GRADIENT INVASION AND ALIGNMENT TAX IN UNIFIED AUTOREGRESSIVE MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified autoregressive multimodal models train a single transformer and a single softmax over a mixed vocabulary of text subwords and visual codes. Prior theory of modality competition was developed for late fusion networks with separate encoders, and it does not determine the share when the bottleneck is one shared partition function. In this paper, we establish an exact last layer decomposition of the squared hidden gradient into a diagonal term and a cross term. On trained models the cross term is negligible, and the diagonal term therefore carries the share. That term is a Rayleigh quotient and yields a gradient share law in the visual token fraction, the residual energies, and an effective embedding scale. The operator norm provides a scale bound on this Rayleigh factor, and that bound does not replace the factor. Mixing a second discrete code stream into the same softmax raises text cross entropy, which we identify as a foreign code tax. A second text stream at the same mix ratio leaves text cross entropy essentially unchanged, while real visual codes and unstructured foreign codes raise it. A paired class caption does not raise it, so the excess is occupancy of the partition function by a foreign code. End layer isolation and inverse entropy reweighting act on different coordinates of an invasion plane, and the two interventions therefore cannot substitute for one another. The same measurements on public unified checkpoints show that gradient share is present for a joint softmax, a dual head, and a continuous visual head, while the scale channel exists only for a discrete shared softmax.

1 INTRODUCTION

A unified autoregressive model treats an image as a foreign language. Vector quantisation maps pixels to a codebook, and text subwords then share one transformer and one softmax with the visual codes (Van Den Oord et al., 2017, Team, 2024). Emu3 shows that next token prediction alone can carry text, images, and video (Wang et al., 2026). Understanding and generation place a further demand on the same template. Janus keeps one transformer but splits the visual pathway, because the two tasks require different granularity (Wu et al., 2025). JanusFlow keeps that split and generates with rectified flow, while Show-o keeps a discrete visual code together with next token text (Ma et al., 2025, Xie et al., 2025). In each case the training signals still meet, in a shared trunk and, for discrete models, in one partition function. Mixed next token training on that shared softmax is the question of this paper.

The measurements describe that training as one tension. Chameleon traces unstable logits to mismatched modality entropies and introduces z-loss (Team, 2024). Holding data and compute fixed, Transfusion finds discrete image tokens slower on both images and text than a continuous visual channel (Zhou et al., 2025). Joint diffusion of image and text likelihoods avoids that vocabulary altogether (Li et al., 2025). Uni-X finds the gradient conflict at both ends of a shared transformer, where the modality statistics still differ, and proposes to separate those ends while sharing the middle (Hao et al., 2026). After pretraining, Janus-Pro still shows nearly orthogonal task gradients, and the authors attribute the magnitude gap to long visual sequences (Rao & Rachuri, 2026). Continual tuning then turns this conflict into forgetting across modalities, which updates confined to one modality reduce (Wei et al., 2026). What remains missing is a softmax quantity that predicts these effects, and a rule for when end isolation outperforms a training time reweight.

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.

Late fusion theory addresses a different object, and later methods inherit it. Extra modalities can lower population risk when they improve the latent estimate (Huang et al., 2021), yet joint training can fail to realise that gain, because on a late fusion network the easier modality captures the encoder and starves the other (Huang et al., 2022). Mismatched overfitting in video classification led to gradient blending (Wang et al., 2020), and encoder gradient modulation pursues the same balance on a discriminative head (Peng et al., 2022). See-Saw shows that a magnitude gap or a misaligned direction blocks balanced convergence (Kwon et al., 2025), while gradient surgery removes the conflicting part of a generic task gradient (Yu et al., 2020). Those arguments assume separate encoders or separate task losses. Autoregressive competition sits inside one partition function, with token count at first order, so a late fusion bound does not determine the share.

We analyse next token training at two resolutions. At the last layer the squared hidden gradient splits into a diagonal term and a cross term. On trained models the cross term is negligible, so the diagonal term carries the share. It is a Rayleigh quotient in the visual token fraction, the residual energies, and an effective embedding scale, and the operator norm bounds that scale without replacing it. A second discrete code in the same softmax raises text cross entropy. At a fixed mix ratio, a second text stream leaves text cross entropy essentially unchanged, while real visual codes and unstructured foreign codes raise it, and a paired class caption lowers it relative to a shuffled pairing. The excess is occupancy of the partition function by a foreign code. Magnitude and cosine are separate coordinates of one invasion plane. Isolation moves the cosine through Jacobian overlap, and inverse entropy reweighting moves the magnitude through the share, so one coordinate cannot tell the two maps apart.

The alignment tax on language tasks is the regression after reinforcement learning from human feedback, and mixing pretraining gradients into the update limits it (Ouyang et al., 2022, Lin et al., 2024). The name is also used when backbones tuned on instructions saturate faster on multimodal data than base models (Li et al., 2026). The feedback tax has a geometric form inside a KL ball, with the rate set by principal angles (Young, 2026). Mixing visual tokens perturbs the pretraining mixture instead, and the geometry reverses. Orthogonal unimodal optima increase the mixture displacement, whereas an orthogonal direction is first order free in the KL ball. Token count, residual energy, and embedding scale trade off in one product, so varying only the mix ratio cannot separate them, and a reported magnitude ratio cannot be inverted for a unique mix. Treating isolation and an inverse entropy weight as two estimates of one scalar is the error that the invasion plane is designed to prevent.

To close these gaps we separate three head geometries, state a last layer identity that holds independently of lazy training, and identify the mix tax with controls that a sweep of the mix ratio cannot supply. The contributions are as follows.

- An exact last layer split of $\|\nabla_h \ell\|^2$ into a diagonal term and a cross term, together with a Rayleigh share law whose operator norm counterpart is a scale bound. Rescaling visual rows at a fixed mix identifies the embedding channel.
- A foreign code identification of the mix alignment tax. A second text stream at the same mix ratio leaves text cross entropy unchanged, while real visual codes and unstructured codes both raise it. The tax is occupancy of a shared partition function by a high entropy outer code.
- An invasion plane on which isolation and inverse entropy reweighting are orthogonal maps, together with parameter matched isolation on Gutenberg text and Janus codes and share measurements on public joint softmax, dual head, and continuous checkpoints.

2 Related work

Late fusion competition concerns a different object. Extra modalities can lower population risk when they improve the latent estimate (Huang et al., 2021). The matching negative result is that joint gradient descent need not realise this gain. On a two layer smoothed ReLU fusion network, the easier modality captures the encoder and starves the other (Huang et al., 2022). Mismatched overfitting rates make joint video classification hard, and gradient blending reweights the modalities to compensate (Wang et al., 2020). Modulating encoder gradients during training pursues that balance on a discriminative head (Peng et al., 2022). See-Saw then shows why the balance can fail. A magnitude gap or a misaligned direction blocks balanced convergence of a vision and language classifier and of fused multitask training (Kwon et al., 2025, Lou et al., 2026). Gradient surgery treats

the directional case for generic tasks by removing the conflicting part of each task gradient, and it does so without a modality entropy (Yu et al., 2020). These arguments assume separate encoders or separate task losses. They do not assume a shared softmax over a mixed vocabulary.

Unified head geometries are separated by where the visual loss sits. Chameleon places text subwords and visual codes in one transformer and one softmax, and it traces unstable logits to mismatched modality entropies (Team, 2024). The first stage uses 2.9T text tokens against 1.9T mixed tokens, so the mix fraction is already near 0.40, the regime in which a small mixed trainer pays a visible text excess. Emu3 carries images and video by the same next token rule (Wang et al., 2026). Show-o keeps the joint transformer and adds discrete diffusion on visual codes, so an official codebook remains inside a joint softmax slot and a probe of those codes measures that slot (Xie et al., 2025, Yu et al., 2024). With data and compute held fixed, Transfusion finds discrete image tokens slower on both images and text once images move to a continuous channel (Zhou et al., 2025). Janus keeps one trunk and splits the visual encoders, JanusFlow keeps that split and generates with rectified flow, and Janus-Pro places the two tasks on a dual head (Wu et al., 2025, Ma et al., 2025, Chen et al., 2025). The scale κ_{eff} exists only inside the joint softmax.

Conflict plots and the alignment tax record a cost that these geometries do not yet explain. Uni-X finds the vision and text conflict strong at both ends and weaker in the middle, takes the mixed baseline as roughly constant in depth, and separates those ends while sharing the trunk (Hao et al., 2026). When the baseline does not change with depth, every shape in $c_g = -(S_{\text{inter}} - S_{\text{base}})$ is a shape in $-S_{\text{inter}}$ and therefore a shape in Jacobian overlap, so isolation forces zero overlap at the ends. Not every released checkpoint need show one interpolation of that overlap. On Janus-Pro the conflict appears as drowning, because near orthogonal task gradients let a large visual magnitude shrink the text step when the optimiser shares a second moment (Rao & Rachuri, 2026). Continual tuning turns the same conflict into forgetting across modalities, and updates confined to one modality reduce it without separating token count, residual energy, and embedding scale (Wei et al., 2026). The alignment tax after human feedback is a regression on language tasks, limited by mixing pretraining gradients into the update (Ouyang et al., 2022, Lin et al., 2024), and a geometric account places it inside a KL ball (Young, 2026). Orthogonal unimodal optima increase the displacement from mixing visual tokens into pretraining, and they reduce the first order feedback tax. The identities below state when isolation and an inverse entropy weight bind.

3 Gradient share at a shared softmax

The share is a property of one joint softmax, and the three head geometries have to be separated before any ratio is read. A joint softmax stacks the text and visual rows in one matrix $E = \begin{bmatrix} E_t \\ E_v \end{bmatrix}$. That is the setting of Chameleon style models and of the transformers trained below. A dual head keeps a language matrix E^{lm} and a generation matrix E^{gen}, as in Janus-Pro. A continuous visual head replaces the visual softmax by a quadratic or projector loss, as in Qwen3-VL and in the Transfusion visual channel (Bai et al., 2025, Zhou et al., 2025). Only the joint softmax has one partition function, and only there does an embedding scale ratio exist.

A run is a triple (θ, E, ρ). The vocabulary is a disjoint union $\mathcal{V} = \mathcal{V}_t \sqcup \mathcal{V}_v$, hidden states $h \in \mathbb{R}^d$ produce logits Eh, and a batch has visual token fraction $\rho \in (0, 1)$. The training objective is

$$L(\theta) = (1 - \rho)L_t(\theta) + \rho L_v(\theta), \qquad L_m = \mathbb{E}[-\log p_\theta(y \mid x) \mid y \in \mathcal{V}_m]. \tag{1}$$

Write $z = Eh$, $p = \text{softmax}(z)$, $\ell = -\log p_y$, and $r = p - e_y$, so that $\nabla_h \ell = E^\top r$. The clean share of modality m is the token weighted second moment $G_m = \rho_m \, \mathbb{E}[\|\nabla_h \ell\|^2 \mid y \in \mathcal{V}_m]$, with $\rho_t = 1 - \rho$ and $\rho_v = \rho$. Residual energy is $R_m = \mathbb{E}[\|r\|^2 \mid y \in \mathcal{V}_m]$. A backward pass that first averages the loss inside a modality is a different object, because that average inserts a factor $1/N_m^2$. Comparing such a mean to G_v/G_t folds token count into the test of the share law.

Assumption 1 (Softmax coupling)**.** *Partition the factors as $E^\top = [E_t^\top \; E_v^\top]$ and $r^\top = [r_t^\top \; r_v^\top]$. The operator scale is $\kappa_{\text{op}} = \|E_v\|_2^2 / \|E_t\|_2^2$, and the Frobenius overlap is $\chi = \|E_t E_v^\top\|_F$.*

Assumption 2 (Lazy quadratic)**.** *In a neighbourhood of initialisation, $L_m(\theta) = \frac{1}{2}(\theta - \theta_m^\star)^\top \Sigma_m (\theta - \theta_m^\star) + H_m$ with $\Sigma_m \succeq \lambda_{\min} I$ (Jacot et al., 2018, Chizat et al., 2019). Weighting and the interpolant use this hypothesis. The last layer identities do not.*

Theorem 1 (Residual energy)**.** *On a calibrated categorical, $R = \mathbb{E}_{y\sim p}\|p - e_y\|^2 = 1 - \|p\|_2^2$. The map $p \mapsto 1-\|p\|_2^2$ is strictly Schur concave on the simplex, and it increases along every temperature family $p_T = \mathrm{softmax}(z_0/T)$ as T increases. On that family, R_m is therefore strictly increasing in the entropy H_m.*

Residual energy is the second moment that the softmax residual actually carries. Entropy may stand in for R_m on a temperature family, because heating the same logits spreads p and raises both quantities together. A trained pair of codebooks is not such a family. The substitution $R = H$ is a construction on that family, and it is not stated as a general identity.

