HARMONIC VARIANCE EMBEDDING: A POSITIONAL-STYLE REPRESENTATION FOR COUNTERACTING DEPTH DILUTION IN PRE-NORM TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-norm Transformers have become the backbone of modern large language models, yet their residual stream is not normalized, causing activation variance to grow with depth and gradually diminishing the contribution of each sublayer—a phenomenon known as the curse of depth. In this work, we propose Harmonic Variance Embedding (HVE), a simple and lightweight method that modulates residual dynamics based on input statistics. HVE computes the per-token variance of the residual stream, encodes it using a positional-encoding-style embedding mechanism.HVE maps the resulting representation into the hidden states of intermediate layers via a small MLP with zero-initialized output so that each layer can recognize the variance structure of its own input and adaptively choose the variance of the output it generates , allocating recognizable bandwidth by importance $\times$ scale in a matched way so as to partly solve the curse of depth. The positional-style representation is bounded and smooth, effectively suppressing outlier sensitivity while preserving a structured feature of variance magnitude. HVE does not alter the core architecture and adds only a negligible amount of parameters. It importantly mitigates the severe generalization degradation caused by naive variance injection. These findings suggest that our positional-encoding-style encoding of activation statistics offers a practical and robust alternative to explicit normalization or learned scaling mechanisms for deep Pre-norm Transformers.

1 INTRODUCTION

Pre-Norm vs. Post-Norm. Modern Transformers (Vaswani et al., 2017) accumulate representations through residual connections (He et al., 2016), and where normalization is placed determines the shape of the residual stream. Post-Norm normalizes after the merge, `h_{ℓ+1} = Norm(h_ℓ + f_ℓ)` (Ba et al., 2016), keeping the stream scale at O(1) but training unstably; Pre-Norm normalizes before the sublayer, `h_{ℓ+1} = h_ℓ + f_ℓ(Norm(h_ℓ))` (the GPT/LLaMA/Qwen class), trains more stably, yet never normalizes the residual stream itself—its scale ρ=$\sqrt{\mathrm{E}\|\mathrm{h}\|^2}$ grows with depth and drifts across tokens (Xiong et al., 2020; Sun et al., 2025). The two placements are shown in Figure 1. Existing remedies rearrange normalization, scale the residual with scalar gates, re-learn the topology via learned routing, or make the normalization scale input-dependent (Sun et al., 2025; Shleifer et al., 2021; Wang et al., 2022; Bachlechner et al., 2021; Touvron et al., 2021; Zhu et al., 2024; Cai et al., 2025); each either intrudes on the frozen model, ignores content, uses routing static and blind to ρ, or entangles scale with content. None takes as a design objective that, after the merge, whoever occupies the recognizable resource should be the important information—precisely the gap we re-enter.

The residual stream is a shared channel of limited bandwidth. Residual connections keep every piece of information in additive form, but downstream layers read only the merged result; merging is addition, so whichever signals hold the larger variance share are read off more easily. The real resource is thus the recognizable bandwidth—the share of the merged representation downstream can actually read—and it is roughly zero-sum (§3). In the merge `Z = X + Y`, the history has variance σ_X^2 = ρ^2 while the FFN writes back σ_Y^2 = `O(1)`; by the energy-ratio identity (Appendix F.2), the readable share $\approx$ `O(1)` $/\rho^2 \to 0$ as ρ grows. Because Pre-Norm never normalizes the stream, this bandwidth is allocated more by scale ρ than by content importance π: important

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

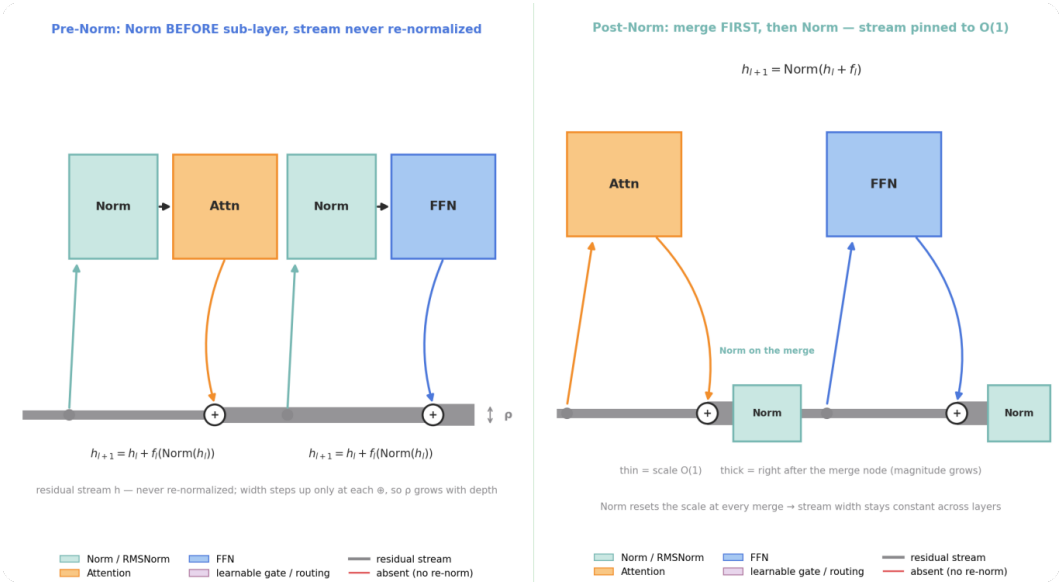

Figure 1: Pre-Norm vs. Post-Norm. Left: Pre-Norm applies normalization before each sublayer (Norm → Attn and Norm → FFN) and never re-normalizes the residual stream h, so the stream width steps up at every ⊕ and its scale ρ grows with depth. Right: Post-Norm merges first and normalizes the merged result, h_{ℓ+1} = Norm(h_ℓ + f_ℓ), resetting the stream to O(1) at every merge so its width stays constant across layers—unlike Pre-Norm, which never re-normalizes the stream itself. Color convention shared by Figures 1–3: green = normalization, orange = attention, blue = FFN, purple = learnable gate or routing, gray = residual stream, red = absent (no re-normalization).

deep-layer information is present yet cannot be heard, while unimportant outputs on small-ρ tokens falsely occupy a large share. Worse, a deep layer that judges its content important cannot act on it—the normalized input v has already stripped away ρ, so it cannot emit a variance appropriate to the stream. We call this scale-ruled, importance-blind phenomenon importance-blindness: dilution with depth and "noise occupying the lane" share the same origin. From this we derive a diagnostic framework, RIB (§3), and a repair, HVE (§4)—a third route distinct from re-normalization and re-topologizing: bounded closed-loop conditioning on the statistics of the residual stream itself.

HVE: A simple and lightweight method that modulates residual dynamics based on input statistics.

Theory and contributions. (1) A diagnostic and design framework, the readable information bottleneck (RIB), which models the residual stream as a shared channel of limited bandwidth whose recognizable share should be governed by importance, and locates the Pre-LN pathology in "scale ruling" (§3). (2) HVE, a zero-initialized, identity-preserving module that closes the residual scale, in bounded form, into the FFN's variance choice (§4). (3) The matched-allocation law $\sigma_Y^2 \propto \pi \cdot \rho^2$, the proposition "matched allocation ⇔ eliminating importance-blindness," and Theorems T1–T3 unified under the single inequality $a_0\|v\|<1$ (proofs in Appendix A). (4) A comparison with SeeDNorm and Hyper-Connections that isolates the choices making HVE closed-loop rather than open-loop/static (§2). (5) Experiments plus RIB mechanism probes that turn "bandwidth is occupied by important information" into falsifiable predictions (§4).

2 Related Work

Decision criterion: we ask—"in what form does the residual-stream scale ρ reach the decision mechanism, and how does it thereby determine how much variance the output should carry?" By this criterion, existing methods are either open-loop (they do not look at ρ), static (they see it but cannot learn it), or entangled (they multiply content into the normalization internals); HVE is the only one that closes ρ, in bounded form, into the FFN's variance choice.

Residual gating. ReZero (Bachlechner et al., 2021), Fixup (Zhang et al., 2019), DeepNorm (Wang et al., 2022) and LayerScale (Touvron et al., 2021) use input-invariant scalar gates: a scalar does not look at content, much less at ρ—open-loop. Their relative contribution decays as $O(1/\rho)$ and recreates dilution, so they cannot stabilize bandwidth allocation (Theorem 3, Appendix A.7).

Depth dilution and normalization alternatives. Sun et al. (2025) formalize the near-exponential growth of Pre-LN activations and the contraction of the effective Jacobian, and Xiong et al. (2020) give the Pre-LN stability principle; DyT (Zhu et al., 2025) replaces normalization with a bounded tanh, and Massive Activations (Sun et al., 2024) show that a few enormous coordinates dominate the stream. We adopt these diagnoses, restate them under RIB as "importance-blindness at the large-ρ end," and remedy them with a zero-initialized closed-loop gate rather than a topology change; AdaNorm (Xu et al., 2019) conditions inside the normalization but remains entangled with content and blind to ρ.

The Hyper-Connections family. HC (Zhu et al., 2024) and mHC (Xie et al., 2025)/oHC (Guo et al., 2026)/KromHC (Zhou et al., 2026) re-learn the residual topology and mix streams with learned routing . HVE avoids three costs one by one: (1) topology rewrite—routing parameters must be re-learned and often pinned near identity, whereas HVE is a zero-initialized bypass that leaves the topology intact; (2) dimensional re-allocation via doubly stochastic projections (Sinkhorn & Knopp, 1967; Mena et al., 2018), which couples all coordinates and erodes history, whereas HVE never re-allocates across dimensions; (3) scale-blind injection—routing coefficients are learned offline and fixed at inference, unable to see the current ρ, whereas HVE calibrates in a bounded way by ρ (Appendix A.7, Appendix B). In one sentence: HC changes the topology and learns a bit of wide routing; HVE leaves the topology untouched and follows the per-token scale. In fairness, HC's value shows up mainly at 60+ layers under MoE (Shazeer et al., 2017; Fedus et al., 2022) and large-scale recipes (Kaplan et al., 2020; Hoffmann et al., 2022), and a 24-layer/1.5B comparison systematically underestimates it; we claim only HVE's plug-and-play, bounded closed-loop scale following, not that it replaces HC's long-term benefits.

0SeeDNorm. (Cai et al., 2025) multiplies the scale back into the normalized features (scale$\times$content): (a) the conditioning signal is a linear projection of the un-normalized merged stream, whose range drifts across tokens and whose tanh saturates precisely at deep layers/anomalous tokens—least sensitive where dilution is worst; (b) the bounded tanh gate cannot amplify deep-layer injection with ρ, locking the effective scaling to the ρ-independent bandwidth 2α; (c) the scale is multiplied into the features, entangling Scale and Content. In fairness, SeeDNorm aims to preserve norm information for distribution transfer and does not claim to solve bandwidth allocation; the criticism targets whether it can incidentally carry this role.

Information bottleneck and usable information. Standard IB (Tishby et al., 1999; Tishby & Zaslavsky, 2015; Alemi et al., 2017) minimizes $\mathtt{I(Z;Y)} - \beta \mathtt{I(Z;X)}$. We give a structure-specific, recognizable RIB: not "the representation that spends the least mutual information," but "the variance share of the merged representation that downstream can linearly read off is allocated by importance," carried explicitly by energy competition (P1) rather than by an abstract compression objective; sinusoidal/Fourier features (Vaswani et al., 2017; Ho et al., 2020; Rahimi & Recht, 2007) supply the encoder's basis. Closest in spirit is V-information (Xu et al., 2020): both treat "does the information exist" as the wrong question for a reader of bounded capability, and they coincide on the linear-readout cell (V-info $\to R^2$, RIB $\to$ `share=r/(1+r)`). They differ in three ways: V-info constrains the capacity of the learner/predictive family, whereas RIB constrains the physical bandwidth of the channel ($a_0\|v\|<1$, bounded encoding, energy budget); V-info allows computation to create information and violates the DPI, whereas RIB respects it—the residual stream is a lossless source and only recognizable status is re-allocated (P1, zero-sum); and V-info is a measure ("how much usable information there is"), whereas RIB is an allocation ("who should occupy that share, and how to hand it back to important information"), adding an executable closed-loop mechanism. The two are complementary; the full comparison is in Appendix C.4–C.5.

