CONNECTORS AS CHANNELS: USABLE INFORMATION BETWEEN FROZEN VISION AND LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision language models couple a frozen visual encoder to a frozen language model through a small trainable connector. The connector family and the number of retained visual tokens are still chosen by ablation. Published comparisons disagree on which of these choices matters, and they still lack a common account of when a given design is appropriate. The information bottleneck is the classical theory of this kind of compression. It prices a learnable code, while a connector sits between two frozen spaces and is read only through the directions that move the next token loss. In this paper, we model the connector as a channel and measure usable information, the mutual information that remains after the frozen language model has read the mapped tokens. On released Qwen3-VL checkpoints, the readable set recovered from last layer hidden states is much thinner than the leading principal subspace of the embedding table. Identifying this set exposes a concrete cost of query mixing, because lengthening the encoder without new signal dilutes the attention mass on a signal set and lowers usable information. We then estimate canonical spectra from teacher forcing hidden states and from caption vectors, and we train linear, two layer, and query connectors on frozen CLIP features that feed a frozen language model. These results account for the large token counts retained by published designs, for the sufficiency of a single matrix when the encoder is already contrastively aligned, and for the behaviour of a query bottleneck as a mixing mechanism rather than an integer token budget.

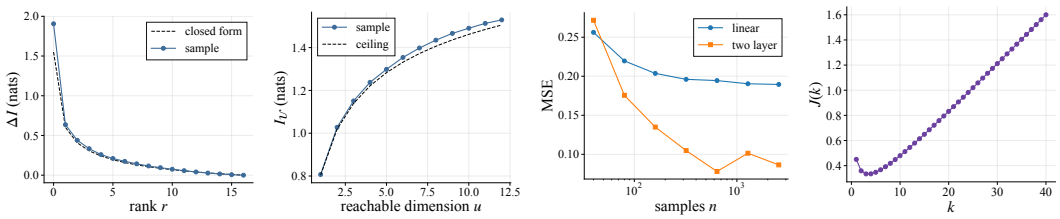
1 INTRODUCTION

Training a large vision language model usually freezes both the visual encoder and the language model. The only piece that is then trained is a connector, which maps visual features to tokens in the language model embedding space (Liu et al., 2023; Li et al., 2023; Alayrac et al., 2022). Because that connector is small, its family and its token count are still chosen by ablation. The resulting comparisons are extensive, and they disagree on which of these choices matters. A change of family can be large for one frozen reader and small for another. The property of that reader that produces the difference is still not isolated. Two studies can then reach different conclusions from the same visual features. In many of these comparisons the data, the resolution, and the frozen reader move together with the connector. The ablation therefore confounds the connector with the rest of the training recipe.

The disagreement follows the order in which the designs accumulated. LLaVA showed that a linear map from a contrastive encoder already supports captioning, and LLaVA-1.5 replaced that map by a two layer network while adding academic visual question data (Liu et al., 2023; 2024). BLIP-2 and Flamingo took the other route and compressed the visual grid into learned queries, after which Honeybee, TokenPacker, and Cambrian-1 kept local structure because that compression had weakened spatial detail (Li et al., 2023; Alayrac et al., 2022; Cha et al., 2024; Li et al., 2025a; Tong et al., 2024). Depth, query length, and local pooling therefore moved together with the data and with the token count. The published record does not separate those changes. A later comparison can then credit the family for a gain that came from the data. Until those factors are held apart, the family, the data, and the token count remain possible causes of the same gain.

Later comparisons begin to separate them. Once token count and resolution are held fixed, MM1 finds that connector type adds little (McKinzie et al., 2024). The type effect returns when Lin et al.

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) Linear residual. (b) Saturation in u. (c) Sample split. (d) Token objective.

Figure 1: Dashed curves are closed forms and markers are Monte Carlo estimates of the linear residual, the reachable dimension, the sample size, and the token objective.

(2024) separate fine grained perception from coarse semantics. Lost in Embeddings and a robustness study link that split to neighbourhood shift, and to the mixing that remains after pooling has matched the token count (Li et al., 2025b; Zhang et al., 2025c). Preserving features leads on spatial detail, and compression is enough for coarse semantics. Neighbourhood shift and the mixing of tokens give two readings of that split between fine grained perception and coarse semantics. What remains open is how many tokens are necessary, when a linear map is enough, and why a two layer network needs more data. It is also open which information is expensive to drop. Matching the token count removes one confound, and separating the tasks removes another.

The information bottleneck is the classical theory of this compression, and it prices a learnable code (Chechik et al., 2003). A connector sits between two frozen spaces, and the language model reads only the directions that move the next token loss. The cost of k tokens is then paid in the context window. The readable set is fixed by the frozen language model, so further tokens spend context without enlarging the set the reader can use. LLaVA-1.5 and BLIP-2 are trained on exactly this frozen slice, and later recipes that unfreeze the encoder can only enlarge the readable set (Liu et al., 2024; Li et al., 2023). Studies usually vary the family, the token count, or the readable directions one at a time. An account of the channel should say which choice binds once the other two are held fixed.

The same account has to name the readable set $\mathcal{U}$ and the task variable Y. Leading principal components of the released LLaVA-1.5 and Qwen3-VL embedding tables do not give $\mathcal{U}$ (Liu et al., 2024; Bai et al., 2025). Those tables are nearly full rank, and last layer hidden states on teacher forced captions occupy a much thinner set. Prefix gradients of the caption loss give a second view of that same set. A caption vector does not give Y. It lives in a retrieval geometry, and generation lives in the hidden trajectory of the next token. A rank count on the embedding table therefore does not limit the number of tokens. A spectrum built from caption vectors can look healthier than the generation task that the language model solves.

We treat the connector as a map $Z \mapsto \widetilde{Z}$ and measure usable information, the mutual information that remains after the language model has read $\widetilde{Z}$. Varying the readable set, the sample size, and the mixing of queries separates these constraints, as shown in Figure 1. A linear map, a two layer network, and a query bottleneck then use the same readable dimensions in different ways. The pair $(\mathcal{U}, Y)$ used by the frozen reader, together with the mixing cost of a query bottleneck, makes the published families comparable. With that pair held fixed, the token count is spent on spatial detail and on the mixing inside a query bottleneck. LLaVA style projectors, BLIP-2 style queries, and the Qwen3-VL spatial merger are regimes of the same channel. The sections below test that reading on released checkpoints and on frozen features.

The main contributions of this work are summarized as follows.

- We identify the reachable set from last layer hidden states and from prefix gradients of the caption loss, and we show that it is not the leading subspace of the embedding table.
- We show that usable information tracks signal mass η when a concentrated query connector is lengthened without new signal.
- We estimate task spectra from teacher forcing representations, train linear, two layer, and query connectors on frozen features, and we read published LLaVA, BLIP-2, Flamingo, and Qwen3-VL choices as regimes of the same channel.

