TOKENIZER DISTORTION AS A FLOOR: COMPUTE-OPTIMAL VISUAL COMPRESSION FOR MULTIMODAL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

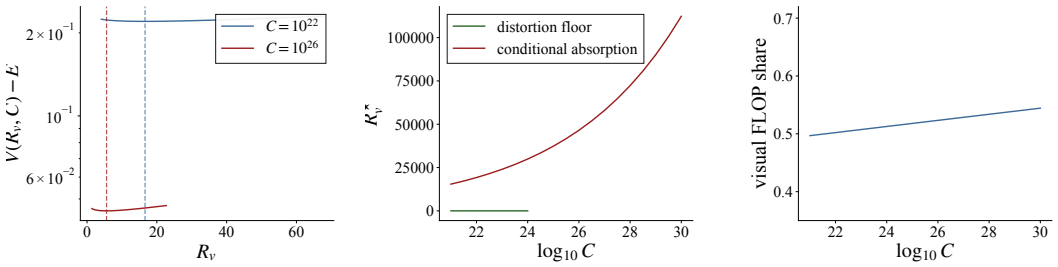
Multimodal scaling laws commonly regard the visual tokenizer as fixed infrastructure and allocate a FLOP budget solely between model parameters and tokens. This convention excludes the visual compression rate from resource allocation. We redefine multimodal pretraining as a joint optimization over parameters, text tokens, visual tokens, and the visual compression rate, with two empirically measured tokenizer curves entering a Chinchilla-style loss. The first curve captures the predictive information carried by each retained visual token. The second curve quantifies the distortion introduced by discarding complementary visual content. The core structural question asks whether larger parameter budgets can absorb this distortion. Across the suite of controlled decoders we train, the residual gap under matched effective visual data remains statistically invariant to model width. This observation identifies tokenizer distortion as a loss floor rather than a term reducible by increased model capacity. Because the floor stationarity condition lacks direct width elasticity, the optimal compression rate relaxes as the compute budget grows. When the efficiency curve is nearly flat, as with CLIP, the interior optimality condition vanishes and the optimum lies at the finest feasible code. A floor model fitted on small compute budgets recovers discrete configurations that remain optimal on held out larger budgets, and it outperforms direct transfer of compression rates from small budget regimes. Generative objectives raise the effective distortion weight, which can push the unconstrained optimal rate below unity so that the feasible optimum settles at the domain boundary. Pretraining optimizes over a static corpus while inference optimizes per query, so the corresponding Lagrangians are not scalar multiples of one another.

1 INTRODUCTION

Vision language pretraining spends a large share of compute on visual tokens. LLaVA 1.5 keeps 576 patches on a fixed grid (Liu et al., 2024), and dynamic resolution can spend tens of thousands of tokens on one image (Wang et al., 2024a). Language scaling splits a FLOP budget between parameters and tokens and leaves the tokenizer outside that split (Kaplan et al., 2020; Hoffmann et al., 2022). Replications then showed that this split moves when the FLOP account changes (Porian et al., 2024; Gadre et al., 2025; Li et al., 2025b). Multimodal laws kept that tokenizer frozen while they refined the remaining allocation. Mixed modal scaling fitted a joint loss at a fixed visual code (Aghajanyan et al., 2023). Native and contrastive scaling then compared fusion, encoders, and datasets at that same fixed rate (Shukor et al., 2025; Tian et al., 2026; Nezhurina et al., 2026). A visual rate copied from a small run is therefore an unexamined share of the budget.

Deployed systems have already chosen that rate. Flamingo, BLIP 2, and LLaVA Mini freeze a small set of visual tokens (Alayrac et al., 2022; Li et al., 2023; Zhang et al., 2025). InternVL and Qwen2 VL instead raise resolution as capacity grows (Chen et al., 2024b; Wang et al., 2024a). These recipes look inconsistent while R_v is treated as infrastructure. Once R_v is a decision, they are two corners of one allocation. One spends the marginal FLOP on language capacity, and the other spends it on resolution. Pruning and nested codes then drop visual tokens from one prompt, which is denser than inference requires (Chen et al., 2024a; Yang et al., 2025; Shang et al., 2025; Alvar et al., 2025; Cai et al., 2025; Dong et al., 2026). Layerwise dropping cuts cost (Xing et al., 2025; Zhao et al., 2025a).

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.



(a) V at two budgets. (b) $R_v^\star$ versus compute. (c) Visual FLOP share.

Figure 1: Optimal visual rate under a distortion floor. The value is U shaped, the rate falls with compute, and the visual budget share rises.

Both lines leave the pretraining rate unset. Given a budget C, what mix of visual rate, language data, and parameters is compute optimal?

We cast that question as a constrained allocation over (N, D_t, D_v, R_v), where $R_v \geq 1$ counts native patches per retained visual token and a larger R_v is a coarser code. Two measured curves enter a Chinchilla loss. The efficiency $\eta(R_v)$ is the predictive information in each retained visual token, and the distortion $\mathcal{D}(R_v)$ is the penalty for discarding the complement. Under the regularity conditions of Section 3, the first order system is strictly convex in logarithmic coordinates, so each fixed rate has a unique allocation. That allocation equates the marginal value of the last text token and the last visual token, and it places R_v at an interior optimum for the budget. Data mixing laws share probability mass across corpora and contain no distortion curve, so they cannot move a visual rate with compute (Ye et al., 2025). Learned tokenizers take the complementary route and improve the code itself (Zhao et al., 2025b; Yue et al., 2026). Here η and $\mathcal{D}$ are measured inputs, and the decision is the budget share of the code.

Relative to Hoffmann et al. (2022), the new objects are the pair $(\eta, \mathcal{D})$ and a rate that can move with compute. Absorption of distortion by extra parameters, through an exponent α_d, is only a conditional extension. In every controlled decoder we train, the residual gap at matched effective visual data is flat in N, so distortion behaves as a floor with no direct elasticity in width. A larger decoder therefore does not justify a coarser code. Along a compute ray the data volume grows, which makes residual quality more valuable than a cheaper coarse code, and the optimal rate falls as C increases. Because the floor does not depend on N, that sign is fixed by the data channel alone. When η is nearly constant, as in contrastive language image pretraining, the interior condition disappears and the optimum lies at the finest feasible code (Radford et al., 2021). Figure 1 summarises this geometry.

At a fixed budget the value in Figure 1(a) is U shaped in R_v, so neither the coarsest nor the finest code is free, and the minimum moves when C grows by four decades. Figure 1(b) shows the floor curve declining, and Figure 1(c) shows the visual share of the budget rising because $\rho = \partial \log R_v^\star / \partial \log C < 0$, until the rate reaches the finest feasible code. A tokenizer chosen on a small budget is therefore too coarse once C grows. We therefore give an allocation theory for (N, D_t, D_v, R_v), a protocol that can accept or reject absorption, and a closed loop test of whether a small budget fit selects the later winner.

The main contributions of this work are summarised below.

- We place R_v inside a Chinchilla allocation over (N, D_t, D_v, R_v), with the measured curves $\eta(R_v)$ and $\mathcal{D}(R_v)$ as inputs. When distortion is a floor, the optimal code becomes finer as C grows and the visual budget share rises with it. A flat efficiency curve removes the interior balance and forces the finest feasible code.
- A matched residual separates that floor from an absorption term with elasticity in N. The remainder is flat on three controlled worlds, so the protocol accepts the floor. The same comparison rejects a three point absorption artefact on the multiplicative world.
- A floor fit at small compute selects the discrete winner on later held out budgets and beats copying the pilot rate. The same balance gives a bounded dual for generation, and it separates the pretraining Lagrangian from the per query inference Lagrangian.

2 Related Work

Compute optimal allocation began as a language model question. Kaplan et al. (2020) and Hoffmann et al. (2022) showed that a fixed FLOP budget should be split between parameters and tokens, and they left the tokenizer outside that split. Replications then refined the exponents and showed that the fitted split moves when the FLOP account changes (Besiroglu et al., 2024; Porian et al., 2024; Gadre et al., 2025; Li et al., 2025b). Data constrained scaling added repetition and overtraining, and it still kept the vocabulary frozen (Muennighoff et al., 2023). Vision scaling treated patch size in the same way, as a property of the code rather than a quantity chosen with compute (Zhai et al., 2022). Each of these steps tightened the parameter and token tradeoff. Across the whole line, the discrete code stayed outside the allocation.

Multimodal scaling inherited that frozen code and studied the loss on the resulting slice. Aghajanyan et al. (2023) fitted a mixed modal loss in parameters and tokens at a fixed visual tokenizer. Native scaling then compared early and late fusion and scaled the vision tower with the language model, again with the compression rate held fixed (Shukor et al., 2025; Tian et al., 2026). Contrastive scaling laws used that same fixed code to compare models and datasets across compute (Nezhurina et al., 2026). These constant FLOP fronts therefore answer how loss falls once the tokenizer is chosen, and they leave open how the tokenizer should move with C. A related line improves the code itself. Learned tokenizers train a better discrete representation (Zhao et al., 2025b; Zhou et al., 2025; Yue et al., 2026), while classical rate distortion and the information bottleneck also treat the code as an object to be learned (Cover & Thomas, 2006; Tishby & Zaslavsky, 2015). Lemma 1 specialises that classical tail to a power law field, which is why a local chart $\mathcal{D} \propto R_v^{\psi}$ is a reasonable operating assumption. Data mixing laws address a different split. They allocate probability mass across text corpora, and they contain neither an encoder surcharge nor tokenizer distortion, so a visual rate cannot move with compute inside that theory (Ye et al., 2025). In the allocation studied here, η and $\mathcal{D}$ are measured inputs, alongside multitask visual fusion (Lou et al., 2026).

Inference compression prices a finished model. The inference optimal account minimises a per query cost in the text and visual token counts of that query, and it therefore prefers fewer visual tokens and more parameters (Li et al., 2025a). Pruning, nested codes, and layerwise dropping carry out that preference on a trained network (Chen et al., 2024a; Yang et al., 2025; Shang et al., 2025; Alvar et al., 2025; Cai et al., 2025; Dong et al., 2026; Xing et al., 2025; Zhao et al., 2025a). Reported reductions of 45% to 94% of LLaVA 1.5 visual tokens, with only a small task change, indicate that a typical prompt is denser than this per query cost requires. Those methods compress a query. The pretraining rate is a separate decision, so a preference for a finer code during training is compatible with heavy pruning at inference time. The same distinction organises public training recipes. Flamingo, BLIP 2, and LLaVA Mini freeze a small set of visual tokens, which is a $\rho = 0$ design (Alayrac et al., 2022; Li et al., 2023; Zhang et al., 2025). InternVL and Qwen2 VL raise resolution with capacity, which is the sign of Theorem 4 (Chen et al., 2024b; Wang et al., 2024a). Unified models keep denser visual codes than understanding models of similar size, which is the direction of Theorem 6 (Team, 2024; Wang et al., 2024b; Zhou et al., 2025). These configurations are the qualitative pattern that an allocation law has to match. Resolution ablations at a frozen N are slices of the value $V(R_v, C)$ at one budget. They can reveal its U shape, and the compute elasticity ρ appears only when C varies, which is the ranking statement of Theorem 5.