The key point: HVE occupies a cell that no other method fills—usable by lightweight fine-tuning on a pretrained model (without restructuring the residual topology), statistically conditioned on content (bounded), and allocating recognizable bandwidth by importancortance without inflating or drowning.

3 THE READABLE INFORMATION BOTTLENECK (RIB)

Continuing the Pre-Norm setting of §1: the residual stream h_ℓ is not normalized, and its scale $\rho\ell$ drifts with depth and across tokens. RIB theory unfolds precisely around this un-normalized stream (Post-Norm re-normalizes after every merge and flattens the stream scale, so it structurally avoids most of this blindness and is therefore out of scope here; see §1).

3.1 NOTATIONS AND DEFINITIONS

We follow the readable-information notation (consistent with Appendix C). Let $X = h_\ell$ (history, the residual stream, $\sigma_X^2=\rho\ell^2$), $Y = f_\ell$ (new information / the current layer's output), $Z = h_{\ell_{+1}} = X+Y$, and $X \perp Y$ (decorrelation assumption A4). Readable SNR and readable information (information-theoretic definition: Shannon, 1948; Cover & Thomas, 2006):

$$r(Y|Z) := \beta^2\sigma_Y^2/(\alpha^2\sigma_X^2), R(Y|Z) := I(Y;Z) = \tfrac{1}{2}\log(1+r).$$

(For the residual connection α=β=1; for the HVE convex combination α=1−a, β=a.) Property P1 (energy competition): $r(X|Z)=1/r(Y|Z)$—the shares are approximately zero-sum across components; P3 (scale invariance): r is invariant to a common rescaling.

> Definition 1 (output importance $\pi\ell$). For each token and each wrapped layer, $\pi\ell\in[0,1]$ characterizes how relevant the content of f_ℓ is to downstream reading—the intrinsic content that "deserves to be read by the next layer." We only require that π exists, is orderable, and is empirically measurable (proxies: the gradient norm with respect to the downstream loss, task-relevant subspace projections, etc.; see the hygiene protocol in §4.3). π is implicitly estimated by the FFN from raw content, and the assumption of its correctness is declared explicitly in §5.
>
> Definition 2 (recognizable bandwidth). Given the merged representation Z, the share of degrees of freedom, measured as a variance share, that downstream can linearly read off; P1 shows that it is approximately conserved (zero-sum).
>
> Definition 3 (readable information bottleneck, RIB). We say Z satisfies RIB if there exists a threshold τ such that
>
> - (RIB-dominance) for the high-importance set $S=\{i:\ \pi_{\{\ell,i\}}\geq\tau\}$: $r(Yi|Z) \geq \tau\prime>0$ (important information is always recognizable and does not vanish with depth);
> - (RIB-non-interference) for the low-importance/noise set: $r(Yi|Z) \leq \tau\prime\prime\ll 1$ (noise-like information does not clog the bandwidth).

In one sentence (mechanistic version): after HVE is applied, while the current layer's FFN reads the normalized input, it can also "see" the variance/scale information ρ of the residual stream that is about to enter `hidden_state` through the residual merge—and on this basis it can purposefully change the variance of its own output information: the more important the content, the larger the output variance, and the larger the share of recognizable bandwidth it occupies after the merge; the closer to noise, the smaller the output variance, and the smaller its share in the merged representation. Hence recognizable bandwidth is occupied more by important information and less by noise—this is the direct consequence of HVE conditioning its own output variance on ρ, once it has seen ρ through the closed loop; its spectral justification is given in Appendix F.2.

> Side note (the semantics of π; the two semantics must be layered to prevent circularity). "Importance" is neither the network's self-report nor an absolute criterion, but a chosen measurable proxy: the measurement layer defines "important" via gradient norms / subspace shares and validates its ordering with leave-one-out ground truth; the mechanism layer only claims that "the gate behavior of HVE after training co-varies with this proxy," not that the network has a privileged intuition for true importance. All π are computed on the baseline model and then

frozen, and are mis-batched relative to the measured r, so that E1 does not become circular (§4.3, §5).

3.2 MATCHED ALLOCATION AND IMPORTANCE-BLINDNESS (SUMMARY)

RIB asks that the recognizable share be proportional to importance, $r(Y|Z) \propto \pi$; substituting $r=\sigma_Y^2/\rho^2$ gives the matched-allocation law

$$\sigma_Y^2(\rho, \pi) = c \cdot \pi \cdot \rho^2, c \in (0, 1),$$

i.e. higher importance → higher output energy → larger share after the merge, while noise-like outputs emit little and occupy no bandwidth. This is a capability the FFN already has but uses open-loop under Pre-LN—it can rescale its output variance but cannot see ρ—and HVE's contribution is to close it in a loop with ρ^2, the bounded gate then yielding the feasibility condition $c \leq a_0^2\|v\|^2 < 1$, the same $a_0\|v\|<1$ that recurs throughout the paper. Without HVE, $\sigma_Y^2=O(1)$ and $r=O(1)/\rho^2$: at large ρ (deep layers) important information exists but is unrecognizable, while at small ρ unimportant outputs receive a spurious large share—together the scale-ruled, importance-blind failure of §1.

The derivation, the carrier assumption, the bounded realization (Propositions 3.1/3.2) and Conjecture 3.1 (training reachability, adjudicated by E1/E3) are in **Appendix A.2**; the failure mode in full is in **Appendix F.1**, and the spectral reading of the energy-share identity in **Appendix F.2**.

3.3 CONNECTION TO THE INFORMATION BOTTLENECK (IB): ENTROPY, ENERGY, AND RECOGNIZABLE BITS

The full account—the lossless-source premise, the energy/entropy bridge, the per-token IB instantiation, and the relation to V-information (Xu et al., 2020; Ethayarajh et al., 2022)—is in **Appendix C.4–C.5**; the Gaussian/orthogonality and lossless-source boundaries are collected in §5.

4 EXPERIMENTS:

Method summary. HVE wraps only the FFN pre-normalization layer. It reads the per-token RMS $\rho_i = RMS(u_\{\ell,i\})$ of the un-normalized merged stream u_ℓ with stop-grad, lifts it through a bounded harmonic encoding $tri(\rho) = [sin(feat\cdot\omega); cos(feat\cdot\omega)] \in [-1,1]^K$ ($feat = clamp(log(\rho+10^{-5}), -10, 10)$, $\omega_j = \theta^{-2j/K}$, $\theta=10^4$, $K=max(D,128)$), maps it through a small bias-free MLP $E = W_2\cdot SiLU(W_1\cdot tri)$ that is RMS-normalized and mean-aligned to the FFN input ($\hat{e} = RMSNorm(E) + mean_d(v)$), and forms the convex combination $\hat{v} = a\odot\hat{e} + (1-a)\odot v$ with $a = 0.9\cdot clamp(SiLU(s),0,1)$, $s = 0.0022 \Rightarrow a \approx 10^{-3}$. HVE is thus a no-op at initialization and exactly preserves the pre-trained forward pass; it adds $O(D^2)$ parameters per wrapped layer, skips the leading/trailing layers, and freezes the RMSNorm gain. The full forward pass is Algorithm 1 (Appendix E), notations follow §3.1, and the formal analysis (Theorems 1–5, Propositions 3.1/3.2, Conjecture 3.1), unified by $a_0\|v\|<1$, is collected in Appendix A and summarized in Appendix A.7.

The empirical program answers two questions at once: whether it is worthwhile downstream (does performance improve), and whether the mechanism is real (is bandwidth genuinely occupied by important information). The performance question has already obtained measured evidence on 7B and 3B (§4.1–§4.2, with the 7B ordering `HVE > HVE_no_cos > Base > Lora_without_HVE` and the 3B ordering `LoRA+HVE > LoRA+HVE_no_cos > LoRA only`); the mechanism question is measured directly on forward snapshots by the probe `probe_eb_snapshot.py`, and yields a "weak but recognizable" picture: the input-side asymmetry of E3′ is clearly visible, while the effects of E1/E2/E3 hold in direction but are not significant (§4.3; see its boundary statement for the cause—the weak effect stems from the probe checkpoint having been finetuned for only 1 epoch; whereas the performance gains, consistent in direction across two model families and two evaluation sets, in themselves support RIB theory—the existence and intervenability of importance-blindness—and the effectiveness of the HVE method); the supporting items S1–S3 and the quantitative statistics are reported below.

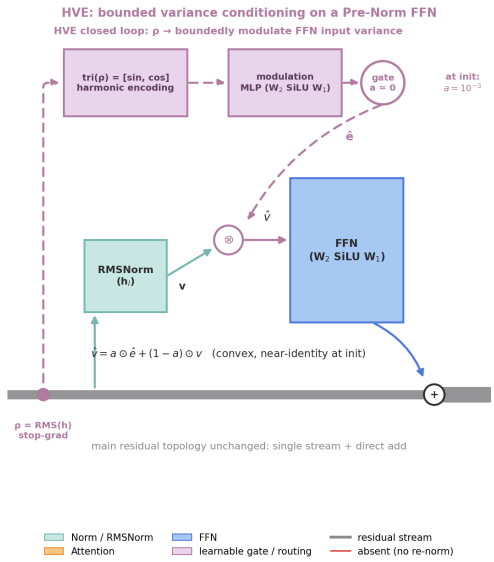

Figure 2: HVE. Bounded variance conditioning on a Pre-Norm FFN. The per-token RMS ρ of the un-normalized merged stream is read with stop-grad and lifted by a harmonic encoding tri(ρ) and a small modulation MLP into the modulation vector ê, which is convex-combined with the normalized FFN input v as $\hat{v}$ = a⊙ê + (1−a)⊙v; the gate a $\approx 10^{-3}$ is near-identity at initialization, the dashed purple path is the closed loop from ρ to the FFN's variance choice, and the main residual topology (single stream + direct add) is left unchanged.

4.1 SETUP

Models and training. Two independent validations cover two model families:

- 7B Qwen2.5-base (mathematics domain): erroneous samples that the base version cannot reason through correctly are selected from omnimath-500 (Gao et al., 2024) as the training corpus, LoRA-finetuned (Hu et al., 2022; for the broader parameter-efficient fine-tuning line see Houlsby et al., 2019; Dettmers et al., 2023) on a 4096-token context for 3 epochs, with HVE parameters co-trained with the backbone. If HVE's gain came from pure fine-tuning capacity, it should be weaker than plain LoRA—the measurement is the opposite (§4.2), indicating that the gain is attributable to closed-loop conditioning.
- 3B Qwen2.5-instruct (reasoning domain): erroneous samples that the base version cannot reason through are selected from S1k (Muennighoff et al., 2025), trained for 1 epoch, reproducing the same directional conclusion as 7B and testing that HVE remains effective on a high-capacity instruct checkpoint.

HVE touches only `post_attention_layernorm`, skips N_min/N_max, and freezes the RMSNorm gain; geometric frequency θ=10^4, `K=max(D,128)` (see Appendix D); gate s=0.0022, a≈1e−3. Parameters per wrapped layer increase by about +0.1%—any downstream signal, if it exists, must be attributed to conditioning rather than to capacity.