2 Related Work

Connector families. Feature preserving maps pass one visual token per patch into the language model. LLaVA does so with a single matrix, and LLaVA-1.5 replaces that matrix by a two layer network while adding academic visual question data (Liu et al., 2023; 2024). The same linear step is already enough for captioning when the encoder carries linguistic supervision, as Merullo et al. show with a frozen reader (Merullo et al., 2023). Query bottlenecks shorten the grid before it enters the language model. BLIP-2 and Flamingo compress visual features into a short sequence of learned queries, as the Perceiver reduces a long input to a small latent set (Li et al., 2023; Alayrac et al., 2022; Jaegle et al., 2021). That compression weakens spatial detail, and Honeybee and Cambrian-1 answer this by keeping local structure while the token count falls (Cha et al., 2024; Tong et al., 2024). The same shortening can also be delayed until the features already exist. VisionZip, PruMerge, and DivPrune then select a support from the encoder output (Yang et al., 2025; Shang et al., 2025; Alvar et al., 2025). PyramidDrop, SparseVLM, and VoCo-LLaMA continue that reduction inside the language model, and LLaVA-Mini carries it to a single visual token (Xing et al., 2025; Zhang et al., 2025b; Ye et al., 2025; Zhang et al., 2025a). Wen et al. ask whether that selection solves the right problem, and Xu et al. show that misalignment between vision and language changes which tokens are safe to drop (Wen et al., 2025; Xu et al., 2026). These designs therefore differ in where the visual sequence is shortened, and they are regimes of the same channel.

Empirical comparisons. Once these families are compared at scale, connector type and token count do not always move together. MM1 finds that type adds little once token count and resolution are fixed (McKinzie et al., 2024). The type effect returns when Lin et al. (2024) separate tasks, since maps that preserve features lead on fine grained perception and compression is enough for coarse semantics, a split shared by local and global fusion (Lou et al., 2026). A depth comparison points the same way. In Honeybee, a two layer projector sits slightly above a linear map, and a six layer projector falls slightly below, which matches the estimation term $wd \log n/n$ (Cha et al., 2024; Ma et al., 2022). Measurement studies explain why matching the token count does not close the gap. Lost in Embeddings links neighbourhood shift and weak patch reconstruction to fine grained failures, and centred kernel alignment is the geometric proxy for that shift (Li et al., 2025b; Kornblith et al., 2019). Compressed projectors also remain easier to attack than uncompressed ones after pooling has matched the token count, so the way tokens are mixed matters beyond how many remain (Zhang et al., 2025c). These studies record which choice moved under a given recipe. They leave open which information a connector can map into a frozen reader, and which of that information is expensive to drop.

Information bottleneck accounts. The information bottleneck is the classical theory of this compression, but it does not describe the connector these models train. For Gaussian variables the optimal map lies in the canonical correlation basis, and the cost is mutual information between a learnable code and its source (Chechik et al., 2003; Hardoon et al., 2004). Multimodal extensions learn a fused code jointly with the downstream model. Mai et al. constrain unimodal and joint features for recognition, and Wu et al. set the bottleneck weight so that task relevant information is retained (Mai et al., 2022; Wu et al., 2025). The number of visual tokens in the context window does not enter that cost. TinyAlign comes closer. Effective mutual information there is the portion that a capacity limited frozen language model can use, and a retrieval connector is built to raise that portion (Hu et al., 2026). Here the connector stays a map between two frozen spaces. The readable set is identified from hidden states and prefix gradients, and query mixing is charged inside that set. This is also why a linear map suffices on a linguistically supervised encoder, as in Merullo et al., while a query bottleneck can look like a token budget when it is a mixing law (Merullo et al., 2023).

3 The Connector Channel

The connector is the only trained map once the visual encoder and the language model are both frozen. This section names the two spaces, identifies the directions the language model can read, and states the mixing, linear, and sample laws used in the experiments. The linear statements specialise canonical correlation analysis to a frozen reader, and the sample law is a Barron bound (Hardoon et al., 2004; Ma et al., 2022). Proofs are placed in the appendix. What is particular to this channel is the readable set, the task variable inside that set, and the signal mass of a query.

3.1 SETUP

An image X is mapped by a frozen encoder g to features $Z = g(X) \in \mathbb{R}^{N \times d_v}$. A connector f_θ, drawn from a class $\mathcal{F}$, sends those features to $\widetilde{Z} = f_\theta(Z) \in \mathbb{R}^{k \times d}$, where d is the embedding width of a frozen language model T. The downstream variable Y is a caption, an answer, or a sufficient statistic of either. When a question is present, the language context is L. Three families cover the designs reviewed above. A linear or feature preserving map keeps one token per patch, with $k = N$ and $\widetilde{Z}_{i\cdot} = Z_{i\cdot} W^\top$. A two layer map of width w uses $\widetilde{Z}_{i\cdot} = V\,\sigma(W Z_{i\cdot} + b)$. A query bottleneck instead learns k queries and forms

$$\widetilde{Z}_{i\cdot} = \sum_{j=1}^{N} \alpha_{ij}\, Z_{j\cdot} W_V^\top, \qquad \alpha_{ij} = \mathrm{softmax}_j\big(q_i^\top W_K Z_{j\cdot}/(\tau\sqrt{d})\big). \tag{1}$$

Mean pooling is the same mixture with the weights held fixed, and a local convolution followed by a fixed pool, as in Honeybee, replaces α_{ij} by a compactly supported kernel (Cha et al., 2024). In both cases the map from keys to the pooled token remains a convex combination, so the mixing bound below still applies.

Assumption 1 (Frozen slice)**.** *g and T are held fixed while f_θ is trained. Updating the encoder or the language model can enlarge $\mathcal{U}$ or change the spectrum of Z.*

Assumption 2 (Task through Z)**.** *$Y \perp X \mid Z$. The connector reads features, and it has no further access to pixels.*

The object of study is therefore the channel

$$X \xrightarrow{g} Z \xrightarrow{f_\theta} \widetilde{Z} \xrightarrow{\Pi_{\mathcal{U}}} T \to Y. \tag{2}$$

Only f_θ is trained, and the language model reads $\widetilde{Z}$ through $\mathcal{U}$. A standard information bottleneck would continue to adapt g or T, and would price mutual information between a learnable code and its source. Here both ends stay frozen, so the free choices are the class of f_θ and the integer k. If the encoder is later unfrozen, the canonical tail of Z can lengthen, and if the language model is adapted, $\mathcal{U}$ can grow, so the bounds below remain bounds on the connector alone. The slice studied here is the one trained by LLaVA-1.5 and BLIP-2 (Liu et al., 2024; Li et al., 2023).

3.2 LANGUAGE REACHABLE INFORMATION

Pretraining calibrates T on word embeddings and on the residual stream those embeddings induce, so a frozen language model reads only part of $\mathbb{R}^d$. Definition 1 separates three constructions of that readable set. The theorems depend on the dimension u and on the projector onto the set, while the choice among the three constructions is left to measurement.