3 The Allocation Problem

A pretraining configuration is a tuple (N, D_t, D_v, R_v). Here N is the decoder parameter count, D_t and D_v are the text and visual token counts, and $R_v \geq 1$ is the visual compression rate, defined as the number of native patches divided by the number of retained visual tokens. A larger R_v is a coarser code, and $R_{v,\min} = 1$ denotes the finest feasible code. The visual encoder enters through the relative cost $\kappa = 1 + N_{\text{enc}}/N$, where N_{enc} is the encoder parameter count. The compute constraint is the standard linear FLOP account

$$C \;=\; 6N\big(D_t + \kappa D_v\big). \tag{1}$$

Quadratic attention adds a higher order term. Appendix D treats that term as a robustness check. Once D_v is given, R_v does not appear in the constraint, so the stationarity condition in the rate is unchanged.

The default loss separates capacity, text, visual tokens, and tokenizer distortion,

$$L \;=\; E + AN^{-\alpha} + B_t D_t^{-\beta} + B_v\big(D_v\,\eta(R_v)\big)^{-\beta} + \gamma\,\mathcal{D}(R_v). \tag{2}$$

The first four terms are a two source Chinchilla law. Here E is the irreducible loss, A is the capacity coefficient, and B_t, B_v are the data coefficients. The fifth term is tokenizer distortion, with weight γ. Extra parameters do not appear in it, so the term is a floor. The effective visual data volume is $\tilde{D}_v = D_v\,\eta(R_v)$. Folding $\mathcal{D}$ into $\tilde{D}_v$ and fitting a single visual data term leaves a systematic residual, which is why the floor is kept separate. A conditional extension replaces the last term by $\gamma\mathcal{D}(R_v)\,N^{-\alpha_d}$ with $\alpha_d \geq 0$. That extension is used only when the protocol of Section 4 accepts it.

Assumption 1 (Regularity)**.** *The four resource terms are additive at second order near the operating point. On that interval, η and $\mathcal{D}$ admit local power charts $\eta = \eta_0 R_v^{\phi}$ with $0 \leq \phi < 1$ and $\mathcal{D} = d_0 R_v^{\psi}$ with $\psi > 0$. At matched $\tilde{D}_v$, any residual gap that remains after the text term and the data term is either constant in N or, when absorption is under test, scales as $N^{-\alpha_d}$. The map from $\log R_v$ to $\log\eta$ is concave, and $\mathcal{D}$ is increasing and log convex on the operating interval.*

Additivity is the working hypothesis that makes the first order system separable. A Kaplan interaction $N^{-\alpha}\tilde{D}_v^{-\beta}$ is common to both arms of a matched comparison, so it cancels in the gap and cannot be mistaken for absorption. A product of capacity and distortion does enter that gap. Section 4 therefore reads the matched residual rather than a single global fit. If the local charts fail away from the operating interval, the measured curves $(\eta, \mathcal{D})$ are used as inputs and the stationarity system is continued numerically. A small multiplicative perturbation of L moves an interior minimum continuously and cannot flip the sign of ρ while the residual remains flat. Discrete vector quantised codes sit outside the charts used for closed forms, and they remain valid as measured inputs to the same continuation.

Lemma 1 (Spectral tail)**.** *Let the image be a Gaussian field on a two dimensional lattice with spectrum $S(\omega) \propto |\omega|^{-\nu}$ and $\nu > 2$. An orthogonal transform code that keeps the leading fraction $1/R_v$ of the modes has reconstruction distortion $\mathcal{D}(R_v) \asymp R_v^{(\nu-2)/2}$. The mutual information per retained token increases with R_v and eventually saturates. A local power chart for η is the first order expansion of that curve on a compact interval away from $R_v = 1$ and from $R_v \to \infty$.*

The exponent $(\nu-2)/2$ is the classical Gaussian tail for a power law spectrum (Cover & Thomas, 2006; Geisler, 2008; Simoncelli & Olshausen, 2001). It is an identity on the synthetic fields of Section 5 and only a local description of natural images, which is why the CLIP and LLaVA charts are measured rather than taken from the formula. The surcharge κ rescales the mix and does not change the sign of ρ. When the vision tower scales in step with the decoder, κ is nearly constant. When the tower is frozen, κ tends to one, the visual stream becomes cheaper, and $D_v^\star/D_t^\star$ rises.

4 Compute Optimal Allocation

The statements below use the floor equation 2 unless a residual has already accepted absorption, and the proofs are in Appendix B. At a fixed rate the inner allocation is unique and has an envelope in R_v. That envelope fixes the mix of the two streams and, when the charts are interior, the rate itself. Along a compute ray the same balance fixes the sign of ρ.

Theorem 1 (Well posedness)**.** *In logarithmic coordinates, minimising equation 2 subject to equation 1 at a fixed R_v is strictly convex. For every $R_v \geq 1$ and every $C > 0$ there is a unique inner minimiser $(N, D_t, D_v)(R_v, C)$. The value $V(R_v, C)$ is continuously differentiable in R_v, and*

$$\frac{\partial V}{\partial R_v} \;=\; \Big[-\beta B_v \tilde{D}_v^{-\beta}\,\tfrac{\eta'(R_v)}{\eta(R_v)} + \gamma\,\mathcal{D}'(R_v)\Big]_{(N^\star, D_t^\star, D_v^\star)}. \tag{3}$$

Strict convexity follows because each resource term is exponential in the log variables, while the compute set is convex through its log sum exp constraint. The inner problem therefore has one minimiser. Differentiability of V is the envelope theorem. Once D_v is given, the constraint does not depend on R_v, so the derivative in the rate contains no multiplier.

Theorem 2 (Equal marginal mix)**.** *At an interior optimum the text and visual streams satisfy*

$$\frac{\beta B_t}{{D_t^\star}^{\beta+1}} \;=\; \frac{\beta B_v\,\eta(R_v^\star)^{-\beta}}{\kappa\,{D_v^\star}^{\beta+1}}\,. \tag{4}$$

When $\kappa \to 1$ and $D_v = 0$, the identity reduces to the two data balance of Hoffmann et al. (2022). When R_v is held fixed, the same reduction holds with visual coefficient $\tilde{B}_v = B_v \eta(R_v)^{-\beta}$.

The identity says that the last text token and the last visual token buy the same reduction in loss, after the encoder surcharge and the information per token have been paid. It does not ask the two streams to have the same raw count. Deeper compression raises η and therefore lowers the visual count that is optimal at a given C. A one to one token mix copied from a language only recipe solves a different first order condition. Whenever $\eta(R_v^\star)$ and κ are not both near one, that copy is a first order error in the mix.

Theorem 3 (Interior sweet spot)**.** *Suppose η is strictly increasing and log concave in $\log R_v$, $\mathcal{D}$ is strictly increasing and convex, and the boundary condition at the finest code*

$$\beta B_v \tilde{D}_v^{-\beta} \frac{\eta'(1)}{\eta(1)} \;>\; \gamma \mathcal{D}'(1) \tag{5}$$

holds. Then $V(\,\cdot\,, C)$ has a unique interior minimiser $R_v^\star(C) \in (1, \infty)$. Under the local charts,

$$R_v^\star \;=\; \Big(\frac{\beta\phi B_v}{\gamma\psi d_0 \eta_0^\beta}\Big)^{1/(\psi+\beta\phi)} D_v^{-\beta/(\psi+\beta\phi)}. \tag{6}$$

In particular $\partial \log R_v^\star / \partial \log N = 0$ and $\partial \log R_v^\star / \partial \log \tilde{D}_v = -\beta/\psi < 0$. If $\phi = 0$, then equation 3 cannot vanish and $R_v^\star = R_{v,\min}$.

The first order condition equates two margins. Keeping fewer but more informative tokens lowers the visual data term, and discarding more of the image raises the floor. Because the floor does not depend on N, a wider decoder does not by itself justify a coarser code. A larger corpus moves the optimum the other way. More data makes the remaining distortion expensive relative to the data term, so the optimum keeps more visual tokens. Equation equation 6 records that comparison, and Figure 1 draws the resulting rate at two budgets.

Theorem 4 (Floor scaling)**.** *Along the compute optimal ray, write $N^\star \propto C^n$, $D_t^\star \propto C^{d_t}$, $D_v^\star \propto C^{d_v}$, and $R_v^\star \propto C^\rho$, where ρ is the elasticity of the optimal rate in compute. If the interior minimum of Theorem 3 is active, then*

$$\rho \;=\; -\frac{\alpha\beta}{\psi(\alpha+\beta)+\alpha\beta\phi} \;<\; 0\,, \quad n \;=\; \frac{\beta}{\alpha+\beta}\,(1+\phi\rho)\,, \quad d_v \;=\; 1-n\,, \quad d_t \;=\; d_v + \frac{\beta\phi}{\beta+1}\,\rho\,. \tag{7}$$

If $\phi = 0$, the rate is already at $R_{v,\min}$, so $\rho = 0$, and the remaining exponents return to the symmetric two data law $n = d_t = d_v = \beta/(\alpha+\beta)$ at $\kappa = 1$.

The sign of ρ is fixed by the data channel of Theorem 3, read along a Chinchilla ray, before any comparison of α_d with α. As C grows, the visual share $s_v^\star = \kappa D_v^\star / (D_t^\star + \kappa D_v^\star)$ therefore rises until the rate reaches the finest code. An equal split that is held fixed while C grows is optimal only when $\phi\rho = 0$.

Absorption can be restored, but only as a corollary and only after the residual has accepted it. If the fifth term is $\gamma\mathcal{D}N^{-\alpha_d}$ and $\alpha_d > \alpha$, the leading balance in ∂_N changes and ρ changes sign with it. The next proposition records that case. It is not the default.

Proposition 1 (Conditional absorption)**.** *Suppose the residual gap at matched $\tilde{D}_v$ scales as $N^{-\alpha_d}$ with $\alpha_d > \alpha$, and that additivity holds at the level required by Lemma 2. Then* $\operatorname{sign}(\rho) = \operatorname{sign}(\alpha_d - \alpha)$*. If the residual is flat in N, absorption is not identified and Theorem 4 remains the law that is used.*

Lemma 2 (Residual decomposition)**.** *Fix rates $R_v^{(A)} < R_v^{(B)}$ and match $\tilde{D}_v$. The observed gap splits as $\Delta L = \Delta_{\text{text}} + \Delta_{\text{data}} + \gamma\Delta\mathcal{D} + \operatorname{Rem}(N)$. A Kaplan interaction cancels in ΔL. If* Rem *has zero slope in* $\log N$*, the floor is identified. If* Rem *decays in N, then α_d is identified. If* Rem *rises in N, neither a nonnegative α_d nor Proposition 1 is in force.*

The protocol is a sequence, not a single regression. The loss equation 2 is fitted first on a cheap IsoFLOP table, and the matched residual is read second. A fit that forces $\alpha_d \geq 0$ reports a boundary zero even when the residual is rising, so the slope is left free. Only a decaying residual licenses Proposition 1, which is the order tested by the controlled decoder in Section 5.

