NEURAL INFORMATION PROCESSING SYSTEMS

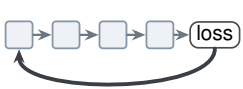
Cumulative-Goodness Free-Riding in Forward-Forward Networks: Real, Repairable, but Not Accuracy-Dominant

Amirhossein Yousefiramandi

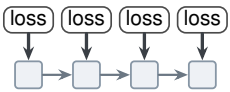
Independent Researcher · amir.usefi75@gmail.com

NeurIPS 2026 · Paris

Forward-Forward: every block learns from its own loss

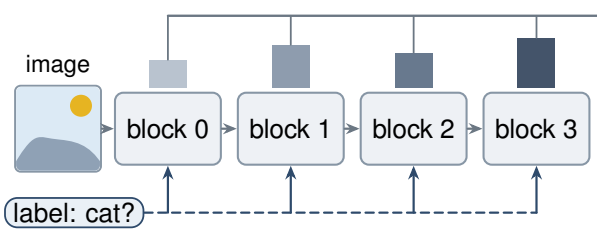


Backpropagation: errors flow back through all layers



Forward-Forward (FF): each block learns from its own loss; no gradient between blocks

goodness: each block's score for (image, label), high for the true label, low for wrong ones



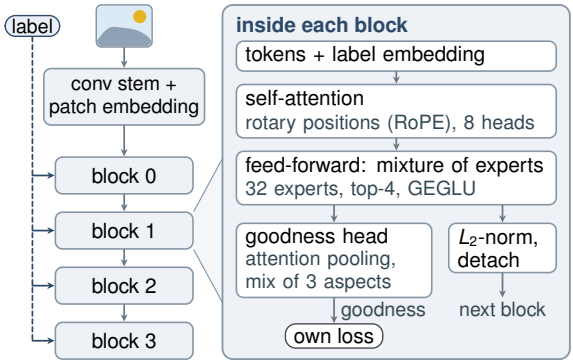
prediction: label with the largest summed goodness



Margin: a block's goodness for the true label minus its goodness for a wrong label.

Forward-Forward (FF): Hinton (2022), arXiv:2212.13345 · bar heights are illustrative schematic of the FF variant studied here: the label enters every block (unlike Hinton 2022; paper Sec. 3.1)

Our testbed: a co-designed FF backbone



Our CIFAR-100 (Results 2–4) and Tiny ImageNet models: dense feed-forward, 9.49 M on CIFAR-100. Tiny ImageNet: labeled validation split.

91.32%

CIFAR-10 test accuracy
with test-time flip · 3 seeds
width 256: 107.2 M parameters,
18.9 M active per token

Single crop, for calibration only

	CIFAR-10	CIFAR-100	Tiny ImageNet
Ours	90.50%	66.37%	48.54%
Prior FF*	90.62% ASGE	65.42% ASGE	35.67%† SCFF
Concurrent‡	91.68% HCL-FF	70.80% BiCovG	50.30% BiCovG

*the strongest prior FF result we list
‡the strongest concurrent (2026) one in B7, as reported
†SCFF (2025): unsupervised, not like-for-like

after paper Fig. 20 · RoPE: Su et al. (2024) · mixture of experts: Shazeer et al. (2017) · GEGLU: Shazeer (2020)
ours: history-free (slide 6), summed goodness · ASGE (Gong et al., 2026): VGG11, linear readout
architectures, readouts, and augmentation differ: calibration, not a leaderboard · paper Sec. 3.1, 4.1–4.3, Tabs. 29, 39

The question we test

Q: If later blocks free-ride on earlier ones, is that why FF lags backpropagation?

In **cumulative-goodness** FF, a block's training score includes the earlier blocks' goodness, so later blocks can inherit a partly solved task and get little signal to improve their own margin: they **free-ride**.

Expected: repair → healthier deep blocks → higher accuracy

If free-riding limited accuracy, repairing it should help. We test that chain.