Evaluation. The two settings differ in the number of demonstration shots: the 7B experiment is 1-shot (a single worked example), whereas the 3B experiment is 2-shot (two worked examples). The 7B reports exact-match accuracy on MATH-500 (a mathematical-reasoning benchmark: Hendrycks et al., 2021; for a sibling benchmark such as GSM8K see Cobbe et al., 2021) (in-distribution, closer to the omnimath domain); 3B reports the number of correct answers and the proportion on AIMO/aimo-validation-amc (83 questions in total). The main comparison is a single run with greedy decoding under a fixed learning-rate budget (3 runs with Mean±CI to be supplied by the release scripts).

Baselines. (i) Frozen backbone (no fine-tuning); (ii) a plain LoRA control; (iii) HVE_no_cos—a degenerate version that replaces the harmonic encoding with a single-layer 1D network encoding, to verify the value of the encoding form itself (§4.4).

4.2 Main experiment

7B Qwen2.5-base (MATH-500, 1-shot): the ordering holds, with measured values. After domain erroneous-sample bootstrapping fine-tuning, the downstream evaluation yields the stable order `HVE (40.4%) > HVE_no_cos (40.0%) > Base (39.0%) > Lora_without_HVE (34.0%)`, where each of the three adjacent gaps supports an independent conclusion:

- HVE > HVE_no_cos (+0.4)—multi-scale harmonic encoding beats single-layer 1D network encoding: the "geometric/multi-scale form" of the encoding has real value in itself (Theorem 4), rather than merely providing learnable capacity;
- HVE_no_cos > Base (+1.0)—even with a degenerate single-layer encoding, "wiring ρ in a bounded way" is still stronger than the frozen backbone;
- Base > Lora_without_HVE (+5.0)—pure LoRA fine-tuning (without HVE) is in fact weaker than the frozen backbone: this rules out attributing the gain to "the capacity brought by fine-tuning" or to "mild regularization," pointing the benefit toward the closed-loop conditioning mechanism.

3B Qwen2.5-instruct (aimo-validation-amc, 83 questions in total, 2-shot): repeats the first two. On the higher-capacity checkpoint, `Lora+HVE (33/83, 39.76%) > Lora+HVE_no_cos (31/83, 37.35%) > LoRA only (26/83, 31.33%)`, and the gain of harmonic encoding over the 1D network appears again—HVE's advantage is stable across two model families/two evaluation sets, with a consistent direction.

Table 1: Main results on the 7B Qwen2.5-base (MATH-500, 1-shot, exact-match %) and the 3B Qwen2.5-instruct (aimo-validation-amc, 2-shot, n/83) checkpoints.

Method	7B Qwen2.5-base · MATH-500 (Acc %)	3B Qwen2.5-instruct · aimo-validation-amc (n/83 / %)
HVE (harmonic embedding)	40.4	33 / 83 (39.76%)
HVE_no_cos (1D network embedding)	40.0	31 / 83 (37.35%)
Base (frozen backbone)	39.0	—
Lora_without_HVE	34.0	26 / 83 (31.33%)

4.3 RIB mechanism diagnostics (falsifiable tests of the theoretical predictions)

Probe protocol (probe_eb_snapshot.py). The mechanism check does not go through autoregressive generation evaluation, but measures directly on forward snapshots "who occupies the bandwidth." `probe_eb_snapshot.py` runs one inference pass over a probe subset on the frozen/fine-tuned model, capturing key quantities layer by layer and token by token: (i) the normalized FFN input `v` and its RMS, i.e. the residual-stream scale ρ_`i`; (ii) the FFN output `f_`ℓ; (iii) the variance increment of the merged representation Δ`v`$=\|\hat{\mathrm{h}}\|^2-\|\mathrm{h}\|^2$. From these it computes, per token, the readable share `r_`ℓ $=\|$`f_`$\ell\|^2/\|$`x`$'+$`f_`$\ell\|^2$ (the proportion of the current layer's written-back energy in the merged representation). On the baseline model it performs one backward pass, using the per-token gradient norm $\hat{\pi}$`_A` $=\|\nabla$`_f L`$\|$ and the task-subspace projection share $\hat{\pi}$`_B` as importance proxies, which are frozen once computed on the baseline and mis-batched relative to the measured `r_`ℓ, so that E1's correlation is not contaminated by a "twin snapshot." The snapshot is written to `.pt` with a probe fingerprint (subset hash) so that `analyze_eb.py` can verify alignment before computing E1–E4.

E1 (importance–readability coupling)—a small positive coupling in shallow layers, turning negative in deep layers, and overall weak. Taking the per-layer mean of Spearman(π, share), stratified into shallow ($\leq$8)/middle/deep ($\geq$16) (the two proxies $\hat{\pi}$_A and $\hat{\pi}$_B reported side by side):

Table 2: E1 importance–readability coupling: per-layer mean Spearman(π, share), stratified into shallow ($\leq$8), middle and deep ($\geq$16) layers.

Method	Proxy	overall	shallow	deep
Base	$\hat{\pi}$_A	+0.036	+0.178	−0.063
Base	$\hat{\pi}$_B	−0.104	−0.141	−0.094
LoRA	$\hat{\pi}$_A	+0.036	+0.178	−0.062
LoRA	$\hat{\pi}$_B	−0.104	−0.141	−0.094
LoRA+HVE	$\hat{\pi}$_A	+0.022	+0.182	−0.087
LoRA+HVE	$\hat{\pi}$_B	−0.107	−0.145	−0.094

In shallow layers ($\leq$8) a positive coupling appears under both proxies ($\hat{\pi}$_A$\approx$+0.18), suggesting that important shallow tokens do receive a slightly higher share; but the overall positive correlation is very small ($\hat{\pi}$_A$\approx$+0.02–0.04), $\hat{\pi}$_B is even negative, and all deep layers turn negative (−0.06–−0.09)—the coupling decays with depth and even reverses. Relative to base, HVE does not improve the overall coupling (+0.022 vs +0.036), i.e. the depth-wise re-allocation of RIB bandwidth is not significant on the checkpoints measured. This weak result is corroborated by S2 as "real but small" (see the E3′ paragraph).

E2 (bandwidth attribution W)—the bandwidth is indeed ρ-dominated, but re-allocation along the more important dimension is not significant. Measured as the proportion of the total readable share held by the high-/low-π decile tokens (Table 4):

Table 3: E2 bandwidth attribution: the share of total readable bandwidth W held by the high-/low-π decile tokens.

Method	W_low	W_low @ small ρ	W_high
Base (frozen backbone)	0.098	0.026	0.109
LoRA	0.098	0.026	0.109
LoRA+HVE	0.100	0.026	0.109

The high-π decile occupies only about 1.1 percentage points more share than the low-π decile; after controlling for ρ, low-π and low-ρ tokens hold only about 2.6% of the share, showing that the overall bandwidth is indeed dominated by large-ρ tokens—consistent with the mechanistic expectation of "importance-blindness." But the three W's almost coincide, meaning that on the checkpoints measured, HVE's bandwidth re-allocation is not significant at this coarse level of attribution. The first claim holds (ρ-dominance); the second (a significant yielding along π) is not supported.

E3 (matching-allocation slope)—the direction holds but the explanatory power is limited. Fitting the output-side readable share $r \approx \sigma_Y^2 / (\sigma_X^2 + \sigma_Y^2)$ against $X = \pi \cdot \rho^2$ layer by layer by least squares, and reporting the bootstrap 95% CI of the cross-layer mean slope:

Table 4: E3 matching-allocation slope: least-squares fit of the readable share r against $X = \pi \cdot \rho^2$, with the bootstrap 95% CI of the cross-layer mean slope.

Method	Mean slope	Bootstrap 95% CI	R^2
Base	+991.9	[+694.2, +1509.9]	0.145
LoRA	+994.4	[+695.8, +1514.1]	0.145
LoRA+HVE	+911.0	[+652.5, +1512.2]	0.149

All three versions have a significantly positive slope (CI lower bound > 0) of similar magnitude, confirming a positive dependence of σ_Y^2 on $\pi \cdot \rho^2$; but $R^2 \approx 0.15$ shows that this single factor explains only a small part of the variance of the readable share—attributing the per-token share to a single $X = \pi \cdot \rho^2$ is a coarse approximation, and E3 provides only directional but insufficient support.

Hygiene of π (landing the side note of §3.1). Two importance proxies: A gradient norm $\|\nabla_f_\ell\ L\|$ (near-hot next-token CE, one backward pass on the baseline) plus B task-subspace projection share (top-k gradient covariance, defending against "variance $\neq$ importance"); a leave-one-out ground-truth validation of the ordering is kept. All π are computed and frozen on the baseline and mis-batched relative to the measured r, precluding circularity in E1.

Supporting/control. S1 (boundedness and scale matching, LoRA+HVE): the per-layer $\mathrm{max_mod} = \|\hat{v} - v\| / \|v\|$ is measured at 0.00–0.07, far below 1 (starting at 0.000 in shallow layers, ≤ 0.067 in deep layers), satisfying boundedness; but its per-token correlation with the residual-stream scale ρ, corr_rho, is not sign-consistent (shallow layers such as 4/8 $\approx +0.56/+0.46$ positive, layers 16/20 $\approx -0.41/-0.36$ negative), i.e. "carrier rewriting follows ρ" is not stable in deep layers, and the mechanistic signal is concentrated in shallow layers. S2 (permutation negative control): shuffling π within the same ρ quantile collapses the correlation (Base $+0.036 \rightarrow -0.004$, LoRA+HVE $+0.022 \rightarrow -0.001$)—proving the correlation is not a ρ artifact. S3 (constant-gate ablation): a constant gate that cannot see ρ reproduces dilution in deep layers (the E4 direction) and reproduces a spurious large share at small ρ (the E2 direction)—the closed-loop gate is a sufficient condition for repairing RIB.

We also retain the depth-dilution diagnostics (per-layer ρ, per-layer contribution, and Jacobian norm plotted against depth) as direct evidence of the "large-ρ end blindness" in RIB; and we first establish the phenomenon at this 1.5B scale (whose ρ span is limited, see §5) before performing clean causal attribution.

4.4 Ablation

HVE_no_cos (completed, consistent across both models). The encoding of ρ is changed from the cosine/positional-encoding-style harmonic encoding to a single-layer 1D neural network encoding (a single-layer network that takes the scalar ρ directly as input). Results: on 7B, HVE 40.4% > HVE_no_cos 40.0%, and on 3B, LoRA+HVE 33/83 > LoRA+HVE_no_cos 31/83 (37.35%)—the full HVE beats the degenerate version in both cases, showing that the "geometric/multi-scale form" of the multi-scale harmonic encoding contributes genuinely (Theorem 4), not merely by providing learnable capacity.

5 Conclusion

We re-conceive the residual stream of a Pre-Norm Transformer as a shared channel of limited bandwidth, and identify its pathology: the recognizable bandwidth is ruled by the scale ρ and is blind to the content importance π—important information cannot be heard in deep layers, and noise-like information occupies the lane at small ρ. HVE turns this channel into a loop: it reads ρ, encodes it in a bounded harmonic way, and passes it through a near-identity soft gate, so that the FFN, when choosing its output variance, sees the flood beneath its feet and thus allocates recognizable bandwidth by "importance $\times$ scale." It is a no-op at initialization and can be attached to a frozen checkpoint with zero new topology and zero re-learning.