Theorem 5 (Rankings do not transfer)**.** *Let $R_v^{(1)} < R_v^{(2)}$ be two feasible rates. Under Theorem 4, the discrete minimiser at the smaller budget is coarser than the optimum at the larger budget. The ranking carries from the smaller budget to the larger one if and only if $\rho = 0$.*

A tokenizer chosen on a pilot run is therefore too coarse for the production budget. The compute needed to leave a factor of two bin around $R_v^\star(C_0)$ is $C_{\text{flip}} = C_0\, 2^{1/|\rho|}$. On the illustration chart that factor is about 65, and on the CLIP chart it is about 43. Both exceed the factor of thirty spanned by the IsoFLOP grid in Section 5, so a short ladder can look flat even though $\rho \neq 0$. A discrete minimiser that jumps among three neighbouring powers of two is compatible with a negative elasticity and with a zero elasticity. The ranking statement that survives inside the window is Theorem 5, and the closed loop in Section 5 asks whether a fit that never sees the later budget can still select the later winner.

Theorem 6 (Bounded dual)**.** *Suppose the objective mixes understanding and generation losses as $L_{\text{und}} + wL_{\text{gen}}$, and that generation puts more weight on distortion. Write γ_u and γ_g for those weights, with $\xi = \gamma_g/\gamma_u > 1$ and $\gamma_{\text{eff}} = \gamma_u(1 + w\xi)$. Relative to the understanding only rate R_v^{und}, the unconstrained unified rate satisfies*

$$\frac{R_v^{\text{uni}}(w)}{R_v^{\text{und}}} \;=\; (1 + w\xi)^{-1/(\psi+\beta\phi)}\,. \tag{8}$$

If the right hand side pushes the rate below $R_{v,\min}$, the bounded solution is $R_v^\star = R_{v,\min}$, and the mix of Theorem 2 is taken at that frozen rate.

A unified model should therefore compress no more than an understanding only model of similar N, and often strictly less. At $w = 1$ and $\xi = 3$ the unconstrained ratio is about 0.29. The weight at which this ratio hits the bound is $w_\star = \xi^{-1}\big((R_v^{\text{und}}/R_{v,\min})^{\psi+\beta\phi} - 1\big)$. On the chart of Section 5 one has $w_\star \approx 3.75$ at $C = 10^{24}$, so a generation weight of one is still interior there. Once C is large enough that $R_v^{\text{und}} \approx 3.2$, the same weight pushes the unconstrained product below one. The design rate is then the boundary, and the unconstrained number is not a feasible code.

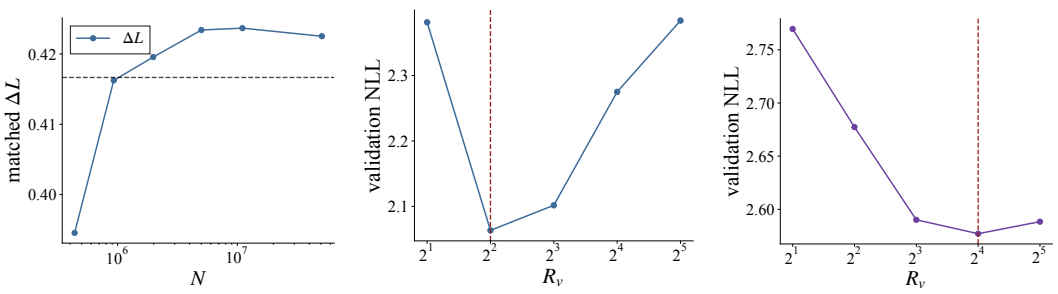
The same bound is why a generation tokenizer cannot be chosen by copying an understanding chart and dividing by a constant. The constant is the closed form ratio, but it multiplies an understanding rate that itself falls with C. At large enough compute the product leaves the feasible set, and the only remaining decision is the mix of Theorem 2 at $R_v = 1$.

Proposition 2 (Pretraining versus inference)**.** *Pretraining prices the corpus (D_t, D_v). A per query budget prices the text and visual token counts (n_t, n_v) of that query through $C_{\text{inf}} \propto N(n_t + n_v)$. The two Lagrangians are not scalar multiples of each other.*

A pretraining optimum can therefore ask for a rising D_v while every deployed query asks for a small n_v (Li et al., 2025a). Pruning tokens at a finished 7B checkpoint does not contradict the pretraining minimum. It solves the inference problem, whose constraint is the prompt rather than the corpus.

The design rule follows the same order as the theorems. Measure $(\eta, \mathcal{D})$ on the frozen tokenizer, and fit $(E, A, B_t, B_v, \alpha, \beta, \gamma)$ on a small IsoFLOP grid at several rates. Read the matched residual next. A flat residual selects Theorem 4, and a residual that decays with $\alpha_d > \alpha$ selects Proposition 1. The continuous prediction is then snapped to the discrete sizes and rates that will actually be trained, and the small grid minimiser is kept only when $\rho \approx 0$. If generation enters the objective, Theorem 6 supplies the rate and the result is clipped at $R_{v,\min}$. If the deployed constraint is latency, Proposition 2 is solved again for n_v. On the illustration chart with $\alpha = \beta = 0.35$, $\phi = 0.3$, $\psi = 1$, and $\kappa = 1.2$, a budget of 10^{24} FLOPs gives $R_v^\star \approx 9.6$, and a budget of 10^{21} FLOPs gives $R_v^\star \approx 22$. Compression relaxes as compute grows.

Because equation 6 depends on D_v and not on N, two runs that share a tokenizer and a token budget but differ in width have the same $R_v^\star$ under the floor. When a public ablation appears to depend on width, the change is in $\tilde{D}_v$ or in the measured $(\eta, \mathcal{D})$, and Lemma 2 separates those cases. A continuous displacement inside a factor of two costs about one percent of reducible loss near $|\log_2(R_v/R_v^\star)| \approx 1.08$. Neighbouring powers of two lie outside that bin, so copying one of them is a first order error on the decoder of Section 5.



(a) Aligned residual. The matched gap is flat in N.

(b) Aligned slice at $C = 3 \times 10^{13}$. The minimum is interior.

(c) Multiplicative slice at $C = 10^{13}$. The U is shallower.

Figure 2: Floor identification and constant FLOP slices. (a) is the aligned matched gap. (b) and (c) are validation NLL. Dashed lines mark the mean gap or the discrete minimum.

5 Experiments

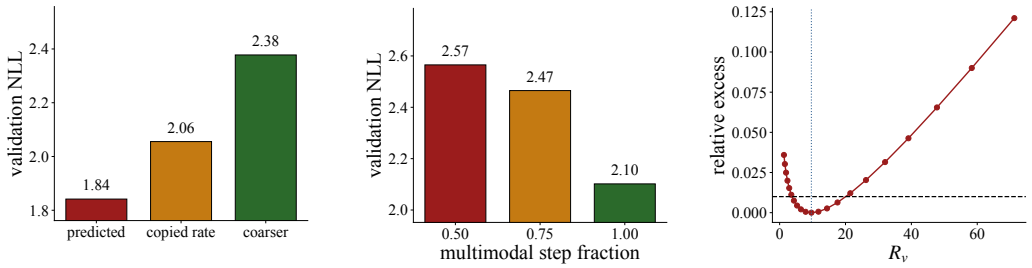
The experiments follow the three claims in order. The controlled decoder asks whether tokenizer distortion is a floor, and whether a fit at small compute can select a later discrete winner. Frozen encoders then ask where a real tokenizer sits on that law.

5.1 Residual floor on a controlled decoder

The controlled decoder is a prefix transformer on an online stream of 16×16 Gaussian field images and 16 attribute tokens. The tokenizer keeps the leading $256/R_v$ transform coefficients, so a larger R_v is a coarser code with a known spectrum. Three worlds change what a wider decoder could hope to recover. In the aligned world the analysis basis matches the latent basis, and discarded coefficients are gone. In the rotated world a fixed orthogonal mix makes the discarded energy linearly readable. In the multiplicative world a hidden high frequency coordinate modulates the low frequency tokens. Width runs from 0.44M to 51M parameters. Training uses Adam with decoupled weight decay, learning rate 10^{-3}, weight decay 0.01, and batch size 128. Compute follows equation 1 at $\kappa = 1$, and validation is mean token negative log likelihood (NLL) on a fresh stream. Each IsoFLOP cell uses at least two seeds, and each residual arm uses three. If extra width could absorb distortion, the rotated or multiplicative remainder would fall in N. The aligned world is the conservative case, because it produces the largest floor, so a flat residual there is not an artefact of an unrecoverable code.

Lemma 2 is applied after the text term and the data matching error are removed. The remainder is then regressed on $\log N$. Figure 2(a) shows the aligned matched gap. On the pair $(4, 16)$ the mean gap is 0.417 nats and the floor term is 0.433. The remainder slope is 0.004, with interval $[-0.003, 0.004]$. The rotated pair has gap 0.303 nats against a floor of 0.259, and its slope 0.005 has interval $[-0.003, 0.003]$. The multiplicative gap is 0.036 nats, and its slope 0.006 has interval $[-0.004, 0.004]$. Every interval includes zero, and the relative error in matched $\tilde{D}_v$ is 10^{-4}. Pair $(2, 8)$ repeats the pattern. The aligned leftover is larger, yet its slope 0.027 has interval $[-0.018, 0.018]$ and does not decline in N. Fits that free α_d return slightly negative estimates, so the protocol reports a floor and does not quote an absorption exponent. A three point multiplicative fit had produced $\alpha_d \approx 1.20$. The six size grid with three seeds replaces that artefact by a flat residual, which is why equation 2 keeps the fifth term independent of N.

The same runs also test the geometry at fixed compute. Figure 2(b) and (c) are constant FLOP slices. The aligned slice is U shaped with an interior minimum, which is the geometry of Theorem 3. The multiplicative slice is shallower and its minimum is coarser, because less of the gap is irreducible floor. Across the grid, 80% of evaluated cells are interior. The discrete winners still jump among 4, 8, and 16. Aligned budgets read 8, 4, 4, and 8. Rotated budgets read 4, 4, 4, and 8. Multiplicative budgets read 16, 16, 8, and 16. A factor of thirty in C is shorter than the horizon on which $R_v^\star$ must leave a factor of two bin, so these jumps do not identify ρ. The ranking statement that fits inside the window is the closed loop below. Remaining worlds are in Appendix C.



(a) Closed loop at $C = 10^{14}$. (b) Mix scan. (c) Excess versus R_v.

Figure 3: Closed loop selection, the mix scan, and continuous excess. (a) trains the predicted cell against a copied rate and a coarser code. (b) varies the multimodal fraction, and (c) moves R_v continuously.