Key idea: training scores add the earlier blocks' goodness

history weight

$$G^{(d)} = g^{(d)} + \gamma \cdot \sum_{j<d} \mathrm{sg}\left(g^{(j)}\right)$$

training only

- $G^{(d)}$: block d's training score
- $g^{(d)}$: its own goodness
- γ: history weight
- $\sum_{j<d} \mathrm{sg}\left(g^{(j)}\right)$: history: earlier blocks' goodness, detached: no gradient flows back (sg = stop-gradient)

$\gamma \in [0, 1]$ weights the history. The history is detached and enters training only: at test time every variant predicts from the plain, unweighted sum of block goodness.

paper Sec. 3.2

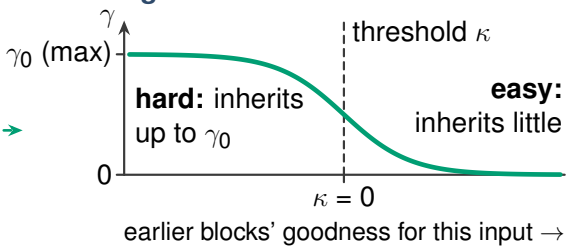
Three history rules: always, never, or gated by hardness

History rules

- **cumulative:** $\gamma = 0.7$ always
- **history-free:** $\gamma = 0$
- **hardness-gated:** this curve

Prediction never uses γ.

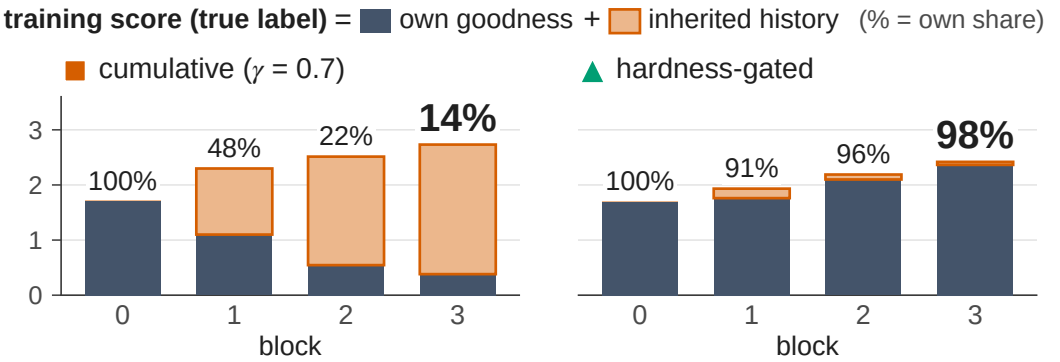
Hardness gate



Hardness gate: block *d* inherits little history where the earlier blocks' goodness for this input is already high (easy examples) and up to γ_0 where it is low (hard ones).

paper Sec. 3.2–3.3

Free-riding: block 3 makes only 14% of its training score itself

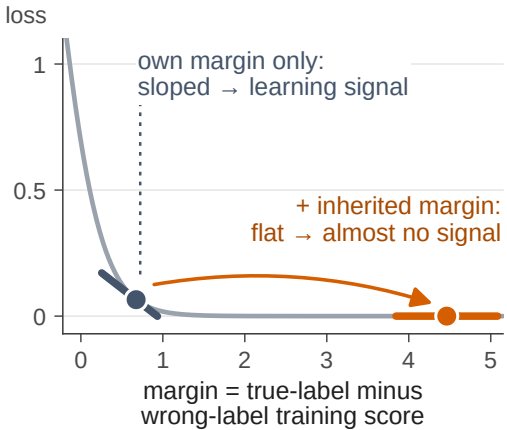


CIFAR-10 · 4 blocks, width 128 · seed 42 · paper Fig. 6

Free-riding: in the cumulative CIFAR-10 model, block 3 makes only 14% of its training score itself; 86% is inherited. With the hardness gate, 98% is its own.

1 **Result 1:** History provably flattens the cumulative loss

Inherited margin flattens the loss



Theorem 3.1 (softplus loss): block *d*'s learning signal from the cumulative loss shrinks exponentially with the earlier blocks' total margin when both it and block *d*'s own margin are positive ($\sim$96% of test images).

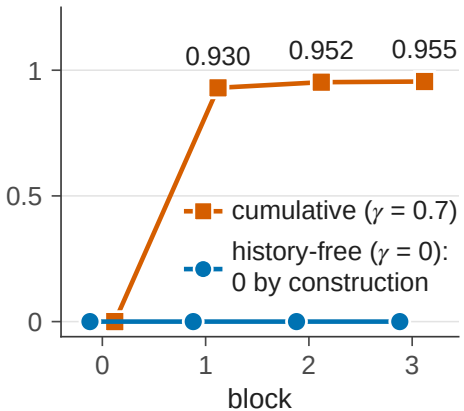
In the picture: the inherited margin moves block *d* to the flat part of the loss, where that loss gives it almost no signal to improve its own margin.