Definition 1 (Language reachable set)**.** *Let E be the embedding matrix of T. $\mathcal{U}_{\mathrm{pca}}$ is the leading principal subspace of the rows of E. $\mathcal{U}_{\mathrm{hid}}$ is the leading principal subspace of last layer hidden states on teacher forced captions. $\mathcal{U}_{\mathrm{jac}}$ is the leading left singular subspace of the Jacobian of caption negative log likelihood (NLL) with respect to a visual prefix in embedding space. Subsequent statements use $u = \dim \mathcal{U}$ and the orthogonal projector $\Pi_{\mathcal{U}}$ onto the chosen set.*

Embedding principal components describe the word table. Hidden states describe the residual stream on the captions that T is asked to produce. The prefix Jacobian describes the embedding directions that change caption NLL at first order. TinyAlign studies a related portion of mutual information that a capacity limited frozen reader can use (Hu et al., 2026). Here the set is read from hidden states and from prefix gradients of a released checkpoint, and Section 4 asks which of the three constructions the reader actually occupies.

Definition 2 (Usable information)**.** *Usable information is $I_{\mathcal{U}}(\widetilde{Z}; Y) = I(\Pi_{\mathcal{U}}\widetilde{Z}; Y)$, and the gap is $\Delta I(f) = I(Z;Y) - I_{\mathcal{U}}(f(Z);Y)$. The language reachable ceiling is $I^\star_{\mathcal{U}} = \sup_f I_{\mathcal{U}}(f(Z);Y)$, which cannot exceed $I(Z;Y)$.*

Data processing gives $\Delta I(f) \geq 0$, so the connector cannot create mutual information that is absent from Z. Coordinates placed in $\mathcal{U}^\perp$ are unread even when they remain inside $\widetilde{Z}$. With k tokens

the readable image therefore has ku real coordinates. If $r_\star$ is the smallest rank whose canonical correlation information meets a target I_{task}, a linear map keeps at most the leading $\min(r_{\text{map}}, ku)$ correlations (Hardoon et al., 2004; Chechik et al., 2003). On released embedding tables, u is comparable to d, and $ku \geq r_\star$ already holds at $k = 1$. That count records the width of the word table. The token budget is then spent on spatial detail and on the mixing law of Section 3.3, which is why MM1 and LLaVA retain many tokens after a linear projector has met a rank count (McKinzie et al., 2024; Liu et al., 2023).

Theorem 1 (Unread gap). *Let $f(Z) = AZ$, and let $W_\star Z$ be linearly sufficient for Y under a jointly Gaussian law. Write $W_\star Z = U + U^\perp$ with $U = \Pi_{\mathcal{U}} W_\star Z$, and parameterise each canonical variate by $\phi_i = \sqrt{1-\rho_i}\, u_i + \sqrt{\rho_i}\, u_i^\perp$ with $u_i \in \mathcal{U}$. Any connector that finishes with $\Pi_{\mathcal{U}}$ retains at most $I(U;Y)$. The unread gap equals*

$$\Delta_\rho = -\frac{1}{2} \sum_i \log \frac{1-\lambda_i^2}{1-\lambda_i^2(1-\rho_i)}, \tag{3}$$

and it vanishes if and only if $\rho_i = 0$ whenever $\lambda_i > 0$.

The gap is the linear loss that appears only because the reader is frozen. It is empty when $\mathcal{U}$ is taken to be the whole embedding space, so the identification in Definition 1 has to precede any rank count. On the Vicuna-7B table used by LLaVA-1.5, and on the Qwen3-VL tables, the embedding width at ninety percent of the variance is a few thousand. Hence $ku \geq r_\star$ holds at one token for every caption spectrum estimated below. The gap becomes informative only after $\mathcal{U}$ is the set that changes the next token loss.

3.3 Query mixing

A query token is a softmax mixture of encoder keys. The mixture is a deterministic function of Z and of the queries, and it therefore cannot raise mutual information with Y above $I(Z;Y)$. What can fall, once the queries have concentrated, is the share of that information which lands on the signal keys.

Theorem 2 (Signal mass). *Fix a signal set S of size m, and let $\eta = \sum_{j \in S} \alpha_j$ for a single query. The pooled token is $\tilde{z} = \eta s + (1-\eta) b$, where s is the renormalised mixture on S and b is the mixture on the complement. Then $I(\tilde{z};Y) \leq I(Z;Y)$. If (Z,Y) is jointly Gaussian and s is more correlated with Y than b is, then $I(\tilde{z};Y)$ is strictly increasing in η. Raising the encoder length N by noise keys, or raising the temperature τ at fixed logits, lowers $\mathbb{E}[\eta]$.*

For m keys of logit Δ and $N - m$ keys of logit 0, the signal mass is $\eta = me^\Delta/(me^\Delta + N - m)$. At $(\Delta, m) = (2, 4)$ this equals 0.71 when $N = 16$ and 0.11 when $N = 256$. Extra keys therefore reduce η when they repeat a correlation that the query has already selected. They raise usable information when they extend the canonical tail of Z, which is the sense in which higher resolution helps. A 2×2 spatial merge, as in the Qwen3-VL vision tower, shortens N by a local pool with fixed weights, and that pool remains a preserve map on a coarser grid. Theorem 2 prices learned queries. A hard selection of keys, as in VisionZip or DivPrune, keeps a support instead of mixing one, so the same statement applies to convex combinations (Yang et al., 2025; Alvar et al., 2025).

The monotone is also distinct from a peakedness score on the attention weights. The substitution $k_{\text{eff}} = k(1-\delta)$, with δ the total variation distance from a one hot vector, restates how concentrated those weights are. Theorem 2 instead links η to $I(\tilde{z};Y)$ after the connector has concentrated. A query that is still close to uniform lies outside that hypothesis. Section 4 therefore trains the queries until η exceeds chance, and only then lengthens N.

Each readable token still faces a coarser capacity bound. With second moment at most P and readout noise σ_0^2, one token carries at most $\frac{u}{2} \log\bigl(1 + P/(u\sigma_0^2)\bigr)$ nats after the power is divided across coordinates, or $u \log(R/\varepsilon) + O(1)$ nats after a ball of radius R is covered at scale ε (Cover et al., 1991). The factor $1/u$ divides a fixed power across readable coordinates, so k tokens supply at most k such capacities. A separate scale balances a linear context price αk against a power law tail $\lambda_i^2 = c\, i^{-\gamma}$, and the stationary point is $k = (\beta c/(2\alpha))^{1/\gamma}$. The expression is a heuristic, since α and β are chosen and are not identified from data. On a CLIP caption spectrum, the price $\alpha = 0.02$ used for grids in the style of LLaVA gives a scale near 8.5, and the price $\alpha = 0.15$ used for query

bottlenecks gives a scale near 2.2. Both lie well below the encoder length and well above the rank count $k_{\min} = 1$. After the dimension count has been met, a token heuristic can therefore still move the design.

3.4 Linear maps and two layer networks

Assumption 3 (Gaussian slice)**.** *(Z, Y) is zero mean and jointly Gaussian, with canonical correlations $\lambda_1 \geq \cdots \geq \lambda_m > 0$.*

Under Assumption 3, mutual information reduces to a sum over canonical pairs (Cover et al., 1991; Chechik et al., 2003).

$$I(Z;Y) = -\frac{1}{2}\sum_{i=1}^{m} \log(1-\lambda_i^2). \tag{4}$$

A linear connector keeps a leading piece of this sum, and the readable width ku limits how many terms remain.

