When Does Margin Clamping Affect Training Variance? Dataset-Dependent Effects in Contrastive Forward-Forward Learning

Anonymous authors
Paper under double-blind review

Abstract

Contrastive Forward-Forward (CFF) learning is a layer-local alternative to backpropagation that trains Vision Transformers using supervised contrastive objectives at each layer independently. In practice, CFF can exhibit substantial seed-to-seed variability, complicating reproducibility and hyperparameter selection. We audit one implementation detail inside the supervised contrastive loss: applying the positive-pair margin via saturating similarity clamping, $\min(s + m, 1)$. We compare this against a post-log-probability subtraction reference that we prove is gradient-neutral under the mean-over-positives reduction (Proposition 4.1), thereby isolating the effect of saturation itself. On CIFAR-10 in a 2×2 factorial ablation (n=7 seeds per cell), the clamped variant exhibits 5.90× higher pooled test-accuracy variance (p=0.003, bootstrap 95% CI [1.62, 15.80]) with no detectable difference in mean accuracy. Clamp activation rates (CAR), layerwise gradient norms, and a reduced-margin dose-response probe jointly indicate that this variance increase is associated with gradient truncation in early transformer layers. However, the effect is *dataset-dependent*: replication on CIFAR-100 (VR = 0.39×), SVHN (VR = 0.25×), and Fashion-MNIST (VR = 0.08×, p=0.029) reveals inverted variance ratios in all three cases. Cross-dataset analysis identifies layer-0 clamp activation rate as a necessary but insufficient condition for variance inflation: CIFAR-10's high L0 CAR (60.7%) co-occurs with the only elevated variance ratio, while CIFAR-100's low L0 CAR (29.0%) and SVHN/Fashion-MNIST's high task accuracy ($> 92\%$) each independently suppress the effect. An SVHN difficulty sweep confirms this interaction: increasing augmentation difficulty on the same dataset drives the variance ratio from 0.25× to 16.73×. These results characterize the conditions under which margin clamping destabilizes CFF training and offer practical guidance for practitioners.

1 Introduction

Reproducibility is a cornerstone of empirical machine learning, yet seed-to-seed variability remains a persistent challenge in modern deep learning (Dodge et al., 2020; Bhojanapalli et al., 2021). When the outcome of training depends strongly on random seed, practitioners face an immediate methodological dilemma: how many independent runs are required for a reliable conclusion, and how should hyperparameter choices be compared under finite compute?

Forward-Forward (FF) learning (Hinton, 2022) replaces end-to-end backpropagation with layer-local learning rules, optimizing each layer against its own objective. Recent work has scaled FF to Vision Transformers through *Contrastive Forward-Forward* (CFF), which trains each layer using a supervised contrastive loss on two augmented views per image (Aghagolzadeh & Ezoji, 2025; Chen et al., 2025). Because each layer is optimized independently, CFF introduces optimization dynamics that are qualitatively different from those of backpropagation-trained Transformers—yet existing CFF work has primarily focused on mean accuracy, treating the underlying loss formulation as a fixed design choice. Consequently, the seed-to-seed stability of CFF training remains uncharacterized: it is not known whether common implementation details inside the contrastive loss can systematically affect training variance.

Focus. We isolate one such implementation detail: applying the positive-pair margin via saturating similarity clamping, $\min(s + m, 1)$. This is the default margin implementation in existing CFF codebases. We compare it against a post-log-probability subtraction alternative and prove that this alternative is gradient-neutral under the mean-over-positives loss reduction (Proposition 4.1), meaning it does not reshape gradients and serves as an effectively no-functional-margin reference. The comparison therefore isolates the effect of saturation-induced gradient truncation on seed-to-seed variance.

Key findings. On CIFAR-10 in a 2×2 factorial ablation (n=7 seeds per cell), the clamped variant exhibits $5.90\times$ higher pooled test-accuracy variance than the subtraction reference (p=0.003), with no detectable difference in mean accuracy. Three diagnostic analyses—clamp activation rates, layerwise gradient norms, and a reduced-margin dose-response probe—provide converging evidence that this variance increase is associated with gradient truncation in early transformer layers.

Critically, the effect does not generalize uniformly. Replication on CIFAR-100, SVHN, and Fashion-MNIST reveals *inverted* variance ratios in all three cases, where clamping is associated with equal or lower variance than the subtraction reference. Cross-dataset analysis identifies two moderating factors: (i) clamp activation rate, which is substantially lower on CIFAR-100 due to reduced positive-pair density; and (ii) task difficulty, which compresses variance on the high-accuracy datasets regardless of truncation levels. An SVHN difficulty sweep provides direct evidence for this interaction on a single dataset.

Contributions.

1. A closed-form specification of the supervised contrastive loss used in CFF training, making the margin variants and their gradient behavior explicit (Section 4).

2. A proof that post-log-probability margin subtraction is gradient-neutral under the mean-over-positives reduction (Proposition 4.1).

3. An empirical audit on CIFAR-10 showing that similarity clamping is associated with substantially higher seed-to-seed variance, supported by mechanistic diagnostics (Sections 5 and 6).

4. A characterization of boundary conditions across four datasets (CIFAR-10, CIFAR-100, SVHN, Fashion-MNIST) and an augmentation difficulty sweep, identifying when the variance effect does and does not manifest (Section 7).

2 Background and Related Work

2.1 Forward-Forward and layer-local learning

Forward-Forward (FF) learning (Hinton, 2022) trains networks using local objectives rather than a single global loss. It belongs to a broader family of greedy layerwise methods that optimize each layer (or block) against a local target (Belilovsky et al., 2019; Nøkland & Eidnes, 2019; Löwe et al., 2019). CFF extends FF to Vision Transformers by applying a supervised contrastive objective at each layer (Aghagolzadeh & Ezoji, 2025; Chen et al., 2025). While greedy layerwise methods have been studied for their representational properties and scalability, their seed-to-seed stability has received comparatively little attention.

2.2 Contrastive and supervised contrastive learning

Self-supervised contrastive objectives learn representations by pulling positive pairs together and pushing negatives apart in a temperature-scaled softmax (van den Oord et al., 2018; Chen et al., 2020; He et al., 2020). Khosla et al. (2020) extended this framework to the supervised setting, where all same-class pairs are treated as positives. The loss used in CFF follows this supervised formulation with the "mean over positives" reduction. Variants of the supervised contrastive loss differ in their reduction over positive pairs and in how margins or similarity modifications are applied (Graf et al., 2021); our work audits one such modification.

2.3 Margins in metric and contrastive learning

Margins are widely used in metric learning to enforce separation between classes. Additive angular and cosine margins in face recognition objectives—SphereFace (Liu et al., 2017), CosFace (Wang et al., 2018), and ArcFace (Deng et al., 2019)—reshape logit contributions via nonlinear operations that can create saturated regions with reduced or truncated gradients. Our setting differs in that (i) the objective is a supervised contrastive loss applied independently at each layer, and (ii) Forward-Forward training enforces layer-local updates via stop-gradients. We document how a saturation-inducing margin manifests in this layer-local regime, with seed-to-seed variance rather than mean accuracy as the primary endpoint.

2.4 Training variance and reproducibility

Seed-to-seed variability in deep learning has been documented across fine-tuning (Dodge et al., 2020), pre-training (Bhojanapalli et al., 2021), and broader model selection (D'Amour et al., 2022). Picard (2021) demonstrated that seed choice alone can substantially affect reported metrics. Sources of variance in backpropagation-trained models—weight initialization, data ordering, dropout masks—are well characterized, but analogous analyses for layer-local training methods are largely absent. Our work addresses this gap for CFF.