Theorem 2 (Last layer share)**.** *Under Assumption 1, the squared hidden gradient splits as*

$$\|\nabla_h \ell\|^2 = \underbrace{\|E_t^\top r_t\|^2 + \|E_v^\top r_v\|^2}_{\text{diag}} + \underbrace{2 r_t^\top (E_t E_v^\top) r_v}_{\text{cross}}. \tag{2}$$

The ratio of modality conditional expectations is then the identity

$$\frac{\mathbb{E}[\|\nabla_h \ell\|^2 \mid v]}{\mathbb{E}[\|\nabla_h \ell\|^2 \mid t]} = \frac{\mathrm{diag}_v + \mathrm{cross}_v}{\mathrm{diag}_t + \mathrm{cross}_t}. \tag{3}$$

If $E_t E_v^\top = 0$, the cross term drops out and the clean share factors as

$$\frac{G_v}{G_t} = \frac{\rho}{1-\rho} \cdot \frac{R_v}{R_t} \cdot \frac{\lambda_{\mathrm{eff}}(E_v, r_v)}{\lambda_{\mathrm{eff}}(E_t, r_t)}, \tag{4}$$

where $\lambda_{\mathrm{eff}}(E_m, r_m) = \mathbb{E}[r_m^\top (E_m E_m^\top) r_m] / \mathbb{E}[\|r_m\|^2]$ is the residual Rayleigh quotient. Let κ_{eff} denote the ratio of those two factors. Each factor is at most the squared operator norm of its own block, so κ_{op} bounds the scale and need not equal κ_{eff}. The gap $\kappa_{\mathrm{op}}/\kappa_{\mathrm{eff}}$ is large when residuals miss the top singular vectors.

Equation 4 is the share once the cross term has been removed, and a product $E_t E_v^\top = 0$ removes that term for every residual. The conditional expectation in equation 3 can vanish under a weaker condition, which uses only the joint second moment of the two residual blocks.

Lemma 1 (Cross energy bound)**.** *Let* cross *denote the summand $2 r_t^\top (E_t E_v^\top) r_v$ in equation 2. Then $|\mathrm{cross}| \le 2\chi \|r_t\|_2 \|r_v\|_2$. If $\mathbb{E}[r_t \otimes r_v] = 0$, then $\mathbb{E}[\mathrm{cross}] = 0$.*

A large overlap χ can therefore leave the expected cross term small, because the bound controls a single residual and the expectation can still cancel. On the trained 36.5M model the measured cross fraction stays below three tenths of a percent, as reported in Section 5. A public joint softmax checkpoint has a large χ and a visual cross term at machine zero. Cross energy and the gap between κ_{op} and κ_{eff} are different objects.

Equation 3 is an identity on any joint softmax whose expectations exist. Late fusion competition allocates hidden units to the easier modality (Huang et al., 2022). Equation 4 allocates a different scarce quantity, namely softmax mass in one partition function. The low entropy modality has small R_m, hence a small raw gradient and a large signal to noise ratio. Raw visual gradients are large when visual tokens are hard, so the gradient that dominates the step is the opposite of the modality that is easy to fit. A large visual gradient can therefore sit next to a visual cross entropy that barely moves.

The operator norm remains a useful diagnostic. It is cheap to compute, it tracks a rescaling of the visual rows exactly, and it is the quantity that z-loss and the embedding initialisation actually move. It is not the factor that multiplies R_v/R_t on the residual of a trained network. Residuals of a mixed trainer concentrate on a subset of the codebook, often far from the top singular pair of $E_m E_m^\top$, and the ratio $\kappa_{\mathrm{op}}/\kappa_{\mathrm{eff}}$ records that concentration. On the 36.5M suite the ratio is about ten. On a public joint softmax checkpoint it is about twenty four and points the other way, because the visual rows are many and individually small while the two residual energies stay comparable. Theorem 2 predicts both observations once the Rayleigh factor is distinguished from the operator scale. Neither observation is a failure of equation 3.

Proposition 1 (Identifiability of the share)**.** *The right hand side of equation 4 is constant on the surface $\rho\, R_v\, \kappa_{\mathrm{eff}} = c(1-\rho)R_t$. A protocol that varies only ρ traces one curve on that surface, and it cannot separate token count from residual energy or from embedding scale. A two point design that also varies κ_{eff}, for instance by rescaling visual rows by s at fixed ρ, is the smallest experiment that identifies the product. Rescaling by s multiplies $\lambda_{\mathrm{eff}}(E_v, r_v)$ by s^2.*

A sweep in ρ is therefore not, by itself, a test of gradient invasion. Visual rows scaled below one place excess softmax mass on visual ids even when the target is text. In the calibrated last layer construction, leak onto $\mathcal{V}_v$ rises from 0.07 at unit scale to 0.52 at visual scale $1/2$. That is the logit drift reported for Chameleon, and z-loss is a regulariser for this small scale regime of equation 4. Scaling the visual rows above one suppresses that leak on text targets and, through the same identity, inflates G_v on visual targets.

Proposition 2 (Discrete versus continuous visual head)**.** *If the visual loss is a quadratic head on a continuous feature, the joint Fisher $E^\top(\mathrm{diag}(p)-pp^\top)E$ is replaced by a unimodal Jacobian $J_v^\top J_v$. The factor κ_{eff} in equation 4 is then absent, and the remaining text excess is a cost of sharing the backbone, with no invasion of a partition function. Any shared discrete softmax brings equation 4 back in full. A dual head replaces that display by a product of two unimodal Rayleigh factors, and it has no cross term $E_t E_v^\top$.*

This is the architectural split in Transfusion, read on the Fisher (Zhou et al., 2025). A continuous visual channel can still share a transformer, and it cannot route residual energy through a joint partition function. Janus-Pro sits on the dual head side of the same split (Chen et al., 2025). The joint softmax identity applies to neither object.

The three geometries are three readings of one residual stream. A shared trunk still couples the modalities through attention when the last layer is split or continuous, so sharing the backbone remains. What disappears is the partition function channel. A continuous head can therefore show a large parameter gradient ratio, as it does on Qwen3-VL, while the embedding scale $\kappa(E)$ is missing. The ratio is a real gap in magnitude. The scale factor is a separate symbol, and it is present only for a discrete shared softmax. A dual head can likewise show nearly orthogonal output gradients, because the two matrices do not share one residual r, even when the trunk is fully shared. Orthogonality at the head is then a fact about two output maps. It is not the statement $E_t E_v^\top = 0$ inside one softmax.

These readings use only a finite softmax. Lazy training, a residual architecture, and a depth profile are not among the hypotheses, so the last layer split and the closed ratio do not change when those modelling choices change. Keeping the cross term leaves equation 3 exact. On the joint softmax checkpoints measured below the term is negligible, and equation 4 is then the working share law, with κ_{op} a bound on the Rayleigh factor. A temperature family allows entropy in place of R. A trained codebook pair is not such a family. Theorem 2 applies to a joint softmax, and a magnitude ratio on a dual head or a continuous head is a measurement on a different map. Feature learning at finite width in the trunk lies outside these identities. The 36.5M and 63M suites check that the symbols can still be read on a trained model.

4 Mix tax and the invasion plane

The share fixes the size of the visual gradient. The tax is the change that this mixture induces in the text loss, and the invasion plane separates the interventions that move that change. Under isotropic curvatures $\Sigma_t = \Sigma_v = I$, the population station of equation 1 is the convex combination $\theta_\rho = (1-\rho)\theta_t^\star + \rho\theta_v^\star$. Along the segment $\theta(\alpha) = (1-\alpha)\theta_t^\star + \alpha\theta_v^\star$, Assumption 2 with $\Sigma_t = I$ makes the text excess quadratic in the path weight,

$$L_t(\theta(\alpha)) - L_t(\theta_t^\star) = \frac{\alpha^2}{2}\|\theta_v^\star - \theta_t^\star\|^2. \tag{5}$$

Mixed training is free to leave the segment, so the display is an envelope for the tax and not a description of the training path. The identifying controls below do not assume laziness.

Definition 1 (Mix alignment tax)**.** *The mix alignment tax of a trained parameter θ is $\Delta L_t = L_t(\theta) - L_t(\theta_t^{\mathrm{ref}})$, where θ_t^{ref} is a text only run with the same backbone, tokeniser, and evaluation pack. The name marks a data mixture counterpart of the alignment tax after human feedback. The excess measured here is not a regression on captions or on instructions.*

Any second stream that occupies a fraction ρ of the tokens reduces the text count. The comparison that identifies the tax holds ρ fixed and changes only the identity of that second stream.

Proposition 3 (Foreign code occupancy)**.** *Consider three second streams at the same ρ, the same text stream, and the same evaluation pack. (i) A second text stream whose ids remain in $\mathcal{V}_t$. (ii) A*

structured visual code stream in $\mathcal{V}_v$. (iii) An unstructured stream drawn uniformly on $\mathcal{V}_v$. If ΔL_t is large for (ii) and (iii) and negligible for (i), the excess is occupancy of a split vocabulary by a foreign code. A paired class caption that lowers L_t relative to a shuffled pairing shows, in addition, that the excess is not produced by the mere presence of an image.

The proposition is an identification design, and it does not supply a closed form for ΔL_t. On the models trained below, a lazy segment tracks the quadratic in equation 5, while mixed stochastic gradient descent finds a path whose excess is strictly smaller. Using $\rho^2(L_t(\theta_v^\star) - L_t(\theta_t^\star))$ as a point predictor of mixed training would overstate the tax. Read as an envelope, the same quantity agrees with the measurements in Section 5.

Proposition 4 (Weighting)**.** *Under Assumption 2 with equal isotropic Hessians, (i) $\max(\Delta L_t, \Delta L_v)$ is minimised at length balance $w_t(1-\rho) = w_v\rho$, (ii) among unbiased linear combinations of per token gradients, $w_m \propto 1/R_m$ minimises the asymptotic covariance, and (iii) equalising continuous time rates at equal curvatures gives $w_m \propto 1/\rho_m$. Inverse entropy weights implement (ii) on top of the batch fraction ρ. They do not implement (i). Treating (ii) as a substitute for (i) starves the high entropy modality at the population station.*

Uni-X records the conflict at both ends of a shared transformer by the cosines

$$S_{\text{inter}} = \cos(g_t^{\text{uni}}, g_v^{\text{uni}}), \qquad S_{\text{base}} = \cos(g_A^{\text{mix}}, g_B^{\text{mix}}), \qquad c_g = -(S_{\text{inter}} - S_{\text{base}}), \tag{6}$$

with unimodal batches for S_{inter} and two independent mixed batches for S_{base} (Hao et al., 2026). Isolation copies the end layers, restricts their attention to keys of one modality, and shares the trunk. On the copied parameters the other modality is absent from the Jacobian, so the construction forces $S_{\text{inter}} = 0$ there.

Theorem 3 (Invasion plane)**.** *To first order, mixed training diagnostics factor through the pair $(\Delta L_t,\ S_{\text{inter}}(\ell))$. Isolation is the map $S_{\text{inter}}(\text{end}) \mapsto 0$ at essentially fixed trunk share equation 4. Inverse entropy and length balance are maps $G_v/G_t \mapsto (w_v/w_t)\, G_v/G_t$ at fixed S_{inter}. A method that reports only a cosine, or only a magnitude ratio, is a coordinate projection of this plane, and it cannot distinguish the two maps. Stacking adds the two maps at first order. A fair isolation comparison has to match parameter count, because extra width at the ends is a third intervention.*

Uni-X is the isolation map, and inverse entropy weights are the reweighting map. Neither replaces the other. A diagnostic that stores only a cosine misses reweighting, and a diagnostic that stores only a gradient norm misses isolation. The zero at an isolated end is structural. If $A_{\text{end}}^{(m)}$ denotes the end layer Jacobian after it has been restricted to modality m, the backward maps factor through a pair with $A_{\text{end}}^{(t)} A_{\text{end}}^{(v)\top} = 0$, and the inner product of the two end gradients vanishes for every trunk state. Layerwise conflict on a shared stream can still be large at the ends and mixed in the middle when the embeddings are block diagonal. Public checkpoints do not share one depth shape, so a U profile is an architectural possibility and not a law. Isolation has to drive $S_{\text{inter}}(\text{end})$ to zero. It does not have to produce a prescribed $\sin^2$ curve.

Under the isotropic quadratic, write $\Delta = \|\theta_v^\star - \theta_t^\star\|^2$ and $\beta = w_v\rho/(w_t(1-\rho) + w_v\rho)$. The two excesses of a lazy interpolant then satisfy

$$\sqrt{2\Delta L_t} = \beta\sqrt{\Delta}, \qquad \sqrt{2\Delta L_v} = (1-\beta)\sqrt{\Delta}, \tag{7}$$

and the pair lies on a quarter circle of radius $\|\theta_v^\star - \theta_t^\star\|$. Length balance is the midpoint $\beta = 1/2$. Inverse entropy selects a different point on the same circle, so the two weights are two points on one front and not two estimates of one scalar. Caption alignment, in which the two optima nearly coincide, sits near the point where the tax vanishes. Generation from a high entropy codebook sits at the opposite end. Understanding and generation on the same codes trace the same circle with the role of ρ exchanged. Pure generation maximises the understanding excess, and any interior mix that uses length balance recovers the population average of the two unimodal optima.

The preceding identities give a practical rule. Estimate residual energy, or a tokenizer proxy, on each stream, and place inverse entropy weights on top of the batch fraction ρ. Multiply by a length balance when visual tokens dominate a sequence, and keep embedding row scales within a small factor of one, or else use z-loss. When end layer S_{inter} is the bottleneck, isolate the ends at matched parameter count. When G_v/G_t is already large at a small S_{inter}, reweight. When a continuous

visual channel is available and text is to be preserved, dropping the discrete softmax removes κ_{eff}. When captions have to survive generation, unpaired codes from a high entropy codebook are the wrong mixture.