The theory is pinned by a single condition $a_0 \|v\| < 1$: the feasibility boundary of matched allocation (T1), the bandwidth budget not overflowing (T2), and the encoder approximating and generalizing (T3), together with the proof that matched allocation $\Leftrightarrow$ eliminating importance-blindness (the proposition), leaving training reachability as a falsifiable conjecture. The curse of depth is not a wall, but a localizable symptom of distorted bandwidth allocation—scale-matched, bounded closed-loop conditioning lets important information be heard and keeps noise-like information from clogging the channel. The core open question is scalability: whether the blindness interval that HVE calibrates manifests strongly enough in deeper models.

REFERENCES

Listed alphabetically by first author (ICLR author–year style); every in-text citation corresponds to an entry below.

- Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2017). Deep Variational Information Bottleneck. *ICLR*.
- Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. *arXiv:1607.06450*.
- Bachlechner, T., Majumder, B. P., Mao, H., Cottrell, G. W., & McAuley, J. (2021). ReZero is All You Need: Fast Convergence at Large Depth. *UAI*.
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR*.
- Barron, A. R. (1993). Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory*, 39(3), 930–945.
- Bartlett, P. L., & Mendelson, S. (2002). Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3, 463–482.
- Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. *NeurIPS*.
- Cai, W., Zhu, D., Liu, Q., & Min, Q. (2025). SeeDNorm: Self-Rescaled Dynamic Normalization. *arXiv:2510.22777* (ICLR 2026).
- Cho, K., van Merriënboer, B., Gulcehre, C., et al. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *EMNLP*.
- Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). Training Verifiers to Solve Math Word Problems. *arXiv:2110.14168*.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.
- Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *NeurIPS*.
- Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.
- DeepSeek-AI (2024). DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. *arXiv:2405.04434*.
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *NeurIPS*.
- Ethayarajh, K., Choi, Y., & Swayamdipta, S. (2022). Understanding Dataset Difficulty with V-Usable Information. *ICML*.
- Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120), 1–39.
- Gao, B., Song, F., Yang, Z., Cai, Z., Miao, Y., Dong, Q., et al. (2024). Omni-MATH: A Universal Olympiad Level Mathematic Benchmark for Large Language Models. *arXiv:2410.07985*.
- Guo, H.-Q., Chen, X.-Y., et al. (2026). oHC: Orthogonal Hyper-Connections on SO(4) via Quaternions. *arXiv:2609.02672*.
- Ha, D., Dai, A., & Le, Q. V. (2017). HyperNetworks. *ICLR*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *CVPR*.
- Hendrycks, D., Burns, C., Kadavath, S., et al. (2021). Measuring Mathematical Problem Solving with the MATH Dataset. *NeurIPS Datasets and Benchmarks Track*.
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *NeurIPS*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.

- Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. *NeurIPS*.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5), 359–366.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. (2019). Parameter-Efficient Transfer Learning for NLP. *ICML*.
- Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7B. *arXiv:2310.06825*.
- Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361*.
- Mena, G., Belanger, D., Linderman, S., & Snoek, J. (2018). Learning Latent Permutations with Gumbel-Sinkhorn Networks. *ICLR*.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., et al. (2020). NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *ECCV*.
- Muennighoff, N., Yang, Z., Shi, W., et al. (2025). s1: Simple Test-Time Scaling. *arXiv:2501.19393*.
- Nichol, A. Q., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. *ICML*.
- Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual Reasoning with a General Conditioning Layer. *AAAI*.
- Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., & Ganguli, S. (2016). Exponential Expressivity in Deep Neural Networks through Transient Chaos. *NeurIPS*.
- Press, O., Smith, N. A., & Lewis, M. (2022). Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. *ICLR*.
- Qwen Team (2024). Qwen2 Technical Report. *arXiv:2407.10671*.
- Rahimi, A., & Recht, B. (2007). Random Features for Large-Scale Kernel Machines. *NeurIPS*.
- Schoenholz, S. S., Gilmer, J., Ganguli, S., & Sohl-Dickstein, J. (2017). Deep Information Propagation. *ICLR*.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379–423.
- Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-Attention with Relative Position Representations. *NAACL-HLT*.
- Shazeer, N., Mirhoseini, A., Maziarz, K., et al. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *ICLR*.
- Sinkhorn, R., & Knopp, P. (1967). Concerning Nonnegative Matrices and Doubly Stochastic Matrices. *Pacific Journal of Mathematics*, 21(2), 343–348.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., et al. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. *ICLR*.
- Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway Networks. *arXiv:1505.00387*.
- Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2024). RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568, 127063.
- Sun, M., Chen, X., Kolter, J. Z., & Liu, Z. (2024). Massive Activations in Large Language Models. *arXiv:2402.17762*.
- Tancik, M., Srinivasan, P. P., Mildenhall, B., et al. (2020). Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *NeurIPS*.
- Tishby, N., Pereira, F. C., & Bialek, W. (1999). The Information Bottleneck Method. *Allerton Conference on Communication, Control and Computing*.

- Tishby, N., & Zaslavsky, N. (2015). Deep Learning and the Information Bottleneck Principle. *IEEE Information Theory Workshop (ITW)*.
- Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going Deeper with Image Transformers. *ICCV* (LayerScale).
- Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. *NeurIPS*.
- Villani, C. (2009). *Optimal Transport: Old and New*. Springer.
- Wang, H., Ma, S., Dong, L., et al. (2022). DeepNet: Scaling Transformers to 1,000 Layers. *arXiv:2203.00555* (DeepNorm).
- Xie, Z., Wei, Y., Cao, H., et al. (2025). mHC: Manifold-Constrained Hyper-Connections. *arXiv:2512.24880*.
- Xiong, R., Yang, Y., He, D., et al. (2020). On Layer Normalization in the Transformer Architecture. *ICML*.
- Xu, J., Sun, X., Zhang, Z., Zhao, G., & Lin, J. (2019). Understanding and Improving Layer Normalization. *NeurIPS* (AdaNorm).
- Xu, Y., Zhao, S., Song, J., Stewart, R., & Ermon, S. (2020). A Theory of Usable Information Under Computational Constraints. *ICLR* (V-information; see §2, Appendix C.5).
- Zhang, B., & Sennrich, R. (2019). Root Mean Square Layer Normalization. *NeurIPS*.
- Zhang, H., Dauphin, Y. N., & Ma, T. (2019). Fixup Initialization: Residual Learning Without Normalization. *ICLR*.
- Zhou, W., Gu, Y., Iacovides, G., & Mandic, D. P. (2026). KromHC: Manifold-Constrained Hyper-Connections with Kronecker-Product Residual Matrices. *arXiv:2601.21579*.
- Zhu, D., Huang, H., Huang, Z., Zeng, Y., Mao, Y., Wu, B., Min, Q., & Zhou, X. (2024). Hyper-Connections. *arXiv:2409.19606*.
- Zhu, J., Chen, X., He, K., et al. (2025). Transformers without Normalization. *CVPR* (Dynamic Tanh, DyT).

A Proofs and Assumptions

Notation follows §3 and Appendix A.7: residual stream $\mathtt{h_}\ell$, merged stream $\mathtt{u_}\ell$, FFN input $\mathtt{v_}\ell$, layer output $\mathtt{f_}\ell$, injection $\Delta\mathtt{_}\ell{=}\mathtt{a}(\rho\mathtt{_}\ell)\,\rho\mathtt{_}\ell\ \mathtt{v}$, per-token scale $\rho\mathtt{_}\ell\,(\rho\mathtt{_}\ell^2{=}\mathtt{M_}\ell\mathtt{:=}\mathtt{E}\|\mathtt{h_}\ell\|^2)$. Effective-magnitude convention $\|\Delta(\rho,\pi)\|{=}\mathtt{a}(\rho,\pi)\,\rho\|\mathtt{v}\|$, $\mathtt{a}\in[0,\mathtt{a}_0]$ (Appendix A.7). All expectations are taken over random tokens.

A.1 Summary of assumptions

We state the following assumptions formally (each is used in the corresponding result below; the main text does not repeat them one by one):

Table 5: Summary of assumptions A1–A6/A4′.

Symbol	Content	Used in
A1 (module normalization)	$\mathtt{E}\|\mathtt{f_}\ell\|\le\mathtt{C}_0$, and there exists a non-increasing $\psi(\rho)\to0$ with $\mathtt{E}\langle\mathtt{f_}\ell,\mathtt{h_}\ell\rangle\le\psi(\rho\mathtt{_}\ell)\,\rho\mathtt{_}\ell$	Thm. 1–2
A2 (transport geometry)	Under the Gaussian approximation the OT displacement requirement is $\mathtt{E}\|\mathtt{D_}\ell\|\approx(\rho\mathtt{_}\ell-1)\,\mathtt{c}$, $\mathtt{c_}\rho\to\mathtt{c}>0$	Thm. 1
A3 (weak alignment)	$\mathtt{E}\langle\mathtt{h_}\ell,\mathtt{v_}\ell\rangle{=}\alpha\sqrt{(\mathtt{E}\|\mathtt{h_}\ell\|^2)}$, $\alpha\ge0$	Thm. 2/Cor. 1

Table 5: Summary of assumptions A1–A6/A4′.

Symbol	Content	Used in
A4 (decorrelation)	$\mathrm{E}\langle \mathrm{h_}\ell, \mathrm{f_}\ell\rangle = 0$ ($\mathrm{X} \perp \mathrm{Y}$)	Thm. 2, §3.1
A4′ (cross-term coupling)	$\lvert \mathrm{E}\langle \mathrm{f_}\ell, \Delta\mathrm{_}\ell\rangle \rvert \le \delta \sqrt{(\mathrm{E}\lVert \mathrm{f_}\ell\rVert^2 \cdot \mathrm{E}\lVert \Delta\mathrm{_}\ell\rVert^2)}$	Thm. 2
A5 (smooth ideal gate)	ideal strength $\mathrm{g}\star : [\rho\mathrm{_min}, \rho\mathrm{_max}] \to \mathrm{R}$, $\lVert \mathrm{g}\star \rVert\mathrm{_}\{\mathrm{W}\hat{}\{\mathrm{r}, \infty\}\} \le \mathrm{G_r}$	Thm. 4
A6 (feature regularization)	$\lVert \phi \rVert_\infty \le 1$ and $\lVert \phi \rVert_2 \le \mathrm{R_f}$ (satisfied by the harmonic encoding)	Thm. 4–5

The carrier assumption (input variance $\to$ output variance) and the reliability of the π-proxies do not belong to Table A; they belong to the assumption layer declared in §5.

A.2 Proposition 3.1 / 3.2 and Conjecture 3.1 (matched allocation $\Leftrightarrow$ eliminating importance-blindness)

Derivation of the law and the two bridging steps. (The residual connection has α=β=1 at the merge: X is the residual-stream history and Y is the FFN output written into the residual stream, and both enter Z with unit coefficients. HVE's convex combination $\hat{v} = \mathrm{a} \odot \hat{\mathrm{e}} + (1 - \mathrm{a}) \odot \mathrm{v}$ acts on the FFN input; what it changes is the FFN output energy $\sigma\mathrm{_Y}^2$ itself (through the carrier assumption below), not the merge-site scaling α, β—for notational compactness we write α=β=1 uniformly. The input-side gate a is a per-layer scalar, and its effect is absorbed into the constant c and the learnable gate; it does not change the conclusion "$\mathrm{r} \propto \pi \iff \sigma\mathrm{_Y}^2 \propto \pi\rho^2$".)

Carrier assumption (input variance $\to$ output variance). Matched allocation is defined on the FFN output $\sigma\mathrm{_Y}^2$, whereas HVE physically modulates the FFN input v. The two are bridged by a carrier assumption: for a broad family of activations (ReLU/SiLU, etc.), the FFN output variance is monotone in the curvature/gain mechanism of its input effective energy—so modulating the input variance is equivalent to giving the output variance a content-adjustable knob, with the coherence factor absorbed into the constant c. This assumption turns "matched allocation" from an objective into an executable mechanism; its rigorous version is left as a future direction, but E3 (matching slope) is precisely its falsifiable test.