CIFAR-10 · 4 blocks, width 128 · seed 42 · dots: block-3 batch means (Tab. 4)
illustration of paper Thm. 3.1 (softplus loss, $\beta = 4$); measured version: paper Fig. 1b

1 **Measured:** history removes most of the loss slope

Measured: free-riding index

share of the loss slope lost to history



Free-riding index: the share of block d's loss slope that history removes, averaged over test images (0 = none).

- **cumulative** ($\gamma = 0.7$): 0.930 / 0.952 / 0.955 at blocks 1–3; 0 at block 0, which has no history
- **history-free** ($\gamma = 0$): 0 by construction

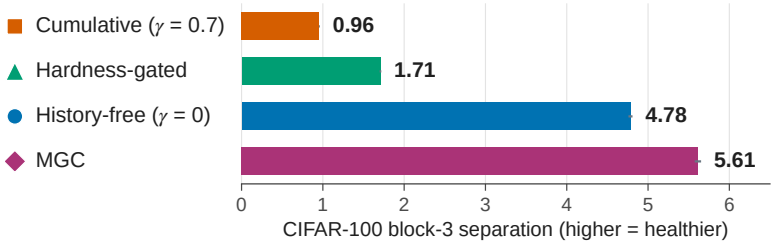
Computed per image from Theorem 3.1 (loss slope with history vs. without); it concerns the cumulative loss only.

CIFAR-10 · 4 blocks, width 128 · seed 42 · first 1024 test images · paper Fig. 1a

2 Result 2: Removing or gating history restores separation

Separation: a block's own margin against the hardest of several sampled wrong labels, averaged.



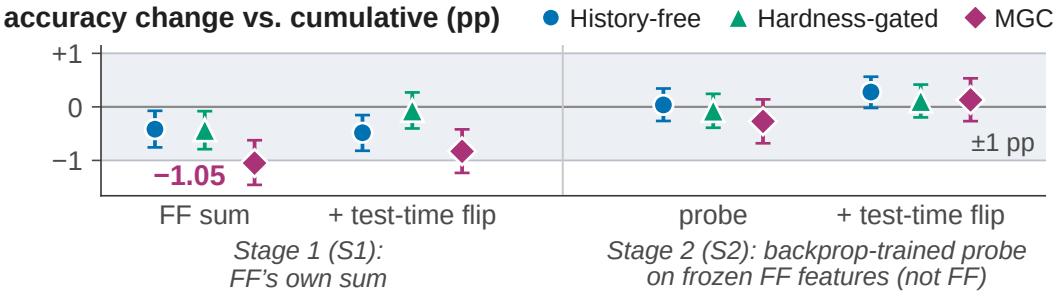


History-free and hardness-gated raise block-3 separation **up to 5**× on CIFAR-100.

CIFAR-100 · 4 blocks, width 256 · 3 seeds · mean ± SD · paper Tab. 3

MGC (missing-gradient-compensated) adds back, approximately, the signal history removes. **Caveat:** MGC replaces the depth-scaled auxiliary that the history rules keep (a loss on each block's own margin, weighted up with depth); in a CIFAR-10 control, 92% of its separation gain came from dropping it, not the compensation. Its CIFAR-100 gain is not decomposed.

3 Result 3: Healthier deep blocks, yet no accuracy gain



CIFAR-100 · paired, 3 seeds × 10,000 test images · flip: scores averaged over the image and its mirror
whiskers: 95% CI from resampling test images, conditional on the trained models · paper Tab. 3

History-free and hardness-gated: every 95% CI inside ±1 percentage point (pp); on the FF sum, slightly below 0. MGC: a consistent FF-sum cost (−1.05 pp) that the probe recovers (−0.27 pp; +0.13 pp with flip).