2.5 Gradient truncation in optimization

Gradient clipping and truncation have been studied as both deliberate regularizers (Pascanu et al., 2013; Zhang et al., 2020) and as unintended consequences of activation saturation. In our setting, similarity clamping introduces a form of implicit gradient truncation that is data-dependent and varies across layers, distinguishing it from explicit gradient clipping strategies.

3 Problem Setup

Let σ denote a random seed that jointly determines weight initialization, data-loader ordering, and stochastic augmentation for a CFF training run. Fix an experimental condition c and let $a_c(\sigma)$ denote the resulting test accuracy on a held-out set after the full two-stage pipeline (Section 5.1).

Primary endpoint. Given n independent seeds $\{\sigma_1, \ldots, \sigma_n\}$, we measure the sample variance

$$\widehat{\mathrm{Var}}_c = \frac{1}{n-1} \sum_{k=1}^{n} \big(a_c(\sigma_k) - \bar{a}_c\big)^2, \tag{1}$$

where $\bar{a}_c$ is the sample mean. We test the null hypothesis of equal population variances between margin types:

$$H_0 \colon \mathrm{Var}(\texttt{clamp}) = \mathrm{Var}(\texttt{subtract}). \tag{2}$$

Experimental factors. We cross two factors in a 2×2 factorial design: (A) margin type (`clamp` vs. `subtract`; Section 4.3), and (B) numerical-stability mode (`detach` vs. `direct`; Eq. (5)). All other pipeline components—architecture, optimizer, learning rate, augmentation distribution, and training duration—are held fixed across cells.

Diagnostic endpoints. To probe the mechanism underlying any observed variance difference, we additionally report: (i) clamp activation rates (CAR), quantifying the frequency of margin-induced saturation per layer; (ii) layerwise gradient ℓ_2 norms; and (iii) a reduced-margin dose-response probe that tests whether lowering the margin schedule attenuates the variance effect.

Symbol	Definition
$\mathcal{D}$	Labeled dataset $\{(x_n, y_n)\}_{n=1}^{N}$ with $y_n \in \{1, \dots, K\}$.
B	Minibatch size (number of images per iteration).
$x_i^{(1)}, x_i^{(2)}$	Two stochastic augmentations of image x_i.
L, ℓ	Number of transformer layers; layer index $\ell \in \{0, \dots, L-1\}$.
$z_{i,\ell}^{(v)} \in \mathbb{R}^d$	ℓ_2-normalized mean-pooled representation of view v at layer ℓ (Eq. (3)).
u, v	Indices over the concatenated view set $\{1, \dots, 2B\}$ (Definition 4.1).
$s_{uv,\ell}$	Cosine similarity: $z_{u,\ell}^{\top} z_{v,\ell} \in [-1, 1]$.
τ	Temperature parameter ($\tau > 0$).
m_ℓ	Nonnegative margin at layer ℓ.
M_{uv}	Positive mask: $M_{uv} = 1$ iff $y(u) = y(v)$ and $u \neq v$.
$\mathcal{P}_u$	Positive set: $\{v : M_{uv} = 1\}$.

Table 1: Core notation. Diagnostic metrics (CAR, gradient norms, VR) are defined in Section 4.4.

4 Methods

4.1 Notation and representations

Table 1 summarizes the main symbols; all definitions are scoped to a single layer ℓ unless otherwise stated.

Representation. Let $h_{i,\ell}^{(v)} \in \mathbb{R}^{T \times d}$ denote the token-level output of layer ℓ for example i, view v, where T is the number of tokens. The contrastive representation is

$$z_{i,\ell}^{(v)} = \frac{\bar{h}_{i,\ell}^{(v)}}{\|\bar{h}_{i,\ell}^{(v)}\|_2}, \qquad \bar{h}_{i,\ell}^{(v)} = \frac{1}{T} \sum_{t=1}^{T} h_{i,\ell,t}^{(v)}, \tag{3}$$

so $z_{i,\ell}^{(v)} \in \mathbb{R}^d$ lies on the unit sphere.

Definition 4.1 (Concatenated view indexing)**.** Given a minibatch of B images with two views each, we form the matrix $Z_\ell = [z_{1,\ell}^{(1)}, \dots, z_{B,\ell}^{(1)}, z_{1,\ell}^{(2)}, \dots, z_{B,\ell}^{(2)}]^{\top} \in \mathbb{R}^{2B \times d}$ and index its rows by $u \in \{1, \dots, 2B\}$, with class label $y(u)$ inherited from the underlying image.

Definition 4.2 (Stop-gradient)**.** For a scalar or tensor x, $\mathrm{sg}(x)$ denotes the identity on the forward pass and zero on the backward pass (i.e., `detach` in autodiff frameworks).

4.2 Layer-local supervised contrastive objective

We write the supervised contrastive loss in closed form to make the margin variants and their gradient behavior explicit.

Base logits. Given modified similarities $\tilde{s}_{uv,\ell}$ (Section 4.3), define

$$b_{uv,\ell} = \frac{\tilde{s}_{uv,\ell}}{\tau}. \tag{4}$$

Numerical-stability shift. Let $\alpha_{u,\ell} = \max_{k \neq u} b_{uk,\ell}$. We shift logits row-wise to prevent overflow:

$$g_{uv,\ell} = \begin{cases} b_{uv,\ell} - \mathrm{sg}(\alpha_{u,\ell}), & (\texttt{detach} \text{ mode}) \\ b_{uv,\ell} - \alpha_{u,\ell}, & (\texttt{direct} \text{ mode}). \end{cases} \tag{5}$$

Both modes yield identical forward values; they differ in whether gradients flow through the max operator (Definition 4.2).

Per-anchor log-probabilities. For anchor u and index $v \neq u$:

$$\log p_{uv,\ell} = g_{uv,\ell} - \log\Big(\sum_{k \neq u} \exp(g_{uk,\ell})\Big). \tag{6}$$

Supervised contrastive loss. Using the mean-over-positives reduction:

$$\mathcal{L}_\ell = -\frac{1}{2B} \sum_{u=1}^{2B} \frac{1}{|\mathcal{P}_u|} \sum_{v \in \mathcal{P}_u} \log p_{uv,\ell}. \tag{7}$$

Forward-Forward locality. Each $\mathcal{L}_\ell$ is optimized independently using only the parameters of layer ℓ; inter-layer gradients are blocked by stop-gradient operations, so layer ℓ updates do not propagate into lower layers.

4.3 Positive-pair margin variants

We compare two strategies that differ in whether they introduce similarity saturation.

4.3.1 Saturating similarity clamping

The clamped variant adds the margin in similarity space and saturates at 1:

$$\tilde{s}_{uv,\ell} = \begin{cases} \min(s_{uv,\ell} + m_\ell,\, 1), & \text{if } M_{uv} = 1, \\ s_{uv,\ell}, & \text{if } M_{uv} = 0. \end{cases} \tag{8}$$

Temperature–margin coupling. Because logits are $\tilde{s}/\tau$ (Eq. (4)), a similarity-space margin m_ℓ corresponds to an effective logit shift of m_ℓ/τ until saturation. The practical strength of clamping therefore depends jointly on m_ℓ and τ.