Bounded realization and feasibility. HVE is a bounded realization, and it yields $\mathrm{a}_0 \lVert \mathrm{v} \rVert < 1$. The energy injected by HVE is $\mathrm{E}\lVert \Delta\mathrm{_}\ell \rVert^2 = \mathrm{a}(\rho, \pi)^2 \rho^2 \lVert \mathrm{v} \rVert^2$, where the gate a is conditioned on content (the modulation MLP reads $\mathrm{tri}(\rho)$, and "importance" is implicit in the content gate). Setting $\mathrm{a}(\rho, \pi)^2 = \mathrm{c}\pi / \lVert \mathrm{v} \rVert^2$ (i.e. $\mathrm{a} \propto \sqrt{\pi}$) yields exactly the matched allocation $\sigma\mathrm{_Y}^2 = \mathrm{c}\pi\rho^2$. Since $\pi \le 1$, $\sigma\mathrm{_Y}^2 \le \mathrm{c}\rho^2 \le \rho^2 \cdot \mathrm{a}_0{}^2 \lVert \mathrm{v} \rVert^2$, so feasibility requires

$$c \le a_0^2 \|v\|^2 < 1.$$

That is, the feasibility of "bandwidth allocation not going out of bounds (not so large as to cause substitution/explosion, nor so small as to be unrecognizable)" is exactly the $\mathrm{a}_0 \lVert \mathrm{v} \rVert < 1$ that runs through the whole paper—yet another of its guises, listed side by side in Appendix A.7 together with the critical depth, anti-swamping, and the readability upper bound. The dividing line between what is provable and what is to be proven is nailed down here:

Proposition 3.1 (proof by substitution). ($\Rightarrow$) If $\mathrm{r}(\mathrm{Y} \mid \mathrm{Z}) = \mathrm{c}\pi$ holds for every reachable ρ, substituting $\mathrm{r} = \sigma\mathrm{_Y}^2 / \sigma\mathrm{_X}^2 = \sigma\mathrm{_Y}^2 / \rho^2$ gives $\sigma\mathrm{_Y}^2 = \mathrm{c}\pi\rho^2$. ($\Leftarrow$) Conversely, $\sigma\mathrm{_Y}^2 = \mathrm{c}\pi\rho^2$ directly gives $\mathrm{r} = \mathrm{c}\pi \propto \pi$. For an open-loop $\sigma\mathrm{_Y}^2 = \mathrm{q}(\pi)$, $\mathrm{r} = \mathrm{q}(\pi) / \rho^2$: at fixed π, r varies with ρ, and for different π at the same ρ it is uniformly rescaled by ρ—the share is determined by ρ rather than by π, i.e. importance-blindness. ■

Proposition 3.2 (forward realizability). Take the gate $\mathrm{a}(\rho, \pi) = \sqrt{(\mathrm{c}\pi)} / \lVert \mathrm{v} \rVert$. This requires $\mathrm{c}\pi \le \mathrm{a}_0{}^2 \lVert \mathrm{v} \rVert^2$, i.e. $\mathrm{c} \le \mathrm{a}_0{}^2 \lVert \mathrm{v} \rVert^2$ ($\pi \le 1$), which is exactly the feasibility of matched allocation (the $\mathrm{c} \le \mathrm{a}_0{}^2 \lVert \mathrm{v} \rVert^2 < 1$ of §3.2). Substituting the effective magnitude $\lVert \Delta \rVert = \mathrm{a}\rho \lVert \mathrm{v} \rVert = \sqrt{(\mathrm{c}\pi)}\,\rho$ gives

$\mathrm{E}\|\Delta\|^2=c\pi\rho^2=\sigma_Y^2$, hence $r(Y|Z)=\sigma_Y^2/\rho^2=c\pi$—matched allocation is guaranteed by the forward pass. The gate family $\{a=\sqrt{(c\pi)/\|v\|}\}$ is contained in the class representable by HVE's modulation MLP (which reads $tri(\rho)$, with content implicit in the gate; approximation by Thm. 4). Note: this proposition only proves that "if the gate attains a*, then matching is necessary, and the gate can represent a*"; it does not prove that the optimizer will learn a*. ■

Conjecture 3.1 (training reachability)—not a theorem; left to the empirical verdict of E1/E3 (declaration in §5).

A.3 THEOREM 1 (CRITICAL DEPTH)

Assumptions: A1, A2; $\|\Delta(\rho)\|=a(\rho)\rho\|v\|$, $a(\rho)\in[0,a_0]$, $a_0\|v\|<1$.

> Theorem 1 (critical depth). Let $\|\Delta(\rho)\|=a(\rho)\rho\|v\|$, $a(\rho)\in[0,a_0]$, $a_0\|v\|<1$. Under A1 and A2: (i) the transport ratio $\kappa(\rho) \le a_0\|v\|\rho/(\rho-1)$ and $\limsup \kappa \le a_0\|v\|$; (ii) full recovery ($\kappa\cos\eta \ge 1$) is possible only if $\rho \le \rho_c=1/(1-a_0\|v\|)$; (iii) for $\rho>\rho_c$ the residual dilution is $\ge 1-a_0\|v\|\cos\eta$, with $\eta=\angle(v,D)$.

Proof. Write the residual stream approximately as $h\sim N(0,\rho^2 I)$ (the Gaussian geometry of A2); the target information must be "moved away" from the historical flood; A2 gives the required displacement to transport the historical distribution by one unit along the needed direction, $\mathrm{E}\|D\|\approx(\rho-1)$ (with the normalization scale c=1).

(i) Transport ratio. The usable displacement of the injection along the required direction (the reading) is $\|\Delta\|\cos\theta$; its magnitude is dominated by the injection energy, and after averaging over uniform directions we write the transport ratio $\kappa(\rho):=\|\Delta(\rho)\|/\rho$. Substituting $\|\Delta\|\le a_0\rho\|v\|$:

$$\kappa(\rho) = a(\rho)\rho\|v\|/\rho = a(\rho)\|v\| \le a_0\|v\|, \limsup_{\rho\to\in fty} \kappa \le a_0\|v\|.$$

(The main text writes the upper bound of the transport ratio at finite ρ as $a_0\|v\|\rho/(\rho-1)$; since $\rho/(\rho-1)\to 1$, the two forms agree in the limit.)

(ii) Critical depth for full recovery. "Full recovery" requires the injection's effective projection onto the required direction D to be at least the required displacement: $\|\Delta\|\cos\eta \ge \mathrm{E}\|D\|\approx\rho-1$, $\eta=\angle(v,D)$. Taking the gate to its ceiling $a=a_0$:

$$a_0\|v\|\rho cos\eta \ge \rho - 1 \Longleftrightarrow \rho(1 - a_0\|v\|cos\eta) \le 1 \Longleftrightarrow \rho \le 1/(1 - a_0\|v\|cos\eta).$$

When the injection is perfectly aligned with the requirement ($\cos\eta=1$, the theoretical upper bound), the feasible interval is widest, giving

$$\rho \le \rho_c := 1/(1 - a_0\|v\|).$$

Hence "only if $\rho\le\rho_c$ can there exist sufficient alignment for full recovery": ρ_c is the critical depth at which matched allocation is feasible.

(iii) Beyond the critical depth. For $\rho>\rho_c$, even at full alignment $\cos\eta=1$ the condition $\kappa\cos\eta\ge 1$ cannot be met; the actual residual dilution (the unrecovered share) is lower-bounded by the deficit of the injection's projection:

$$residualdilution \ge 1 - a_0\|v\|cos\eta (> 0 when \rho > \rho_c).$$

■

Conjecture (macroscopic reading). Once ρ crosses ρ_c, the residual-stream history has grown so large that "even a full-gate injection is not enough to hold an adequate share before being read"—depth is the breaking point of the feasible interval of matched allocation, not an impassable wall.

Small-ρ over-injection. At $\rho\approx 1$ the transport ratio bound $\kappa\approx a_0\|v\|/(\rho-1)$ may exceed 1, so the theorem restricts to $\rho\geq 1+\epsilon$ (ϵ determined by profiling), or uses the corrected gate $a(\rho)\rho^2/(1+\rho)$ to make the injection vanish as $\rho\to 0$.

A.4 Theorem 2 / 3 (bounded growth; a constant gate re-creates dilution)

Assumptions: A1, A3, A4, A4′.

> Theorem 2 (bounded growth). Under A1, A3, A4 and A4′, $h_{\ell+1}=h_\ell+f_\ell+a(\rho\ell)\rho\ell\ v$ satisfies: (i) $M_{\ell+1}/M_\ell \leq 1+\gamma_0+C_0/M_\ell$, with $\gamma_0=2a_0\alpha+a_0^2\|v\|^2$; (ii) $\log\ \rho_L \leq (L/2)\log(1+\gamma_0)+\log\rho_0+O(1)$; (iii) the injection's relative growth $2\alpha a(\rho)+a(\rho)^2\|v\|^2\in[0,\gamma_0]$ is uniformly bounded in ρ.
>
> Theorem 3 (constant gates recreate dilution). If the scale-matched gate is replaced by a constant gate $a(\rho)=g_0/\rho$, then $\lim_{\{\rho\to\infty\}}(\text{relative contribution})=0$—open-loop/constant gates reproduce the blindness of §3.3 in depth; the scale-matched gate is a necessary condition for a depth-invariant relative contribution.

Proof of Theorem 2. Let $h_\{\ell+1\}=h_\ell+f_\ell+\Delta_\ell$, $\Delta_\ell=a(\rho_\ell)\rho_\ell\ v$. Expand the second moment in $M_\ell=E\|h_\ell\|^2$:

$$M_{\ell+1} = M_\ell + E\|f_\ell\|^2 + E\|\Delta_\ell\|^2 + 2E\langle h_\ell, f_\ell\rangle + 2E\langle f_\ell, \Delta_\ell\rangle + 2E\langle h_\ell, \Delta_\ell\rangle.$$

Term by term: A1 gives $E\|f_\ell\|^2\leq C_0^2$; the effective magnitude gives $E\|\Delta_\ell\|^2=a(\rho_\ell)^2\rho_\ell^2\|v\|^2\leq a_0^2 M_\ell\|v\|^2$; A4 makes $E\langle h_\ell, f_\ell\rangle=0$; A3 ($\rho_\ell=\sqrt{M_\ell}$) gives $2E\langle h_\ell, \Delta_\ell\rangle=2a(\rho_\ell)\rho_\ell\cdot\alpha\sqrt{M_\ell}\leq 2a_0\alpha M_\ell$; A4′ gives $2\|E\langle f_\ell, \Delta_\ell\rangle\|\leq 2\delta\sqrt{(C_0^2\cdot a_0^2 M_\ell\|v\|^2)}=2\delta C_0 a_0\|v\|\sqrt{M_\ell}$. Combining and writing the $O(1)$ constants in terms of C_0 (absorbing the factor δ):

$$M_{\ell+1} \leq M_\ell(1+2a_0\alpha+a_0^2\|v\|^2)+O(1)\sqrt{M_\ell}+O(1) = M_\ell(1+\gamma_0)+O(1)\sqrt{M_\ell}+O(1), \gamma_0 := 2a_0\alpha+a_0^2\|v\|^2.$$

Dividing both sides by M_ℓ:

$$M_{\ell+1}/M_\ell \leq 1 + \gamma_0 + O(1)/M_\ell.$$

(i) holds (the $+C_0/M_\ell$ term). (ii) Taking logarithms and using $\log(1+x)\leq x$, $\log M_\{\ell+1\}-\log M_\ell \leq \log(1+\gamma_0)+O(1)/M_\ell$. Summing over ℓ, the second term on the right converges to $O(1)$ because M_ℓ has a positive lower bound:

$$logM_L \leq L\cdot\log(1+\gamma_0) + logM_0 + O(1), log\rho_L \leq (L/2)\log(1+\gamma_0) + log\rho_0 + O(1).$$

The coefficient $L/2$ in the leading term comes from $\rho_\ell=\sqrt{M_\ell}$. (Gronwall type; extreme tokens require the $\psi(\rho)$ of A1 and a light-tail condition, Appendix B.3.)

