochs, label smoothing 0.1, and mixed precision; (ii) **Full model training**: all layers unfrozen, base lr 5×10^{-4} with cosine annealing over 50 epochs, weight decay 0.05, gradient clipping 1.0, batch size 128, AdamW optimizer [Loshchilov and Hutter, 2017](eps 1×10^{-8}, betas $[0.9, 0.999]$), and label smoothing 0.1.

Replacement Schedules. For DCR (aggr20), α transitions from 1.0 to 0.0 over the first 20% of training: Phase 1 (0–10%) $\alpha = 1.0 \to 0.3$; Phase 2 (10–20%) $\alpha = 0.3 \to 0.0$; Phase 3 (20–100%) $\alpha = 0.0$. Stochastic Theseus variants follow the inverse probability p: Phase 1 (0–10%) $p = 0.1 \to 0.7$; Phase 2 (10–20%) $p = 0.7 \to 1.0$; Phase 3 (20–100%) $p = 1.0$. Constant 0.7 and 0.5 schedules for Theseus, as well as linear 0.1–1.0 over 50% of steps was attempted for Theseus as suggested by their paper and extrapolated to our training setup, but aggr20 was found to outperform for our experiments. DFG showed best results with matched schedule to aggr20.

Additional Details. Key settings and terms used in experiments:

- **Gumbel**: Theseus-Gumbel stochastic replacement with temperature $\tau = 1.0$.
- **KL Distillation**: Teacher-student soft-target guidance with fixed temperature 4.
- **DFG (Deep Feature Guidance)**: Auxiliary L2 loss on student-teacher intermediate outputs, controlled by weight λ.
- **Student Initialization**: Reinitialized attention modules using Kaiming initialization.
- **Hyperparameter Search**: Optimal settings determined via preliminary student-only cross-entropy training.

A.2 Theoretical Results and Proofs

This section contains the full statements and proofs of all theoretical propositions referenced in the main text.

A.2.1 Proposition 1: Variance Decomposition for Theseus

Proposition 1 (Variance decomposition for Theseus). If a hard gate $z \sim \mathrm{Bernoulli}(p)$ selects student vs. teacher (independent of the data), then

$$\nabla_{\theta_\ell} L \;=\; z\, a, \qquad \mathbb{E}[\nabla_{\theta_\ell} L] = p\, \mathbb{E}[a],$$

$$\mathrm{Var}[\nabla_{\theta_\ell} L] \;=\; p\, \mathrm{Var}[a] \;+\; p(1-p)\, \|\mathbb{E}[a]\|^2 \;\le\; p\, \mathrm{Var}[a] + p(1-p)\, \mathbb{E}\|a\|^2,$$

where $a := J_{G_\ell}^\top \frac{\partial S_\ell}{\partial \theta_\ell} \in \mathbb{R}^{\dim(\theta_\ell)}$.

Proof. By independence, $\mathbb{E}[za] = \mathbb{E}[z]\mathbb{E}[a] = p\, \mathbb{E}[a]$. For the variance, by the law of total variance,

$$\begin{aligned} \mathrm{Var}[za] \;&=\; \mathbb{E}[\mathrm{Var}(za \mid z)] + \mathrm{Var}(\mathbb{E}[za \mid z]) \\ &= \mathbb{E}[z^2]\, \mathrm{Var}[a] + \mathrm{Var}(z\, \mathbb{E}[a]) \\ &= p\, \mathrm{Var}[a] + p(1-p)\, \|\mathbb{E}[a]\|^2. \end{aligned}$$

Use $\|\mathbb{E}[a]\|^2 \le \mathbb{E}\|a\|^2$ for the inequality. □

A.2.2 Proposition 2: Deterministic Gate Removes Gate-Induced Variance

Proposition 2 (Deterministic gate removes gate-induced variance). Let $Y_\alpha := (1-\alpha)\, T_\ell + \alpha\, S_\ell$ and define $a(y) := J_{G_\ell}(y)^\top \frac{\partial S_\ell}{\partial \theta_\ell}$, so that the DCR gradient is

$$\nabla_{\theta_\ell} L_{\mathrm{DCR}} \;=\; \alpha\, a(Y_\alpha),$$

whereas under Theseus (hard gate $z \sim \mathrm{Bernoulli}(p)$) the student gradient is $\nabla_{\theta_\ell} L_{\mathrm{Th}} = z\, a(S_\ell)$. Let X denote the minibatch. Then the *gate-induced* component of gradient variance,

$$\mathbb{E}\big[\mathrm{Var}\big(\nabla_{\theta_\ell} L \,\big|\, X\big)\big]\,,$$

is zero for DCR and equals $p(1-p)\,\mathbb{E}\|a(S_\ell; X)\|^2$ for Theseus. Hence,

$$\mathbb{E}\big[\mathrm{Var}\big(\nabla_{\theta_\ell} L_{\mathrm{Th}} \,\big|\, X\big)\big] \;-\; \mathbb{E}\big[\mathrm{Var}\big(\nabla_{\theta_\ell} L_{\mathrm{DCR}} \,\big|\, X\big)\big] \;=\; p(1-p)\,\mathbb{E}\|a(S_\ell; X)\|^2 \;\geq\; 0.$$