4.3.2 Gradient-neutral subtraction reference

The subtraction baseline computes log-probabilities from unmodified similarities ($\tilde{s} = s$) and subtracts the margin post hoc:

$$\log \tilde{p}_{uv,\ell} = \log p_{uv,\ell} - m_\ell\, M_{uv}. \tag{9}$$

Proposition 4.1 (Post-log-probability subtraction is gradient-neutral)**.** *Let $m_\ell \geq 0$ be a fixed scalar (not a function of model parameters). Under the mean-over-positives loss* (7)*, replacing $\log p_{uv,\ell}$ with $\log \tilde{p}_{uv,\ell}$ from* (9) *shifts each per-anchor term by the constant m_ℓ and leaves all gradients with respect to model parameters unchanged.*

Proof. For a fixed anchor u:

$$-\frac{1}{|\mathcal{P}_u|} \sum_{v \in \mathcal{P}_u} \log \tilde{p}_{uv,\ell} = -\frac{1}{|\mathcal{P}_u|} \sum_{v \in \mathcal{P}_u} (\log p_{uv,\ell} - m_\ell) = -\frac{1}{|\mathcal{P}_u|} \sum_{v \in \mathcal{P}_u} \log p_{uv,\ell} \;+\; m_\ell.$$

The term m_ℓ is constant with respect to all model parameters (by assumption), so it vanishes under differentiation. □

Interpretation. The subtraction baseline does not reshape gradients; it serves as a gradient-neutral reference. Clamping (Eq. (8)), by contrast, can truncate gradients for any positive pair where saturation occurs.

Table 2: **Hyperparameter configuration.**

Parameter	Value
Architecture	ViT (d=128, H=4, L=8)
Patch size	4×4
Batch size	512
Stage 1 (repr. learning)	600 epochs, AdamW (η=4×10^{-3}, β=(0.9, 0.999), wd=10^{-4})
Stage 2 (linear probe)	50 epochs, AdamW (η=5×10^{-4})
Temperature τ	0.15
Standard margin schedule	m_0=0.4 $\rightarrow$ m_{L-1}=0.1 (linear across layers)
Low margin schedule	m_0=0.2 $\rightarrow$ m_{L-1}=0.1
Augmentation	`RandomCrop`(32, pad=12), `HFlip`, channel normalization

4.4 Diagnostic metrics

Variance ratio (primary endpoint).

$$\mathrm{VR} = \frac{\widehat{\mathrm{Var}}(\texttt{clamp})}{\widehat{\mathrm{Var}}(\texttt{subtract})}, \tag{10}$$

where $\widehat{\mathrm{Var}}$ is the sample variance of test accuracies across seeds (Eq. (1)).

Clamp Activation Rate (CAR). Let $\mathcal{P}_\ell = \{(u, v) : M_{uv} = 1\}$ be the set of positive pairs at layer ℓ. The CAR is

$$\mathrm{CAR}_\ell = \frac{1}{|\mathcal{P}_\ell|} \sum_{(u,v)\in\mathcal{P}_\ell} \mathbb{1}[s_{uv,\ell} + m_\ell > 1]. \tag{11}$$

Gradient norm. For each layer ℓ with parameter set Θ_ℓ:

$$\|\nabla_\ell\| = \sqrt{\sum_{p\in\Theta_\ell} \left\|\nabla_p \mathcal{L}_\ell\right\|_2^2}, \tag{12}$$

i.e., the ℓ_2 norm of the full gradient vector of $\mathcal{L}_\ell$ with respect to all parameters in the ℓ-th encoder block.

5 Experiments

5.1 Experimental setup

Datasets. The primary analysis uses CIFAR-10 with the standard 50k/10k train/test split; we hold out 5k training images for validation, yielding 45k train, 5k validation, and 10k test images. Generalization experiments (Section 7) replicate the variance audit on CIFAR-100 (100 classes), SVHN (10 classes, street-view house numbers), and Fashion-MNIST (10 classes, grayscale). Dataset-specific details appear in Section 7.

Architecture and hyperparameters. Table 2 summarizes the full configuration. We use a Vision Transformer (Dosovitskiy et al., 2021) with embedding dimension $d = 128$, 4 attention heads, and $L = 8$ encoder layers. Input images (32×32) are divided into 4×4 patches, producing $T = 64$ tokens per image.

Data augmentation (CIFAR-10 and CIFAR-100). Each image undergoes `RandomCrop(32, padding=12)` and `RandomHorizontalFlip`, followed by per-channel normalization ($\mu = (0.491, 0.482, 0.447)$, $\sigma = (0.202, 0.199, 0.201)$). No RandAugment, AutoAugment, or color jitter is applied. All conditions share the same augmentation pipeline; augmentation draws are controlled by the random seed.

Data augmentation (SVHN and Fashion-MNIST). SVHN uses the same 12-pixel crop padding as CIFAR-10 with dataset-specific normalization ($\mu = (0.438, 0.444, 0.473)$, $\sigma = (0.198, 0.201, 0.197)$). Fashion-MNIST images (28×28, grayscale) are converted to 3-channel and resized to 32×32, normalized with $\mu = \sigma = (0.5, 0.5, 0.5)$, and augmented with `RandomCrop(32, padding=4)` and `RandomHorizontalFlip`. The reduced crop padding (4 vs. 12 pixels) reflects Fashion-MNIST's lower spatial complexity relative to natural images.

SVHN difficulty sweep augmentation. The difficulty sweep (Section 7.4) varies augmentation intensity while holding the dataset and architecture fixed: (i) *Easy* (standard): `RandomCrop(32, padding=12)` and `HFlip`; (ii) *Medium*: `RandomCrop(32, padding=4)`, `RandomRotation(10)`, and `ColorJitter(0.2, 0.2, 0.2)`; (iii) *Hard*: `RandomCrop(32, padding=6)`, `HFlip(0.5)`, `RandomRotation(15)`, `RandomErasing(p=0.5, scale=(0.1, 0.3))`, and `ColorJitter(0.4, 0.4, 0.4, 0.1)`.

Training protocol. CFF training proceeds in two stages:

1. *Representation learning.* Each of the L=8 layers is trained against its own supervised contrastive loss (Eq. (7)) for 600 epochs.
2. *Linear probe.* A linear classifier is trained for 50 epochs on frozen representations. The reported test accuracy is the checkpoint achieving the highest validation accuracy during this stage.

Random seed control. Each seed σ jointly determines weight initialization, data-loader shuffling order, and stochastic augmentation draws. All hyperparameters are held fixed across seeds and conditions.

Margin schedule. Margins decrease linearly across layers:

$$m_\ell = m_0 + (m_{L-1} - m_0)\,\frac{\ell}{L-1}, \quad \ell \in \{0, \ldots, L{-}1\}. \tag{13}$$

Factorial design. Under the standard margin schedule, we cross two factors: (A) margin type (`clamp` vs. `subtract`) and (B) stability mode (`detach` vs. `direct`), with n=7 independent seeds per cell (28 runs total).

Dose-response probe. To test whether reducing the margin attenuates the variance effect, we run a low-margin schedule (m_0=0.2 $\rightarrow$ m_{L-1}=0.1) with 7 seeds each for `clamp_detach` and `clamp_direct` (14 clamped runs). The subtraction reference reuses the 7 `subtract_detach` seeds from the standard-margin factorial; this is valid because the subtraction baseline is gradient-neutral (Proposition 4.1), so the margin schedule value does not affect the subtract condition's gradients or trained model. The resulting pooled comparison is unbalanced (n=14 clamp vs. n=7 subtract); we conserve compute based on the factorial's finding that stability mode does not affect variance ($p > 0.82$; Table 4).