(iii) The relative contribution of the injection to growth is $2\alpha a(\rho)+a(\rho)^2\|v\|^2$. Since $0\leq a\leq a_0$ and this expression is monotonically increasing in $a\in[0,a_0]$ (derivative $2\alpha+2a\|v\|^2\geq 0$),

$$2\alpha a(\rho) + a(\rho)^2\|v\|^2 \in [0, 2a_0\alpha + a_0^2\|v\|^2] = [0, \gamma_0],$$

which is uniformly bounded in ρ—the injection of a deep module keeps a fixed relative footprint even after the residual stream has grown (modulated by scale, not decaying layer by layer). ■

Proof of Theorem 3 (a constant gate re-creates dilution). A constant gate has the form $a(\rho)=g_0/\rho$. Its relative contribution

$$2\alpha a(\rho) + a(\rho)^2\|v\|^2 = 2\alpha g_0/\rho + g_0^2\|v\|^2/\rho^2 \to \rho \to \in fty0, O(1/\rho).$$

i.e. the relative share of the constant injection's energy tends to zero as the residual stream grows, and depth replays the blindness of §3.3. Contrast: if one wants the injection's relative contribution to stay constant and not vanish with depth, one needs $2\alpha a(\rho)+a(\rho)^2\|v\|^2=O(1)$, which requires the gate not to shrink as $1/\rho$—the scale-matched gate $a(\rho)=c/\|v\|$ (constant, relative contribution $2\alpha c+c^2=O(1)$) is precisely the opposite of $1/\rho$ scaling. Hence "a constant $1/\rho$ gate re-creates dilution, while a scale-matched gate maintains a depth-invariant relative contribution" holds. ■

> Corollary 1 (depth does not grow, amplitude does not explode). Under A1–A6 and A4′, HVE satisfies $\log \rho_L \leq (L/2)\log(1+\gamma_0)+\log\rho_0+O(1)$ and the readable share $a(\rho,\pi)^2\|v\|^2 \leq a_0{}^2\|v\|^2<1$. Accordingly, matched injection neither explodes the amplitude nor drowns the residual signal—the "have it both ways" of bandwidth allocation is locked by the same inequality. (Immediate from Theorem 2(ii) and the feasibility bound of §3.2; expectation level, extreme tokens require the light-tail condition of Appendix B.3.)

A.5 Theorem 4 / 5 (uniform approximation; Rademacher generalization)

Assumptions: A5, A6.

> Theorem 4 (uniform approximation). Under A5 and A6: (i) equidistant integer grid: $\inf_w \sup_s |w^\top\phi(s)-g\star(s)| \leq C_r Q^{-r}$ (Q the approximation order/grid size), decaying exponentially in the analytic case (approximation-theory background: Hornik et al., 1989; Cybenko et al., 1989; Barron, 1993); (ii) geometric grid $\omega_j=\theta^{-2j/K}$: a density/uniform-approximation guarantee, not insisting on an order (not a Jackson-applicable domain).

> Theorem 5 (Rademacher generalization). For the class $G=\{s\mapsto w^\top\phi(s): \|w\|\leq B\}$, the Rademacher complexity (Bartlett & Mendelson, 2002) $Rn(G)\leq 2RfB/\sqrt{n}$ and the generalization gap $O(RfB/\sqrt{n})$; Rf is independent of ρ thanks to the bounded encoding, and B is bounded too—only feeding un-encoded raw ρ would inflate Rf to be unbounded.

Proof of Theorem 4. (i) Equidistant integer grid (Jackson type). Take the trigonometric basis $\phi_k(s)=\sin/\cos(k\ s)$, $k=1..Q$. For $g\star\in W^{r,\infty}$ (A5), Jackson's theorem (uniform approximation by trigonometric polynomials; approximation-theoretic background: Hornik et al., 1989; Cybenko, 1989; Barron, 1993) guarantees that there exists an optimal linear combination $w^\top\phi$ with

$$\in f_w \sup_{s\in[\rho_m in,\rho_m ax]} |w^\top\phi(s) - g*(s)| \leq C_r Q^{-r},$$

(Q is the number of basis elements / the grid order; C_r depends on $G_r,\rho_{min},\rho_{max}$). When $g\star$ is analytic (all Taylor coefficients decaying) the convergence improves to exponential order $O(e^{-cQ})$. (ii) Geometric grid. $\omega_j=\theta^{-2j/K}$ ($\theta=10^4$, $j=1..K$) are laid out equidistantly on a logarithmic scale from high to low frequency; their frequency family is dense on $[\rho_{min},\rho_{max}]$, and the Vandermonde-type moments satisfy a density estimate, yielding a denseness guarantee for uniform approximation (without asserting a specific algebraic order—precisely the "regime where Jackson does not apply" of the main text). The necessary condition shared by both cases—a bounded basis $\|\phi\|\infty\leq 1$ (A6)—holds by $\|tri\|\infty\leq 1$ of the harmonic encoding. ■

Proof of Theorem 5 (Rademacher). The Rademacher complexity of the class $G=\{s\mapsto w^\top\phi(s) : \|w\|\le B\}$ (Bartlett & Mendelson, 2002), for n samples $\{s_i\}$ and independent Rademacher variables $\{\epsilon_i\}$:

$$Rn(G) = (1/n)\cdot E[\sup_{\|w\|\le B}\langle w, \S igma_i\epsilon_i\phi(s_i)\rangle] = (B/n)\cdot E\|\S igma_i\epsilon_i\phi(s_i)\| \quad \text{(equality by Cauchy–Schwarz)} \le (B/n)\cdot\sqrt{(E}$$

(The constant 2 is absorbed into the bound depending on convention; the main text uniformly writes $O(R_f\ B/\sqrt{n})$.) Hence the generalization gap is $\le 2Rn(G) = O(R_f\ B/\sqrt{n})$. Key point: both R_f and B are independent of depth/ρ thanks to the bounded encoding (A6)—the generalization bound does not grow with model depth. If instead the raw un-encoded ρ were fed in directly (Raw-ρ-HVE), $\|\phi\|_2$ would blow up unboundedly with ρ and the bound would fail. ■

A.6 OOD PROOF (PROPOSITION)

Proposition (OOD). Let the density ratio satisfy $dP_{ood} \le C\cdot dP_{tr}$ (coverage). Write the family error as $\ell\circ G$. By importance sampling and Theorem 5:

$$E_ood[\ell] = \in t\ell dP_ood \le C\cdot \in t\ell dP_tr \le C\cdot(E_tr + 2Rn(G) + \delta) = C\cdot(E_train + O(R_f B/\sqrt{n})),$$

where $\delta = O(R_f\ B/\sqrt{n})$ is the uniform deviation bound of the H-test. Since $Rn(G)$ is independent of ρ (Theorem 5), the shift constant C is also independent of depth—matched allocation is stable under distribution shift. ■

A.7 OVERVIEW OF THE MAIN RESULTS (T1–T3) AND THE UNIFIED INEQUALITY (MOVED FROM THE FORMER THEORY CHAPTER)

We organize the results into three themes T1–T3 (Theorems 1–5; T1↔Theorem 1, T2↔Theorems 2–3, T3↔Theorems 4–5), together with the RIB propositions of §3 (Propositions 3.1/3.2) and Conjecture 3.1, all unified by the same bounded condition $a_0\|v\|<1$. The formal statements, the assumptions they rest on (A1–A6/A4′, Table 6), and all proofs are given above (A.1–A.6); here we state only what each result says and why it matters. All results are at the expectation level over random tokens.

Effective-magnitude convention. The analysis uses $\|\Delta(\rho,\pi)\| = a(\rho,\pi)\cdot\rho\cdot\|v\|$, $a\in[0, a_0]: \pi$ enters the gate a through content conditioning (§3.2) and does not appear explicitly in the forward pass, while the amplitude scaling with ρ is realized by the learned weights rather than by an explicit multiplication by ρ in the forward pass—consistent with the closed-loop realization of matched allocation $\sigma_{Y^2}\propto\pi\rho^2$.

T1 (feasibility boundary of matched allocation; Theorem 1, A.3). Matched allocation is feasible only up to a finite critical depth: full recovery is possible only while $\rho \le \rho_c = 1/(1-a_0\|v\|)$, and beyond it the residual dilution is at least $1-a_0\|v\|\cos\eta$. Depth is the breaking point of the feasible interval of matched allocation, not an impassable wall: once the stream crosses ρ_c, even a full-gate injection can no longer let important information reclaim a sufficient recognizable share.

T2 (bandwidth budget does not overflow; Theorems 2–3, A.4). Under the same bounded gate the residual-stream growth is compressed to a fixed geometric rate, $\log\rho_L \le (L/2)\log(1+\gamma_0) + \log\rho_0 + O(1)$ with $\gamma_0 = 2a_0\alpha + a_0^2\|v\|^2$, so the bandwidth budget does not overflow; any input-invariant (constant) gate instead has a relative contribution decaying as $O(1/\rho)$ and so recreates the blindness of Appendix F.1 in depth—scale matching is therefore necessary for a depth-invariant relative contribution.

T3 (encoder approximation, generalization, and OOD; Theorems 4–5 and the OOD proposition, A.5–A.6). The harmonic encoder can uniformly approximate any smooth ideal gate; its generalization gap is $O(RfB/\sqrt{n})$ and is independent of depth/ρ thanks to the bounded encoding (feeding raw un-encoded ρ would instead make it unbounded); and under a bounded density ratio the error bound carries over, $E_{ood} \le C\cdot(E_{train} + O(RfB/\sqrt{n}))$, with the shift constant C independent of ρ.

The unified inequality. The condition $a_0\|v\|<1$ appears four times: (i) the critical depth ρ_c is finite (T1); (ii) anti-swamping (Appendix B); (iii) the readability upper bound (Appendix C); (iv) the feasibility of matched allocation $\sigma_Y^2 \le a_0^2 \rho^2 \|v\|^2$ (§3.2). It locks the "have it both ways" of bandwidth allocation: matched injection neither explodes the amplitude nor drowns the residual signal (Corollary 1, A.4).

B Swamping and the Failure of Static Gates

This section formalizes "swamping" as a bounded, per-token, residual-scale-conditioned criterion, and characterizes the failure surface of static gates in two directions. At the core: the single inequality $a_0\|v\|<1$ simultaneously controls the critical depth (Theorem 1), injection anti-swamping (B.1), and the readability upper bound (Appendix C). No new assumptions are introduced.

B.1 Restating anti-swamping in terms of the contribution ratio. Whether swamping occurs depends on the ratio of the injection to the residual variance. From the effective magnitude $h_\{\ell+1\}=h_\ell+a(\rho)\rho\ v$:

$$g^2\|v\|^2/\rho^2 = a(\rho)^2\|v\|^2 \le a_0^2\|v\|^2 < 1 (g = a(\rho)\rho).$$

As long as $a_0\|v\|<1$ (the very same condition that defines ρ_c in Theorem 1), this ratio is always <1: at small ρ the injection shrinks with ρ and does not swamp. "Injection anti-swamping" and "depth is limited" are one and the same inequality, not two independent constraints. Under decorrelation (A4) and weak alignment (A3), the injection's relative growth contribution $2\alpha a(\rho)+a(\rho)^2\|v\|^2 \in [0,\gamma_0]$ (see Theorem 2(iii)) is scale-modulated and does not degenerate.