Theorem 3 (Linear sufficiency)**.** *Let $f(Z) = AZ$ with $\operatorname{rank}(A) \leq r$, followed by $\Pi_{\mathcal{U}}$, and set $r_{\text{eff}} = \min(r, ku, m)$. Among linear images of rank at most r, $I(AZ;Y)$ is maximised by the leading r_{eff} canonical directions on the Z side, and the maximiser is unique whenever $\lambda_{r_{\text{eff}}} > \lambda_{r_{\text{eff}}+1}$. The residual from the sum equation 4 is $\Delta I = -\frac{1}{2}\sum_{i>r_{\text{eff}}} \log(1-\lambda_i^2) \geq \frac{1}{2}\sum_{i>r_{\text{eff}}} \lambda_i^2$.*

The maximum is the variational characterisation of canonical correlation analysis, in the form used by Gaussian information bottleneck theory (Hardoon et al., 2004; Chechik et al., 2003). Merullo et al. showed that one matrix already supports captioning when the encoder carries linguistic supervision (Merullo et al., 2023). Theorem 3 is the information identity behind that observation. If the same matrix places mass in $\mathcal{U}^{\perp}$, Theorem 1 adds the unread gap on top of the discarded tail. A two layer network is the next class when the target alignment is nonlinear, and the sample size then decides whether that class improves on a matrix.

Assumption 4 (Barron target)**.** *Each coordinate of a target alignment $g_\star$ has finite Fourier first moment C_f (Ma et al., 2022). The observation model is $Y = g_\star(Z) + \xi$ with independent Gaussian noise.*

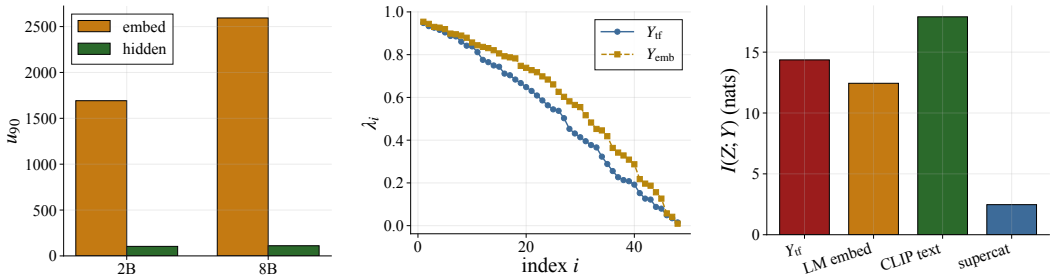
Theorem 4 (Two layer tradeoff)**.** *A two layer network of width w attains $\mathbb{E}\|f_w(Z) - g_\star(Z)\|^2 \leq C_f^2/w$. Empirical risk minimisation on n samples has risk $O(C_f^2/w) + O(wd\log n/n)$. For fixed width and excess risk ε, the sample size must therefore satisfy $n = \tilde{O}(wd/\varepsilon^2)$. If $g_\star$ is nonlinear, every matrix leaves a strictly positive information gap under Assumption 4.*

Greater width lowers the approximation term and raises the estimation term, so a linear map can have the lower risk when n is small. LLaVA-1.5 therefore pairs a two layer network with more instruction data, which is the joint change the bound requires (Liu et al., 2024). The six layer projector in Honeybee raises the effective width and inflates the same estimation term (Cha et al., 2024). If a linear probe residual already matches the tail of equation 4, the missing information is linear, and further depth does not remove it. The two layer class is supported once a nonlinear residual remains and n covers $wd\log n/n$.

Corollary 1 (Two regimes)**.** *If $g_\star$ is linear and n is large, every connector in the linear family attains the same canonical correlation ceiling. If $g_\star$ is nonlinear, a two layer network attains strictly more usable information than any matrix, once $n = \tilde{O}(wd/\varepsilon^2)$.*

Contrastive encoders already bring caption alignment close to a linear map, so Corollary 1 places a matrix at the ceiling (Radford et al., 2021). A two layer network remains appropriate when Y is the teacher forced loss of a frozen reader, because the crossing in the corollary belongs to a nonlinear teacher. Which visual information the language context can replace is a separate ranking, and it uses the redundancy lattice only to order truncation costs (Williams & Beer, 2010; Cover et al., 1991).

Proposition 1 (Truncation ranking)**.** *Let L be observed at test time, and let the downstream loss be bounded. Removing only redundancy R leaves excess risk 0. Removing unique vision U_Z leaves excess risk at least the Fano transform of the lost unique information. Removing a synergistic factor S reduces $I(Y;\widetilde{Z},L)$ to $I(Y;L)$.*



(a) Readable dimension. (b) Task spectra. (c) Information in Y.

Figure 2: Readable dimension and task spectra on released Qwen3-VL checkpoints. Hidden state principal components are compared with those of the embedding table. The same vision features are scored against teacher forcing states, caption vectors, and a discrete label.

Language context can replace a redundant copy of itself, and it cannot restore a visual coordinate that was never passed forward or complete a pairing that needs both factors. Fine grained and spatial tasks therefore load U_Z and S, whereas coarse semantic tasks load R (Lin et al., 2024; Cha et al., 2024; Li et al., 2025b). Proposition 1 organises those ablations, and the experiments report only the induced order.

4 Experiments

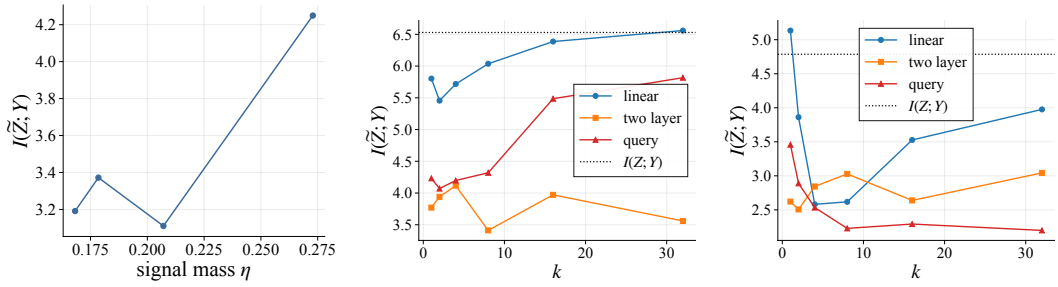
The measurements below test the three claims that can fail once the reader and the task are fixed. We identify $\mathcal{U}$ on released checkpoints, estimate spectra of one vision tower under several choices of Y, and train linear, two layer, and query maps on frozen features. The closed forms in Figure 1 check the Gaussian identities. What can still fail is the identification of $\mathcal{U}$, the dependence of the spectrum on Y, and the fall of usable information with signal mass after a query has concentrated.

4.1 The readable set and the task

A rank count on the embedding table does not limit the number of tokens, because that table is nearly full rank. On Vicuna-7B the leading 256 components hold 24% of the variance, against 6.3% under isotropy, and ninety percent of the variance still occupies 3015 dimensions. Qwen3-VL-4B, Qwen3-VL-2B, and Qwen3-VL-8B give the same picture, with ninety percent widths 2116, 1692, and 2594. Any caption spectrum whose canonical rank lies below a few thousand is then covered by one token. The count would begin to bind only if the readable width were the thin internal dimension of the reader.