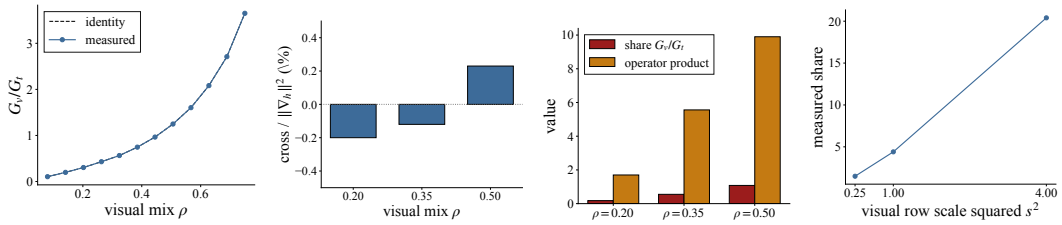
5 EXPERIMENTS

5.1 LAST LAYER SHARE

The identities above are exact for a joint softmax, and the trained model is where we test whether the cross term can be dropped. We use a 36.5M causal transformer on interleaved Gutenberg text and Janus codes, with 10 layers, width 384, 6 heads, and sequences of length 256. Text uses a byte pair encoding of 8000 ids, and vision uses 16384 Janus codes drawn from 598 images. Every cross entropy is scored on a held out pack of one modality. On 119 held out images the visual cross entropy stays near 5.7 nats, so the visual stream is not memorised. Text and visual unigram entropies are 6.63 and 6.27 nats. Their ratio is near one, and it cannot account for the measured share.

Theorem 2 becomes exact once the unembedding is block diagonal. Figure 1a shows a calibrated head with equal vocabularies of size 32, and the median relative error of the residual identity is 1.4×10^{-16} across twelve mixes. Residual energy rises with label entropy, from $R = 0.22$ at $H = 0.66$ nats to $R = 0.97$ at $H = 3.43$ nats, so R is the second moment that a temperature family actually moves. The trained 36.5M model is not block diagonal. Even so, the closed ratio in equation 3 matches the measured share at $\rho \in \{0.20, 0.35, 0.50\}$, with geometric error 1.000. As shown in Figure 1b, the cross fraction is -0.2%, -0.1%, and $+0.2\%$. The working law is therefore the diagonal Rayleigh quotient.

The operator norm only bounds that quotient. Figure 1c shows that a product built from κ_{op} overshoots the clean share by a factor of 9.2 to 10.1, and restoring the cross term in equation 2 does not remove the gap. At $\rho = 0.35$, $\kappa_{\text{eff}} \approx 1.18$ while $\kappa_{\text{op}} = 11.86$, because residuals sit away from the top singular vectors. A sweep in ρ cannot separate embedding scale from token count, so Figure 1d holds the mix fixed and multiplies visual rows by $s \in \{0.5, 1, 2\}$. The share moves from 1.49 through 4.40 to 20.4, in line with s^2. Collapsing the codebook to 256 or 32 ids barely moves R_v, so $R = H$ is not identified. The tax at the same three scales is 0.539, 0.331, and 0.417. Share rises with s, and the largest tax is at $s = 0.5$, the leak regime that Chameleon z-loss is built to resist.



(a). Last layer identity. (b). Cross fraction. (c). Operator bound. (d). Scale channel.

Figure 1. Share law. (a) On a calibrated block diagonal unembedding the residual identity of Theorem 2 holds to machine precision. (b) On a trained 36.5M model the cross term in equation 2 stays below three tenths of a percent. (c) The operator product overestimates the clean share by a factor of about ten. (d) At fixed $\rho = 0.35$, rescaling visual rows by $s \in \{0.5, 1, 2\}$ moves share as s^2.

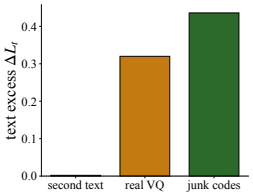
5.2 FOREIGN CODE OCCUPANCY

The share fixes the size of the visual gradient. The tax is the change that gradient induces in the text loss. Proposition 3 identifies the excess as occupancy of the shared partition function. If a smaller text count were the whole mechanism, a second text stream at the same ρ would pay the same excess. Figure 2 holds $\rho = 0.35$ on the 36.5M backbone. The second text stream stays inside the text vocabulary and leaves $\Delta L_t = 0.002$ across two seeds. Real Janus codes raise the excess to 0.320. Unstructured ids on $\mathcal{V}_v$ raise it to 0.436, and they move visual cross entropy from 5.69 to 10.04. The same fraction of the batch is harmless as text and costly as a foreign code.

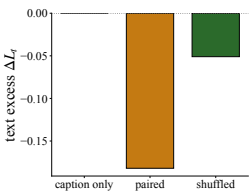
Caption pairing tests whether the presence of an image is sufficient for that tax. CIFAR-100 is packed at the same mix, either with the class name or with a shuffled name, and every run is scored on one caption only pack. Caption only training reaches $L_t = 2.649$. Paired names reach 2.467.

Shuffled names reach 2.597. As shown in Figure 2, ordinary caption alignment lowers text cross entropy, and a shuffled name returns most of that gain. The positive tax of Gutenberg against unpaired codes is therefore the foreign code effect. Paired class names move text cross entropy in the other direction.

The same words have been used for a regression after a preference step and for an angle inside a KL ball (Ouyang et al., 2022, Young, 2026). Here the update is the insertion of a foreign block into one partition function, and a paired caption does not insert that block. A lazy interpolant between a text only station and a vision only station is only an envelope of equation 5. Along $\alpha \in \{0.20, 0.35, 0.50, 0.75\}$ the excess is roughly quadratic, with median geometric error 1.42 against $\alpha^2(L_t(\theta_v^\star) - L_0)$. Mixed training at $\rho = 0.35$ sits at 0.331, against an interpolant value of 1.824. Stochastic gradient descent leaves the segment that joins the two unimodal optima, so the quadratic describes the interpolant and is not a training identity.

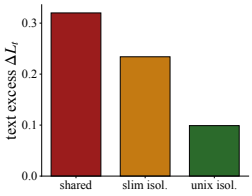


(a). Dilution control.

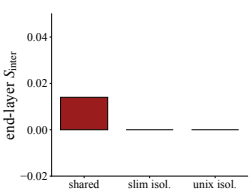


(b). Caption pairing.

Figure 2. Identifying the mix tax at $\rho = 0.35$. Second text is free. Real and junk codes raise it.



(a). Isolation tax.



(b). End cosine.

Figure 3. Parameter matched isolation. Slim isolation drops tax from 0.320 to 0.234 and zeros end S_{inter}.

5.3 Invasion plane

Theorem 3 places isolation and reweighting on different coordinates, and Figure 3 together with Table 1 tests the first of those maps at matched size. On Gutenberg and Janus codes, slim isolation keeps the parameter count of the shared net, to within 3072 floats, and drops the tax from 0.320 to 0.234, a factor of 1.37. End layer S_{inter} falls from 0.014 to zero. Full isolation adds width at the ends and reaches 0.099, so part of that further drop is extra capacity. A Markov net with matched width shows the same sign, with shared tax 2.228 against isolated tax 0.246. Table 1 repeats that sign on a 63M model of width 512 and depth 12, where slim isolation lowers the tax from 0.693 to 0.548 before Gutenberg is memorised. Runs with final text cross entropy below 0.1 nats are discarded, because the text has been memorised.

Table 1. Isolation on Gutenberg and Janus codes at $\rho = 0.35$. Tax is text CE minus a text only run with $L_0 = 3.564$ on the 36.5M suite and $L_0 = 4.304$ at 63M mid training. Slim isolation matches the shared parameter count to 0.008% and 0.0065%. End layer S_{inter} is zero on both isolated nets. Two seeds on 36.5M. Entries marked n/a were not measured.

run	params	ΔL_t	end S_{inter}	visual CE
36.5M shared	36.54M	0.320	0.014	5.69
36.5M slim isol.	36.54M	0.234	0.00	5.70
36.5M matched width	38.90M	0.182	≈ 0	5.60
36.5M Uni-X isol.	40.08M	0.099	0.00	5.72
63M shared, mid	62.94M	0.693	n/a	5.66
63M slim isol., mid	62.94M	0.548	n/a	5.79

Figure 4 shows why that drop is a Jacobian split at the ends. On the shared net the overlap is small but nonzero at the first layer and again near the last, while the middle stays mixed. Slim isolation copies one layer at each end and lets those keys see only their own modality. The ends sit at zero by construction, and the trunk stays of the same order as the shared trunk, so the split does not delete cross modal features from the residual stream. The depth curve is not a $\sin^2$ interpolation of the end value, which is why isolation remains the right map when a released profile is not a U.

The same plane organises the remaining maps. As shown in Figure 5, slim isolation moves vertically toward zero cosine, while extra end width continues toward a smaller tax and is a third map. Inverse entropy is the horizontal map. Figure 6 shows a Markov net at $\rho = 0.40$ whose visual stream is much more entropic than the text, and weights $w_v = H_t/H_v = 0.22$ cut the tax from 2.23 to 1.79. Length balance equalises token counts and leaves the tax essentially unchanged, as Proposition 4 predicts when the imbalance sits in residual energy. Matching the source entropies collapses the tax from

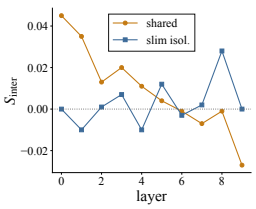
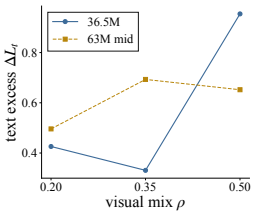
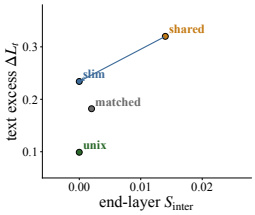


Figure 4. Layerwise. Slim ends are zero. Trunk stays mixed.



(a). Tax versus mix.



(b). Invasion plane.

Figure 5. Tax versus mix and the invasion plane. Isolation moves $(S_{\text{inter}}, \Delta L_t)$ toward zero cosine.

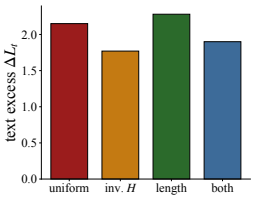
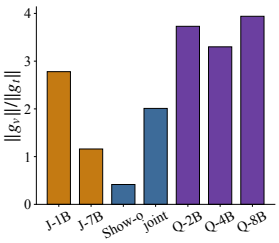


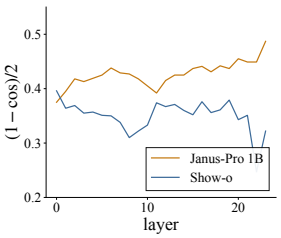
Figure 6. At $\rho = 0.40$. Inverse entropy binds. Length balance does not.

about 2.1 nats to 0.26, and stacking isolation on that weight reaches 0.22, close to isolation alone, because end overlap is the bottleneck. A score that stores only c_g, or only $\|g_v\|/\|g_t\|$, therefore sees only one of the two maps.

Public checkpoints carry the same share at a much larger width. Figure 7a reports parameter gradient ratios of order one, and token counts do not recover them. Janus-Pro 1B has $N_v/N_t \approx 15$ against a magnitude of 2.78, while the 7B model sits at 1.16. Show-o keeps official codes inside the joint softmax and has $N_v/N_t \approx 6$ against a magnitude of 0.42, with head overlap near 0.02. A public checkpoint in the Chameleon style gives 2.01, with head overlap near -0.05. Qwen3-VL uses a continuous visual head and gives 3.73, 3.30, and 3.94 at 2B, 4B, and 8B, so $\kappa(E)$ is absent, as Proposition 2 predicts. On that joint softmax the closed ratio in equation 3 still holds, with clean share 2.83 against an operator product of 0.12, because the visual rows are many and individually small. As shown in Figure 7b, the depth profile is architectural in the same sense. Janus-Pro 1B rises toward the last layer, Show-o is flatter, and the Chameleon style checkpoint stays near $1/2$ at every depth. None of these profiles is a $\sin^2$ curve.



(a). Public share.



(b). Layer conflict.

Figure 7. Neither conflict profile is a $\sin^2$ U.

Table 2 isolates that partition function on one Markov backbone at $\rho = 0.40$. The streams, the mix, and the text loss stay fixed, and only the visual head changes. A discrete softmax pays a text excess of 1.88. A mean squared error head onto the next visual embedding pays 0.13. The gap is Proposition 2 on a single trunk. Public continuous heads already show that $\kappa(E)$ can be missing at a much larger width, and the controlled swap is the comparison that holds the trunk fixed while the head changes. Chameleon Stage 1 already spends a mixed fraction near 0.40 (Team, 2024). The 36.5M tax is 0.33 at $\rho = 0.35$ and 0.95 at $\rho = 0.50$, so the same object is visible at a width where every symbol can be read. Two seeds agree in sign. Shared tax stays near 0.32, slim isolation near 0.23, and a second text stream near zero, while unstructured codes push held out visual cross entropy from about 5.7 to 10.