Compute. All runs were conducted on NVIDIA V100-SXM2-32GB GPUs. Each CIFAR-10 run requires approximately 7 GPU-hours. The full experimental campaign (CIFAR-10 factorial, dose-response, and generalization datasets) totals approximately 88 runs and $\sim$550 GPU-hours.

5.2 Results: CIFAR-10

5.2.1 Standard-margin factorial results

Table 3 reports test accuracy statistics for each cell of the 2×2 factorial under the standard margin schedule. Per-seed accuracies are listed in Appendix A.

Factorial variance tests. To assess whether variability depends on margin type, stability mode, or their interaction, we apply Levene and Brown–Forsythe procedures (two-way ANOVA on absolute deviations from cell means and medians, respectively). Table 4 reports the results. Neither test finds evidence that stability

Table 3: **CIFAR-10 factorial results (standard margin $0.4 \to 0.1$).** Test accuracy across n=7 seeds per cell.

Margin type	Stability	Mean (%)	Std (%)	Var
`clamp`	`detach`	78.52	0.927	0.8590
`clamp`	`direct`	78.44	1.158	1.3408
`subtract`	`detach`	78.73	0.467	0.2178
`subtract`	`direct`	78.30	0.224	0.0500

Table 4: **Factorial variance sensitivity tests (standard margin).**

Test	Margin type	Stability mode	Interaction
Levene	0.061	0.824	0.058
Brown–Forsythe	0.058	0.856	0.137

mode affects variance (both $p > 0.82$). The margin-type effect is marginal ($p \approx 0.06$) but directionally consistent with the pooled comparison below. The interaction is not supported under Brown–Forsythe ($p = 0.137$); the marginal Levene interaction ($p = 0.058$) appears driven by the higher variance of the `clamp_direct` cell.

Pooled primary comparison. Because stability mode does not affect variance, we pool across stability modes for the primary analysis, yielding n=14 seeds per margin type (Table 5).

Mean comparison. A Welch two-sample t-test finds no evidence of a mean difference between margin types ($t(19.4) = 0.10$, $p = 0.92$; 95% CI for $\mu_{\texttt{clamp}} - \mu_{\texttt{subtract}}$: $[-0.64, 0.59]$). The variance difference therefore cannot be attributed to a shift in central tendency.

Variance comparison. An F-test for equality of variances yields $F(13, 13) = 5.90$, $p = 0.003$, rejecting H_0 (Eq. (2)). We additionally report:

- Levene's test: $p = 0.058$ (marginal, consistent in direction).
- Bootstrap 95% CI for VR: $[1.62, 15.80]$ (10,000 percentile-bootstrap resamples of the ratio of sample variances).

Shapiro–Wilk tests do not reject normality for either pooled group ($W = 0.951$, $p = 0.58$ for clamp; $W = 0.952$, $p = 0.59$ for subtract; Table 23 in Appendix D), supporting the F-test assumption.

Figure 1 displays the per-seed accuracy distributions.

5.2.2 Dose-response probe

If the variance effect is mediated by saturation, reducing the starting margin should attenuate it. Table 6 compares the two margin schedules.

Halving the starting margin from $m_0 = 0.4$ to $m_0 = 0.2$ reduced the variance ratio from 5.90× to 2.98×, consistent with a dose-response relationship. The low-margin ratio does not reach significance individually ($F(13, 6) = 2.98$, $p = 0.19$; bootstrap 95% CI for VR: $[0.79, 30.77]$), likely reflecting reduced power from the smaller effect size and unbalanced design (n=14 vs. n=7). We therefore interpret this probe as directional evidence supporting the saturation-mediated mechanism, not as a standalone significance claim.

A Welch t-test on the low-margin pooled means yields $t(18.4) = -1.59$, $p = 0.13$ (95% CI for $\mu_{\texttt{clamp}} - \mu_{\texttt{subtract}}$: $[-1.03, +0.14]$ pp), indicating no significant mean difference, though the trend is less cleanly null than under the standard margin ($p = 0.92$).

Table 5: **Pooled test accuracy by margin type (standard margin, n=14 per group).**

Condition	n	Mean (%)	Std (%)	Var
Clamp (pooled)	14	78.48	1.008	1.0170
Subtract (pooled)	14	78.51	0.415	0.1724

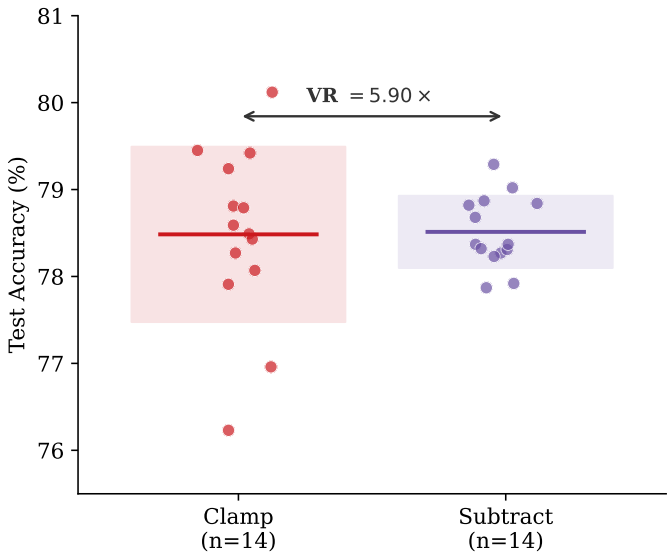

Figure 1: **Per-seed test accuracy by margin type (standard margin, pooled).** Each dot is one seed (n=14 per group). Horizontal lines show group means; shaded bands show ±1 standard deviation.

This probe has two limitations. First, because τ is fixed, lowering m_0 simultaneously reduces the effective logit shift m_0/τ, confounding changes in saturation frequency (Eq. (11)) with overall logit-scale changes. Second, the bootstrap CI for the low-margin VR includes 1.0, so the data are also consistent with no variance difference under this reduced schedule.

6 Mechanism Analysis

This section presents diagnostic evidence bearing on why the clamped variant exhibits higher seed-to-seed variance on CIFAR-10. We emphasize that the diagnostics establish necessary ingredients for the proposed mechanism but do not constitute a causal proof.

6.1 Gradient truncation under saturation

The derivative of $f(x) = \min(x, 1)$ is $f'(x) = 1$ for $x < 1$ and $f'(x) = 0$ for $x > 1$. By Eq. (8), for any positive pair (u, v) at layer ℓ with $s_{uv,\ell} + m_\ell > 1$, the modified similarity $\tilde{s}_{uv,\ell} = 1$ is constant as a function of $s_{uv,\ell}$, so the *direct* (numerator) gradient contribution of that pair to anchor u's loss term is zero.

Denominator pathway. Gradient truncation along the direct path does not imply that the full derivative $\partial \mathcal{L}_\ell / \partial s_{uv,\ell}$ is zero. The clamped similarity $\tilde{s}_{uv,\ell} = 1$ still enters the softmax denominators of *other* anchors $u' \neq u$ via Eq. (6). Gradients can therefore flow through this indirect (denominator) pathway. The net effect of saturation is a partial, not complete, gradient truncation—one that is data-dependent, varying with which pairs happen to exceed the saturation threshold on each minibatch.