5.2 Closed loop selection

The floor is fitted only on runs with $C \leq 2 \times 10^{13}$, then scored on the discrete grid at larger budgets. Each predicted cell is trained against the pilot rate $R_v = 8$ and a coarser neighbour. On the aligned world the prediction is $N \approx 0.93$M and $R_v = 4$ at $C = 3 \times 10^{13}$, with NLL 1.91 against 2.08 and 2.38, and $N \approx 2.0$M and $R_v = 4$ at $C = 10^{14}$, with NLL 1.84 against 2.06 and 2.38. Both cells are the trained minimum. Rank correlations are 0.85 and 0.76, and the rotated world at $C = 10^{14}$ repeats the hit at 2.15 against 2.31 and 2.54. Figure 3 shows that aligned comparison. Copying the pilot rate costs 0.16 to 0.22 nats.

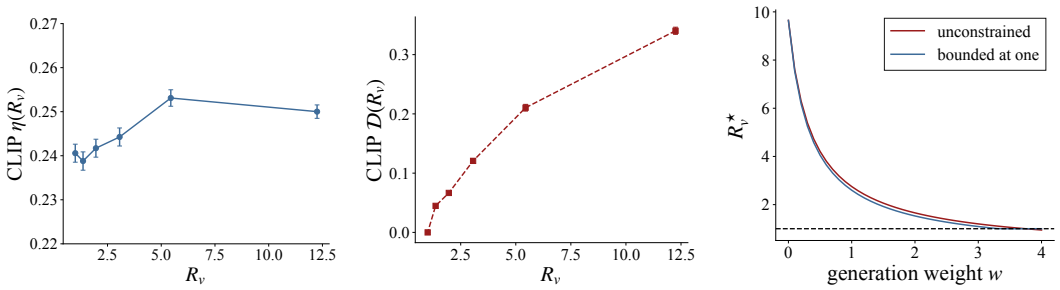
The mix scan asks whether text and visual tokens have an interior balance once the rate is fixed. The multimodal step fraction is the only control that moves D_t/D_v independently of R_v in this generator, and validation remains fully multimodal. At fractions 0.50, 0.75, and 1.00, aligned NLL is 2.57, 2.47, and 2.10, while rotated NLL is 2.61, 2.53, and 2.38. Figure 3(b) shows that both minima sit at fraction one. Theorem 2 is still the stationarity condition. This generator has no interior mix, because each sample already binds the visual attributes, and language only steps raise the loss. Figure 3(c) is the continuous counterpart of the discrete penalty in Table 1. Relative excess crosses one percent near $|\log_2(R_v/R_v^\star)| \approx 1.08$, so a deviation inside a factor of two is cheap. The neighbouring power of two sits outside that bin, and Table 1 is where that cost appears. Copying $R_v = 8$ raises trained NLL by 0.16 to 0.22 nats on every complete comparison, including the rotated run at $C = 10^{14}$. The rotated cell at $C = 3 \times 10^{13}$ is absent because the prediction was not on the trained grid. The table and the scan are therefore different margins. One changes the discrete code, and the other changes only the share of multimodal steps at a fixed code. Reusing the pilot rate pays the gap recorded in the table. Replacing multimodal steps with language only steps pays the penalty in Figure 3(b). Both costs are already visible on the budgets that were trained, so the comparison does not need a longer compute ladder.

Table 1: Held out closed loop cells. $\widehat{L}$ is predicted loss and L is trained NLL. Bold marks the discrete minimiser.

World	C	N	R_v	$\widehat{L}$	L
Aligned	3×10^{13}	**0.93M**	**4**	1.89	**1.91**
Aligned	3×10^{13}	0.93M	8	1.94	2.08
Aligned	3×10^{13}	2.0M	4	2.02	2.06
Aligned	3×10^{13}	2.0M	8	1.96	2.10
Aligned	3×10^{13}	2.0M	32	2.20	2.38
Aligned	10^{14}	**2.0M**	**4**	1.65	**1.84**
Aligned	10^{14}	2.0M	8	1.73	2.06
Aligned	10^{14}	5.2M	4	1.74	1.92
Aligned	10^{14}	5.2M	8	1.72	2.08
Aligned	10^{14}	5.2M	32	2.00	2.38
Rotated	10^{14}	**2.0M**	**4**	1.94	**2.15**
Rotated	10^{14}	2.0M	8	2.07	2.31
Rotated	10^{14}	5.2M	4	2.07	2.24
Rotated	10^{14}	5.2M	32	2.12	2.54

5.3 Frozen charts and production slices

Real encoders supply $(\eta, \mathcal{D})$ rather than a synthetic spectrum. CLIP ViT B/32 is read at six resolutions on 200 COCO validation images with 24 text queries (Radford et al., 2021; Lin et al., 2014). Distortion is one minus the cosine to the native 224 feature, and efficiency is the cosine between text and image features. Figure 4 shows both charts. The distortion slope is $\psi = 0.94$ with interval $[0.91, 0.97]$, close to the moderately decaying tail of Lemma 1. The efficiency slope is $\phi = 0.024$ with interval $[0.017, 0.031]$. Because ϕ is nearly zero, Theorem 3 places the optimum at the finest



(a) CLIP cosine versus rate. Slope $\phi = 0.024$ lies in $[0.017, 0.031]$.

(b) CLIP distortion versus rate. Slope $\psi = 0.94$ lies in $[0.91, 0.97]$.

(c) Unified rate versus weight. The bounded solution stays feasible.

Figure 4: CLIP charts and the bounded dual. (a) and (b) use CLIP ViT B/32 on 200 COCO images. (c) evaluates Theorem 6 at $C = 10^{24}$.

feasible code. LLaVA 1.5 repeats the pair on 100 COCO images. Distortion at $R_v = 9$ is 0.397 and falls to a numerical zero at $R_v = 1$, with $\phi \approx 0.026$ and $\psi \approx 2.29$. CLIP accuracy falls from 1.00 at $R_v = 1$ to 0.31 at $R_v = 12.3$, and retrieval Recall@1 falls from 0.88 to 0.58. Both track $\mathcal{D}$. The two encoders therefore sit in the flat efficiency corner of Theorem 3. Patch merging changes how many tokens are kept, and the information in each retained token stays nearly constant. Raising resolution with scale therefore pays to lower the floor, and freezing a small set of latent queries sets $\rho = 0$ by architecture.

The same charts determine the generation dual. At $C = 10^{24}$ and $w = 1$ the unconstrained ratio is 0.285, and the bounded rate is still above one. At $C = 10^{28}$ the understanding only rate has fallen to 3.25, the unconstrained unified rate is 0.93, and the bounded solution is exactly one, as shown in Figure 4(c). Unified models that keep denser visual codes sit on that sign (Zhou et al., 2025; Team, 2024; Wang et al., 2024b). Frozen Qwen3-VL Instruct at 2B, 4B, and 8B traces a U shaped pooled NLL in resolution on 100 COCO images (Bai et al., 2025). Absolute levels are not comparable across sizes, so the slice does not identify α_d. The discrete minimum moves from 168 at 2B to 448 at 4B and 8B. This is a decoder side reading of Theorem 3, not a pretraining optimum. Released token budgets in the Flamingo, BLIP 2, Qwen, LLaVA, InternVL, and Emu families do not fall with calendar year. Fixed bottleneck models sit near $\rho = 0$ by design, and families that raise tokens per image sit on the sign of Theorem 4. A continuation of equation 7 matches the closed form tails to a few thousandths, with measured $(n, \rho) = (0.490, -0.118)$ against $(0.482, -0.120)$, and the details are in Appendix D. Taken together, the residual protocol identifies the floor, the constant FLOP slices show an interior U at fixed C, and Table 1 shows that a cheap fit selects the later winner. CLIP and LLaVA occupy the flat efficiency corner, where that winner is the finest feasible code. A flat residual means the matched gap is stable in width, so the floor in equation 2 is the right default on this decoder. The external charts then occupy the remaining corner, where efficiency is too flat for an interior rate and a generation weight moves the choice onto the boundary. At $C = 10^{24}$ that weight still leaves an interior code, and the step to $C = 10^{28}$ is what pins the bounded solution at one.

6 Conclusion

Vision language pretraining spends one budget on parameters, text tokens, visual tokens, and the visual compression rate. The allocation extends a Chinchilla split by two measured tokenizer curves, efficiency and distortion. When distortion is a floor, the optimal code becomes finer as compute grows, and the visual share of the budget rises with it. A flat efficiency curve removes the interior balance and places the tokenizer at the finest feasible code. The floor is independent of width, so a larger decoder keeps the same rate. On controlled decoders a matched residual identifies that floor, and a fit at small compute selects the discrete winner on later budgets, ahead of copying the pilot rate. A generation term raises the weight on distortion and clips the feasible rate at the finest code. Inference prices a single query rather than the corpus, so pruning tokens at deployment is compatible with a pretraining run that keeps a finer code. The resulting rule is to measure the two curves, fit the floor on a cheap IsoFLOP grid, and snap the prediction to the sizes and rates of the target run.

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

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in neural information processing systems*, volume 35, pp. 23716–23736, 2022.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

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.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *International Conference on Learning Representations*, pp. 46254–46272, 2025.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. ISBN 978-0-471-24195-9.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multimodal coverage maximization for efficient inference of vlms. In *International Conference on Learning Representations*, pp. 48075–48099, 2026.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. In *International Conference on Learning Representations*, pp. 67661–67682, 2025.

Wilson S. Geisler. Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59(1):167–192, 2008.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PmLR, 2023.

Kevin Li, Sachin Goyal, Joao D Semedo, and Zico Kolter. Inference optimal vlms need fewer visual tokens and more parameters. In *International Conference on Learning Representations*, volume 2025, pp. 96066–96083, 2025a.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (mis)fitting scaling laws: A survey of scaling law fitting techniques in deep learning. In *International Conference on Learning Representations*, pp. 51134–51165, 2025b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. In *European conference on computer vision*, pp. 740–755, 2014.

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.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50358–50376, 2023.

Marianna Nezhurina, Tomer Porian, Giovanni Puccetti, Tommie Kerssies, Romain Beaumont, Mehdi Cherti, and Jenia Jitsev. Scaling laws for robust comparison of open foundation language-vision models and datasets. In *Advances in Neural Information Processing Systems*, volume 38, pp. 153979–154018, 2026.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100535–100570, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763, 2021.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–23. IEEE, 2025.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

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

Changyao Tian, Hao Li, Gen Luo, Xizhou Zhu, Weijie Su, Hanming Deng, Jinguo Zhu, Jie Shao, Ziran Zhu, Yunpeng Liu, et al. Navil: Rethinking scaling properties of native multimodal large language models under data constraints. In *Advances in Neural Information Processing Systems*, volume 38, pp. 85618–85646, 2026.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2015.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Conical visual concentration for efficient large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14593–14603. IEEE, 2025.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations*, pp. 82263–82287, 2025.

Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, Yi Wang, Limin Wang, and Yali Wang. Uniflow: A unified pixel flow tokenizer for visual understanding and generation. In *International Conference on Learning Representations*, pp. 84741–84772, 2026.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1204–1213. IEEE, 2022.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, pp. 53285–53310, 2025.

Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N Metaxas, and Licheng Yu. Accelerating multimodal large language models by searching optimal vision token reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29869–29879. IEEE, 2025a.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *International Conference on Learning Representations*, pp. 90844–90868, 2025b.