Table 2. Text excess at $\rho = 0.40$. Tax is text CE minus $L_0 = 0.989$.

head	text CE	ΔL_t
discrete softmax	2.87	1.88
continuous MSE	1.12	0.13

6 Conclusion

This paper gives an exact account of gradient invasion and of the mix alignment tax in a unified autoregressive model. At the last layer the squared hidden gradient splits into a diagonal term and a cross term, and the working share law is a Rayleigh product in the visual token fraction, the residual energies, and an effective embedding scale, while the operator norm only bounds that scale. At a fixed mix, a second text stream leaves text cross entropy unchanged, while real visual codes and unstructured codes raise it, and a paired class caption lowers it relative to a shuffled pairing. Isolation drives the end layer overlap to zero and, at matched parameter count, reduces the tax by a factor of 1.37. Inverse entropy reweighting moves the gradient share and leaves that overlap in place. Public checkpoints show a gradient share of order one, with the embedding scale present for a joint softmax.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments, and assisting with translation. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, implementing methods, supporting qualitative and thematic data analysis, interpreting results. Besides, generating synthetic data sets, helping develop theoretical models or conceptual frameworks, cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for modifying scientific figures or images, editing a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in neural information processing systems*, volume 32, 2019.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. Uni-x: mitigating modality conflict with a two-end-separated architecture for unified multimodal models. In *International Conference on Learning Representations*, pp. 123737–123758, 2026.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, volume 34, pp. 10944–10956, 2021.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 9226–9259, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, volume 31, 2018.

JuneHyoung Kwon, Mihyeon Kim, Eunju Lee, Juhwan Choi, and YoungBin Kim. See-saw modality balance: See gradient, and sew impaired vision-language balance to mitigate dominant modality bias. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4364–4378, 2025.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2779–2790. IEEE, 2025.

Ziran Li, Qiang Wang, Zhengyu Chen, Shanglin Lei, Borun Chen, Jingang Wang, and Xunliang Cai. What transfers from text to vision? capability scaling laws and transfer dynamics for vlms. *arXiv preprint arXiv:2608.00013*, 2026.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 580–606, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Yiwei Lou, Yuanpeng He, Rongchao Zhang, Yongzhi Cao, Hanpin Wang, and Yu Huang. Multitasks-based deep evidential fusion network for blind image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7512–7520, 2026.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. IEEE, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, volume 35, pp. 27730–27744, 2022.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8237. IEEE, 2022.

Abinav Rao and Sujan Rachuri. Do understanding and generation fight? a diagnostic study of dpo for unified multimodal models. *arXiv preprint arXiv:2603.17044*, 2026.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in neural information processing systems*, volume 30, 2017.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12692–12702. IEEE, 2020.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, et al. Multimodal learning with next-token prediction for large multimodal models. *Nature*, 650(8101):327–333, 2026.

Xiwen Wei, Mustafa Munir, and Radu Marculescu. Mitigating intra-and inter-modal forgetting in continual learning of unified multimodal models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 151991–152019, 2026.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, pp. 28240–28264, 2025.

Robin Young. What is the alignment tax? *arXiv preprint arXiv:2603.00047*, 2026.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *International Conference on Learning Representations*, pp. 765–783, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in neural information processing systems*, volume 33, pp. 5824–5836, 2020.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, pp. 6446–6469, 2025.

APPENDIX

A Notation 14

B Proofs of the main statements 14

C Additional lemmas on Fisher, Rayleigh, and cross energy 16

D Objects, protocol, and measurement conventions 18

E Last layer constructions, share, and the scale channel 18

F Dilution, pairing, and the interpolant envelope 21

G Isolation, invasion plane, and the Markov toy 22

H Public checkpoints and head geometries 26

A NOTATION

The main text reports three measured objects. Share uses the clean second moment G_m, the mix tax ΔL_t is text cross entropy on a held out pure text pack relative to a text only reference, and isolation is the end layer cosine $S_{\mathrm{inter}}(\mathrm{end})$ at matched parameter count. Table 3 lists the remaining symbols used with those objects.

Table 3. Symbols used in the main text and in this appendix.

symbol	meaning
$\mathcal{V}_t, \mathcal{V}_v$	text and visual blocks of the mixed vocabulary
ρ	visual token fraction in a mixed pack
E, E_t, E_v	unembedding and its two blocks
h, z, p, r	hidden state, logits, softmax, residual $p - e_y$
R_m	residual energy $\mathbb{E}[\|r\|^2 \mid m]$
H_m	conditional entropy of modality m
G_m	clean share $\rho_m \mathbb{E}[\|\nabla_h \ell\|^2 \mid m]$
$\hat{G}_m$	mean CE backward, includes $1/N_m^2$
λ_{eff}	residual Rayleigh quotient
$\kappa_{\mathrm{eff}}, \kappa_{\mathrm{op}}$	Rayleigh scale and operator scale
$\kappa(E)$	embedding scale of a joint unembedding
κ_F	Frobenius scale $\|E_v\|_F^2 / \|E_t\|_F^2$
χ	Frobenius overlap $\|E_t E_v^\top\|_F$
ΔL_t	mix alignment tax on a held out text pack
$S_{\mathrm{inter}}, S_{\mathrm{base}}, c_g$	Uni-X cosines and conflict
$\theta_t^\star, \theta_v^\star, \theta_\rho$	unimodal optima and the interpolant station
w_m	per modality loss weight
N_m	token count of modality m in a sequence

B PROOFS OF THE MAIN STATEMENTS

The last layer identities are finite dimensional linear algebra together with the softmax Jacobian. The weighting argument uses a convex quadratic, and the isolation argument uses a residual Jacobian whose end attention stays inside one modality. The lazy quadratic is used only in the statements that name it.

B.1 SOFTMAX JACOBIAN AND RESIDUAL ENERGY

Let $z \in \mathbb{R}^V$, $p = \mathrm{softmax}(z)$, and $\ell = -\log p_y$. The Jacobian of softmax is the Fisher matrix of a categorical,

$$F_z = \nabla_z^2 \ell = \mathrm{diag}(p) - pp^\top, \tag{8}$$

which is positive semidefinite of rank $V - 1$ with kernel $\mathrm{span}\{1\}$. The gradient in logit space is $\nabla_z \ell = p - e_y$. Through $z = Eh$,

$$\nabla_h \ell = E^\top (p - e_y), \qquad \nabla_h^2 \ell = E^\top F_z E. \tag{9}$$

Let $r = p - e_y$ and denote R as $\mathbb{E}\|r\|^2$. Expanding,

$$\|r\|^2 = \|p\|_2^2 - 2p_y + 1. \tag{10}$$

If $y \sim p$, then $\mathbb{E}[p_y] = \|p\|_2^2$, hence $R = 1 - \|p\|_2^2$.

Proof of Theorem 1. The identity $R = 1 - \|p\|_2^2$ is the previous display. The squared Euclidean norm $\|p\|_2^2$ is strictly Schur convex on the simplex, so $1 - \|p\|_2^2$ is strictly Schur concave. Along the temperature family $p_T = \mathrm{softmax}(z_0/T)$, raising T mixes the distribution toward uniformity. If $0 < T < T'$, then $p_{T'}$ is majorised by p_T, and strict Schur concavity gives $R(T') > R(T)$. Entropy $H(p_T)$ is also strictly increasing in T, so R is strictly increasing in H along that family. □

Near a peaked distribution, $1 - \|p\|_2^2 \sim 2(1 - p_{\max})$ and $H(p) \sim -(1 - p_{\max}) \log(1 - p_{\max})$. Thus R and H vanish at the same leading rate, up to a logarithm, as p concentrates. This local comparison is why a temperature family can replace R by H, while an arbitrary pair of categoricals does not admit the same replacement.

B.2 GRADIENT SHARE

Partition $E^\top = [E_t^\top \; E_v^\top]$ and $r^\top = [r_t^\top \; r_v^\top]$. Then

$$\nabla_h \ell = E_t^\top r_t + E_v^\top r_v \tag{11}$$

and

$$\|\nabla_h \ell\|^2 = \|E_t^\top r_t\|^2 + \|E_v^\top r_v\|^2 + 2 r_t^\top (E_t E_v^\top) r_v, \tag{12}$$

which is the split in equation 2. The first two terms are the diagonal energy, and the third term is the cross energy. Taking the conditional expectation given modality m and forming the ratio of the two conditionals gives equation 3. Both sides are the same real number whenever the expectations exist, so the geometric error of equation 3 against $\mathbb{E}[\|\nabla_h \ell\|^2 \mid v]/\mathbb{E}[\|\nabla_h \ell\|^2 \mid t]$ is identically one.

If $E_t E_v^\top = 0$, the cross term vanishes. Then

$$\mathbb{E}[\|\nabla_h \ell\|^2 \mid m] = \mathbb{E}[r_m^\top (E_m E_m^\top) r_m] = \lambda_{\text{eff}}(E_m, r_m)\, R_m. \tag{13}$$

Multiplying by ρ_m and forming G_v/G_t yields equation 4. The operator norm $\|E_m\|_2^2$ is the largest eigenvalue of $E_m E_m^\top$, so

$$\lambda_{\text{eff}}(E_m, r_m) \le \|E_m\|_2^2, \tag{14}$$

with equality if and only if the residual second moment is supported on a top eigenspace. The ratio $\kappa_{\text{op}}/\kappa_{\text{eff}}$ therefore measures how far the residual sits from that top eigenspace, while equation 3 remains an identity.

On a calibrated block diagonal unembedding $E = \text{blkdiag}(I, sI)$ one has $E_t E_v^\top = 0$ and

$$\|\nabla_h \ell\|^2 = \|r_t\|^2 + s^2 \|r_v\|^2 \tag{15}$$

exactly, and therefore $G_v/G_t = (\rho/(1-\rho))(R_v/R_t)s^2$ with no remainder. This is the last layer construction in Figure 1a. The median relative error 1.4×10^{-16} is floating point error on that identity.

Proof of Proposition 1. The map $(\rho, R_v, R_t, \kappa_{\text{eff}}) \mapsto (\rho/(1-\rho))(R_v/R_t)\kappa_{\text{eff}}$ is constant on each level set $\rho R_v \kappa_{\text{eff}} = c(1-\rho)R_t$. A curve that moves only ρ stays on one level set once the other three coordinates compensate. Rescaling the visual rows by s replaces E_v by sE_v, and therefore replaces $\lambda_{\text{eff}}(E_v, r_v)$ by $s^2 \lambda_{\text{eff}}(E_v, r_v)$. The point leaves the level set unless some other coordinate compensates. Holding ρ and the residual energies fixed therefore identifies the embedding channel. □

Proof of Proposition 2. A continuous visual loss $\frac{1}{2}\|f_v(h) - u\|^2$ has hidden gradient $J_v^\top (f_v(h) - u)$. There is no visual block of E and no joint Fisher matrix. The ratio G_v/G_t still exists as a pair of squared gradient energies, but it does not factor through $\kappa_{\text{eff}}(E)$. A dual head applies E^{lm} on text positions and a generation map on visual positions. The two residuals live in different output spaces, so there is no product $E_t E_v^\top$ and no overlap χ. When the generation map is a multilayer perceptron, a single matrix E^{gen} is already an approximation to that map. □

B.3 FOREIGN CODE OCCUPANCY AND THE INTERPOLANT

Proof of Proposition 3. Let $L_t(\theta) = \mathbb{E}[-\log p_\theta(y \mid x),\, y \in \mathcal{V}_t]$ on a fixed text pack. A second stream that occupies a fraction ρ of the training tokens but remains inside $\mathcal{V}_t$ changes the text token count and does not add a new block of the partition function. A second stream in $\mathcal{V}_v$ does add such a block. When the measured ΔL_t is negligible in the first case and large in the second, at matched ρ, matched text data, and matched evaluation, the excess is the new block. An unstructured stream on $\mathcal{V}_v$ is the extreme of that block. A paired caption stream that lowers L_t relative to a shuffled pairing, on a caption only pack, shows that ordinary visual grounding is not the source of a positive excess. □

Along the interpolant $\theta(\alpha) = (1-\alpha)\theta_t^\star + \alpha\theta_v^\star$, equation 5 is the case $\Sigma_t = I$ of Assumption 2. The display describes the interpolant. Mixed training solves a different objective and can leave the segment. For a general $\Sigma_t \succeq \lambda_{\min} I$ the same calculation gives

$$L_t(\theta(\alpha)) - L_t(\theta_t^\star) = \frac{\alpha^2}{2}(\theta_v^\star - \theta_t^\star)^\top \Sigma_t (\theta_v^\star - \theta_t^\star) \ge \frac{\alpha^2}{2}\lambda_{\min}\|\theta_v^\star - \theta_t^\star\|^2. \tag{16}$$

The interpolant is therefore a conservative envelope whenever mixed training stays closer to $\theta_t^\star$ than the segment does.

B.4 WEIGHTING

Proof of Proposition 4. Under Assumption 2 with $\Sigma_t = \Sigma_v = I$, the stationary point of $L_w = w_t(1-\rho)L_t + w_v\rho L_v$ is the convex combination with weights $w_t(1-\rho)$ and $w_v\rho$. The two excesses are equation 7, where $\beta = w_v\rho/(w_t(1-\rho)+w_v\rho)$ and $\Delta = \|\theta_v^\star - \theta_t^\star\|^2$. The pair traces a quarter circle. The equal excess point is $\beta = 1/2$, which is claim (i). For claim (ii), a per token gradient has covariance proportional to R_m under the linearised softmax, and the minimum variance combination of two unbiased estimators with those covariances uses weights $1/R_m$. For claim (iii), the continuous time rate of modality m is $\eta w_m \rho_m \lambda_m$. Equalising the two rates at $\lambda_t = \lambda_v$ forces $w_m \propto 1/\rho_m$. □

Length balance and inverse entropy therefore select two different points on the same circle, and the two points answer different questions. Figure 8 draws that circle, with length balance in orange and inverse entropy on the toy entropies in green.