Table 6: **Dose-response probe.** Reducing the margin schedule from $0.4 \to 0.1$ to $0.2 \to 0.1$ reduces the observed variance ratio. The low-margin comparison is unbalanced (n=14 clamp vs. n=7 subtract; see Section 5.1).

Schedule	**Clamp var**	**Subtract var**	**VR**	**F-test p**
$0.4 \to 0.1$ (standard)	1.0170	0.1724	5.90×	0.003
$0.2 \to 0.1$ (low)	0.6498	0.2178	2.98×	0.19

Table 7: **Clamp Activation Rate (CAR) by layer (CIFAR-10).**

Layer	**Margin (std / low)**	**CAR (std)**	**CAR (low)**	**Reduction**
0	0.40 / 0.20	60.7%	40.3%	33.7%
1	0.36 / 0.19	58.3%	43.7%	24.9%
2	0.31 / 0.17	54.0%	42.5%	21.3%
3	0.27 / 0.16	49.6%	40.5%	18.3%
4	0.23 / 0.14	45.8%	38.6%	15.6%
5	0.19 / 0.13	41.6%	36.6%	12.0%
6	0.14 / 0.11	37.5%	34.5%	7.8%
7	0.10 / 0.10	32.3%	32.3%	0.2%

6.2 Clamp Activation Rate by layer

Table 7 reports CAR (Eq. (11)) across all 8 layers for both margin schedules, computed at the final Stage-1 epoch (epoch 600) by averaging over minibatches and seeds.

Higher starting margins produce substantially higher CAR in early layers, with the gap narrowing monotonically and vanishing at layer 7 where both schedules share $m = 0.1$. Under the low-margin schedule, CAR increases slightly from layer 0 (40.3%) to layer 1 (43.7%) despite a lower margin at layer 1; this likely reflects higher within-class cosine similarity at layer 1 (i.e., more clustered representations), though we do not analyze this further.

6.3 Gradient norms by layer

Table 8 reports gradient ℓ_2 norms (Eq. (12)) at the final training epoch for layers 0–3. We restrict to these layers because CAR drops below 50% by layer 3 under the standard schedule, and layers 4–7 show negligible differences between conditions ($< 5\%$ relative difference in gradient norms; not tabulated).

At layer 0, where CAR is highest (60.7%), clamping reduces gradient norms by a factor of 4.0× relative to the subtraction baseline. The reduction narrows to 2.1× at layer 1 and is absent by layer 2. This spatial correspondence between CAR and gradient norm reduction is consistent with saturation-induced truncation being the dominant cause of the gradient difference at early layers.

Figure 2 visualizes both profiles jointly.

6.4 Mechanistic interpretation

The diagnostic evidence is consistent with the following qualitative mechanism. Under the standard schedule, over 60% of positive pairs at layer 0 exceed the clamp threshold (Table 7); reducing the starting margin halves this rate (Table 7, Reduction column). This frequent saturation truncates gradients at early layers: gradient norms at layer 0 are 4.0× lower under clamping than under the subtraction baseline (Table 8), and the reduction tracks the CAR profile spatially (Figure 2).

The truncation is not uniform: which positive pairs saturate depends on the realized cosine similarities, which vary with initialization and data ordering. Different seeds therefore experience different patterns of gradient

Table 8: **Mean gradient norms by layer (final epoch, CIFAR-10).** Values are mean ± std across seeds.

Layer	Clamp	Subtract	Ratio (subtract / clamp)
L0	0.400 ± 0.028	1.614 ± 0.045	4.0×
L1	0.115 ± 0.008	0.243 ± 0.008	2.1×
L2	0.048 ± 0.002	0.049 ± 0.008	1.0×
L3	0.023 ± 0.003	0.018 ± 0.007	0.8×

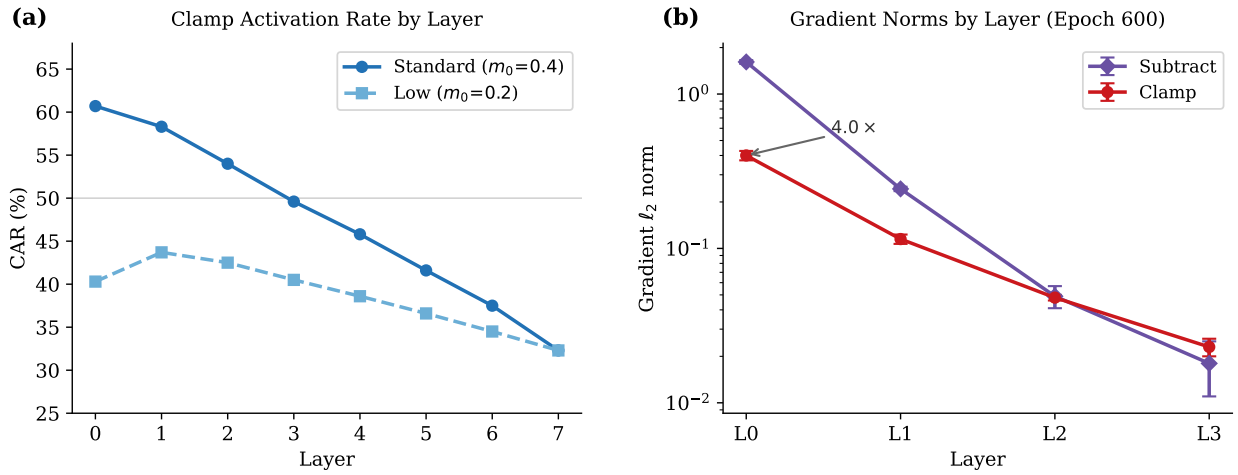

Figure 2: **Diagnostic profiles by layer (CIFAR-10).** (a) CAR under standard and low-margin schedules. (b) Gradient ℓ_2 norms at the final epoch. The gradient norm gap between conditions tracks the CAR profile, closing by layer 2.

zeroing across minibatches, causing optimization trajectories to diverge more than under the non-truncating baseline. The 5.90× variance ratio (Table 5) is the aggregate outcome; halving the margin reduces VR to 2.98× (Table 6), consistent with less saturation producing less divergence.

The first two links in this chain—frequent saturation and reduced gradient norms—are directly verified by our measurements. The subsequent links, from stochastic truncation to trajectory divergence and from divergence to accuracy variance, are inferred from the joint pattern of evidence rather than demonstrated causally. In particular, we do not track CAR or gradient norms over training (only at epoch 600), nor do we measure per-seed variation in CAR, which would more directly test whether seeds with higher truncation rates produce more extreme final accuracies.

7 Cross-Dataset Generalization

The CIFAR-10 results establish that clamping can inflate variance when CAR is high. To characterize the conditions under which this effect manifests, we replicate the variance audit on three additional datasets that differ in number of classes, positive-pair density, task difficulty, and visual domain. Table 9 summarizes the key results; we discuss each dataset below.

7.1 CIFAR-100: Low positive-pair density

With 100 classes and batch size 512, each minibatch contains approximately 2,580 positive pairs—roughly 10× fewer than CIFAR-10's ∼25,700. We ran a 2×2 factorial (n=5 seeds per cell; architecture and training identical to CIFAR-10 except K=100).

Table 9: **Cross-dataset summary.** Variance ratio (VR), layer-0 CAR, approximate positive pairs per minibatch (batch size 512), and mean clamp accuracy. CIFAR-10 is the only dataset exhibiting VR > 1.