Last layer hidden states on the same captions supply that thinner width. As shown in Figure 2, the ninety percent hidden width is 104 on Qwen3-VL-2B and 111 on Qwen3-VL-8B, and the leading 64 components hold 84% of that variance. The same count holds only 9.2% of the 2B embedding table and 7.7% of the 8B table. Overlap between the leading 32 embedding directions and the hidden subspace is 0.040 against chance 0.016 on the 2B reader, and 0.009 against chance 0.008 on the 8B reader. Overlap with the prefix gradient of caption loss is 0.016 on the 2B reader, against a random baseline of 0.015. The gradient span, fit on eighty captions, is too noisy to provide the width u. The overlap with the embedding table is already at chance, which is enough to set that table aside. Hidden states are therefore the readable set used below. A released LLaVA projector still places 1.74 times as much energy on the leading embedding components as a matched random map, so the geometry of training and the set that T reads are different objects.

Once that set is fixed, the spectrum is still a property of the pair (Z, Y). On four hundred COCO validation images the Qwen3-VL-2B merger yields features of shape 64×2048. Against teacher forcing hidden states the Gaussian formula gives 14.36 nats and canonical rank 24. Embeddings of the same captions inside the language model give 12.44 nats and rank 23, while CLIP text of those captions gives 17.90 nats and rank 26. A global CLIP feature against CLIP text lies nearby, at 16.26 nats and rank 25. The ranks stay in a narrow band, so a coarse preserve or compress reading is unchanged when a caption vector replaces teacher forcing. Mutual information still moves, by



(a) Mass and usable information. (b) CLIP ceiling. (c) Qwen ceiling.

Figure 3: Signal mass after concentration, and held out mutual information at a matched token count. Noise keys are appended once the queries have concentrated. Linear, two layer, and query maps are compared on CLIP patches and on Qwen merger tokens.

about one fifth between teacher forcing and CLIP text, and the caption vector is therefore a different geometry from the generation task. Figure 2 also shows a twelve way COCO supercategory on the same images, which collapses the spectrum to 2.47 nats and rank 6. The collapse shows that Y belongs in the object being estimated, and that a retrieval embedding can look richer than the task the reader is trained to solve.

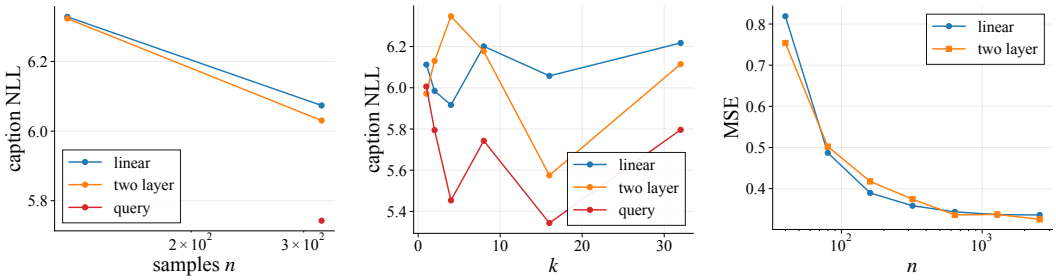
4.2 Signal mass after concentration

Theorem 2 applies after a query has concentrated, and a query that is still uniform does not test it. On real CLIP patches a random query is already close to uniform. We therefore plant the task in two of sixteen tokens and train two queries until signal mass reaches 0.27, above twice the chance level 0.125. Only after that peak do we append noise keys. Figure 3 shows that a longer encoder, with no new signal, lowers both signal mass and usable information, with rank correlation 0.40. A higher temperature at fixed length also lowers signal mass, while the Gaussian estimate of mutual information rises, so that sweep is set aside. The comparison that matches the theorem is the addition of noise keys after a peak exists. Extra keys help when they extend the canonical tail of Z. When they only repeat a mode that the query has already selected, signal mass falls and usable information falls with it. A high resolution recipe that keeps a linear projector leaves each key in its own token, so the mixing cost stays at zero.

The same order appears when the signal is not planted. On CLIP patches a hard spatial pool at 32 tokens reaches the held out mutual information of about 6.53 nats, and a trained query map remains 0.71 nats lower, as shown in Figure 3. On Qwen merger tokens the held out value is 4.79 nats, and the linear, two layer, and query maps reach 3.98, 3.04, and 2.20. In both cases the query map is the weakest reader of Y at a matched token count, which is the mixing cost in Theorem 2. On the Qwen tokens none of the three maps meets the ceiling, so those features are less linearly aligned than CLIP patches, while the order among families stays the same. When attention is diffuse, the effective token count stays far below k, in line with the concentration account of Zhang et al. (2025c).

4.3 Three families on one channel

Training only the connector puts the three families under the loss that the language model actually pays. CLIP ViT-B/32 patches and the Qwen3-VL-2B reader stay frozen, and the connector is fit for teacher forced caption loss on four hundred images, with eighty held out (Radford et al., 2021; Bai et al., 2025; Lin et al., 2014). At eight tokens the linear loss is 6.329 with 128 samples and 6.074 with 320 samples, while the two layer loss is 6.324 and 6.030. The two layer map is already slightly lower at the smaller sample size, and the gap at 320 samples is 0.044 nats. Corollary 1 allows a gap of that size when the alignment is nearly linear and the sample size has not yet paid the estimation term $wd \log n/n$. Query maps reach the lowest loss on the grid, 5.344 at sixteen tokens. Figure 4 places that minimum next to the scales 8.5 and 2.2 from the cheap and dear token prices in Section 3.3. Those scales have the same order of magnitude as the empirical minimum. Rank one retrieval on this slice stays between 0.01 and 0.04, so the caption loss is the quantity we read.



(a) Loss against sample size. (b) Loss against token count. (c) Nonlinear teacher.

Figure 4: Teacher forced caption loss of a frozen Qwen3-VL reader for linear, two layer, and query connectors. (a) shows the loss against sample size. (b) shows the loss against token count. (c) reports mean squared error under a high gain $\tanh$ teacher.

Caption loss of the frozen reader is a different score from mutual information against a CLIP text vector. A query map can sit below the CLIP ceiling and still reduce the loss of T, because the two scores use different coordinates. On CLIP caption vectors the linear map is best at every sample size in the grid, which is the linear teacher in Corollary 1. The crossing appears once the teacher is nonlinear. Figure 1 shows a $\tanh$ teacher overtaking the linear map at eighty samples, and Figure 4 shows that a high gain $\tanh$ teacher keeps only a small two layer advantage out to 2560 samples. A quadratic readout of trailing CLIP components still shows no two layer win at 320 samples. The comparison among families is therefore a statement about the teacher, and about whether the sample size covers the estimation term.

The same measurements organise the published connectors. LLaVA keeps a large token count with a linear map, and LLaVA-1.5 keeps that count with a two layer map and more instruction data, so both sit in the preserve family on a cheap token (Liu et al., 2023; 2024). BLIP-2 and Flamingo compress the grid into 32 and 64 queries, which the channel reads as a bottleneck under a dearer token price (Li et al., 2023; Alayrac et al., 2022). The Qwen3-VL merger is a two layer map after a 2×2 spatial pool, from 256 tokens down to 64, so it remains a preserve map on a coarser grid (Bai et al., 2025). Theorem 2 prices learned queries, and a fixed spatial pool lies outside that cost.