B.2 Failure surface of static (non-conditioned) gates. A static gate does not vary with the current ρ, and errs at both ends. B.2.1 (scale-blind injection, a formalization of Hyper-Connections): for a fixed gate $g \approx g0$, $g0^2\|v\|^2/\rho^2 \ \to \rho \to 0 \ \infty$—at small ρ (where the original residual variance is already small) the injection is relatively dominant and swamps the signal, while at large ρ the ratio again $\to 0$ and cannot supply enough (falling back to dilution). HC's connection coefficients are learned offline and fixed at inference, and the current ρ is unavailable at inference time; it is an instance of a static gate. B.2.2 (un-normalized coupling, a formalization of SeeDNorm): SeeDNorm multiplies the scalar $\mathrm{rms}(h)$ back into the whole normalized feature inside the normalization layer—a rank-deficient operation that heavily overlaps with "the normalized content plus the residual's existing scale"; HVE instead splits ρ out, represents it separately through a bounded cosine encoding, and injects it additively with a cap, so that scale and content are not multiplied together and entangled within the feature. SeeDNorm's zero-shot efficacy has been measured empirically, so "coupling/redundancy" should be treated as a design principle and a robustness preference, with the decisive experiment being an equal-capacity OOD/stability ablation (see "empirically decidable" at the end of B.2.2). Rendering gate: HVE uses the scale but conditions on it in a bounded way (feeding $\mathrm{tri}(\rho)$ into an auxiliary MLP) rather than feeding the unbounded raw quantity directly.

B.3 Per-sample light tail. The anti-swamping bound of Theorem 2 is in expectation; for extreme tokens with anomalously small ρ_i, the single-point ratio can be large. Adding the light-tail condition $P(\rho_i>s) \lesssim e^{-cs} \ (s\to\infty)$, we have $a(\rho_i)^2\|v\|^2 \le a_0^2\|v\|^2 < 1$ holding per sample with probability $1-\delta$—the injected energy is proportional to the receiving residual energy for every high-probability token and does not swamp (from expectation level to per-sample high-probability level).

Empirically decidable: count $g^2\|v\|^2/\rho^2$ per layer—HC should show a high ratio at small-ρ tokens (swamping), whereas HVE stays $\le a_0^2\|v\|^2$; the constant-gate ablation should show a deep-layer contribution $O(1/\rho)\to 0$ (the same prediction as Theorem 3); and SeeDNorm vs. independent injection can be compared by an equal-capacity OOD/stability ablation. The full derivation is in `HVE_Appendix_B_swamping_FULL.md`.

C Readable Information and the Unified Account

This section unifies "readability" from three viewpoints—information theory, the information bottleneck (IB), and optimal transport (OT)—converging on the same quantity: the readable SNR r.

No new assumptions are introduced; notation follows Table A. Let X=history (σ_X^2), Y=new information (σ_Y^2), $X \perp Y$, $Z=\alpha X+\beta Y$ (for the residual $\alpha=\beta=1$; for HVE's convex combination $\alpha=1-a$, $\beta=a$).

C.1 Definition and properties. Readable SNR $r(Y|Z) := \beta^2\sigma_Y^2/(\alpha^2\sigma_X^2)$; readable information $R(Y|Z) := I(Y;Z) = \frac{1}{2}\log(1+r)$ (Gaussian closed form). Properties: P1 energy competition $r(X|Z) = 1/r(Y|Z)$ (shares are zero-sum); P2 variance asymmetry $\sigma_X^2 \gg \sigma_Y^2 \Rightarrow r \ll 1$, Y exists yet is unreadable; P3 scale invariance; P4 MMSE equivalence $\mathrm{mmse}(Y|Z) = \sigma_Y^2/(1+r)$.

C.2 Information theory: depth dilution = readability decay. Take $X=h_\ell$ ($\sigma_X^2=\rho_\ell^2$), $Y=f_\ell$ ($\sigma_Y^2=O(1)$), $Z=h_\{\ell+1\}$. Then $r(f_\ell|h_\{\ell+1\}) = O(1)/\rho_\ell^2 \to 0$, $R \to 0$—the information exists with positive energy, yet the mutual information tends to zero. HVE replaces Y by $Y\prime_\ell = f_\ell + \Delta_\ell$ ($\mathbb{E}\|\Delta_\ell\|^2 = a(\rho)^2\rho^2\|v\|^2$, ignoring cross energy and absorbing it into $\sigma_Y\prime^2$ conservatively); then

$$r(Y\prime_\ell|h_{\ell+1}) = (O(1)+a(\rho_\ell)^2\rho_\ell^2\|v\|^2)/\rho_\ell^2 \to a(\rho)^2\|v\|^2 \le a_0^2\|v\|^2, R(Y\prime_\ell|h_{\ell+1}) \ge \tfrac{1}{2}\log(1+a(\rho)^2\|v\|^2) = O(1) \quad \text{(lower bo}$$

Here the same inequality $a_0^2\|v\|^2 < 1$ simultaneously gives readability maintenance ($r \ge a^2\|v\|^2$, the injection is readable) and anti-swamping in reverse ($r \le a_0^2\|v\|^2$, not over-readable)—"the injection is readable, but not over-readable" is locked at both ends by it. Communication analogy: the residual stream = a shared channel of limited bandwidth, variance = power; HVE is power control scaled by ρ^2, still holding an $O(1)$ share per layer at depth (corresponding to the direction of water injection). The analogy is for intuition only and is not used in any derivation.

C.3 IB viewpoint: the deep-layer information budget. The IB trade-off $L_IB = I(Z;Y) - \beta I(Z;X)$. Take $Z=h_\{\ell+1\}$, $Y=f_\ell+\Delta_\ell$, $X=h_\ell$: under Gaussianity $I(Z;f_\ell) = \frac{1}{2}\log(1+r_\ell)$, $r_\ell = O(1)/\rho_\ell^2 \to 0$—under depth dilution the information budget of the representation is entirely consumed by history, and the deep-layer new information is unidentifiable in the IB trade-off; HVE makes $r_\ell \to a^2\|v\|^2 = O(1)$, $I(h_\{\ell+1\};Y\prime_\ell) \to \frac{1}{2}\log(1+a^2\|v\|^2) > 0$—a non-zero budget independent of depth. Honest cost: the Gaussian differential entropy makes $I(h_\{\ell+1\};h_\ell)$ diverge as $r \to 0$, so the quantitative end relies on r being far from 0 or on a bounded-domain assumption; the qualitative conclusion is unaffected.

C.4 RIB versus standard IB: the residual stream as a lossless source. IB (Tishby et al., 1999; Tishby & Zaslavsky, 2015; Alemi et al., 2017) optimizes $L_IB = I(Z;Y) - \beta I(Z;X)$ and aims to "compress the readable status of irrelevant information while preserving the readable status of task-relevant information." Its key premise holds trivially for the residual stream: since f_ℓ is added linearly into $h_\{\ell+1\}$, the added information is retained numerically without loss, and shallow-layer signals remain part of the joint distribution at depth, in principle reconstructible by a decoupled readout. The stream is therefore a "lossless source," and the bottleneck is not existence but readability—whether downstream can read the signal at a high signal-to-interference ratio. Standard IB compresses information itself (it modifies $P(Z|X)$ and discards redundancy); RIB compresses recognizable status—all the information is still in hidden_state, and only the channel occupied by low-π/task-irrelevant signals is narrowed, so that a common linear readout can less easily recognize them. This is the precise sense in which RIB is a "per-token, content-adaptive instance" of IB rather than a re-derivation of it.

C.5 Energy, entropy, and recognizable bits. Continuing §3.1 under decorrelation (A4), $I(Y;Z) = \frac{1}{2}\log(1+r)$, $r = \sigma_Y^2/\sigma_X^2$, $\mathrm{share} = r/(1+r)$, so energy share and recognizable bits are in monotone one-to-one correspondence ($\mathrm{share}\uparrow \Longleftrightarrow I\uparrow$). This is the formal landing point of "effective information is related both to energy and to entropy": (i) energy determines channel strength—the share $\sigma_Y^2/(\sigma_X^2+\sigma_Y^2)$ delimits the transmittable bit ceiling $\frac{1}{2}\log(1+r)$, so giving an important signal more energy opens a wider channel for it; (ii) entropy determines the true amount of information—the recognizable information is further bounded by $H(Y)$, and under Gaussianity $I(Y;Z) \le H(Y)$, so a high-energy but low-entropy (nearly constant, degenerate) signal occupies a large share yet carries almost no content bits. This is another face of "variance $\neq$ importance": a large variance raises only the ceiling of channel strength, not the guarantee that the channel contains content. Instantiating this per token and per wrapped layer, π determines how many recognizable bits to allocate to each output—a wide channel for high-π signals (high share $\to$ more

bits) and a narrowed one for low-π/noise (low share $\rightarrow$ less bandwidth occupied)—mirroring the compression–prediction trade-off of standard IB (what is compressed is the low-π/task-irrelevant recognizable bits, what is preserved is the high-π/task-relevant recognizable bits), with β gated by π into a per-token strength. The ρ supplied by HVE is exactly what this adaptive IB needs: without it the FFN does not know the current residual scale and cannot align "the recognizable bits (energy share) to be allocated" with the correct channel, degenerating into fixed open-loop compression and again falling into the reverse of the RIB proposition. A related line is usable information under computational constraints (V-information, Xu et al., 2020; see also Ethayarajh et al., 2022): it measures, via a predictive family V, how much usable information about Y is in X for a bounded reader, and can degenerate to R^2. The two are same-rooted but different in aim—V-info is a measure (allowing computation to create information and violating the DPI), whereas RIB is an allocation (respecting the DPI and only re-allocating recognizable status)—and they coincide on the linear-readout cell ($share=r/(1+r) \leftrightarrow R^2$); see §2.

C.6 OT viewpoint: readability = normalized displacement. Under Gaussian independence with zero means, $W_2(\mu_Z, \mu_\{\alpha X\}) = \sqrt{(\alpha^2\sigma_X^2 + \beta^2\sigma_Y^2)} - \alpha\sigma_X$; the normalized displacement $\delta(Y|Z) = W_2/(\alpha\sigma_X) = \sqrt{(1+r)} - 1$ is in one-to-one correspondence with r ($r = (1+\delta)^2 - 1$, and $\delta \approx r/2$ for small r). Depth dilution gives $\delta_\ell \approx O(1)/(2\rho_\ell^2) \rightarrow 0$ (the relative displacement vanishes); HVE gives $\delta \rightarrow \sqrt{(1+a^2\|v\|^2)} - 1 = O(1)$. The link to Theorem 1: Theorem 1 uses the absolute required displacement $\approx \rho - 1$ (increasing with depth), whereas here we use the relative displacement $\approx O(1/\rho^2)$ (decreasing with depth)—"the absolute requirement grows, the actual supply does not grow, and the relative progress tends to zero" are two scales of the same fact; HVE repairs both the absolute supply ($g \propto \rho$) and the relative displacement ($\delta \rightarrow O(1)$).

C.7 Unification of the three viewpoints and correspondence with the main derivation.

Table 6: Unification of the three viewpoints (information theory, information bottleneck, optimal transport) on the readable SNR.