Dataset	K	**Pos. pairs**	**L0 CAR**	**Acc. (%)**	**VR**
CIFAR-10	10	∼25,700	60.7%	78.5	5.90×
CIFAR-100	100	∼2,580	29.0%	51.8	0.39×
SVHN	10	∼30,500	51.0%	96.7	0.25×
Fash-MNIST	10	∼25,700	—	92.6	0.08×

Table 10: **CIFAR-100 pooled results (n=10 per group).**

Condition	**n**	**Mean (%)**	**Std (%)**	**Var**
Clamp (pooled)	10	51.82	0.505	0.255
Subtract (pooled)	10	51.38	0.814	0.663

The variance ratio is 0.39×—inverted relative to CIFAR-10—and non-significant ($F(9, 9) = 0.39$, $p = 0.17$). CAR measurements on CIFAR-100 reveal the mechanism: L0 CAR is only 29.0% (Table 11), compared to 60.7% on CIFAR-10, and average CAR across layers is 19.7% versus 47.5%. With fewer positive pairs per batch, fewer similarity values approach the saturation boundary, so the gradient truncation pathway that drives variance inflation on CIFAR-10 is largely inactive.

Layerwise gradient norms confirm that low CAR translates to an absence of gradient truncation (Table 12). At layer 0, the clamp-to-subtract norm ratio is 0.58×—clamping produces *smaller* gradients, not larger ones. This is the opposite of CIFAR-10, where the same ratio is 0.25× (clamp norms are 4.0× lower, meaning subtract norms dominate). On CIFAR-100, the two conditions produce similar gradient landscapes, consistent with the absence of a variance effect.

7.2 SVHN: High accuracy, high CAR

SVHN is a 10-class digit recognition dataset with non-uniform class frequencies, yielding ∼30,500 positive pairs per minibatch. We ran n=5 seeds each for `clamp_direct` and `subtract_direct`.

The variance ratio is 0.25× (inverted; $F(4, 4) = 0.25$, $p = 0.21$). SVHN's CAR profile is qualitatively different from CIFAR-10's: CAR *increases* with depth (L0: 51.0%, L7: 78.3%), whereas CIFAR-10's decreases monotonically (L0: 60.7%, L7: 32.3%). Average CAR is 70.6%—substantially higher than CIFAR-10's 47.5%. Despite this, the variance ratio is inverted.

This dissociation between CAR level and variance inflation shows that high CAR is *not sufficient* for the variance effect. We hypothesize that SVHN's high accuracy (∼97%) compresses the space for seed-to-seed divergence: when the task is nearly solved, all seeds converge to similar optima regardless of gradient truncation.

7.3 Fashion-MNIST: High accuracy, moderate difficulty

Fashion-MNIST is a 10-class grayscale dataset. We ran n=5 seeds each for clamp and subtract conditions.

The variance ratio is 0.08×—the strongest inversion across all datasets—and is significant ($F(4, 4) = 0.08$, $p = 0.029$). Fashion-MNIST's high accuracy (∼92%) places it in the same regime as SVHN: task difficulty is low enough that seeds converge to similar optima despite gradient truncation. CAR was not measured for Fashion-MNIST; based on the SVHN pattern (high accuracy co-occurring with high CAR due to well-clustered representations), we would expect moderate-to-high CAR that nonetheless does not produce variance inflation.

Table 11: **CAR comparison: CIFAR-10 vs. CIFAR-100 (standard margin).**

Layer	CAR (CIFAR-10)	CAR (CIFAR-100)	Difference
0	60.7%	29.0%	−31.7 pp
1	58.3%	27.8%	−30.5 pp
2	54.0%	24.2%	−29.8 pp
3	49.6%	21.4%	−28.2 pp
4	45.8%	18.4%	−27.4 pp
5	41.6%	15.5%	−26.1 pp
6	37.5%	12.3%	−25.2 pp
7	32.3%	8.7%	−23.6 pp
Mean	**47.5%**	**19.7%**	−27.8 pp

Table 12: **Mean gradient norms by layer (CIFAR-100, single seed).** Contrast with CIFAR-10 (Table 8), where clamping reduces L0 norms by 4.0×.

Layer	Margin	Clamp	Subtract	Ratio (C/S)
L0	0.400	1.265 ± 0.099	2.170 ± 0.279	0.58×
L1	0.357	0.262 ± 0.019	0.358 ± 0.032	0.73×
L2	0.314	0.080 ± 0.004	0.073 ± 0.004	1.09×
L3	0.271	0.046 ± 0.002	0.029 ± 0.002	1.60×

7.4 SVHN difficulty sweep

To directly test whether task difficulty moderates the clamping-variance relationship, we repeated the SVHN experiment with progressively more aggressive augmentation that reduces accuracy. This holds the dataset, architecture, and number of classes constant while varying only the optimization difficulty.

The sweep reveals a dramatic transition: on the same dataset with the same architecture, increasing augmentation difficulty transforms the variance ratio from 0.25× to 16.73×. Under hard augmentation, the clamp condition exhibits bimodal behavior, with some seeds converging (∼35% accuracy) while others fail (∼18%), producing extreme variance. This provides the strongest evidence that task difficulty—not CAR alone—determines whether clamping destabilizes training.

7.5 Synthesis: When does clamping affect variance?

The cross-dataset evidence suggests that clamping inflates variance when two conditions co-occur:

1. **High L0 clamp activation rate.** CIFAR-100's low L0 CAR (29.0%) prevents the truncation pathway from activating. High CAR is necessary but not sufficient.

2. **Intermediate task difficulty.** SVHN and Fashion-MNIST achieve $> 92\%$ accuracy, leaving little room for trajectories to separate. At the other extreme, very low accuracy under hard augmentation produces extreme variance (Table 15).

CIFAR-10 uniquely satisfies both conditions: high L0 CAR (60.7%) and intermediate accuracy (∼78%). The SVHN difficulty sweep confirms that these conditions interact on a single dataset: the same architecture produces VR = 0.25× at 97% accuracy and VR = 16.73× at 25% accuracy.

Table 13: **SVHN results (n=5 per group).**

Condition	n	Mean (%)	Std (%)	Var
Clamp	5	96.70	0.430	0.185
Subtract	5	95.29	0.853	0.727

Table 14: **Fashion-MNIST results (n=5 per group).**

Condition	n	Mean (%)	Std (%)	Var
Clamp	5	92.57	0.113	0.0127
Subtract	5	91.87	0.406	0.1645

8 Discussion

Variance without mean cost. On CIFAR-10, the clamped variant produces 5.90× higher test-accuracy variance with no detectable difference in mean accuracy ($p = 0.92$; Table 5). This implies that switching to the subtraction reference reduces noise at no observed cost to the central estimate in this setting.

Practical consequences for seed budgets. The standard error of the sample mean scales as $\text{Std}/\sqrt{n}$. To illustrate: under clamping on CIFAR-10 (Std $\approx$ 1.0 pp), achieving a standard error of $\pm$0.3 pp requires $n \approx 11$ seeds; under subtraction (Std $\approx$ 0.42 pp), $n \approx 2$ suffices. This difference can determine whether a hyperparameter comparison is adequately powered under finite compute.