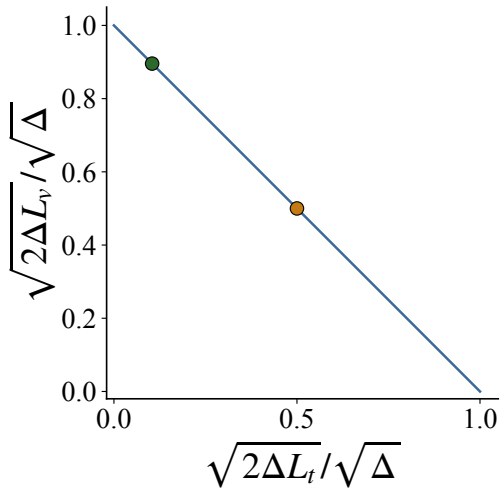


Figure 8. Isotropic Pareto circle of the two excesses. The orange point is length balance, and the green point is inverse entropy on the toy entropies.

B.5 INVASION PLANE

Proof of Theorem 3. Let $g_m^{(\ell)}$ denote the unimodal backward gradient at layer ℓ. Then $S_{\text{inter}}(\ell) = \cos(g_t^{(\ell)}, g_v^{(\ell)})$ depends on the two Jacobians from the loss to that layer. Replacing an end layer by two modality specific maps, each attending only to keys of its own modality, removes the other modality from those Jacobians. On those parameters $\langle g_t, g_v\rangle = 0$ identically, and therefore $S_{\text{inter}}(\text{end}) = 0$. The shared trunk still sees a mixed residual stream and still obeys the last layer share of Theorem 2. Inverse entropy weights multiply one modality loss by a scalar and therefore scale G_v/G_t without changing the two Jacobian subspaces, so S_{inter} is unchanged to first order. A diagnostic that stores only S_{inter}, or only G_v/G_t, is a coordinate projection of the plane and cannot separate the two maps. Extra end layer width changes the trunk capacity and is a third map. Matching the parameter count isolates the Jacobian map. □

Let $A_\ell = I + \mathrm{D}F_\ell$ denote the residual Jacobian at layer ℓ. Isolation at an end layer replaces A_{end} by a block diagonal pair $(A_{\text{end}}^{(t)}, A_{\text{end}}^{(v)})$ with $A_{\text{end}}^{(t)} A_{\text{end}}^{(v)\top} = 0$ on the attention keys. The backward maps then factor as $J_t = A_{\text{end}}^{(t)} J_t^{\text{trunk}}$ and $J_v = A_{\text{end}}^{(v)} J_v^{\text{trunk}}$, so

$$\langle g_t, g_v\rangle = \langle A_{\text{end}}^{(t)} J_t^{\text{trunk}} \nabla L_t,\ A_{\text{end}}^{(v)} J_v^{\text{trunk}} \nabla L_v\rangle = 0 \tag{17}$$

on the isolated parameters. The inner product is zero for every trunk state, so the end cosine is a structural zero.

C ADDITIONAL LEMMAS ON FISHER, RAYLEIGH, AND CROSS ENERGY

Lemma 2 (Fisher kernel)**.** *F_z annihilates the all ones vector and is positive definite on $1^\perp$. Consequently $\nabla_h^2 \ell$ has a kernel that contains every direction h for which Eh is constant on $\mathcal{V}$.*

Proof. $F_z 1 = p - (1^\top p)p = 0$. For $v \perp 1$, $v^\top F_z v = \sum_i p_i v_i^2 - (\sum_i p_i v_i)^2$, which is the variance of v under p and vanishes only if v is constant on the support of p. □

Lemma 3 (Rayleigh variational form)**.** *Let $S_m = E_m E_m^\top$. Then $\lambda_{\text{eff}}(E_m, r_m) = \mathbb{E}[r_m^\top S_m r_m]/\mathbb{E}[\|r_m\|^2]$ lies in $[\lambda_{\min}^+(S_m), \lambda_{\max}(S_m)]$, where $\lambda_{\min}^+$ is the smallest positive eigenvalue. Equality with $\lambda_{\max}$ holds if and only if the residual second moment is supported on a top eigenspace.*

Proof. The Rayleigh quotient of a positive semidefinite matrix, evaluated on a nonzero vector, lies between the extreme eigenvalues on the support of that vector. A ratio of expectations preserves the bound, because $\mathbb{E}[r^\top S r] = \mathbb{E}[\|r\|^2 \cdot (r^\top S r/\|r\|^2)]$ and the pointwise bound applies inside the expectation. □

Proof of Lemma 1. $|r_t^\top (E_t E_v^\top) r_v| \le \|E_t E_v^\top\|_2 \|r_t\|_2 \|r_v\|_2 \le \chi \|r_t\|_2 \|r_v\|_2$. The factor of two comes from the polarisation identity. If $\mathbb{E}[r_t \otimes r_v] = 0$, then $\mathbb{E}[r_t^\top M r_v] = \langle M, \mathbb{E}[r_t r_v^\top] \rangle_F = 0$ for every matrix M. □

The same bound is Lemma 1 in Section 3, and it controls the conditional share once the cross term is retained. The 36.5M cross fractions and the public checkpoint are reported in Section 5.

Lemma 4 (Mean CE backward)**.** *Let L_m be a mean cross entropy over N_m tokens of modality m. Then $\nabla_h L_m = N_m^{-1} \sum_{i \in m} \nabla_h \ell_i$, so $\mathbb{E}[\|\nabla_h L_m\|^2]$ differs from $\mathbb{E}[\|\nabla_h \ell\|^2 \mid m]$ by a factor N_m^{-2} times a token covariance. A comparison of $\hat{G}_v/\hat{G}_t$ with $(\rho/(1-\rho))(R_v/R_t)\kappa_{\text{op}}$ therefore folds N_v^{-2}/N_t^{-2} into the test.*

Proof. The factor is immediate from the definition of a mean. If the tokens are independent, $\mathbb{E}[\|\nabla_h L_m\|^2] = N_m^{-2} \sum_i \mathbb{E}[\|\nabla_h \ell_i\|^2]$. The clean object G_m multiplies the per token second moment by ρ_m, and not by N_m^{-2}. □

Figure 9 compares the two ratios on the same checkpoints. At $\rho = 0.20$ the mean cross entropy ratio is 12.0, while the clean share is 0.179. Every comparison reported in the main text uses the clean share.

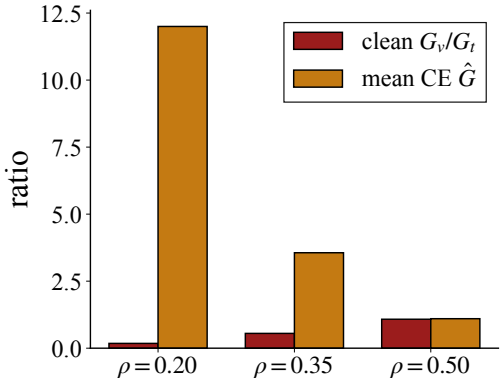


Figure 9. Clean share and a mean cross entropy backward on the same three mixes. The two ratios are different objects.

Lemma 5 (Linear Gaussian companion)**.** *In the lazy quadratic $y = x^\top \theta_m^\star + \sqrt{H_m}\,\varepsilon$, with $x \sim \mathcal{N}(0, I/d)$ and ε of unit variance, a single token gradient $g = (x^\top \theta - y)x$ satisfies $\mathbb{E}\|g\|^2 = H_m(1 + O(d^{-1})) + \|\theta - \theta_m^\star\|^2 \cdot O(d^{-1})$. In the noise dominated regime the leading monomials are H_m and, after mixing, $\rho/(1-\rho)$.*

Proof. Conditioning on x gives $\mathbb{E}[\|g\|^2 \mid x] = ((x^\top(\theta - \theta_m^\star))^2 + H_m)\|x\|^2$. Taking $\mathbb{E}\|x\|^2 = 1$ and $\mathbb{E}(x^\top v)^2 = \|v\|^2/d$, the Gaussian fourth moments contribute $O(d^{-1})$. At $\theta = 0$ and $H_m \gg d^{-1}$ the ratio of expected squares is $(H_v/H_t)(1 + o(1))$. □

The companion shows why a quadratic proxy and a last layer softmax agree on the monomials (ρ, R, E). It does not replace κ_{eff} by H_v/H_t on a trained transformer.

D OBJECTS, PROTOCOL, AND MEASUREMENT CONVENTIONS

A joint softmax uses one shared matrix E. The transformers trained in this paper, the last layer constructions, the Chameleon style public checkpoint, and Show-o with official MAGViTv2 codes in the joint slot are joint softmax models. Janus-Pro is a dual head model, while Qwen3-VL and the mean squared error visual head on the Markov toy are continuous visual heads. Applying equation 4 to a dual head or a continuous head is not a test of Theorem 2.

The clean share G_m is an expectation of a per token squared hidden gradient. A mean cross entropy backward that already averages inside a modality includes a factor $1/N_m^2$, and it is labelled as such whenever it is reported. Share numbers in the main text are G_v/G_t, or a parameter gradient ratio $\|g_v\|/\|g_t\|$. A run is marked memorised, and is left out of the reported tax, when held out text cross entropy falls below 0.1 nats or held out visual cross entropy falls below 0.05. The final 63M Gutenberg checkpoints are memorised, while the 36.5M suite is not. Visual cross entropy near 5.7 nats on held out Picsum codes is the unmemorised band used throughout. Text evaluation always uses a pure text pack that contains no visual ids, and visual evaluation uses a pure visual pack, so the mix tax is measured outside the mixed training batch. Pairing experiments use a caption only pack, so a class name cannot be stored as an extra text stream.

The 36.5M backbone is a 10 layer causal transformer of width 384 with 6 heads and sequence length 256, with text vocabulary 8000 and visual vocabulary 16384. Text is a Gutenberg byte pair encoding of $545{,}582$ tokens, and the visual codes are Janus vector quantised codes of 598 images, with 119 images held out. The mean visual length is 128 ids. Mix ratios $\rho \in \{0.20, 0.35, 0.50\}$ are packed at the token level, and optimisation uses AdamW in bfloat16. Two seeds are reported wherever a second seed was trained. The 63M replica uses width 512, depth 12, and sequence length 384. Slim isolation reduces the isolated feed forward width so that the parameter count matches the shared net to 0.008% on 36.5M and to 0.0065% on 63M. Public probes freeze the released weights and backpropagate unimodal text and visual packs separately on the shared residual stream. The magnitude ratio is a parameter gradient ratio, except where a last layer hidden probe is named. Show-o uses official MAGViTv2 codes in the documented codebook slot. Official Chameleon-7B weights are not released, and the joint softmax public row is a Chameleon style checkpoint with one image and text softmax. Table 4 records which identity is defined on which object.

Table 4. Identities by head geometry. A mark means the identity is defined on that object. A missing mark means the symbol does not exist, or that the identity is not a test of the statement.

object	equation 3	equation 4	κ_{eff}	χ
joint softmax, trained transformer	yes	Rayleigh	yes	yes
joint softmax, last layer toy	yes	exact at s^2	yes	0
joint softmax, public	yes	Rayleigh	yes	yes
dual head	n/a	n/a	two maps	none
continuous visual head	n/a	n/a	missing	none

Tokenizer unigram entropy on the packed Gutenberg stream is 6.63 nats, and visual unigram entropy is 6.27 nats, so a direct substitution of H_v/H_t is near one and does not account for the measured share. A LLaVA style text unigram entropy of 4.50 nats, against a Zipf codebook of size 8192 at half usage with entropy 6.07 nats, gives $\sqrt{H_v/H_t} \approx 1.16$ (Liu et al., 2024). That factor of order one is a tokeniser proxy, and it is not a substitute for κ_{eff}.

E LAST LAYER CONSTRUCTIONS, SHARE, AND THE SCALE CHANNEL

The calibrated last layer object, the trained share measurements, and the scale intervention form one identification. Residual energy and leak are the construction, and closure together with the factor s^2 is the trained check. The numerical tables below report the same symbols at machine precision and after training.

E.1 RESIDUAL ENERGY AND LEAK

The calibrated construction uses a block diagonal unembedding with $|\mathcal{V}_t| = |\mathcal{V}_v| = 32$, and the targets are drawn from a temperature family. Figure 10 records R against H, and Figure 11 records

the softmax leak onto $\mathcal{V}_v$ for a calibrated text hidden state as the visual rows are scaled. Leak is 0.518 at $\kappa = 0.25$, 0.226 at $\kappa = 0.56$, 0.074 at unit scale, and 0.010 at $\kappa = 1.96$. This is the small κ_{eff} regime associated with Chameleon z-loss, read on a single hidden state.