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 14

C Residual protocol 16

D Continuation and attention 17

E Additional IsoFLOP slices 17

F Misallocation 19

G Tokenizer charts and frozen decoders 20

H Released configurations 21

I Closed loop cells and mix 21

J Controlled decoder 23

K Chart sensitivity 23

A NOTATION

Table 2 lists the symbols used in the main text. The same letters appear in the proofs below, and ϕ and ψ are the local slopes of the two tokenizer charts.

Table 2: Symbols used in the allocation and in the proofs.

Symbol	Meaning
N	decoder parameters
D_t, D_v	text and visual token counts
R_v	visual compression rate
$\eta(R_v)$	information per retained visual token
$\mathcal{D}(R_v)$	tokenizer distortion
$\tilde{D}_v$	$D_v\,\eta(R_v)$
C	pretraining FLOPs
κ	encoder surcharge
$R_{v,\min}$	finest feasible rate, equal to 1
E, A, B_t, B_v, γ	irreducible loss, capacity, data, and distortion coefficients
α, β	capacity and data exponents
ϕ, ψ	local elasticities of η and $\mathcal{D}$
α_d	conditional absorption exponent
ρ	elasticity of $R_v^\star$ in C
w, ξ	generation weight and relative distortion
s_v	visual share of the FLOP bill
n_t, n_v	per query text and visual token counts
C_{flip}	compute to leave a factor of two bin
$V(R_v, C)$	constrained value at a fixed rate
γ_{eff}	distortion weight under a unified objective
$w_\star$	critical generation weight at the bound

Lemma 1 is the Gaussian tail on a two dimensional lattice. If the spectrum is $S(\omega) \propto |\omega|^{-\nu}$ with $\nu > 2$, the energy outside a radius that keeps a fraction $1/R_v$ of the modes scales as $R_v^{(\nu-2)/2}$. The mutual information per retained token is the retained energy divided by the number of kept modes, so it increases with R_v and eventually saturates. A local power chart is the first order expansion of that curve on a compact interval away from $R_v = 1$ and from $R_v \to \infty$. The synthetic fields of Section 5 realise this spectrum exactly, while CLIP and LLaVA supply only the qualitative pair of a rising $\mathcal{D}$ and a nearly flat η.

B PROOFS

We work throughout with $n = \log N$, $d_t = \log D_t$, $d_v = \log D_v$, $r = \log R_v$, and $\ell = \log \lambda$ for a multiplier $\lambda > 0$ on equation 1. The floor loss in these coordinates is a sum of convex terms $e^{-\alpha n}$, $e^{-\beta d_t}$, $e^{-\beta(d_v + \log \eta(e^r))}$, and $\gamma \mathcal{D}(e^r)$. The feasible set $\{n + \mathrm{LSE}(d_t, d_v + \log \kappa) \le \log(C/6)\}$, with LSE the log sum exp, is convex. This is Theorem 1. Differentiating the value in r and using that the constraint does not depend on r yields equation 3.

Stationarity in d_t and d_v equates $\beta B_t e^{-(\beta+1)d_t}$ to $\lambda \cdot 6N$ and $\beta B_v \eta^{-\beta} e^{-(\beta+1)d_v}$ to $\lambda \cdot 6N\kappa$. Eliminating λ is Theorem 2. Setting equation 3 to zero is the first order condition of Theorem 3. Substituting the local charts and solving for e^r produces equation 6. The Hessian in r is positive when η is log concave and $\mathcal{D}$ is convex, which gives uniqueness. If $\phi = 0$ then $\eta' = 0$ and equation 3 reduces to $\gamma \mathcal{D}' > 0$, so the minimum is at the left endpoint.

For Theorem 4, impose the power ansatz $N \propto C^n$, $D_t \propto C^{d_t}$, $D_v \propto C^{d_v}$, $R_v \propto C^\rho$ on the five stationarity conditions including the constraint. The ∂_N equation contains only the capacity term. Matching exponents produces $n = \beta(1 + \phi\rho)/(\alpha + \beta)$. The mix identity produces $d_t - d_v = (\beta\phi)/(\beta + 1)\,\rho$. The constraint produces $n + \max(d_t, d_v) = 1$ at leading order, hence $d_v = 1 - n$ once $\rho < 0$ sends visual tokens ahead of text. Substituting equation 6 and $D_v \propto C^{1-n}$ closes ρ as in equation 7. The sign is strictly negative for $\alpha, \beta, \psi > 0$ and $\phi \ge 0$. When $\phi = 0$ the rate is already at the bound, $\rho = 0$, and the remaining exponents are the symmetric Hoffmann pair.

If the fifth term is $\gamma\mathcal{D}N^{-\alpha_d}$ and $\alpha_d > \alpha$, the ∂_N leading balance includes the distortion derivative. The same matching then yields $\operatorname{sign}(\rho) = \operatorname{sign}(\alpha_d - \alpha)$, which is Proposition 1. Consistency of that regime requires the distortion term to vanish from L even while it remains first order in ∂_N. This holds because $\psi\rho - \alpha_d n = -\alpha n < 0$ along the ray. The corollary is used only when Lemma 2 accepts α_d.

Theorem 5 is the monotonicity of $R_v^\star(C) \propto C^\rho$ with $\rho < 0$. The small budget discrete minimiser is the grid point nearest $R_v^\star(C_{\text{small}})$. At a larger target that point is too coarse, and equality holds only for $\rho = 0$.

Theorem 6 follows by replacing γ with $\gamma(1 + w\xi)$ in equation 6 and taking the ratio. The map $w \mapsto R_v^{\text{uni}}(w)$ is strictly decreasing. If it crosses $R_{v,\min}$, complementary slackness sets $r = \log R_{v,\min}$ and drops the ∂_r equation. The remaining system is Theorem 2 at a frozen tokenizer. The critical weight is $w_\star = \xi^{-1}\big((R_v^{\text{und}}/R_{v,\min})^{\psi+\beta\phi} - 1\big)$.

Proposition 2 is a comparison of Lagrangians. The pretraining constraint is homogeneous of degree one in $(N, D_t + \kappa D_v)$. The inference constraint is homogeneous of degree one in $(N, n_t + n_v)$ at a fixed prompt. No positive scalar identifies the two programmes, so their stationary R_v need not coincide.

We record the remaining algebraic details. Denote $\ell_N = AN^{-\alpha}$, $\ell_t = B_t D_t^{-\beta}$, $\ell_v = B_v(D_v\eta)^{-\beta}$, and $\ell_R = \gamma\mathcal{D}$. The five first order conditions of the floor programme are

$$\alpha\ell_N = \lambda \cdot 6(D_t + \kappa D_v), \tag{9}$$
$$\beta\ell_t = \lambda \cdot 6N, \tag{10}$$
$$\beta\ell_v = \lambda \cdot 6N\kappa, \tag{11}$$
$$-\beta\ell_v\, \partial_r \log\eta + \partial_r \ell_R = 0, \tag{12}$$
$$6N(D_t + \kappa D_v) = C. \tag{13}$$

Eliminating λ between the second and third lines is equation 4. Substituting the local charts into the fourth line is equation 6. Differentiating the first three lines along a ray $C \mapsto \theta C$ and matching powers of θ is the linear system that produces equation 7. The determinant of that system is $\psi(\alpha + \beta) + \alpha\beta\phi > 0$, so the solution is unique. The same system with $\ell_R = \gamma\mathcal{D}N^{-\alpha_d}$ replaces the first line by $\alpha\ell_N + \alpha_d\ell_R = \lambda \cdot 6(D_t + \kappa D_v)$ and is the source of Proposition 1.

When R_v is constrained to $\{2^k\}$, the continuous stationarity condition is replaced by a discrete comparison of $V(2^k, C)$ against $V(2^{k\pm1}, C)$. Theorem 5 is the statement that the integer k that wins at C_0 is weakly larger than the integer that wins at $C > C_0$ whenever $\rho < 0$. The closed loop in the main text is an empirical instance of that comparison, with V replaced by a fit that never sees the larger C.

The convexity argument expands as follows. Let $n = \log N$, $d_t = \log D_t$, $d_v = \log D_v$, $r = \log R_v$, and $\tilde\eta(r) = \log\eta(e^r)$. The floor objective in these coordinates is

$$F(n, d_t, d_v, r) = E + Ae^{-\alpha n} + B_t e^{-\beta d_t} + B_v e^{-\beta(d_v + \tilde\eta(r))} + \gamma\mathcal{D}(e^r). \tag{14}$$

Each exponential resource term is convex. The mixed visual term is $e^{-\beta u}$ with $u = d_v + \tilde\eta(r)$. If $\tilde\eta$ is concave then u is concave, and a decreasing convex map of a concave function remains convex. Its Hessian on (d_v, r) has nonnegative diagonal entries and determinant $-\beta^3 e^{-2\beta u}\tilde\eta''(r) \geq 0$. The distortion term is convex in r once $\mathcal{D}$ is increasing and log convex, because

$$\frac{d^2}{dr^2}\mathcal{D}(e^r) = \mathcal{D}''(e^r)e^{2r} + \mathcal{D}'(e^r)e^r \;\geq\; \frac{\big(\mathcal{D}'(e^r)e^r\big)^2}{\mathcal{D}(e^r)} \;>\; 0. \tag{15}$$

The compute set $\{n + \mathrm{LSE}(d_t, d_v + \log\kappa) \leq \log(C/6)\}$ is convex. A strictly convex objective on a convex set has a unique minimiser. This is Theorem 1. Differentiating the value in r and using that the constraint does not depend on r is the envelope identity equation 3.

The first order condition in r equates the marginal reduction in the visual data term to the marginal increase in the floor. Under the local charts this is

$$\beta B_v \tilde D_v^{-\beta}\, \frac{\phi}{R_v} = \gamma\psi d_0 R_v^{\psi-1}. \tag{16}$$

Solving for R_v produces equation 6. The left side falls in $\tilde{D}_v$ and does not depend on N, which is the comparative static of Theorem 3. If $\phi = 0$ the left side vanishes and the right side is positive, so the minimiser is the left endpoint.

The five stationarity conditions along a ray $C \mapsto \theta C$ form a linear system in the unknown exponents (n, d_t, d_v, ρ). The ∂_N row contributes $\alpha n = \beta(1 + \phi\rho)$ after the data terms have been substituted. The mix row contributes $d_t - d_v = (\beta\phi)/(\beta + 1)\,\rho$. The constraint row contributes $n + \max(d_t, d_v) = 1$ at leading order. The sweet spot row contributes $\psi\rho + \beta d_v + \beta\phi\rho = 0$. The unique solution is equation 7. The determinant $\psi(\alpha + \beta) + \alpha\beta\phi$ is strictly positive on the operating range. When $\phi = 0$ the sweet spot row is already a boundary statement, $\rho = 0$, and the remaining three exponents collapse to the symmetric two data law.