Because the embedding width makes the rank count equal to one in every case, LLaVA and BLIP-2 differ by the token price and by whether the map preserves the grid or compresses it. At large sample size on a nearly linear alignment, Corollary 1 places every linear family member on one ceiling. That is the regime in which MM1 finds connector type nearly idle once token count and resolution are fixed (McKinzie et al., 2024). On the same features, treated as redundant, unique, and synergistic bits, dropping a copy of the language costs nothing, while a unique visual bit costs 0.495 nats and a synergistic bit costs 0.47 nats. The order matches Proposition 1, and the values move across seeds, so only the order is a result. The readable set is a property of the frozen reader, the spectrum is a property of (Z, Y), and the query cost appears only after concentration. A contrastive encoder already meets the retrieval ceiling with a linear map, while caption loss of the frozen reader is the score that separates the three families. The small gap on that loss is the finite sample term in Corollary 1, and it grows into a two layer advantage only after the teacher is nonlinear.

5 Conclusion

This paper treats the connector as the channel that carries usable information from a frozen visual encoder into a frozen language model. The readable set is recovered from last layer hidden states and from prefix gradients of the caption loss, and on released Qwen3-VL checkpoints it is much thinner than the embedding table. The task spectrum comes from teacher forcing states. Linear maps, two layer networks, and query bottlenecks are then regimes of one channel. A linear map meets the canonical correlation ceiling on a nearly linear alignment, and a two layer network buys the nonlinear case at the sample cost paid by LLaVA-1.5. A query bottleneck spends signal mass, and usable information tracks that mass once the queries have concentrated. BLIP-2 and Flamingo occupy this compressed regime, and the Qwen3-VL spatial merger remains a preserve map, so a generous token budget keeps a matrix stable until the task is fine grained or the queries mix.

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

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*, 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.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13817–13827. IEEE, 2024.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for Gaussian variables. In *Advances in Neural Information Processing Systems*, volume 16, pp. 1–8, 2003.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*. wiley, 1991. ISBN 978-0471062592.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Yuanze Hu, Xinyu Wang, Zhichao Yang, Gen Li, Ye Qiu, Zhaoxin Fan, Yifan Sun, Wenjun Wu, Jin Dong, and Xiaotie Deng. TinyAlign: Boosting lightweight vision-language models by mitigating modal alignment bottlenecks. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4585–4597, 2026.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMlR, 2019.

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, 2023.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133(10):6794–6812, 2025a.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulic, and Anders Søgaard. Lost in embeddings: Information loss in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22676–22693, 2025b.

Junyan Lin, Haoran Chen, Dawei Zhu, and Xiaoyu Shen. To preserve or to compress: An in-depth study of connector selection in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5666–5680, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36, pp. 34892–34916, 2023.

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*, 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.

Chao Ma, Lei Wu, et al. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2022.

Ahmed Masry, Juan A Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-André Noël, et al. Alignvlm: Bridging vision and language latent spaces for multimodal understanding. In *ICLR 2025 Workshop on Representational Alignment (Re-Align)*, pp. 1–21, 2025.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *International Conference on Learning Representations*, 2023.

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*, pp. 22857–22867. IEEE, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356, 2024.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15537–15549, 2025.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *International Conference on Machine Learning*, pp. 1–39, 2025.

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*, pp. 14593–14603. IEEE, 2025.

Rui Xu, Yunke Wang, Yong Luo, and Bo Du. Rethinking visual token reduction in LVLMs under cross-modal misalignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27323–27331, 2026.

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.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. VoCo-LLaMA: Towards vision compression with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29836–29846. IEEE, 2025.

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*, volume 2025, pp. 53285–53310, 2025a.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*, pp. 74840–74857, 2025b.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Jiansheng Chen, Zhanhui Kang, Di Wang 0052, and Yu Wang. The security threat of compressed projectors in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20397–20407, 2025c.

APPENDIX

A Notation 14

B Linear identities 14

C Proof of the mixing theorem 16

D Two layer maps and token scale 17

E Truncation ranking and relaxations 18

F Identification of the readable set 20

G Task spectra and the choice of Y 22

H Query mixing after concentration 23

I Frozen connector training 25

J Gaussian checks and finite sample information 27

A NOTATION

Table 1 lists the symbols used in the theorems and the numerical results. The same list covers the readable set, the task variable, and the signal mass.

Table 1: Symbols used in the theorems and the numerical results.

symbol	meaning
$Z \in \mathbb{R}^{N \times d_v}$	frozen visual features
$\widetilde{Z} \in \mathbb{R}^{k \times d}$	connector output
d	language model embedding width
$\mathcal{U}, u$	language reachable set and its dimension
$\Pi_{\mathcal{U}}$	orthogonal projector onto $\mathcal{U}$
$I_{\mathcal{U}}(\widetilde{Z}; Y)$	$I(\Pi_{\mathcal{U}} \widetilde{Z}; Y)$
$\Delta I(f)$	$I(Z;Y) - I_{\mathcal{U}}(f(Z);Y)$
$I^{\star}_{\mathcal{U}}$	$\sup_f I_{\mathcal{U}}(f(Z);Y)$
λ_i	canonical correlations of (Z, Y)
r_{eff}	$\min(r_{\text{map}}, ku, m)$
$r_{\star}(I_{\text{task}})$	least rank meeting an information target
η	attention mass on a designated signal set
s, b	signal mixture and complement mixture
L	language context when a question is present
R, U_Z, S	PID redundancy, unique vision, synergy
P, σ_0^2	second moment and readout noise
α, β	chosen token price and residual weight

Three constructions of $\mathcal{U}$ appear in the numerical results. $\mathcal{U}_{\text{pca}}$ is the leading principal subspace of the embedding table, and it is kept as a control. $\mathcal{U}_{\text{hid}}$ is the leading principal subspace of last layer hidden states on teacher forced captions. $\mathcal{U}_{\text{jac}}$ is the leading left singular subspace of the Jacobian of caption negative log likelihood with respect to a visual prefix. The operational choice in the main text is $\mathcal{U}_{\text{hid}}$, since the Jacobian span is noisy at the available sample size.

Three constructions of Y are used on the same captions. Y_{tf} is the last layer hidden state under teacher forcing, Y_{lm} is the language model embedding of the caption string, and Y_{clip} is the CLIP text embedding of that string. A twelve class COCO supercategory indicator is included once, as a discrete task on the same images.