4 **Result 4:** The sum hides where the evidence is

Same network, goodness re-split

goodness moved between blocks



block 3: 14% of sum

block 3: 98% of sum

same sum for every label

0 / 10,000 predictions change

Retrained with a repair (probe + flip)



∼13% of predictions change

between any two history rules;
about as many fixed as broken

MGC vs. cumulative: 22.7%

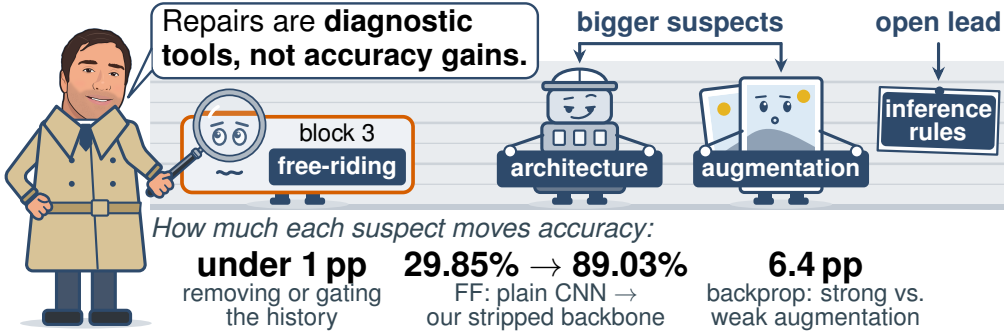
CIFAR-100 · re-split: cumulative model, seed 42, two arbitrary splits · retrained: 3 seeds · paper Prop. 3.2, App. B.7

Re-splitting the sum changes no prediction (Prop. 3.2). Repairs are not exact re-splits: they change what the network learns. That accuracy stays put on these readouts (MGC's Stage-1 cost aside) is an empirical finding, not a consequence of Prop. 3.2.

Is free-riding why FF lags backpropagation?

Real and repairable, but not the main bottleneck in our settings.

Observed: repair → healthier deep blocks ✓ → **higher accuracy ✗**



How much each suspect moves accuracy:

under 1 pp
removing or gating the history

29.85% → 89.03%
FF: plain CNN → our stripped backbone

6.4 pp
backprop: strong vs. weak augmentation

Sources: Result 3 and paper Sec. 1 (repairs); paper Table 2, CIFAR-10 (3 seeds; 89.03%: 1 seed, FF's augmentation)

A bounded negative result: scope and open questions

Scope

- softplus loss; squared hinge in one CIFAR-10 setting
- one goodness statistic
- our co-designed backbones
- CIFAR-10/100 and Tiny ImageNet
- recipe controls reduce, not remove, a recipe confound

Open

- *ImageNet scale*
- *classic FF architectures*
- *do repairs matter at greater depth?*
- *how do history and the depth-scaled auxiliary (slide 10) interact?*

paper Sec. 5

Thank you. Questions welcome

Cumulative-goodness free-riding is real and repairable but, in the settings we study, not the dominant accuracy bottleneck.

Code
github.com/amirhossein-yousefi/ff-free-riding

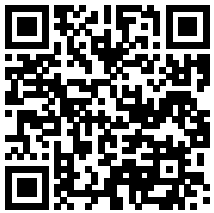
Paper
arxiv.org/abs/2605.06240

Checkpoints
huggingface.co/Amirhossein75/ff-free-riding

Project page
amirhossein-yousefi.github.io/ff-free-riding

Contact
amir.usefi75@gmail.com

Code



Paper



Backup slides

For questions at the poster. **Dissociation** here means: block health (separation) improves, accuracy does not.

B1 where MGC's separation gain comes from (a CIFAR-10 control)

B2 a second loss: the squared hinge

B3 depth: 8 blocks

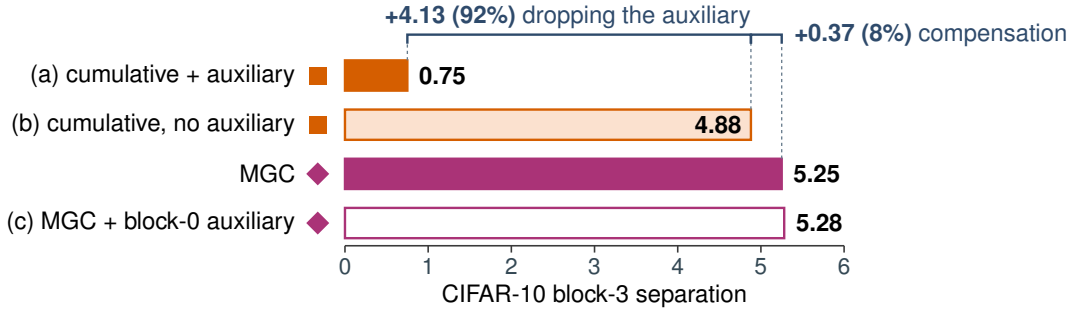
B4 could the recipe, not the mechanism, cause the dissociation?