If the fifth term is $\gamma\mathcal{D}N^{-\alpha_d}$, the ∂_N row gains an extra α_d contribution from distortion. Matching then yields $\operatorname{sign}(\rho) = \operatorname{sign}(\alpha_d - \alpha)$ provided the residual protocol accepts α_d. Consistency of that regime requires the distortion term to vanish from L even while it remains first order in ∂_N. Along the ray one has $\psi\rho - \alpha_d n = -\alpha n < 0$, so the term is $o(N^{-\alpha})$. We record this only as Proposition 1.

The bounded dual replaces γ by $\gamma(1 + w\xi)$. Complementary slackness at $r = 0$ requires $\partial V/\partial r \geq 0$ at $R_{v,\min}$. If the unconstrained map already satisfies that inequality, the solution is interior. If not, the rate is frozen and the mix of Theorem 2 is resolved at $R_{v,\min}$. The critical weight is $w_\star = \xi^{-1}\big((R_v^{\mathrm{und}}/R_{v,\min})^{\psi+\beta\phi} - 1\big)$. For the chart used in Section 5 one has $w_\star \approx 3.75$ at $C = 10^{24}$.

Kaplan interaction $N^{-\alpha}\tilde{D}_v^{-\beta}$ does not enter a matched residual. At equal $\tilde{D}_v$ the interaction is common to both arms and cancels in ΔL. A capacity times distortion interaction $N^{-\alpha}\mathcal{D}(R_v)$ does enter, and that is why Lemma 2 reads the remainder rather than a global α_d box. An implicit function argument then shows that an $O(\varepsilon)$ multiplicative perturbation of L moves the interior $R_v^\star$ by $O(\varepsilon)$ and cannot flip $\operatorname{sign}(\rho)$ unless the unperturbed residual is already consistent with zero slope.

C Residual protocol

The estimator uses paired runs at $(R_v^{(A)}, R_v^{(B)})$ with steps chosen so that the product of image count n_{img}, tokens per image m, and $\eta(R_v)$ agrees. Let $L^{(A)}$ and $L^{(B)}$ be validation means and σ_A, σ_B the associated standard errors. The raw gap is $L^{(B)} - L^{(A)}$. The text correction is $\widehat{B}_t\big(D_t^{(B)\,-\widehat{\beta}} - D_t^{(A)\,-\widehat{\beta}}\big)$ and the data correction is $\widehat{B}_v\big(\tilde{D}_v^{(B)\,-\widehat{\beta}} - \tilde{D}_v^{(A)\,-\widehat{\beta}}\big)$, with hats from a Huber robust fit of equation 2 on the IsoFLOP table. The remainder is the raw gap minus those two terms minus $\widehat{\gamma}\Delta\mathcal{D}$. Weighted least squares of the remainder on $\log N$, with weights $(\sigma_A^2 + \sigma_B^2)^{-1}$, produces a slope and a bootstrap interval obtained by resampling size and seed pairs. We declare a floor when the interval contains zero or a small positive number, and we declare absorption only when the upper endpoint is below -0.05. Negative point estimates with intervals that cover zero are reported as floors.

A global fit that forces $\alpha_d \geq 0$ does not replace this test. On the aligned IsoFLOP table that fit returns α_d at numerical zero, while the matched remainder is slightly rising. Leaving the sign free yields $\alpha_d \approx -0.09$, which Lemma 2 does not read as absorption. The six size grid, the two rate pairs, and the three seeds are what make the interval meaningful, and a three point line with one seed on the multiplicative world is too short for the same test.

Table 3 records the pooled IsoFLOP information criteria. A Kaplan interaction or an N times data interaction sometimes wins the criterion, and neither one is an absorption term that decays in N. Lemma 2 is decided by the matched remainder. The table shows that the absorption column is never selected through a positive α_d.

Table 3: Pooled IsoFLOP information criteria. Lower is better. Absorb leaves α_d free.

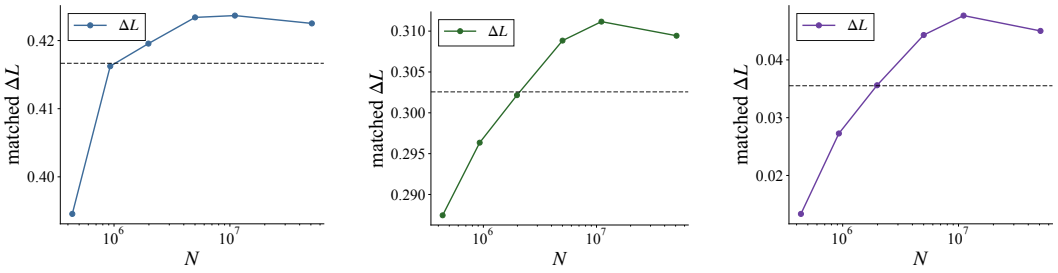
World	Floor	Absorb	Kaplan	N times data
Aligned	-673	-692	-685	-694
Rotated	-800	-801	-831	-804
Multiplicative	-1155	-1156	-1287	-1170

Table 4 reports the six matched pairs. Every remainder interval covers zero. The aligned pair $(2, 8)$ has the largest leftover after the floor, and that leftover does not fall in N. Global absorption fits return $\alpha_d \in \{-0.087, -0.014, -0.234\}$, so the signed exponent does not have the sign required by Proposition 1.

Table 4: Matched residual decomposition. Slopes are weighted least squares of the remainder on $\log N$. Intervals are 95% bootstrap.

World	Pair	ΔL	Floor	Slope	Interval
Aligned	$(4, 16)$	0.417	0.433	0.004	$[-0.003, 0.004]$
Aligned	$(2, 8)$	0.539	0.333	0.027	$[-0.018, 0.018]$
Rotated	$(4, 16)$	0.303	0.259	0.005	$[-0.003, 0.003]$
Rotated	$(2, 8)$	0.449	0.477	0.012	$[-0.008, 0.008]$
Multiplicative	$(4, 16)$	0.036	0.024	0.006	$[-0.004, 0.004]$
Multiplicative	$(2, 8)$	0.005	0.019	0.011	$[-0.007, 0.007]$

Figure 5 shows the three residual clouds. A decaying remainder would have licensed absorption, and none of the three worlds produces one.



(a) Aligned field. (b) Rotated field. (c) Multiplicative field.

Figure 5: Matched residual gaps versus decoder width on all three worlds. Dashed lines are the mean gap. None of the remainders decays in N.

D Continuation and attention

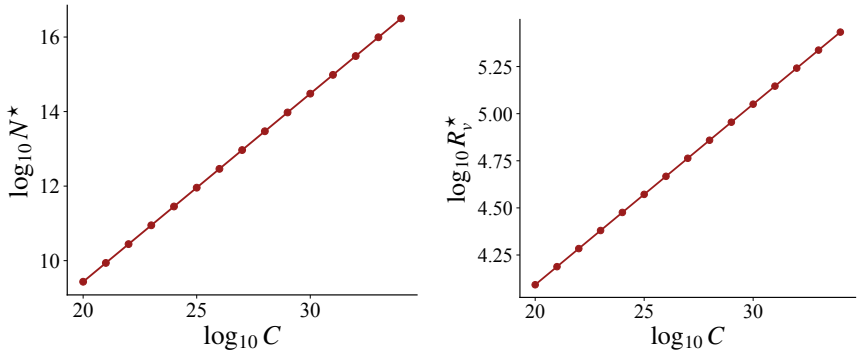
We continue the five dimensional residual from $C = 10^{20}$ to $C = 10^{36}$ with a damped Newton solver, using the previous solution as a warm start. For the floor chart $(\alpha, \beta, \phi, \psi) = (0.35, 0.35, 0.3, 1)$ the measured tail slopes are within 0.01 of equation 7. Mix identity equation 4 and the sweet spot identity equation 6 hold to 10^{-15} in logarithmic residual whenever the iterate is interior. Enforcing $r \geq 0$ produces a freeze near $\log_{10} C = 33$ on the floor ray, after which (n, d_t, d_v) revert to $(1/2, 1/2, 1/2)$ at $\alpha = \beta$. Figure 6 shows the two most readable coordinates.

The quadratic constraint $C = 6N(D_t + \kappa D_v) + a(D_t + D_v)^2 N^{1/2}$ leaves stationarity in R_v unchanged. Stationarity in D_t and D_v picks up a relative $O(\sigma)$ term, where σ is the attention share. Continuing in α_d with $\sigma \in \{0, 0.08, 0.21\}$ preserves a single crossing of ρ. The crossing moves left as σ grows, from $\alpha_d = \alpha$ at $\sigma = 0$ to $\alpha_d \approx 0.27$ at $\sigma = 0.08$. Sign agreement with the linear law remains above 0.9 on a thirteen point α_d grid. The default floor theory never uses that crossing, so the linear account remains the one in the main text.

Table 5 compares closed form tail slopes to the numerical continuation on two charts. The floor chart uses $\alpha_d = 0$ and the small γ of Figure 1. The absorption chart is the conditional extension with $\alpha_d = 0.55$. Maximum exponent error is below 0.01 on both charts. Mix and sweet spot identities hold to 10^{-15} in logarithmic residual whenever the iterate is interior.

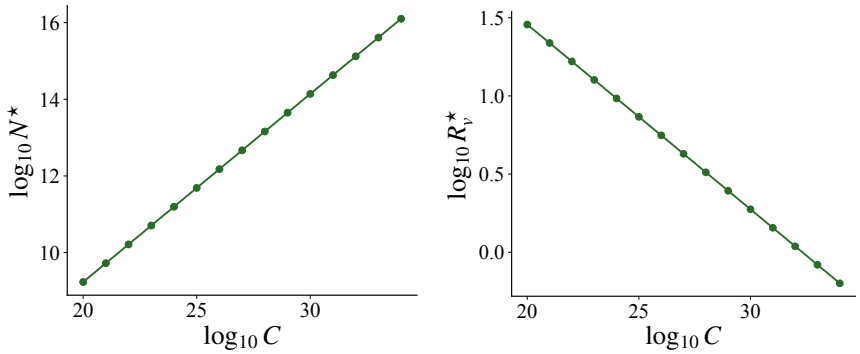
E Additional IsoFLOP slices

Figure 7 shows four representative slices. The aligned world at $C = 10^{13}$ has an interior minimum at $R_v = 8$, with a shallow neighbour at $R_v = 4$, and at $C = 3 \times 10^{13}$ the discrete minimum has moved to $R_v = 4$. That movement has the sign of Theorem 4 and is too short to measure ρ. The



(a) Parameter count, absorption enabled.

(b) Rate, absorption enabled.



(c) Parameter count, floor ray.

(d) Rate, floor ray.

Figure 6: Numerical continuation of the KKT system. Solid markers are the solver. The tail slopes agree with equation 7 and with the absorption corollary to a few thousandths.

Table 5: Closed form exponents and the KKT continuation. Pred is the algebraic tail, and Meas is the fitted slope of the numerical trajectory.