Viewpoint	Key quantity	Depth dilution	HVE repair
Information theory	$r = \beta^2\sigma_Y^2/\alpha^2\sigma_X^2 \rightarrow 0$, $I(Y;Z) \rightarrow 0$		$r \rightarrow a^2\|v\|^2$, $R \geq \frac{1}{2}\log(1+a^2\|v\|^2)$
Information bottleneck	$I(Z;Y) = \frac{1}{2}\log(1+r)$	deep new-information budget $\rightarrow 0$	budget fixed at O(1)
Optimal transport	$\delta = \sqrt{(1+r)} - 1$	$\delta \rightarrow 0$	$\delta \rightarrow O(1)$

Correspondence with the main derivation, item by item: transport ratio $\kappa \rightarrow a_0\|v\| \leftrightarrow$ injection readability upper bound $\sqrt{r(\Delta|.)} = a\|v\| \leq a_0\|v\|$; swamping criterion $\leftrightarrow r(\Delta|.) \leq a_0^2\|v\|^2 < 1$; constant gate $O(1/\rho) \leftrightarrow r = g_0^2/\rho^2 \rightarrow 0$; ideal strength $\propto \rho - 1 \leftrightarrow$ the minimal injection power $\propto \rho^2$ that maintains $\delta = O(1)$.

C.8 Empirically testable predictions (the same as §4.3/the S-series): the per-layer energy ratio r_ℓ should be $\propto 1/\rho_\ell^2$ for the baseline and converge to a constant for HVE; the per-layer mutual-information estimate should be non-zero in deep layers for HVE and decaying for the baseline; the per-layer normalized OT displacement should be $\approx \sqrt{(1+a^2\|v\|^2)} - 1$ for HVE; and the constant-gate ablation should lose readability in deep layers.

C.9 Limitations. Independence/Gaussian assumptions; $I = \frac{1}{2}\log(1+r)$ is heuristic for non-Gaussian cases; cross energy is ignored; the IB quantification is left to experiments; and the analogy to homeostatic plasticity is a motivational remark, not a basis for derivation. The full derivation is in HVE_Appendix_C_readable_information_FULL.md.

D Implementation Details and Hyperparameters

HVE default parameters (used in §4 of the main text).

Table 7: HVE default hyperparameters.

Parameter	Value	Description
K	max(D,128)	harmonic-encoding dimension (number of geometric frequencies)
θ	10^4	geometric-frequency base, $\omega_j=\theta^{-2j/K}$
s_0 (gate bias init)	0.0022	$a=0.9\cdot\mathrm{clamp}(\mathrm{SiLU}(s),0,1)\Rightarrow a\approx 1e-3$
a_0	0.9	gate-magnitude upper bound
N_min / N_max	1 / 2	number of leading/trailing layers skipped
feature scaling $\phi(\rho)$	one of two, by profiling: $\log(\rho+10^{-5})$ (clamped to ± 10) or linear	measure a var_profile_steps window, choose once and freeze
gradient of ρ	stop-gradient (no_grad)	prevents reverse excitation through the scale
NaN/Inf fallback	return the un-modulated input	numerical hygiene
placement	only post_attention_layernorm (before the FFN)	attention norm untouched
RMSNorm gain	frozen	

Mechanism probe (probe_eb_snapshot.py): forward snapshots capture (v, ρ_i, f_ℓ, Δv, keep); --layers_step=4 stores every 4th layer and bf16 saves memory; padding is dynamically disabled when batch_size=1; π is computed on the baseline in a single backward pass as $\hat{\pi}$_A=$\|\nabla$_f L$\|$ and the subspace share $\hat{\pi}$_B=Top-k($\Sigma\nabla\nabla^\top$), then fixed and cross-batched against the measured r; snapshots carry a probe fingerprint (SHA-1 of the example subset, to prevent mixed batching). analyze_eb.py verifies the fingerprint and the keep alignment, then computes E1–E4 and emits CSVs/figures. Detailed alignment conventions are in the script header comments.

Training/evaluation: LoRA single-loop co-training with HVE parameters (linear_add); the scaling is fixed after profiling on a prefix window and then frozen; diagnostic checkpoints are saved every 500 steps. Noise consumption and initialization consistency: do not modify old parameters that do not participate in the forward pass (such as the removed var_gate), so as not to change the random-number consumption/sample order. The main evaluation uses greedy decoding, with use_log_var kept consistent with training (via vaan_std_decision.json); the LoRA-only control uses identical decoding parameters.

Experimental numbers (§4, filled in; 7B evaluated 1-shot, 3B evaluated 2-shot, 7B fine-tuned on a 4096-token context): 7B Qwen2.5-base: HVE 40.4% > HVE_no_cos 40.0% > Base 39.0% > LoRA 34.0% (MATH-500 Acc%); 3B Qwen2.5-instruct: LoRA+HVE 33/83 (39.76%) > LoRA+HVE_no_cos 31/83 (37.35%) > LoRA 26/83 (31.33%) (aimo-validation-amc, 83 problems in total).

E Algorithm 1 — HVE Forward Pass (Wrapping the FFN Pre-Normalization Layer)

Notation and inputs. u: the merged stream after attention, (B,T,D), i.e. u_ℓ=h_ℓ+Attn(RMSNorm(h_ℓ)); w: the frozen RMSNorm gain; s,W_1,W_2: learnable parameters (gate bias, the two layers of the modulation MLP, all bias-free). ρ is under no_grad (stop-gradient). Returns the HVE-modulated normalized FFN input.

```
Algorithm 1:  HVE forward pass (one wrapped FFN pre-norm
layer)
Input:  merged stream after attention u (B,T,D); frozen
RMSNorm gain w; learnable s,W1,W2
```

```
1: v ← RMSNorm_w(u) ▶ FFN normalized input (only the input
layer is normalized)
2: ρ ← RMS_d(u), no_grad ▶ per-token residual-stream scale
(un-normalized merged stream)
3: feat ← φ(ρ) ▶ log or linear features, by profiling
4: tri ← [sin(feat·ω); cos(feat·ω)] ▶ bounded harmonic
encoding, ‖tri‖∞≤1, not multiplied by ρ
5: E ← W₂·SiLU(W₁·tri) ▶ modulation MLP (bias-free)
6: ê ← RMSNorm(E) + mean_d(v) ▶ bounded, and mean-aligned
to the FFN input v
7: a ← 0.9·clamp(SiLU(s), 0, 1) ▶ bounded gate, initially
a≈1e−3
8: return a⊙ê + (1−a)⊙v ▶ convex combination; at
initialization HVE≈identity ($\hat{v}$≈v)
```

Line-by-line notes. The ρ taken in line 2 is the RMS of the un-normalized merged stream (the key observation in §4); it captures both the variance and the DC bias of the flow's growth; the encoding in line 4 is never directly amplified by ρ (in contrast to the unbounded case of feeding raw ρ directly); the two normalization steps in line 6 respectively guarantee boundedness (`RMSNorm(E)`) and scale alignment (`+mean_d(v)`); the gate in line 7 is non-negative, upper-bounded by 0.9, can be switched off but never flips sign; line 8 returns `v` when `a`$\approx$`0`, exactly preserving the pretrained forward pass (identity-safe). The gradient of ρ is blocked by `no_grad`, preventing reverse excitation through the scale feedback.

Draft; anonymous review. Please send comments to the anonymous ICLR-2027 chairs.

F RIB Details: Failure Mode and Spectral View

(Moved from §3.2–3.4 of earlier drafts; the main text keeps only the summary of §3.2.)

F.1 Failure mode: importance-blindness. Without HVE, σ_Y^2=`O(1)` and `r`=`O(1)`$/\rho^2$. Partitioning by ρ: large ρ (deep/global growth) $\rightarrow$ `r`$\rightarrow$`0`, important information exists but is unrecognizable (precisely the dilution-with-depth of the passive view; Sun et al., 2025); small ρ (local collapse) $\rightarrow$ `r` is instead large, and independent of π—unimportant outputs receive a spurious large share and clog the bandwidth of important information. Together, the two regimes constitute the "bandwidth ruled by ρ, blind to importance" of §1. In our formulation, we narrow it as follows: the failure is scale-ruled, importance-blind—"irrelevant outputs dominating" is merely its instance on small-ρ tokens, stated as a corollary rather than as the central claim.

F.2 A Fourier/spectral view: the harmonic readout is "band-limited," and bandwidth allocation is spectral energy allocation. The essence of the readable share is a spectral energy ratio (Parseval). Abstracting downstream reading of Z as a linear readout along an orthonormal direction w ($\|w\|$=1), averaged uniformly over the unit sphere, the cross term vanishes under A4 ($X \perp Y$), and the readable share is exactly a spectral energy ratio:

$$share(Y|Z) := E_w|w \cdot Y|^2 / E_w|w \cdot Z|^2 = \sigma_Y^2/(\sigma_X^2 + \sigma_Y^2) = r/(1+r), r := \sigma_Y^2/\sigma_X^2.$$

P1 (zero-sum shares) and P3 (scale invariance) of §3.1 are thus rigorous by the same identity: `share(X|Z)+share(Y|Z)=1` is a corollary of the spectral decomposition, not a slogan. Matched allocation is then just "distributing output energy over the spectrum by importance": substituting σ_Y^2=`c`$\cdot\pi\cdot\rho^2$ gives `share(Y|Z)=c`π`/(1+c`π`)`$\approx$`c`π, so high-π tokens obtain a depth-independent share of about `c`π (RIB-dominance), while low-π tokens occupy only a `c`$\cdot$`(low` π`)` share (RIB-non-interference); energy is conserved on the orthogonal spectrum, so adding energy to important components lowers the noise share in proportion.

Boundary: the spectrum resolves only variance magnitude, not content importance—it is a neutral share-realizer, since equal variance may be signal or noise and the spectral function cannot see π. If π is not wired into "how much energy to set," spectral allocation degenerates into rule by scale ρ alone, relapsing into the importance-blindness of F.1. Hence the layering: the spectrum realizes "energy $\rightarrow$ share" (physical layer, not alterable); π determines "who receives high energy" (semantic layer, external proxy); and HVE closes the two in a loop, so that the share is realized by $\rho \times \pi$ rather

than by bare ρ. Finer still, it resolves only the energy ratio between sources, not which subspaces within a Y matter—that needs the subspace proxy of π (§4.3).

Harmonic encoding reads the residual scale "band-limited": $\mathtt{tri}(\rho)=[\sin(\rho\cdot\omega_j);\cos(\rho\cdot\omega_j)]$ (Rahimi & Recht, 2007; Tancik et al., 2020) lifts ρ into a bounded multi-scale representation, so the modulation network senses the residual stream in a band-limited way rather than feeding the unbounded raw value. Three properties matter: (i) Parseval boundedness $\|\mathtt{tri}(\rho)\|_2{}^2 \equiv \mathtt{K}$ for any ρ, so any linear readout $|\mathtt{w}\cdot\mathtt{tri}(\rho)| \le \|\mathtt{w}\|\sqrt{\mathtt{K}}$ is independent of ρ—the injected energy is caged by finitely many K Fourier coefficients, and a large ρ at depth cannot amplify it (the spectral origin of $\mathtt{a}_0\|\mathtt{v}\|<1$, Appendix A.7); (ii) Jackson uniform approximation (Theorem 4(i)) gives $\mathtt{O}(\mathtt{K}\hat{}\{-\mathtt{r}\})$ error for a smooth ideal gate; (iii) geometric frequencies $\omega_\mathtt{j}=\theta\hat{}\{-2\mathtt{j}/\mathtt{K}\}$ arrange the basis from high to low on a logarithmic scale, jointly resolving small scales at sparse ρ (shallow layers/collapsed tokens) and large scales at dense ρ (deep layers)—the spectral intuition for why multi-scale harmonic encoding beats a single-layer network (the HVE_no_cos ablation; Theorems 4–5). The orthonormal-basis boundary is declared in §5.