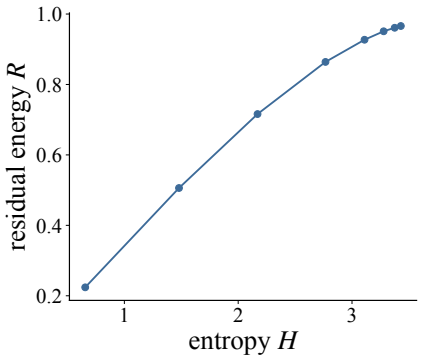


Figure 10. Residual energy against entropy on a temperature family. On that family R is strictly increasing in H, and R is not a general substitute for H.

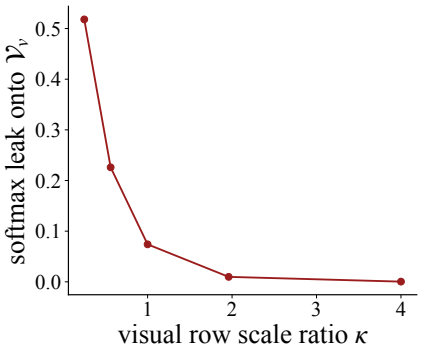


Figure 11. Softmax leak onto visual ids for a text target, as a function of visual row scale. Visual rows scaled below one capture mass that belongs to $\mathcal{V}_t$.

Scaling the visual rows below one places excess softmax mass onto $\mathcal{V}_v$ even when the target lies in $\mathcal{V}_t$. Scaling them above one suppresses that leak and inflates G_v on visual targets through equation 4. Keeping the two blocks within a small factor of unit scale is the embedding counterpart of inverse entropy weights. The z-loss penalises a large $\log \sum_y e^{z_y}$, and it is therefore a regulariser for the same small scale regime (Team, 2024). In this paper z-loss is a named control of κ_{eff}.

E.2 Share measurements and the scale channel

Table 5 records the closure of equation 3 and the operator residual on the 36.5M mid training checkpoints. The closed geometric error is 1.000 at every mix, and the operator product is larger than the clean share by a factor between 9.2 and 10.1. Adding the measured cross term does not close that gap. Share in the table is the clean ratio G_v/G_t, the operator product uses κ_{op}, and the closed column is equation 3.

Table 5. Last layer closure on 36.5M mid training checkpoints. Share is the clean ratio G_v/G_t. The operator product uses κ_{op}, and the closed column is equation 3.

ρ	R_v/R_t	κ_{op}	share	operator	closed geom.	cross fraction
0.20	0.58	11.68	0.179	1.70	1.000	-0.2%
0.35	0.87	11.86	0.552	5.56	1.000	-0.1%
0.50	0.97	10.17	1.080	9.90	1.000	$+0.2\%$

At $\rho = 0.35$ one has $\kappa_{\text{eff}} \approx 1.18$ against $\kappa_{\text{op}} = 11.86$. The ratio of those two numbers is the operator residual, and it is distinct from the overlap χ. Table 6 records the scale intervention at the same mix. Visual rows are multiplied by s after training, and share is then measured on the same pack, while the tax column comes from matched training runs that apply the same scale throughout. Share tracks s^2, residual energy barely moves, and the tax is largest at $s = 0.5$.

Table 6. Identifying the embedding channel at fixed $\rho = 0.35$. Visual rows are multiplied by s after training, and share is then measured on the same pack. Tax comes from matched training runs that apply the same scale throughout.

s	s^2	measured share	operator product	ΔL_t
0.5	0.25	1.49	1.64	0.539
1.0	1.00	4.40	6.32	0.331
2.0	4.00	20.4	29.0	0.417

Collapsing the codebook to 256 or 32 visual ids moves the measured share from 4.40 to 4.34 and 4.64, while residual energy barely moves. An entropy substitution $R = H$ is not identified on this axis, and the embedding channel is. On the 63M replica at mid training, before Gutenberg is memorised, the geometric errors of hidden share against an operator product are 5.71, 1.68, and 1.45 at $\rho \in \{0.20, 0.35, 0.50\}$. The same pattern appears as on the 36.5M suite. The middle mix

lies closer to the operator product, and the two ends do not, which is residual alignment that depends on ρ. Equation equation 3 remains closed. Figure 12 plots the tax column of Table 6. Share is monotone in s, and the tax is not.

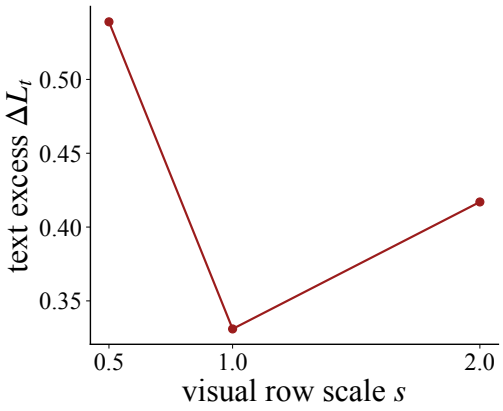


Figure 12. Mix tax as a function of visual row scale at $\rho = 0.35$. Share is monotone in s, and the tax is not.

E.3 Last layer mix identity, numerical table

Table 7 records the calibrated block diagonal construction used in Figure 1a. Measured share equals the residual identity to machine precision, with median relative error 1.4×10^{-16}. Twelve mixes from $\rho = 0.080$ to $\rho = 0.750$ match on both sides of $(\rho/(1-\rho))(R_v/R_t)$. Residual energies stay in a narrow band, with R_t near 0.76 and R_v near 0.93, so the ratio is a mix identity. This is Theorem 2 on a construction whose unembedding is block diagonal by design. The ratio column is the measured G_v/G_t, and the identity column uses $(\rho/(1-\rho))(R_v/R_t)$.

Table 7. Calibrated block diagonal unembedding. The ratio is the measured G_v/G_t, and the identity uses $(\rho/(1-\rho))(R_v/R_t)$.

ρ	R_t	R_v	measured ratio	identity
0.080	0.768	0.928	0.105	0.105
0.141	0.770	0.932	0.199	0.199
0.202	0.765	0.921	0.304	0.304
0.263	0.771	0.930	0.430	0.430
0.324	0.791	0.932	0.564	0.564
0.385	0.779	0.930	0.746	0.746
0.445	0.770	0.926	0.965	0.965
0.506	0.760	0.925	1.248	1.248
0.567	0.760	0.929	1.603	1.603
0.628	0.756	0.932	2.082	2.082
0.689	0.764	0.934	2.710	2.710
0.750	0.764	0.930	3.651	3.651

Table 8 records residual energy against entropy on the same temperature family, with $R = 1 - \|p\|_2^2$ on a calibrated categorical.

Table 8. Temperature family on a calibrated categorical, with $R = 1 - \|p\|_2^2$.

T	H	R	leak onto $\mathcal{V}_v$ at unit scale
0.6	0.656	0.224	0.074
0.8	1.480	0.506	0.074
1.0	2.170	0.716	0.074
1.3	2.767	0.864	0.074
1.7	3.110	0.927	0.074
2.2	3.278	0.951	0.074
3.0	3.375	0.961	0.074
4.5	3.429	0.966	0.074

Leak in the last column of Table 8 is measured at unit visual scale and does not depend on T in this construction. Table 9 records the same leak as a function of scale at fixed $T = 1$. Together with

Table 7, these tables are closed forms for the last layer construction, and the 36.5M suite is the finite width check that the same symbols remain readable after training.

Table 9. Softmax leak onto $\mathcal{V}_v$ for a text target as visual rows are scaled.

visual scale s	$\kappa = s^2$	leak
0.50	0.25	0.518
0.75	0.56	0.226
1.00	1.00	0.074
1.40	1.96	0.010
2.00	4.00	0.0004

E.4 Embedding scale after training

Proposition 1 identifies the embedding channel at fixed ρ and fixed residual energy. Table 10 is the trained counterpart on the 36.5M shared net at $\rho = 0.35$, with $L_0 = 3.564$. Visual rows are multiplied by $s \in \{0.5, 1, 2\}$ at initialisation, and the mix is then trained. The tax is largest at the under scaled station $s = 0.5$, which agrees with softmax leak onto $\mathcal{V}_v$ on text targets. The over scaled station $s = 2$ lies above unit scale and below $s = 0.5$. Held out visual cross entropy stays in the unmemorised band on all three runs.

Table 10. Trained embedding scale on the 36.5M suite at $\rho = 0.35$. $L_0 = 3.564$. Visual rows are multiplied by s at initialisation.

s	text CE	visual CE	ΔL_t
0.5	4.103	5.732	0.539
1.0	3.894	5.660	0.331
2.0	3.980	5.872	0.417

Last layer leak at unit temperature, from Table 9, falls from 0.518 at $s = 0.5$ to 0.0004 at $s = 2$. Leak and the trained tax do not share a monotone map, because leak is a last layer occupancy and the tax is a finite width trajectory. Both still pass through κ_{eff}, which is why row scales are kept near one, or why z-loss is applied. The tax in Table 10 does not invert against s^2.

E.5 Share decomposition at mid training

Table 5 is the last layer split of Theorem 2 at step 250. The geometric errors of the operator product against the true share are 9.53, 10.08, and 9.17 at $\rho \in \{0.20, 0.35, 0.50\}$. Those factors are why the main text treats κ_{op} as a scale bound. Using κ_{op} in place of κ_{eff} overstates visual share on this suite by about one decade.

F Dilution, pairing, and the interpolant envelope

Table 11 expands Figure 2 at $\rho = 0.35$ on the 36.5M suite, with $L_0 = 3.564$ and held out visual cross entropy. The two seeds agree in sign and in magnitude. A second text stream leaves the tax near zero, while unstructured ids on $\mathcal{V}_v$ raise it above the real vector quantised row.

Table 11. Dilution control at $\rho = 0.35$ on the 36.5M suite, with $L_0 = 3.564$. Visual cross entropy is held out.

second stream	seed	ΔL_t	visual CE	verdict
real VQ	0	0.331	5.66	foreign code
real VQ	1	0.309	5.71	foreign code
second text	0	0.004	3.59	dilution null
second text	1	0.000	3.57	dilution null
unstructured on $\mathcal{V}_v$	0	0.429	10.05	foreign code
unstructured on $\mathcal{V}_v$	1	0.443	10.03	foreign code

Table 12 expands the caption control on CIFAR-100. Evaluation uses one shared caption only pack, with $L_0 = 2.649$. The set has four hundred images, three hundred and twenty for training and eighty for evaluation, and the captions are class names in a fixed template. When the second stream is a class name aligned with the image, text cross entropy on that pack falls. When the class name is shuffled, the fall is smaller.

Table 12. Caption pairing on CIFAR-100. Evaluation uses one shared caption only pack, with $L_0 = 2.649$. Four hundred images are split into three hundred and twenty for training and eighty for evaluation. Captions are class names in a fixed template.

training stream	caption L_t	ΔL_t
caption only	2.649	0
image and class name, paired	2.467	−0.182
image and class name, shuffled	2.597	−0.051

Table 13 records the interpolant $\theta(\alpha) = (1-\alpha)\theta_t^\star + \alpha\theta_v^\star$ and the mixed training path. The gap is $L_t(\theta_v^\star) - L_0 = 14.89$, and the median interpolant geometric error is 1.42. Figure 13 is that interpolant, and Figure 14 compares mixed training with the same envelope at three mix ratios. Mixed training does not lie on the segment.

Table 13. Interpolant $\theta(\alpha) = (1-\alpha)\theta_t^\star + \alpha\theta_v^\star$ and mixed training. The gap is $L_t(\theta_v^\star) - L_0 = 14.89$, and the median interpolant geometric error is 1.42. Mixed training does not lie on the segment.

weight or mix	measured excess	α^2 envelope	geom.
$\alpha = 0.20$	0.685	0.595	1.15
$\alpha = 0.35$	1.386	1.824	1.32
$\alpha = 0.50$	2.073	3.722	1.80
$\alpha = 0.75$	5.506	8.374	1.52
mixed $\rho = 0.20$	0.426	0.595	1.40
mixed $\rho = 0.35$	0.331	1.824	5.51
mixed $\rho = 0.50$	0.954	3.722	3.90

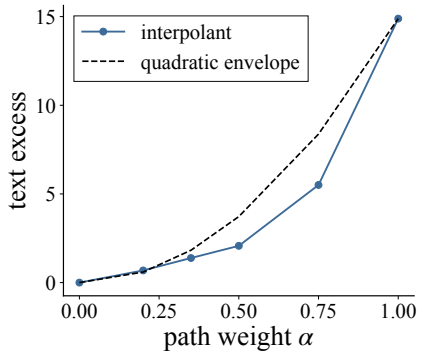


Figure 13. Text excess along the interpolant. The dashed curve is the quadratic envelope.

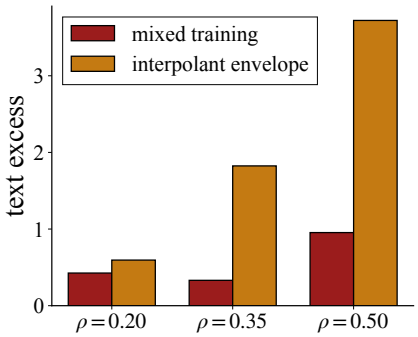


Figure 14. Mixed training and the interpolant envelope at three mix ratios.