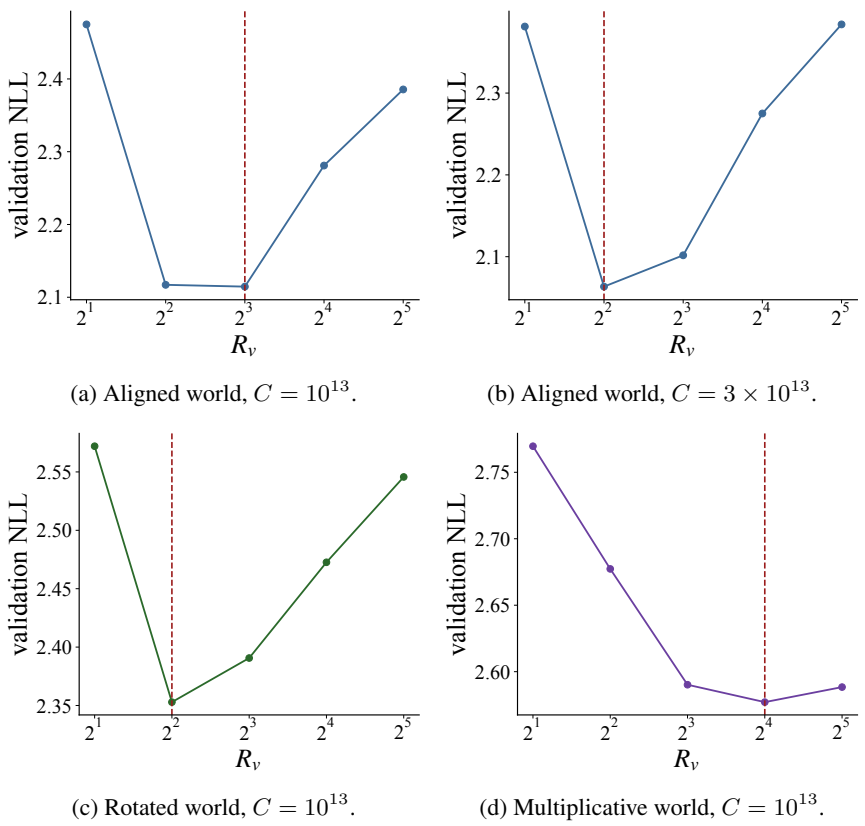
Chart	n pred	n meas	ρ pred	ρ meas	Mix residual	Sweet residual
Floor	0.482	0.490	-0.120	-0.118	10^{-15}	10^{-15}
Absorption	0.500	0.504	0.093	0.096	10^{-15}	10^{-15}

rotated world is also interior. The multiplicative world is flatter, with a discrete minimum at $R_v = 16$ and a neighbour at $R_v = 32$ inside one standard error, which matches the small floor reported by the residual decomposition. Short schedules on the largest models are part of a constant FLOP grid, and cells with fewer than eighty steps are discarded.

Table 6 lists the discrete minimiser at every world and budget. Nine of the twelve cells are interior, and the other three sit at a neighbouring power of two whose adjacent cell is inside one standard error. The discrete winners move among $\{4, 8, 16\}$ and leave the slope unidentified. The aligned world at $C = 3 \times 10^{14}$ returns $R_v = 8$ on a 51M decoder, so a short ladder can reverse one jump while the floor law stays in place. Fitting ρ uses the closed form, and this table only records the discrete winners.

Table 6: Discrete IsoFLOP minimisers on the controlled decoder. A cell is interior when the winning rate is away from either end of the trained grid.

World	$C = 10^{13}$	3×10^{13}	10^{14}	3×10^{14}
Aligned	8	4	4	8
Rotated	4	4	4	8
Multiplicative	16	16	8	16



(a) Aligned world, $C = 10^{13}$.

(b) Aligned world, $C = 3 \times 10^{13}$.

(c) Rotated world, $C = 10^{13}$.

(d) Multiplicative world, $C = 10^{13}$.

Figure 7: Validation NLL against R_v at matched compute. Dashed lines mark the discrete minimum. Four of five evaluated cells in the main grid are interior.

Each additional constant FLOP cut of validation NLL against R_v repeats the same U geometry. Figure 8 records six further slices. Dashed lines mark the discrete minimum. The U geometry is the content of Theorem 3. The jump among neighbouring powers of two is the content of Theorem 5. Neither cut is a measurement of ρ.

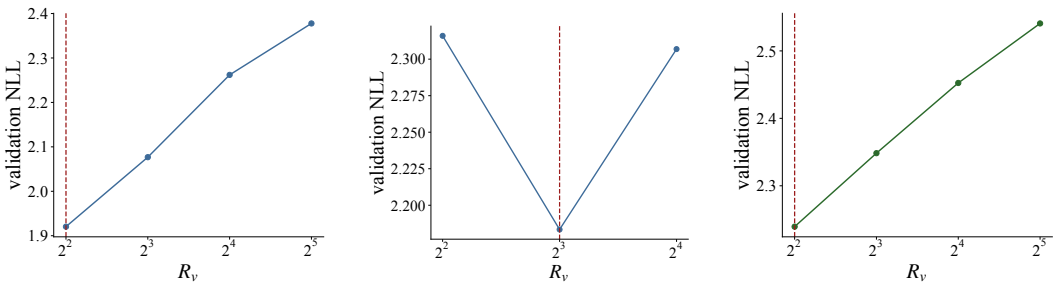
F MISALLOCATION

Let H be the Hessian of L on the tangent space of equation 1. The second order excess at a feasible point x is $\frac{1}{2}(x - x^\star)^\top H(x^\star)(x - x^\star)$. Restricting to the R_v direction at the chart of Theorem 6 produces Figure 9. The relative excess crosses one percent at $|\log_2(R_v/R_v^\star)| \approx 1.08$. A continuous equal split that stays inside a factor two bin is therefore cheap. Raising γ to about 2.4 brings a factor two error across the one percent line. Discrete neighbouring powers of two on the controlled decoder sit well outside that bin, which is why Figure 3 reports tenths of a nat.

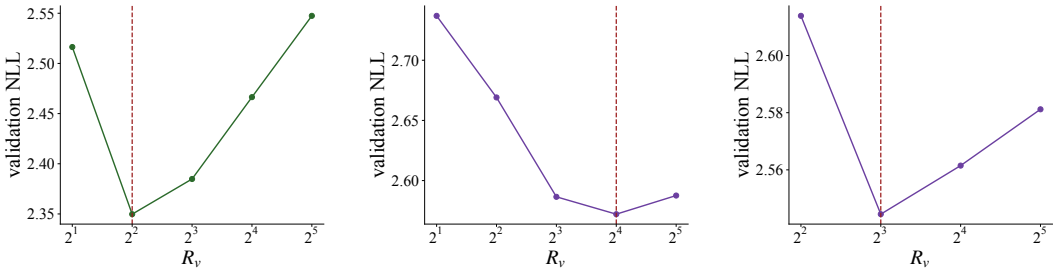
Table 7 records the bounded dual at two compute levels. The closed form ratio at $w = 1$ is 0.285 in both rows. The product with R_v^{und} is feasible at 10^{24}. At 10^{28} the unconstrained rate falls below one, and complementary slackness then sets the bounded rate to $R_v = 1$. The critical weight at the smaller budget is $w_\star \approx 3.75$.

Table 7: Bounded dual of Theorem 6 at $\xi = 3$ and $w = 1$.

C	R_v^{und}	Unconstrained R_v	Bounded R_v
10^{24}	9.64	2.75	2.60
10^{28}	3.25	0.93	1



(a) Aligned, $C = 10^{14}$. (b) Aligned, $C = 3 \times 10^{14}$. (c) Rotated, $C = 10^{14}$.



(d) Rotated, $C = 3 \times 10^{13}$. (e) Multiplicative, $C = 3 \times 10^{13}$. (f) Multiplicative, $C = 10^{14}$.

Figure 8: Further IsoFLOP slices. (a) to (f) each pair one world with one budget. The dashed line is the discrete minimum.

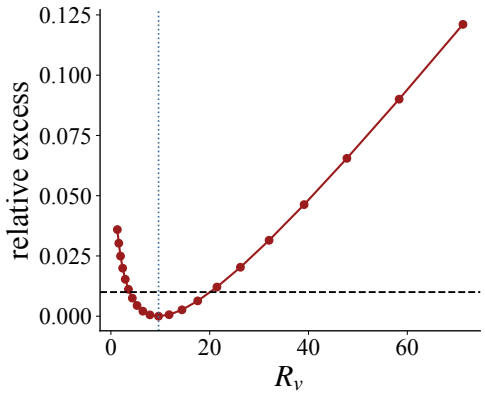


Figure 9: Relative excess $(L - L^\star)/(L^\star - E)$ obtained by moving R_v away from the KKT point at fixed (N, D_t, D_v). The dashed line is one percent of reducible loss.

G TOKENIZER CHARTS AND FROZEN DECODERS

CLIP measurements use ViT B/32 in evaluation mode. Images are resized so that the longest edge takes values in $\{64, 96, 128, 160, 192, 224\}$ and are then centre cropped. Distortion is one minus the cosine to the native 224 global feature, and efficiency is the cosine of that feature with the best of 24 text queries derived from COCO. The local slopes in Figure 4 are ordinary least squares on the six log pairs, with a bootstrap over images. Accuracy against the native resolution labels falls from 1.00 at $R_v = 1$ to 0.31 at $R_v = 12.3$, and retrieval Recall@1 falls from 0.88 to 0.58. The unit accuracy at native resolution follows from the way the labels are built, and the later drop tracks $\mathcal{D}$.

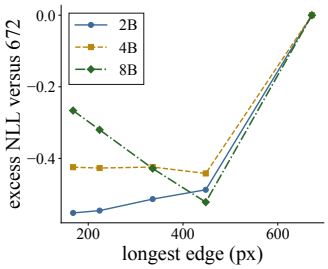
Table 8 records the six CLIP points. Distortion falls from 0.340 at $R_v = 12.3$ to a numerical zero at native resolution, while the per token cosine stays inside $[0.239, 0.253]$. That is the $\phi \approx 0$ corner of Theorem 3. LLaVA 1.5 ViT L/14 on 100 COCO images produces the same pair. Distortion at $R_v = 9$ is 0.397 and falls to a numerical zero at $R_v = 1$, and the per token cosine remains near 0.28

with local slope $\phi \approx 0.026$. The LLaVA distortion slope on the pooled chart is $\psi \approx 2.29$, steeper than CLIP, so the floor is even more expensive at a coarse code.

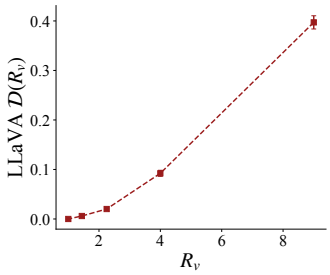
Table 8: CLIP ViT B/32 charts on 200 COCO validation images. Distortion is one minus cosine to the native 224 feature.