B LINEAR IDENTITIES

B.1 GAUSSIAN CCA

Let (Z, Y) be zero mean jointly Gaussian in $\mathbb{R}^{d_Z} \times \mathbb{R}^{d_Y}$. Let $\Sigma_Z = \mathbb{E}[ZZ^\top]$, $\Sigma_Y = \mathbb{E}[YY^\top]$, $\Sigma_{ZY} = \mathbb{E}[ZY^\top]$, all assumed full rank after restricting to the supports. The CCA generalised eigenproblem

$$\Sigma_Z^{-1} \Sigma_{ZY} \Sigma_Y^{-1} \Sigma_{YZ}\, a_i = \lambda_i^2 a_i$$

yields orthonormal directions a_i, b_i after whitening, with

$$\phi_i = a_i^\top Z, \qquad \psi_i = b_i^\top Y, \qquad \mathbb{E}\phi_i\phi_j = \delta_{ij}, \quad \mathbb{E}\psi_i\psi_j = \delta_{ij}, \quad \mathbb{E}\phi_i\psi_j = \lambda_i \delta_{ij}.$$

Different pairs (ϕ_i, ψ_i) are independent, and each scalar pair is a bivariate Gaussian with correlation λ_i. The Gaussian mutual information therefore splits as a sum over pairs (Cover et al., 1991),

$$I(Z;Y) = \sum_{i=1}^{m} I(\phi_i; \psi_i) = -\frac{1}{2} \sum_{i=1}^{m} \log(1 - \lambda_i^2). \tag{5}$$

Scalar formula. The joint covariance of (ϕ_i, ψ_i) is

$$\begin{pmatrix} 1 & \lambda_i \\ \lambda_i & 1 \end{pmatrix}, \qquad \det = 1 - \lambda_i^2.$$

Differential entropy of a p dimensional Gaussian is $\frac{1}{2}\log\big((2\pi e)^p \det\Sigma\big)$, so

$$h(\phi_i, \psi_i) - h(\phi_i) - h(\psi_i) = \frac{1}{2}\log(1-\lambda_i^2).$$

The identity $I = -h(\phi_i \mid \psi_i) + h(\phi_i)$ then recovers equation 5 for one pair.

Relation to Gaussian IB. Gaussian information bottleneck (Chechik et al., 2003) uses the same eigenvectors and rescales them by the bottleneck multiplier. A rank constraint with no additive compression noise is the large multiplier limit on a subspace of prescribed dimension, which is ordinary CCA truncation, and that truncation is the linear connector of Theorem 3. The quantities that belong to a frozen reader are the reachable set $\mathcal{U}$ and the token count k. The bottleneck prices $I(T;Z)$ for a learnable code, and the connector prices k tokens in a frozen context window. The mutual information is then read after $\Pi_{\mathcal{U}}$.

Lemma 1 (Interlacing)**.** *If $S \subset \mathbb{R}^{d_Z}$ is an r dimensional subspace and $\mu_1(S) \geq \cdots \geq \mu_r(S)$ are the canonical correlations between $\Pi_S Z$ and Y, then $\mu_j(S) \leq \lambda_j$ for each j, with equality throughout if and only if S contains the leading r CCA directions whenever those λ_j are simple.*

This is the variational characterisation of CCA (Hardoon et al., 2004). Because $t \mapsto -\log(1-t^2)$ is increasing on $[0,1)$, Lemma 1 maximises equation 5 over r dimensional linear images by keeping the leading r coefficients.

Lemma 2 (Unused coordinates)**.** *If A has rows in $\operatorname{span}\{a_1,\dots,a_r\}$ and $A^\perp$ completes an orthonormal basis of $\mathbb{R}^{d_Z}$, then $A^\perp Z \perp (AZ, Y)$ under Assumption 3.*

This is the Gaussian characterisation of uncorrelatedness. It is why discarding the trailing CCA directions cannot be repaired by a later nonlinear map of AZ alone.

Lemma 3 (Elementary logarithm)**.** *For $t \in [0,1)$, $t \leq -\log(1-t) \leq t/(1-t)$.*

The left inequality is the series $\sum_{m\geq 1} t^m/m \geq t$, and the right inequality is $\log(1-t) \geq -t/(1-t)$ on the same interval. Both convert CCA tails into comparable scales, while the mutual information itself remains equation 5.

B.2 Proofs of the linear statements

Rank. $\operatorname{rank}(A) \leq r$ implies that AZ is supported on an r dimensional subspace almost surely. Composing with $\Pi_{\mathcal{U}}$ and stacking k tokens yields at most ku real coordinates, so the effective dimension satisfies $r_{\text{eff}} \leq \min(r, ku, m)$. On released embedding tables u is comparable to d, and $ku \geq r_\star$ already holds at $k = 1$ for every caption spectrum in Section G. The count therefore describes the word table, and token count is spent on spatial detail and on mixing.

Optimum. On any r dimensional subspace S,

$$I(\Pi_S Z; Y) = -\frac{1}{2}\sum_{j=1}^{r}\log\big(1-\mu_j(S)^2\big).$$

Lemma 1 and monotonicity of $-\log(1-t^2)$ give the unique maximum at the leading CCA subspace, which is Theorem 3.

Loss. The residual is the tail of the CCA series. Lemma 3 yields $\Delta I \geq \frac{1}{2}\sum_{i>r_{\text{eff}}}\lambda_i^2$. A matching series $\sum_{m\geq 1}\lambda_i^{2m}/(2m)$ is available if a tighter expansion is wanted. $\Delta I = 0$ if and only if every discarded λ_i vanishes.

Unread gap. Let $W_\star$ achieve $I(W_\star Z; Y) = I(Z;Y)$. Decompose $W_\star Z = U + U^\perp$ with $U = \Pi_{\mathcal{U}} W_\star Z$. Joint Gaussianity of $(U, U^\perp, Y)$ and the chain rule yield

$$I(W_\star Z; Y) = I(U;Y) + I(U^\perp; Y \mid U).$$

Any connector that finishes with $\Pi_{\mathcal{U}}$ retains at most $I(U;Y)$, so the gap is $\Delta_\rho = I(U^\perp; Y \mid U)$. Parameterise each canonical variate as

$$\phi_i = \sqrt{1-\rho_i}\,u_i + \sqrt{\rho_i}\,u_i^\perp,$$

with $(u_i, u_i^\perp)$ orthonormal and u_i in $\mathcal{U}$. The conditional covariance of ψ_i given u_i is $1-\lambda_i^2(1-\rho_i)$, while the conditional covariance given ϕ_i is $1-\lambda_i^2$. A bivariate Gaussian calculation therefore gives

$$I(\phi_i;\psi_i) - I(u_i;\psi_i) = -\frac{1}{2}\log\frac{1-\lambda_i^2}{1-\lambda_i^2(1-\rho_i)}.$$

Summing over i gives equation 3. The gap vanishes if and only if $\rho_i = 0$ whenever $\lambda_i > 0$. The energy weighted scalar $\rho = \sum_i \lambda_i^2\rho_i / \sum_i \lambda_i^2$ recovers the first order expansion $\Delta_\rho = \frac{1}{2}\sum_i \lambda_i^2\rho_i + O(\|\rho\|^2)$, and that expansion stays outside the theorem statements.