A direct ρ^2 check on mixed training, tax/ρ^2 at $\rho \in \{0.20, 0.35, 0.50\}$, takes the values 10.66, 2.70, and 3.82, with relative spread 0.61. Mixed tax is therefore not a quadratic in ρ, while the interpolant is, which is why equation 5 is an envelope for the mixed path. When the second stream is an unpaired vector quantised code or an unstructured code, text cross entropy rises. The mix alignment tax of Definition 1 is that occupancy. The paired caption drop is a separate comparison.

G Isolation, invasion plane, and the Markov toy

Figure 15 collects the four isolation conditions on the 36.5M suite at $\rho = 0.35$. Slim isolation matches the shared parameter count. Layerwise S_{inter} on the shared net and on slim isolation is Figure 4 in Section 5.

Table 14 expands Table 1 by seed on Gutenberg and Janus codes at $\rho = 0.35$, with $L_0 = 3.564$. Final 63M text cross entropy below 0.1 is marked memorised and is omitted. Table 15 is the 63M mid training snapshot at step 400, with $L_0 = 4.304$, and slim isolation matches the shared count to 0.0065%.

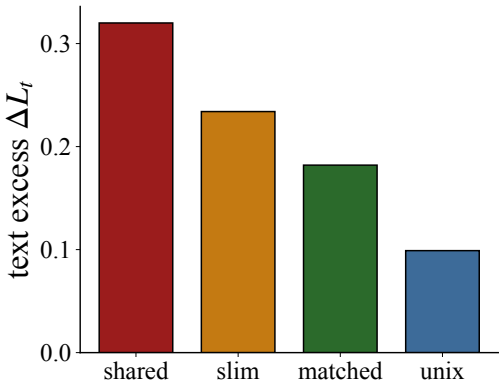


Figure 15. Four isolation conditions on the 36.5M suite at $\rho = 0.35$. Slim matches the shared parameter count.

Table 14. Isolation on Gutenberg and Janus codes at $\rho = 0.35$, both seeds. $L_0 = 3.564$.

run	seed	params	ΔL_t	end S_{inter}
shared	0	36,536,064	0.331	0.009
shared	1	36,536,064	0.309	0.018
slim isolation	0	36,539,136	0.230	0.00
slim isolation	1	36,539,136	0.238	0.00
matched width	0	38,895,360	0.179	≈ 0
matched width	1	38,895,360	0.184	≈ 0
Uni-X isolation	0	40,078,080	0.104	0.00
Uni-X isolation	1	40,078,080	0.093	0.00

Table 15. 63M mid training at step 400. $L_0 = 4.304$. Slim isolation matches the shared count to 0.0065%.

run	params	L_t	ΔL_t
shared, $\rho = 0$	62.94M	4.304	0
shared, $\rho = 0.20$	62.94M	4.800	0.496
shared, $\rho = 0.35$	62.94M	4.998	0.693
shared, $\rho = 0.50$	62.94M	4.957	0.652
slim isolation, $\rho = 0.35$	62.94M	4.852	0.548
matched width, $\rho = 0.35$	68.18M	4.954	0.649
Uni-X isolation, $\rho = 0.35$	69.24M	4.804	0.499

Figure 16 collects the four isolation conditions on that replica at mid training and $\rho = 0.35$. The signs match the 36.5M suite, including the slim isolation drop at matched parameter count. Figure 17 is the Markov toy isolation control. The width matched shared tax is 2.224 at 106816 parameters, and Uni-X isolation is 0.246 at 107072 parameters.

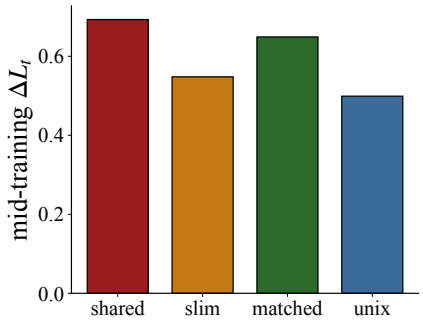


Figure 16. Four isolation conditions on the 63M replica at mid training, $\rho = 0.35$. Signs match the 36.5M suite.

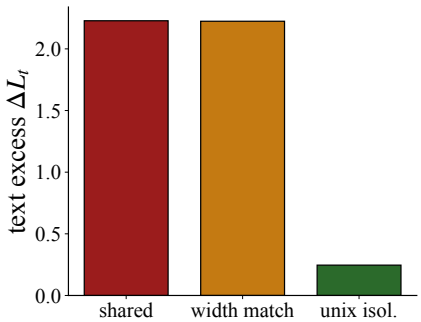


Figure 17. Markov toy isolation. Width matched shared tax is 2.224 at 106816 parameters. Uni-X isolation is 0.246 at 107072 parameters.

Table 16 records reweighting on the interleaved Markov model at $\rho = 0.40$. Tax is text cross entropy minus a matched text only run, and isolation uses one separated layer at each end. Inverse entropy moves the magnitude, and isolation moves the cosine.

Table 16. Interleaved Markov model at $\rho = 0.40$. Tax is text cross entropy minus a matched text only run. Isolation uses one separated layer at each end.

run	params	text CE	ΔL_t	mean c_g
text only	82k	0.97	0	n/a
shared	82k	3.12	2.23	1.01
Uni-X isolation	107k	1.14	0.25	0.63
width matched shared	107k	3.11	2.22	n/a
inverse entropy	82k	2.68	1.79	1.13
isolation and inverse entropy	107k	1.11	0.22	0.63

Matching the two source entropies lowers the tax from about 2.1 nats to 0.26. Table 17 and Figure 18 record the mix sweep across three seeds. Tax is text cross entropy minus the matched text only run, and the standard deviation is taken across seeds. Text excess is monotone in ρ on held out pure text. Inverse entropy at the same mix is Figure 6 in Section 5, and the discrete softmax against the continuous visual head on this backbone is Table 2 in the same section.

Table 17. Interleaved Markov model, three seeds. Tax is text cross entropy minus the matched text only run. The standard deviation is taken across seeds.

ρ	mean ΔL_t	std.	visual CE
0.12	0.619	0.017	7.05
0.24	1.461	0.039	6.71
0.36	2.188	0.011	6.20
0.48	2.606	0.209	5.87
0.60	3.235	0.174	5.22

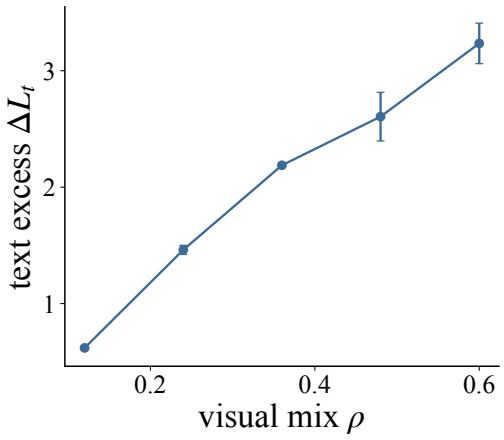


Figure 18. Interleaved Markov model, three seeds. Text excess is monotone in ρ on held out pure text.

On the Uni-X protocol replica, S_{base} stays in $[0.56, 0.70]$ on both the shared net and the isolated net. When the mixed batch baseline is nearly constant in depth, every shape in c_g is a shape in $-S_{\text{inter}}$. Isolation sets $S_{\text{inter}}(\text{end}) = 0$ and therefore sets $c_g(\text{end})$ equal to $S_{\text{base}}(\text{end})$. Inverse entropy leaves S_{inter} negative and can raise the mean of c_g. This is Theorem 3 on a diagnostic that stores only the cosine.

G.1 Layerwise isolation on the 36.5M suite

Table 18 and Table 19 record S_{inter} and the conflict $(1 - \cos)/2$ by layer on the shared net and on slim isolation at $\rho = 0.35$, seed 0. The isolated ends are identically zero, and the trunk remains mixed and of the same order as the shared trunk. Isolation therefore acts at the ends and leaves the trunk in place. The two tables are the finite width check of Theorem 3. Magnitude share on the two nets remains of the same order, while the end cosine does not, so a magnitude ratio groups the two runs together and the end cosine separates them. Held out visual cross entropy is 5.69 on the shared net and 5.70 on slim isolation, in agreement with Table 1. Neither run has memorised the visual stream, so the tax comparison is a comparison of two unmemorised stations at matched parameter count.

Table 18. Shared 36.5M net, $\rho = 0.35$, seed 0. S_{inter} and conflict $(1 - \cos)/2$ by layer.

layer	S_{inter}	conflict
0	0.045	0.477
1	0.035	0.483
2	0.013	0.494
3	0.020	0.490
4	0.011	0.494
5	0.004	0.498
6	−0.001	0.500
7	−0.007	0.504
8	−0.001	0.500
9	−0.027	0.514

Table 19. Slim isolation, 36.5M, $\rho = 0.35$, seed 0. Ends are structurally zero.

layer	S_{inter}	conflict
0	0.000	0.500
1	−0.010	0.505
2	0.001	0.499
3	0.007	0.496
4	−0.010	0.505
5	0.012	0.494
6	−0.003	0.501
7	0.002	0.499
8	0.028	0.486
9	0.000	0.500

G.2 Markov toy layerwise overlap

The width matched isolation control is a six layer Markov model. Table 20 records S_{inter} by block at $\rho = 0.40$. The columns are shared and Uni-X attention and MLP blocks, and Uni-X isolation copies one layer at each end. Shared tax is 2.228 at 82240 parameters, width matched shared tax is 2.224 at 106816 parameters, and Uni-X tax is 0.246 at 107072 parameters. The isolated ends are again structurally zero. Head cosine on the shared net is positive, and it is near zero after Uni-X isolation.

Table 20. Layerwise S_{inter} on the Markov toy at $\rho = 0.40$. Columns are shared and Uni-X attention and MLP blocks. Uni-X isolation copies one layer at each end. Shared tax is 2.228 at 82240 parameters. Width matched shared tax is 2.224 at 106816 parameters. Uni-X tax is 0.246 at 107072 parameters.

block	shared attn	shared mlp	Uni-X attn	Uni-X mlp
0	−0.470	−0.235	0.000	0.000
1	−0.481	−0.122	0.009	0.011
2	−0.546	−0.244	−0.047	0.044
3	−0.634	−0.223	−0.058	−0.008
4	−0.652	−0.015	−0.091	0.031
5	−0.839	0.193	0.000	0.000
head	0.641	n/a	−0.028	n/a

The tax drop by a factor of about nine survives the width match, so extra end width is not the source of that drop. Shared attention cosine becomes more negative with depth, and isolation sets the ends to zero while leaving a small trunk residue. The factor 1.37 on Gutenberg and the factor of about nine on the toy are two stations of the same map.

G.3 Layerwise overlap on the 63M replica

Table 21 records S_{inter} and the conflict $(1 - \cos)/2$ by layer on the 63M shared net at mid training, $\rho = 0.35$, step 400. Gutenberg is not yet memorised. The profile sits near orthogonality at every depth, and it is not a $\sin^2$ interpolation of end isolation. Widening the residual stream from 384 to 512 and the depth from 10 to 12 does not produce a U shaped curve. The same snapshot is the source of the 63M tax row in Table 15. Isolation remains an end map, and the table shows that this

map does not force a U on a wider residual stream. Public checkpoints in Section H.1 make the same point at a larger scale, while the 63M suite makes it at a scale where tax and isolation can still be trained, before Gutenberg is memorised.

Table 21. Shared 63M net at mid training, $\rho = 0.35$, step 400. S_{inter} and conflict $(1 - \cos)/2$ by layer.

layer	S_{inter}	conflict
0	-0.057	0.528
1	-0.054	0.527
2	-0.064	0.532
3	-0.042	0.521
4	-0.021	0.511
5	-0.026	0.513
6	-0.027	0.514
7	-0.040	0.520
8	-0.038	0.519
9	-0.029	0.515
10	-0.010	0.505
11	-0.037	0.519

H PUBLIC CHECKPOINTS AND HEAD GEOMETRIES

Table 22 collects the frozen public probes. Magnitude is a parameter gradient ratio. Official Chameleon-7B weights are not released, and Show-o uses official MAGViTv2 codes. Figures 19, 20, 21, 22, and 23 are the depth profiles. None of them is a $\sin^2$ interpolation of end isolation. Isolation remains the map that applies when the end Jacobian is the bottleneck, because it sets $S_{\text{inter}}(\text{end}) = 0$ by construction.

Table 22. Frozen public probes. Magnitude is a parameter gradient ratio. Official Chameleon-7B weights are not released. Show-o uses official MAGViTv2 codes.

model	head	$\|g_v\|/\|g_t\|$	N_v/N_t	head S_{inter}
Janus-Pro 1B	dual, VQ	2.78	15.2	n/a
Janus-Pro 7B	dual, VQ	1.16	5.3	0
Show-o	joint softmax	0.416	6.0	0.02
Chameleon style 7B	joint softmax	2.01	4.4	-0.05
Qwen3-VL-2B	continuous	3.73	n/a	0.04
Qwen3-VL-4B	continuous	3.30	n/a	0.07
Qwen3-VL-8B	continuous	3.94	n/a	0.10

Figures 19 and 20 are the joint softmax and Janus-Pro 7B depth profiles. The joint softmax profile sits near $1/2$, and conflict on Janus-Pro 7B rises toward the last layers.