Proof. Condition on X. Under Theseus, $\nabla_{\theta_\ell} L_{\mathrm{Th}} = z\, a(S_\ell; X)$ with $z \perp X$, so $\mathrm{Var}(\nabla_{\theta_\ell} L_{\mathrm{Th}} \mid X) = p(1-p)\|a(S_\ell; X)\|^2$ and taking expectation over X gives the stated value. Under DCR there is no gate randomness given X, so $\mathrm{Var}(\nabla_{\theta_\ell} L_{\mathrm{DCR}} \mid X) = 0$. □

A.2.3 Proposition 3: Curvature Bias Bound

Proposition 3 (Curvature bias bound). Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable with $\sup_{y \in \mathrm{seg}(T_\ell, S_\ell)} \|\nabla^2 \psi(y)\|_{\mathrm{op}} \leq M$ (segment between T_ℓ and S_ℓ), and set $\mu = (1-p)\,T_\ell + p\,S_\ell$, $\Delta = S_\ell - T_\ell$. For Theseus with $Y = z\,S_\ell + (1-z)\,T_\ell$ where $z \sim \mathrm{Bernoulli}(p)$, we have

$$\big|\, \mathbb{E}[\psi(Y)] - \psi(\mu) \,\big| \;\leq\; \tfrac{M}{2}\, p(1-p)\, \|\Delta\|^2.$$

Proof. Second-order Taylor around μ gives $\psi(Y) = \psi(\mu) + \nabla\psi(\mu)^\top (Y-\mu) + \frac{1}{2}(Y-\mu)^\top \nabla^2 \psi(\xi_Y)\,(Y-\mu)$ for some ξ_Y on the segment between Y and μ. Take expectations: the linear term vanishes ($\mathbb{E}[Y-\mu] = 0$), and $\mathbb{E}\|Y-\mu\|^2 = \mathbb{E}[(z-p)^2]\|\Delta\|^2 = p(1-p)\|\Delta\|^2$. □

Corollary (Deterministic path avoids mixing bias). For DCR, $Y_\alpha = (1-\alpha)\,T_\ell + \alpha\,S_\ell$ is deterministic, so $\mathbb{E}[\psi(Y_\alpha)] = \psi(Y_\alpha) = \psi(\mathbb{E}[Y_\alpha])$; no stochastic mixing bias arises. Theseus pays a curvature-dependent penalty scaling with $p(1-p)\|\Delta\|^2$.

A.2.4 Proposition 4: Smooth Loss Path

Proposition 4 (Smooth loss path). Let $f(y) := (L \circ G_\ell)(y)$. If f is L_y-Lipschitz on the segment between T_ℓ and S_ℓ, then along $y(\alpha) = (1-\alpha)\,T_\ell + \alpha\,S_\ell$,

$$\big|\, f(y(\alpha)) - f(y(0)) \,\big| \;\leq\; L_y\, \alpha\, \|S_\ell - T_\ell\| \;\leq\; L_y\, \alpha\, D_\ell \quad \text{for any } D_\ell \geq \|S_\ell - T_\ell\|.$$

Proof. $y(\alpha) - y(0) = \alpha(S_\ell - T_\ell)$ and Lipschitz continuity give the bound. □

A.3 DCR Implementation Details

Algorithm 1 DCR training step with global gate $\alpha(t)$ (pre-norm residual).

Require: batch $(x, y^\star)$; frozen teachers $\{T_\ell\}_{\ell \in \mathcal{I}}$ (`eval()`, no-grad); trainable students $\{S_\ell\}_{\ell \in \mathcal{I}}$; global gate $\alpha(t)$

$x_1 \leftarrow x$
for $\ell = 1$ **to** L **do**
  if $\ell \notin \mathcal{I}$ **then**
    $x_{\ell+1} \leftarrow$ original block forward
  else
    $h_\ell \leftarrow \mathrm{LN}(x_\ell)$
    without gradients: $t_\ell \leftarrow T_\ell(h_\ell)$
    $s_\ell \leftarrow S_\ell(h_\ell)$
    $x_{\ell+1} \leftarrow x_\ell + \alpha(t)\, t_\ell + (1 - \alpha(t))\, s_\ell$
  end if
end for
$\hat{y} \leftarrow$ task head on x_{L+1}
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{task}}(\hat{y}, y^\star)$
Backprop (student params only); optimizer step; skip computing t_ℓ once $\alpha(t) = 0$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions of the paper, including the proposed method (DCR/DFG/GUM/etc.), its advantages over baselines, and the scope of experiments. These claims are supported by the theoretical analysis and experimental results presented in the paper. The limitations and assumptions, such as dataset choices and model configurations, are either explicitly mentioned or can be inferred, ensuring that the claims do not overstate the results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the limiations of the paper are discussed in the final section of the paper alongwith its conclusions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, each theoretical result alongwith experimental results are based on the the full set of assumptions that were formulated in the problem formulation section.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the training setup and experiments process has been explicitly mentioned along-with the hyper parameters used in the experiments appendix section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The dataset we used is open source. Also, we aim to create the open source code once our full limitations are resolved further.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, detailed explanation of the hyper parameters are mentioned in the appendix experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We ran a single seed due to compute limits, so we cannot report statistical significance. We flag this as a limitation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All the essential details are mentioned in the Appendix experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: No such study has been performed in this work.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risks have been observed so far. But we are welcome to suggestions.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, clearly cited and acknowledged.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No needed here.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No Human study performed in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer:[NA]

    Justification: No such study performed in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Except for the grammatically purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.