B5 Tiny ImageNet

B6 prior FF results: calibration, not a leaderboard

B7 concurrent FF work (2026): calibration, not a leaderboard

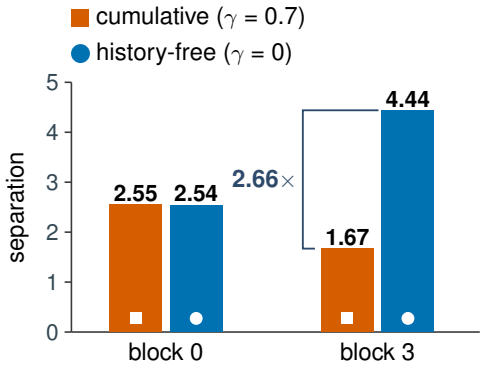
MGC's separation gain: 92% from dropping the auxiliary



CIFAR-10 · 4 blocks, width 128 · 3 seeds per arm · accuracy: every paired 95% CI inside ±1 pp · paper App. B.1.1, Tab. 6
GPUs and software differ across runs, so (b) − (a) is confounded with them; the separation effects are large against this

MGC's gain of 4.50 splits into 4.13 (92%) from dropping the **depth-scaled auxiliary** (a loss on each block's own margin, weighted up with depth, which MGC replaces) and 0.37 (8%) from compensation. The CIFAR-100 gain is not decomposed.

Squared-hinge loss: free-riding and the dissociation again



With a squared-hinge loss the inherited margin shrinks the signal linearly, to exactly zero. Cumulative training again starves block 3 (separation 1.67 vs. 4.44; block 0 unchanged), yet accuracy moves at most 0.16 pp.

87.08% vs. 86.93%

probe + flip accuracy: cumulative vs. history-free

CIFAR-10 · 4 blocks, width 128 · 3 seeds per arm · paper Thm. 3.1 remark, App. B.2, Tab. 7

loss $\max(0,\ 2 - u)^2$, with u the training margin (true-label minus wrong-label training score)

At 8 blocks the dissociation holds; deeper is open

8 blocks, width 128, 16 experts	block-7 separation	accuracy (probe + flip)
▲ hardness-gated	6.94	86.84%
○ $\gamma = 1$, no auxiliary	0.15	86.79%
8 blocks, width 128, 32 experts (mixture of experts)		
■ cumulative ($\gamma = 0.7$) + depth-scaled auxiliary		89.61%
● history-free ($\gamma = 0$)		89.22%
○ $\gamma = 1$, no auxiliary		89.20%

At 8 blocks, hardness gating raises block-7 separation (6.94 vs. 0.15 with $\gamma = 1$ and no auxiliary), yet accuracy moves +0.05 pp. With 32 experts, the depth-scaled repair is best, but all three stay within 1 pp on one seed: open.

CIFAR-10 · one seed (42) per row · paper Sec. 4.5, App. C, Tab. 15, Sec. 5

Could our recipe, not the mechanism, cause the dissociation?

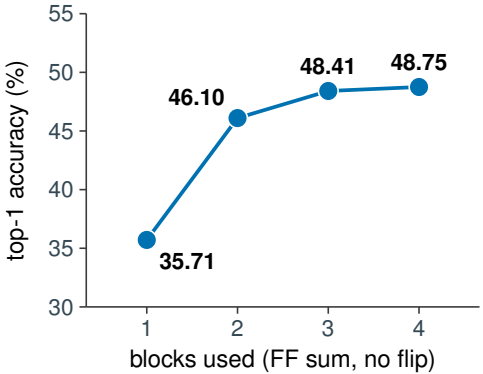
Four controls bear on it:

- ✓ the CIFAR-100 comparison uses dense blocks: no mixture of experts, no SAM
- ✓ removing SAM or SupCon, or setting $\gamma = 0$ with no auxiliary and no depth-order term, each changes probe + flip accuracy by +0.25 pp (CIFAR-10, seed 42)
- ✓ FF on our backbone stripped of its mixture of experts, memory, contrastive, and reconstruction terms still reaches 89.03% (CIFAR-10, one seed)
- ✓ a squared-hinge loss reproduces the dissociation (B2)

These controls mitigate, but do not eliminate, the concern.