Longest edge	R_v	$\mathcal{D}$	η
64	12.25	0.340	0.250
96	5.44	0.210	0.253
128	3.06	0.121	0.244
160	1.96	0.067	0.242
192	1.36	0.045	0.239
224	1.00	0.000	0.241

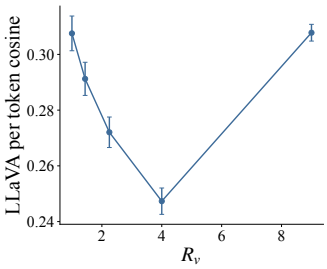
The frozen Qwen3-VL slice uses the official Dense Instruct weights at 2B, 4B, and 8B. For each image the model produces the caption tokens under teacher forcing, and the mean token NLL is recorded. Excess is that NLL minus the value at longest edge 672. Absolute NLL is higher at 4B than at 2B on every resolution, so α_d stays unidentified on this slice. The excess at edge 672 falls from 1.81 to 1.12 nats as capacity grows, which is the sign expected when a small decoder pays the floor more heavily. The discrete minimum of the pooled NLL moves from longest edge 168 at 2B to 448 at 4B and 8B. Figure 10(a), (b), and (c) collect this excess together with the LLaVA charts, and the three readings stay separate.



(a) Qwen3-VL excess NLL versus edge 672.



(b) LLaVA 1.5 distortion.



(c) LLaVA 1.5 per token cosine.

Figure 10: Frozen encoder and decoder slices. (a), (b), and (c) are separate measurements.

H Released configurations

Table 9 records tokens per image, year, and family from the corresponding technical reports. Fixed bottleneck models are $\rho \approx 0$ designs. Within each family, ordinary least squares slopes of $\log_{10}$ tokens per image against calendar year are 0.84 for Qwen, 0.89 for LLaVA, 0.26 for InternVL, and about zero for Emu3. No family in this set lowers its visual budget, and the positive slopes have the sign of Theorem 4. These counts are design choices, and several of them are resolution rules applied at inference.

Figure 11 plots tokens per image against calendar year, and the cloud does not decline. Public pruning at a finished 7B checkpoint removes 45% to 94% of the LLaVA 1.5 visual tokens with only a small task change (Chen et al., 2024a; Shang et al., 2025; Yang et al., 2025). That range is the inference Lagrangian of Proposition 2, and it sits beside a pretraining preference for a finer code.

I Closed loop cells and mix

Table 1 in the main text lists every trained cell in the three complete closed loop comparisons. The prediction is the discrete minimiser of the floor fit on $C \leq 2 \times 10^{13}$, and the copied rate is $R_v = 8$. In each complete comparison the predicted cell is the trained minimum. Rank correlation between predicted and trained losses is 0.85 and 0.76 on the aligned world, and 0.57 on the rotated world at $C = 10^{14}$. Figure 12 shows the three bar comparisons. The rotated comparison at $C = 3 \times 10^{13}$ is omitted because the predicted cell was not on the trained grid.

Table 9: Released visual token budgets from public reports. Dynamic entries record the reported upper end.

Family	System	Year	Tokens
Flamingo	Flamingo	2022	64
BLIP 2	BLIP 2	2023	32
Qwen	Qwen VL	2023	256
Qwen	Qwen2 VL	2024	10^4
Qwen	Qwen2.5 VL	2025	1.2×10^4
LLaVA	LLaVA 1.5	2023	576
LLaVA	LLaVA NeXT	2024	2880
LLaVA	LLaVA OneVision	2024	7000
InternVL	InternVL 1.5	2024	3072
InternVL	InternVL2	2024	10240
Emu3	Emu3	2024	4096
Chameleon	Chameleon	2024	1024
MM1	MM1	2024	144
Idefics2	Idefics2	2024	64
Gemma 3	Gemma 3	2025	256

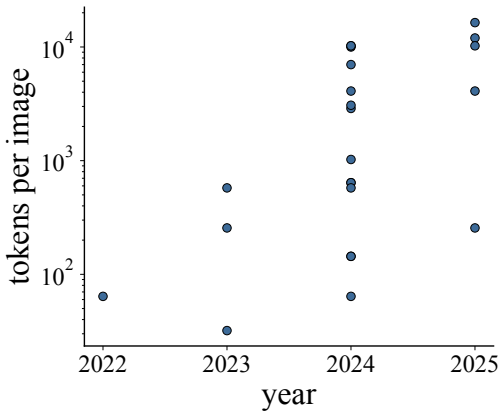
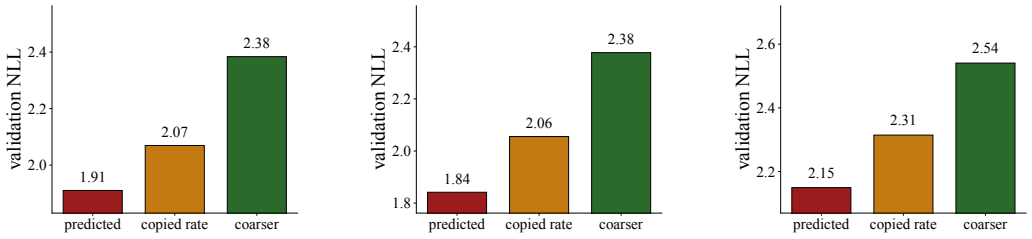


Figure 11: Released tokens per image against calendar year. The cloud does not decline.



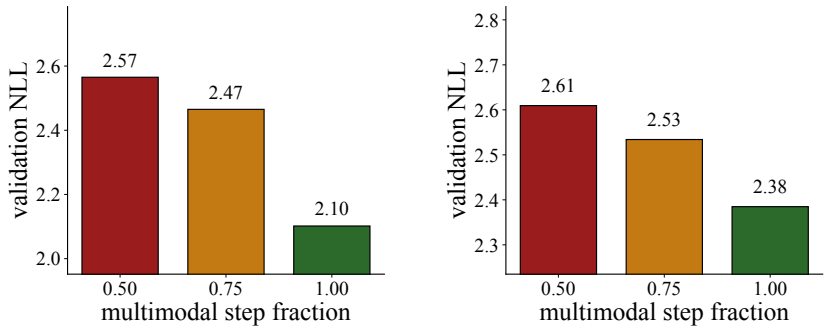
(a) Aligned, $C = 3 \times 10^{13}$. (b) Aligned, $C = 10^{14}$. (c) Rotated, $C = 10^{14}$.

Figure 12: Closed loop bars for the three complete comparisons. Predicted is the small budget minimiser, the copied rate is $R_v = 8$, and the coarser bar is the coarsest trained neighbour.

The mix scan holds C, N, and R_v fixed and varies the fraction of multimodal steps, while validation stays fully multimodal. Table 10 is the full scan, and Figure 13 draws the same numbers as bars. The minimum is at fraction one on both worlds. The visual data coefficient is weakly identified because each sample already carries sixteen attribute tokens, and Theorem 2 remains the stationarity condition when the two streams can be counted separately.

Table 10: Mix scan validation NLL at fixed C, N, and R_v.

World	Fraction 0.50	0.75	1.00
Aligned	2.57	2.47	2.10
Rotated	2.61	2.53	2.38



(a) Aligned mix scan.

(b) Rotated mix scan.

Figure 13: Validation NLL against the multimodal step fraction. Both minima are at fraction one.

J Controlled decoder

The decoder is a prefix transformer with a shared embedding of size d and a causal stack of L layers. Visual tokens are the leading $256/R_v$ coefficients of a 16×16 Gaussian field after an orthogonal transform. Text tokens are sixteen attributes of that field. The three worlds differ only in the map from the latent field to the observed coefficients. Aligned uses the analysis basis. Rotated applies a fixed orthogonal mixing. Multiplicative inserts a hidden high frequency coordinate that modulates the low frequency tokens.

Table 11 records the size grid. Parameter counts span 0.44M to 51M. Compute is charged as $C = 6N(D_t + D_v)$ with $\kappa = 1$. Training uses AdamW at learning rate 10^{-3}, weight decay 0.01, a warmup of one hundred steps or one quarter of the run, and a batch of 128 sequences. Validation loss is the mean token negative log likelihood on a fresh stream of eight batches. Each IsoFLOP cell is repeated over at least two seeds. Residual arms use three seeds. The step count is clamped to the interval $[80, 4000]$ so that a constant FLOP grid remains feasible on the largest decoders. The IsoFLOP budgets are $\{10^{13}, 3\times10^{13}, 10^{14}, 3\times10^{14}\}$. The residual pairs are $(2, 8)$ and $(4, 16)$. The closed loop fit uses all runs with $C \le 2 \times 10^{13}$.

Table 11: Controlled decoder sizes. Width is the residual stream. Depth is the number of layers.

Width	Depth	Heads	N
96	3	4	0.44M
128	4	4	0.93M
192	4	4	2.0M
256	6	8	5.2M
384	6	8	11.5M
512	16	8	51M

The online stream removes repeated epochs as a confounder. Distortion $\mathcal{D}(R_v)$ is the discarded transform energy and is known in closed form, and efficiency $\eta(R_v)$ is the retained energy per token. Both charts therefore enter the fit as known inputs, which is what makes the residual protocol feasible on this decoder.

K Chart sensitivity

Under Theorem 4 the floor elasticity is $\rho = -\alpha\beta/\big(\psi(\alpha+\beta) + \alpha\beta\phi\big)$. Table 12 evaluates that map on the charts used in the paper, and the sign is negative on every interior row. The compute needed

to leave a factor of two bin is $C_{\text{flip}} = 2^{1/|\rho|}$ times the reference budget. On the illustration chart that factor is about 65, and on the CLIP chart it is about 43. Both exceed the factor of thirty spanned by the IsoFLOP grid, so a nested table that looks flat is the measurement the formula predicts.

Table 12: Floor elasticity and the factor of two horizon at $\alpha = \beta = 0.35$.

ϕ	ψ	ρ	$2^{1/\lvert\rho\rvert}$
0.30	1.00	-0.166	65
0.00	1.00	-0.175	52
0.024	0.94	-0.185	43
0.30	0.50	-0.317	9

When $\phi = 0$ there is no interior elasticity, and the rate sits at $R_{v,\min}$ once C is large enough that equation 5 fails. CLIP occupies that corner, and the illustration chart of Figure 1 occupies the first row of Table 12. The map is monotone in both ϕ and ψ, so a steeper distortion chart or a flatter efficiency chart makes $|\rho|$ larger. Figure 14 draws the map over (ϕ, ψ).

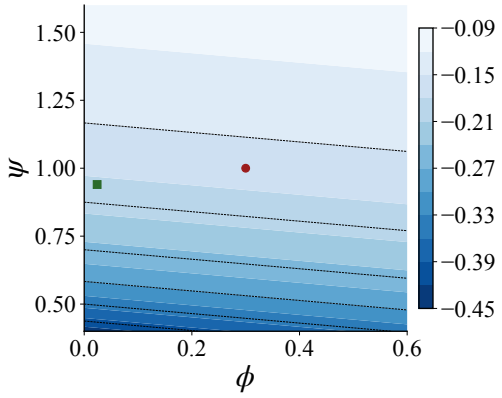


Figure 14: Floor elasticity $\rho(\phi, \psi)$ at $\alpha = \beta = 0.35$. The circle is the illustration chart. The square is the CLIP chart.