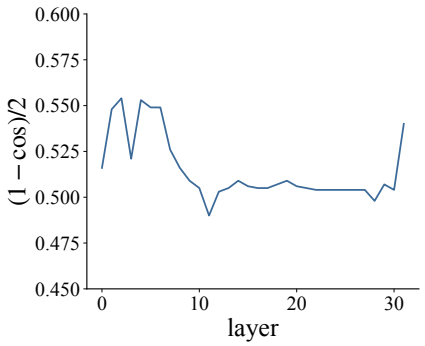


Figure 19. Layerwise conflict on the public Chameleon style joint softmax checkpoint. The profile sits near $1/2$.

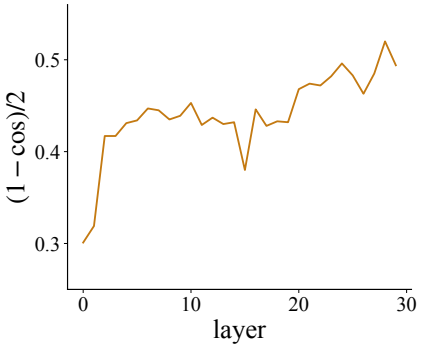


Figure 20. Layerwise conflict on Janus-Pro 7B, parameter gradients. Conflict rises toward the last layers.

Figures 21 and 22 are the two smaller continuous widths. The visual head is a continuous projector, so $\kappa(E)$ is missing on both.

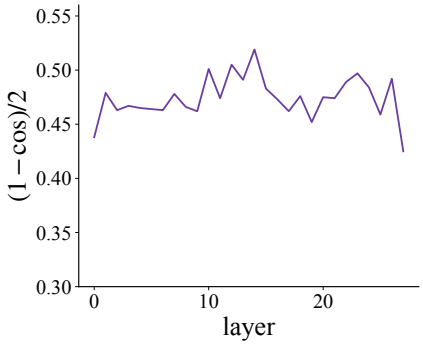


Figure 21. Layerwise conflict on Qwen3-VL-2B. The visual head is continuous, and $\kappa(E)$ is missing.

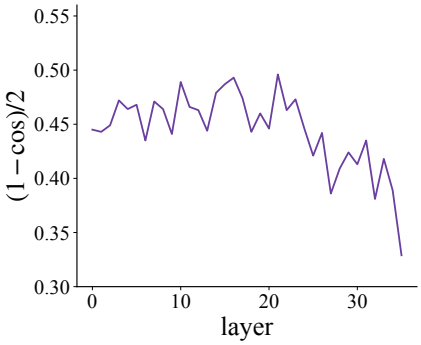


Figure 22. Layerwise conflict on Qwen3-VL-4B. The projector geometry matches the 2B model, and $\kappa(E)$ remains missing.

Figure 23 continues the same continuous projector at 8B, and $\kappa(E)$ is again missing. Figure 24 collects the parameter gradient magnitudes on Qwen3-VL 2B, 4B, and 8B. All three stay of order one, and the embedding scale channel remains missing.

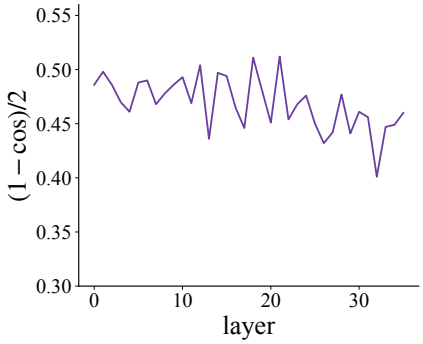


Figure 23. Layerwise conflict on Qwen3-VL-8B. The projector is continuous, and $\kappa(E)$ is missing.

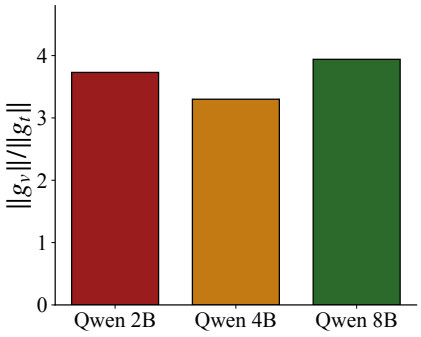


Figure 24. Parameter gradient magnitude on Qwen3-VL 2B, 4B, and 8B. All three stay of order one, and the embedding scale channel remains missing.

A last layer probe that uses κ_{op} on the Chameleon style joint softmax gives an operator product 0.117 against a clean share 2.83. The closed identity equation 3 still holds, with geometric error 1.000 and a visual cross term at machine zero. The residual and operator numbers are $R_v = 1.000$, $R_t = 1.080$, $\kappa_{\mathrm{op}} = 0.029$, Frobenius scale $\kappa_F = \|E_v\|_F^2/\|E_t\|_F^2 = 0.261$, and $\chi = 1119$. The operator bound is loose in the opposite direction from Table 5, because the visual rows are numerous and individually small while residual energy is comparable across the two blocks. Lemma 3 accounts for that gap.

A product that treats the last linear map of the Janus generation network as E^{gen} misses the measured ratio by a factor of about 23. On that probe, $\rho = 0.91$, $R_v/R_t \approx 0.99$, $\kappa_{\mathrm{op}} = 0.020$, and the mean cross entropy ratio is 0.0086. The generation head is a projection, a Gaussian error linear unit, and a vision head. The row shows that a dual head is a different object from a joint softmax, and it is not a numerical test of equation 4.

Token counts do not recover the public magnitude ratios, as already noted in Section 5. Rao and Rachuri report a gap between generation and understanding of 11 to 14 on a different packing and a different diagnostic (Rao & Rachuri, 2026). That band is not inverted for ρ, R, or κ_{eff}. Proposition 1 is the reason a single ratio cannot be inverted. Figure 25 repeats the Janus-Pro 1B and Show-o pair already used in the main text. Neither profile is a $\sin^2$ curve.

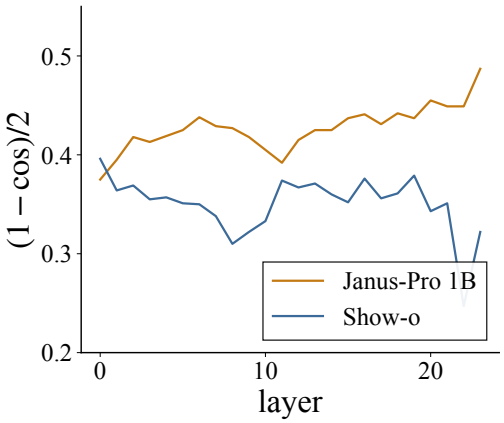


Figure 25. Janus-Pro 1B and Show-o with official MAGViTv2 codes. Neither profile is a $\sin^2$ curve.

H.1 Layerwise public profiles

Tables 23, 24, and 25 record the conflict $(1 - \cos)/2$ on frozen public models. They are existence profiles. Janus-Pro 1B rises toward the last layers, Show-o is flatter and slightly lower at the end, and the joint softmax checkpoint sits near $1/2$ at every depth. Isolation remains defined on each profile.

Table 23. Janus-Pro 1B, parameter gradient conflict $(1 - \cos)/2$ by layer.

layer	conflict	layer	conflict	layer	conflict
0	0.375	8	0.427	16	0.441
1	0.395	9	0.418	17	0.431
2	0.418	10	0.405	18	0.442
3	0.413	11	0.392	19	0.437
4	0.419	12	0.415	20	0.455
5	0.425	13	0.425	21	0.449
6	0.438	14	0.425	22	0.449
7	0.429	15	0.437	23	0.487

Table 24. Show-o with official MAGViTv2 codes, conflict $(1 - \cos)/2$ by layer.

layer	conflict	layer	conflict	layer	conflict
0	0.396	8	0.310	16	0.376
1	0.364	9	0.322	17	0.356
2	0.369	10	0.333	18	0.361
3	0.355	11	0.374	19	0.379
4	0.357	12	0.367	20	0.343
5	0.351	13	0.371	21	0.351
6	0.350	14	0.360	22	0.247
7	0.338	15	0.352	23	0.322

Table 25. Chameleon style joint softmax checkpoint, conflict $(1 - \cos)/2$ by layer.

layer	conflict	layer	conflict	layer	conflict
0	0.516	11	0.490	22	0.504
1	0.548	12	0.503	23	0.504
2	0.554	13	0.505	24	0.504
3	0.521	14	0.509	25	0.504
4	0.553	15	0.506	26	0.504
5	0.549	16	0.505	27	0.504
6	0.549	17	0.505	28	0.498
7	0.526	18	0.507	29	0.507
8	0.516	19	0.509	30	0.504
9	0.509	20	0.506	31	0.540
10	0.505	21	0.505		

Table 26 continues the Janus-Pro family at 7B, and conflict again rises toward the last layers. Tables 27, 28, and 29 record Qwen3-VL 2B, 4B, and 8B. The three widths sit in a similar band, the visual head is continuous, and none of the profiles is a $\sin^2$ curve. The isolation map does not require that shape.

Table 26. Janus-Pro 7B, parameter gradient conflict $(1 - \cos)/2$ by layer.

layer	conflict	layer	conflict	layer	conflict
0	0.301	10	0.453	20	0.468
1	0.319	11	0.429	21	0.474
2	0.417	12	0.437	22	0.472
3	0.417	13	0.430	23	0.482
4	0.431	14	0.432	24	0.496
5	0.434	15	0.380	25	0.483
6	0.447	16	0.446	26	0.463
7	0.445	17	0.428	27	0.485
8	0.435	18	0.433	28	0.520
9	0.439	19	0.432	29	0.494

Table 27. Qwen3-VL-2B, conflict $(1 - \cos)/2$ by layer. The visual head is continuous.

layer	conflict	layer	conflict	layer	conflict
0	0.438	10	0.501	20	0.475
1	0.479	11	0.474	21	0.474
2	0.463	12	0.505	22	0.489
3	0.467	13	0.491	23	0.497
4	0.465	14	0.519	24	0.484
5	0.464	15	0.483	25	0.459
6	0.463	16	0.473	26	0.492
7	0.478	17	0.462	27	0.425
8	0.466	18	0.476		
9	0.462	19	0.452		

Table 28. Qwen3-VL-4B, conflict $(1 - \cos)/2$ by layer. Same continuous projector protocol as the 2B row.

layer	conflict	layer	conflict	layer	conflict
0	0.445	12	0.463	24	0.446
1	0.443	13	0.444	25	0.421
2	0.449	14	0.479	26	0.442
3	0.472	15	0.487	27	0.386
4	0.464	16	0.493	28	0.409
5	0.468	17	0.474	29	0.424
6	0.435	18	0.443	30	0.413
7	0.471	19	0.460	31	0.435
8	0.464	20	0.446	32	0.381
9	0.441	21	0.496	33	0.418
10	0.489	22	0.463	34	0.389
11	0.466	23	0.473	35	0.329

Table 29. Qwen3-VL-8B, conflict $(1 - \cos)/2$ by layer. Same continuous projector protocol as the 2B and 4B rows.

layer	conflict	layer	conflict	layer	conflict
0	0.486	12	0.504	24	0.476
1	0.498	13	0.436	25	0.450
2	0.486	14	0.497	26	0.432
3	0.470	15	0.494	27	0.442
4	0.461	16	0.465	28	0.477
5	0.488	17	0.446	29	0.441
6	0.490	18	0.511	30	0.461
7	0.468	19	0.481	31	0.456
8	0.478	20	0.451	32	0.401
9	0.486	21	0.512	33	0.447
10	0.493	22	0.454	34	0.449
11	0.469	23	0.468	35	0.460

H.2 Dual head and joint softmax probes

The three public geometries are different objects. A joint softmax last layer has a single matrix E and a single residual r. A dual head has a language model unembedding on text positions and a generation map on visual positions. A continuous head has a regression Jacobian and no κ_{eff}. Table 30 records one last layer probe on each object. The joint softmax row is a valid test of equation 3. The dual head row is a different contraction, and the continuous row has no embedding scale channel. Joint softmax share in the table is the clean residual object. The dual head hidden ratio is $\|\nabla_h L_v\|^2/\|\nabla_h L_t\|^2$ through a generation network, and the operator product is not an inversion on either row.

Table 30. Last layer probes on frozen public weights. Joint softmax share is the clean residual object. The dual head hidden ratio is $\|\nabla_h L_v\|^2/\|\nabla_h L_t\|^2$ through a generation network. The operator product is not an inversion on either row.

model	head	clean or hidden share	operator product	geom.
Chameleon style, 7B	joint softmax	2.83	0.12	24
Janus-Pro 1B	dual MLP	0.0086	0.20	23
Show-o MAGViTv2	joint softmax	0.416	n/a	n/a
Qwen3-VL 2B	continuous	3.73	missing	n/a
Qwen3-VL 4B	continuous	3.30	missing	n/a
Qwen3-VL 8B	continuous	3.94	missing	n/a

The dual head hidden ratio 0.0086 and the parameter ratio 2.78 are two different contractions of the same Janus-Pro 1B head, and neither inverts equation 4. Show-o token counts are $N_v = 48$ and $N_t = 8$ on the probe pack, and they do not recover the parameter ratio 0.416. Figure 24 repeats the three Qwen3-VL magnitudes. Isolation remains defined on all three geometries, because it is a statement about the Jacobian and does not require a joint matrix E.