Why layer-local training may amplify the effect. In end-to-end backpropagation, gradient truncation at one layer is modulated by the full chain of downstream gradients, and compensating signals from other layers can partially offset local artifacts. In CFF, each layer optimizes independently with no inter-layer gradient flow, so saturation-induced truncation at layer 0 has no compensating mechanism from deeper layers. This architectural difference may make CFF more sensitive to margin-induced gradient truncation than end-to-end training.

Dataset dependence and the role of positive-pair density. The inverted variance ratio on CIFAR-100 (VR = 0.39×) is explained by its 10× lower positive-pair density (100 classes vs. 10), which produces correspondingly lower CAR at every layer (Table 11). This confirms that the gradient truncation mechanism depends on a sufficient density of positive pairs to create widespread saturation. Layerwise gradient norms corroborate this picture: on CIFAR-100, the clamp-to-subtract gradient norm ratio at layer 0 is 0.58× (1.27 vs. 2.17), indicating that clamping does *not* amplify gradient magnitudes when CAR is low—in contrast to CIFAR-10, where the analogous ratio is 4.0× higher under clamping (Table 8). Practitioners working with many-class problems are unlikely to encounter the variance inflation documented on CIFAR-10.

Task difficulty as a moderating factor. SVHN and Fashion-MNIST demonstrate that high CAR is not sufficient for variance inflation: both achieve $> 92\%$ accuracy, limiting seed-to-seed spread regardless of truncation levels. The SVHN difficulty sweep (Table 15) provides the most direct evidence for this interaction, showing on a single dataset that reducing accuracy from 97% to 25% transforms the variance ratio from 0.25× to 16.73×. The clamping-variance effect thus appears most pronounced in an intermediate-difficulty regime where optimization trajectories are sensitive to gradient perturbations but the task is not so easy that all seeds converge regardless.

Connection to margin saturation in metric learning. Hard margins in face recognition objectives (SphereFace, CosFace, ArcFace) produce saturated logit regions by design. Our finding suggests that variance audits—not just mean-accuracy comparisons—may be informative when evaluating margin implementations in those settings, though end-to-end gradient flow and different loss reductions limit the directness of the analogy.

Table 15: **SVHN difficulty sweep.** Increasing augmentation difficulty on the same dataset drives VR from 0.25× (easy, ∼97% accuracy) to 16.73× (hard, ∼25% accuracy).

Difficulty	n	Acc. (%)	Clamp var	Subtract var	VR	Pattern
Easy (standard)	5	96.7	0.185	0.727	0.25×	Inverted
Medium	5	26.8	37.57	17.23	2.18×	Elevated
Hard	5	25.0	85.45	5.11	16.73×	Extreme

Practical guidelines. Based on the cross-dataset evidence: (i) In settings resembling CIFAR-10 (moderate accuracy, high positive-pair density), switching to the subtraction reference provides a substantial variance reduction at no observed mean cost. (ii) In high-accuracy settings ($> 90\%$) or many-class settings (low positive-pair density), clamping does not appear to inflate variance and may not require modification. (iii) When in doubt, measuring L0 CAR provides a lightweight diagnostic: values substantially below 50% suggest the truncation pathway is largely inactive.

9 Limitations and Threats to Validity

Single architecture and training protocol. All results use a fixed ViT configuration (d=128, H=4, L=8), training schedule (600 Stage-1 epochs), and augmentation pipeline. We do not characterize how the variance effect depends on model scale, depth, or alternative training schedules.

Cross-dataset design imbalances. The CIFAR-10 primary analysis uses the full 2×2 factorial (n=7 per cell, n=14 pooled). CIFAR-100 uses n=5 per cell (n=10 pooled). SVHN and Fashion-MNIST use n=5 per condition without the stability-mode factor. These design differences reflect compute allocation prioritizing the primary dataset; the generalization experiments have correspondingly lower statistical power.

SVHN design limitations. SVHN experiments used only the `direct` stability mode due to a bug in the `detach` condition discovered post hoc. The CIFAR-10 factorial establishes that stability mode does not affect variance ($p > 0.82$), so this limitation is unlikely to affect conclusions, but we note the asymmetry.

Missing Fashion-MNIST CAR. CAR was not measured for Fashion-MNIST due to checkpoint availability. The significant variance inversion (VR $= 0.08\times$, $p = 0.029$) at 92.6% accuracy is consistent with the task-difficulty moderation pattern observed on SVHN, but without CAR data we cannot confirm whether the truncation pathway is active or inactive on this dataset.

Temperature–margin confound. Clamping (Eq. (8)) is applied before temperature scaling (Eq. (4)), so its effective logit-space strength is m_ℓ/τ until saturation. In the dose-response probe, we vary m_0 while holding $\tau = 0.15$ fixed, confounding changes in saturation frequency with overall logit-scale changes.

SVHN difficulty sweep confounds. The difficulty sweep varies augmentation intensity, which simultaneously changes task difficulty, the distribution of pairwise similarities, and the effective positive-pair informativeness. We cannot disentangle which of these factors drives the VR transition.

Augmentation sensitivity. Our augmentation pipeline uses 12-pixel padding for random crops on 32×32 images, which is more aggressive than the standard 4-pixel padding. Because augmentation intensity affects the distribution of pairwise cosine similarities—and hence the frequency of clamp saturation—the magnitude of the variance effect may differ under milder augmentation. The within-experiment comparison remains valid (all conditions share identical augmentation), but the absolute CAR values and variance ratios should not be assumed to transfer to other pipelines.

Two-stage pipeline. The reported test accuracy reflects both representation learning (Stage 1) and a linear probe on frozen representations (Stage 2). Our mechanistic analysis focuses entirely on Stage 1 (CAR,

gradient norms), but probe-stage variance could also contribute to the observed seed-to-seed differences. We do not isolate the two stages' contributions to total variance.

Diagnostic timing. CAR and gradient norms are reported at the final Stage-1 epoch only. We do not track how these quantities evolve over training.

Per-seed mechanistic link. CAR and gradient norms are reported as cross-seed averages. We do not measure per-seed variation in these diagnostics, which would more directly test whether seeds experiencing more frequent truncation produce more extreme final accuracies.

Statistical power. The standard-margin factorial uses $n = 7$ seeds per cell ($n = 14$ pooled per margin type). While the F-test rejects variance equality at $p = 0.003$, the bootstrap CI for VR is wide ($[1.62, 15.80]$), so the *magnitude* of the effect is less precisely constrained than its *direction*. The generalization datasets use even smaller samples ($n = 5$), further limiting power for individual dataset-level significance tests.

Reduction scope. Proposition 4.1 establishes gradient-neutrality under the mean-over-positives reduction used throughout this paper. Under alternative reductions (e.g., sum over positives), the gradient-neutrality result would not hold.

10 Conclusion

We presented a formal specification of the supervised contrastive loss used in Contrastive Forward-Forward training and proved that a common margin variant—post-log-probability subtraction—is gradient-neutral under the mean-over-positives reduction (Proposition 4.1), establishing it as a true no-margin reference. Using this reference, we isolated the effect of the alternative implementation: saturating similarity clamping via $\min(s + m, 1)$.

On CIFAR-10 (2×2 factorial, n=7 seeds per cell), clamping is associated with nearly six-fold higher pooled test-accuracy variance (F-test $p = 0.003$) at statistically indistinguishable mean accuracy. Three diagnostic analyses support a saturation-mediated account: clamp activation rates exceed 60% at early layers, gradient norms at layer 0 are 4.0× lower under clamping, and halving the starting margin reduces the variance ratio from 5.90× to 2.98×.