SAM = sharpness-aware minimization · SupCon = supervised contrastive loss · paper Sec. 4.5, Tab. 2, App. E Tab. 26

Tiny ImageNet: block health again does not track accuracy



History-free ($\gamma = 0$) backbone tuned for the dataset (200 classes, 64×64): 52.32% top-1 with probe + flip (3 seeds).

Separation peaks at block 1 (blocks are numbered 0–3; the x-axis counts blocks used) and declines deeper, a depth effect outside Theorem 3.1 (cause open), yet accuracy rises with every block.

Tiny ImageNet · 4 blocks, width 256 · labeled validation split · 52.32 ± 0.34% over 3 seeds · per-block accuracy: seed 42
cross-dataset evidence, not an architecture-controlled transfer · paper Sec. 4.3, App. H.7

Prior FF results: calibration, not a leaderboard

Method	Architecture (CIFAR-10)	CIFAR-10	CIFAR-100	Tiny ImageNet
FF (Hinton, 2022)	locally connected, 2–3 hidden layers	59	n/r	n/r
Layer Collaboration (2024)	MLP	48.4	n/r	n/r
CwC (2024)	CNN, 4 layers in 2 blocks	78.11	51.23	n/r
Trifecta (2024)	VGG-like CNN, 12 layers	∼84	35.79	n/r
CFFM (2025)	ViT, 7 layers	85.37	n/r	n/r
Distance-Forward (2024)	CNN, 10 layers	88.15	59.01	n/r
DeeperForward (2025)	ResNet-like CNN, 17 layers	86.22	53.09	n/r
SFF (2025)	CNN, 3 blocks	81.38	55.34	n/r
SCFF (2025), unsupervised	3-layer CNN	80.8	n/r	35.67
HFF (2026)	3-layer CNN	83.08	54.34	n/a
ASGE (2026)	VGG11 CNN	90.62	65.42	n/a
Ours, single crop	transformer, 4 blocks*	90.50	66.37	48.54
Ours, with flip	same	91.32	66.93	49.09

top-1 %, single crop as each work reports it (no listed prior work reports test-time flip) · n/r: not reported
ASGE, CFFM, CwC, and SCFF train a classifier on their features; ours: summed goodness · concurrent FF work (2026): B7
*CIFAR-10: 32 experts per block, 107.2 M parameters (18.9 M active); dense on CIFAR-100 (9.49 M) and Tiny ImageNet
our Tiny ImageNet numbers: labeled validation split · n/a: no comparable number · paper App. Tabs. 25, 28–29

Concurrent FF work (2026): calibration, not a leaderboard

Method	Model; readout	CIFAR-10	CIFAR-100	Tiny ImageNet
BiCovG (2026)	VGG-8 / VGG-16; n/r / fused layer logits	n/r	70.80	50.30
HCL-FF (2026)	17-layer residual CNN; chosen layers	91.68	70.09	48.46
Ghader et al. (2026)	17-layer ResNet; weighted layers	89.37	63.52	49.87
DTG-FF (2026)	VGG8; backprop-trained head	91.33	67.28	48.17
Ours, single crop	transformer, 4 blocks; summed goodness	90.50	66.37	48.54

top-1 %, as each work reports it (paper App. Tab. 29, note p) · n/r: not reported · protocols: paper App. G
BiCovG: VGG-8 on CIFAR-100, VGG-16 on Tiny ImageNet · HCL-FF's Tiny ImageNet entry: last layer only
Ghader et al.: model named for CIFAR-10 only · DTG-FF: checkpoint chosen by test accuracy
runs: HCL-FF 5 (CIFAR-10/100), BiCovG 3 (Tiny ImageNet), DTG-FF 1; not given for the other entries
none of them analyzes cumulative goodness or free-riding, and none describes test-time augmentation
on each dataset, at least one of them is above our backbone: calibration, not a leaderboard