Critically, this effect does not generalize uniformly. Replication on CIFAR-100, SVHN, and Fashion-MNIST reveals inverted variance ratios in all three cases. Cross-dataset analysis identifies two moderating factors: positive-pair density (which determines CAR) and task difficulty (which limits trajectory separation at high accuracy). An SVHN difficulty sweep confirms this interaction on a single dataset, with the variance ratio transitioning from 0.25× at 97% accuracy to 16.73× at 25% accuracy.

The practical implication is conditional: in the regime exemplified by CIFAR-10—moderate accuracy, many same-class pairs per batch—switching to the gradient-neutral subtraction reference substantially reduces training variance at no observed cost. In high-accuracy or many-class settings, the effect is absent or inverted, and clamping may not require modification. We recommend that CFF practitioners measure L0 CAR as a lightweight diagnostic and report the chosen margin implementation explicitly to support reproducibility.

Broader Impact Statement

This work audits seed-to-seed variance arising from a margin implementation choice in a contrastive loss. It introduces no new model capabilities, datasets, or deployment-facing systems, and we identify no specific pathway to negative societal impact.

References

Hossein Aghagolzadeh and Mehdi Ezoji. Contrastive forward-forward: A training algorithm of vision transformer. *Neural Networks*, 192:107867, 2025. doi: 10.1016/j.neunet.2025.107867.

Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Greedy layerwise learning can scale to ImageNet. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 583–593, 2019.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10231–10241, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Xing Chen, Dongshu Liu, Jérémie Laydevant, and Julie Grollier. Self-contrastive forward-forward algorithm. *Nature Communications*, 16:5978, 2025. doi: 10.1038/s41467-025-61037-0.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 3821–3830, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6738–6746, 2017.

Sindy Löwe, Peter O'Connor, and Bastiaan Veeling. Putting an end to end-to-end: Gradient-isolated learning of representations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Arild Nøkland and Lars Hiller Eidnes. Training neural networks with local error signals. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 4839–4850, 2019.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 1310–1318, 2013.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision, 2021.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5265–5274, 2018.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*, 2020.

A Per-Seed Results

Tables 16 and 17 report individual test accuracies for all CIFAR-10 runs. Seeds are numbered S1–S7 consistently within each condition; the same seed index across conditions corresponds to the same random seed value. The `subtract_detach` rows are identical in both tables because the subtraction baseline is gradient-neutral (Proposition 4.1): the margin schedule value does not affect the trained model.

Table 16: **Per-seed CIFAR-10 results (standard margin $0.4 \to 0.1$).**

Condition	S1	S2	S3	S4	S5	S6	S7
`clamp_detach`	78.49	78.27	78.43	76.96	78.81	78.59	80.12
`clamp_direct`	78.07	77.91	79.42	76.23	79.24	78.79	79.45
`subtract_detach`	78.82	79.29	78.87	79.02	77.87	78.37	78.84
`subtract_direct`	78.27	78.31	78.68	78.23	78.37	78.32	77.92

Table 17: **Per-seed CIFAR-10 results (low margin $0.2 \to 0.1$).** The `subtract_detach` row is identical to Table 16; see text.

Condition	S1	S2	S3	S4	S5	S6	S7
`clamp_detach`	78.97	78.49	79.04	78.15	78.46	78.49	77.99
`clamp_direct`	80.16	77.36	78.33	77.74	76.88	77.55	78.35
`subtract_detach`	78.82	79.29	78.87	79.02	77.87	78.37	78.84

B SVHN CAR Profile

Table 18: **CAR comparison: CIFAR-10 vs. SVHN (standard margin).** SVHN's CAR *increases* with depth, opposite to CIFAR-10.

Layer	CAR (CIFAR-10)	CAR (SVHN)	Difference
0	60.7%	51.0%	−9.7 pp
1	58.3%	65.3%	+7.0 pp
2	54.0%	75.0%	+21.0 pp
3	49.6%	78.3%	+28.7 pp
Mean (L0–L7)	**47.5%**	**70.6%**	+23.1 pp

C Per-Seed Results: Generalization Datasets

Tables 19–22 report individual test accuracies for all cross-dataset experiments. Seeds are numbered S1–S5 consistently within each condition.

Table 19: **Per-seed CIFAR-100 results (standard margin $0.4 \to 0.1$, n=5 per cell).**

Condition	S1	S2	S3	S4	S5
`clamp_detach`	51.98	52.06	52.32	51.62	52.36
`clamp_direct`	50.84	51.60	51.62	52.43	51.36
`subtract_detach`	50.85	52.25	52.31	50.17	50.90
`subtract_direct`	50.78	52.52	51.98	50.81	51.23

Table 20: **Per-seed SVHN results (n=5 per group, `direct` mode only).**

Condition	S1	S2	S3	S4	S5
`clamp`	97.16	96.42	97.11	96.17	96.66
`subtract`	95.09	96.11	95.36	93.95	95.92

Table 21: **Per-seed Fashion-MNIST results (n=5 per group).**

Condition	S1	S2	S3	S4	S5
`clamp`	92.39	92.69	92.63	92.55	92.57
`subtract`	92.00	91.32	92.06	92.36	91.62

Table 22: **Per-seed SVHN difficulty sweep results (n=5 per condition).** Augmentation details in Section 5.1.

Difficulty	Condition	S1	S2	S3	S4	S5
Medium	`clamp`	28.71	21.05	36.37	23.89	23.92
	`subtract`	22.01	30.19	28.50	21.22	28.02
Hard	`clamp`	18.11	35.75	19.02	34.41	17.62
	`subtract`	23.67	25.93	21.98	21.65	19.97

D Normality Tests

Table 23 reports Shapiro–Wilk tests for the pooled groups used in the primary variance comparisons.

Table 23: **Shapiro–Wilk normality tests for pooled groups.**

Group	n	W	p-value
Clamp (standard margin)	14	0.9513	0.5816
Subtract (standard margin)	14	0.9516	0.5853
Clamp (low margin)	14	0.9608	0.7355
Subtract (low margin)	7	0.9074	0.3780

E Reproducibility Details

Software. All experiments use PyTorch 2.6.0 with CUDA 12.4 (system driver CUDA 12.8) on NVIDIA V100-SXM2-32GB GPUs. The CFF implementation is based on the codebase of Aghagolzadeh & Ezoji (2025) with modifications to support the subtraction margin variant and diagnostic logging.

Seed values. CIFAR-10 experiments use seven random seeds: 1, 2, 3, 4, 5, 6, 7. Cross-dataset experiments (CIFAR-100, SVHN, Fashion-MNIST) use seeds 1–5. Each seed is passed to `torch.manual_seed`, `numpy.random.seed`, and `random.seed`, and `torch.cuda.manual_seed_all` for full determinism.

Bootstrap procedure. All bootstrap confidence intervals use 10,000 percentile resamples. For the variance ratio, each resample draws n values with replacement from each group independently, computes the sample variance of each resampled group, and takes their ratio. The reported 95% CI is the 2.5th and 97.5th percentiles of the resulting distribution.

Diagnostic computation. CAR values (Eq. (11)) are computed at the final Stage-1 epoch (epoch 600) by evaluating Eq. (11) on each minibatch of the training set and averaging across minibatches, then averaging across seeds. Gradient norms (Eq. (12)) are computed on a single minibatch at the final epoch and averaged across seeds.