Bivariate calculation. Let $(u, u^\perp, \psi)$ be jointly Gaussian, with u and $u^\perp$ standard and orthogonal, and $\psi = \lambda(\sqrt{1-\rho}\,u + \sqrt{\rho}\,u^\perp) + \sqrt{1-\lambda^2}\,\xi$ for an independent standard ξ. Then $\mathrm{Corr}(\phi,\psi) = \lambda$ with $\phi = \sqrt{1-\rho}\,u + \sqrt{\rho}\,u^\perp$, while $\mathrm{Corr}(u,\psi) = \lambda\sqrt{1-\rho}$. The two mutual informations are $-\frac{1}{2}\log(1-\lambda^2)$ and $-\frac{1}{2}\log(1-\lambda^2(1-\rho))$. Their difference is the summand of equation 3. If $\rho = 0$, the unread coordinate is independent of ψ and the gap vanishes. If $\rho = 1$, the entire canonical variate lies in $\mathcal{U}^\perp$ and the reader sees none of that pair.

The gap is nonzero only when $\mathcal{U}$ is the set that T reads. If $\mathcal{U}$ is the whole embedding space, then $\rho_i = 0$ identically and equation 3 is zero. Definition 1 therefore identifies the set before the gap is evaluated. Embedding PCA is nearly full rank, so it remains the control, and the theorem object is the thinner set that the reader uses.

Distribution free ceiling. Data processing gives $I(\Pi_{\mathcal{U}} f(Z);Y) \le I(Z;Y)$, and the bound $I(\cdot;Y) \le H(Y)$ is the Shannon ceiling. A Euclidean ball of radius R in dimension ku has ε covering number at most $(1+2R/\varepsilon)^{ku}$ (Cover et al., 1991). Quantising to the net produces a discrete random variable of entropy at most $ku\log(1+2R/\varepsilon)$, and the mutual information of the unquantised variable differs by a rate distortion remainder of $O(1)$ nats from the quantisation noise, uniformly on compact sets. When the linear family leaves a gap to $I(Z;Y)$, that residual is the information in a nonlinear sufficient statistic of Z for Y, which is the starting point of Theorem 4.

C Proof of the mixing theorem

Data processing. Softmax mixing is a deterministic function of Z and of the queries, so $I(\tilde{z};Y) \le I(Z;Y)$. Dropout at training time is additional noise and can only lower the mutual information. With frozen attention weights the map is linear of rank at most k, and Theorem 3 applies with $r_{\mathrm{map}} \le k$.

Mass and correlation. Let $\tilde{z} = \eta s + (1-\eta)b$ with $\eta = \sum_{j\in S}\alpha_j$, where s is the renormalised mixture on a signal set S of size m and b is the mixture on the complement. Under Assumption 3, $(\tilde{z}, Y)$ is jointly Gaussian after an affine reparameterisation, and

$$\mathrm{Corr}(\tilde{z},Y) = \eta\,\mathrm{Corr}(s,Y) + (1-\eta)\,\mathrm{Corr}(b,Y)$$

along each aligned coordinate. Whenever $\mathrm{Corr}(s,Y) > \mathrm{Corr}(b,Y)$, the correlation is strictly increasing in η, and so is $-\frac{1}{2}\log(1-\mathrm{Corr}^2)$. That monotone is Theorem 2.

Mass formula. Let m keys have logit Δ and $N-m$ keys have logit 0. The softmax mass on the signal set is

$$\eta = \frac{me^{\Delta}}{me^{\Delta} + N - m}.$$

This decreases in N at fixed (Δ, m) and decreases in τ when logits are scaled as Δ/τ. For $(\Delta, m) = (2, 4)$ one has $\eta = 0.71$ at $N = 16$ and $\eta = 0.11$ at $N = 256$. For $(\Delta, m) = (2, 2)$, as in the planted signal experiment, $\eta = 0.88$ at $N = 4$, $\eta = 0.58$ at $N = 16$, and $\eta = 0.13$ at $N = 256$ in the noiseless mass formula. Training moves Δ, so the realised $\eta = 0.27$ at $N = 16$ lies below the noiseless value and remains about twice chance. Appending keys whose features are independent of Y leaves $\mathrm{Corr}(b,Y)$ below the signal correlation, so both η and $I(\tilde{z};Y)$ fall. When the appended keys carry fresh signal, both η and $I(\tilde{z};Y)$ can rise, which is the high resolution case separated from pure dilution in the main text. A 2×2 spatial merge shortens N by a local pool with fixed weights, so the cost above applies to learned queries.

Peakedness and signal mass. Let $\delta = 1 - \mathbb{E}\|\alpha - e_{j_\star}\|_{\mathrm{TV}}$ be the total variation distance from a one hot vector, and set $k_{\mathrm{eff}} = k(1-\delta)$. The factor $1-\delta$ renames attention peakedness, and multiplying a slot capacity by that factor restates the same peakedness. Theorem 2 is the monotone from η to $I(\tilde{z}; Y)$ after a connector has concentrated. A uniform random query leaves η at chance, so the lengthening cut in Section H is run only after training has raised η above chance.

Hard slot capacity. Interaction information can be negative, so $I(\Pi_{\mathcal{U}}\widetilde{Z}; Y) \le \sum_i I(\Pi_{\mathcal{U}}\widetilde{Z}_i; Y)$ need not hold. The valid chain for discrete variables is $I(\Pi_{\mathcal{U}}\widetilde{Z}; Y) \le H(\Pi_{\mathcal{U}}\widetilde{Z}) \le \sum_{i=1}^{k} H(\Pi_{\mathcal{U}}\widetilde{Z}_i)$, and the analogous statement with differential entropy replaces the discrete entropy by a noise floor. Adding independent readout noise of variance σ_0^2 in each of the u coordinates of each token gives $h(\Pi_{\mathcal{U}}\widetilde{Z}_i \mid \text{noise}) \ge \frac{u}{2}\log(2\pi e\sigma_0^2)$. A second moment constraint $\mathbb{E}\|\Pi_{\mathcal{U}}\widetilde{Z}_i\|^2 \le P$ maximises differential entropy at a Gaussian, with $h(\Pi_{\mathcal{U}}\widetilde{Z}_i) \le \frac{u}{2}\log(2\pi e \cdot P/u)$. Subtracting these bounds yields at most $\frac{u}{2}\log(1 + P/\sigma_0^2)$ after the isotropic split P/u. A Euclidean ball of radius R in $\mathbb{R}^u$ has covering number at most $(1 + 2R/\varepsilon)^u$, hence entropy at most $u\log(1 + 2R/\varepsilon)$ after quantisation. When the k queries are perfectly correlated, the bound of k times one slot is loose, and it remains a valid necessary count.

Distances among samples. If the k queries are treated as random linear projections of Z, pairwise distances among n samples are preserved once $k = O(\varepsilon^{-2}\log n)$ (Dasgupta & Gupta, 2003). The statement concerns the Gram matrix of Z. The mixing monotone is the separate claim about $I(\widetilde{Z}; Y)$ after concentration.

D Two layer maps and token scale

D.1 Proof of the two layer tradeoff

Approximation. For each scalar coordinate, a width w superposition of sigmoids attains an $L^2(\mu)$ error of order $C_{f,j}^2/w$ when the Barron norm is finite (Ma et al., 2022). Summing the d coordinates and writing $C_f^2 = \sum_j C_{f,j}^2$ is the first claim of Theorem 4. The same $O(1/w)$ rate holds for ReLU networks on Barron spaces, up to an absolute factor.

Estimation. The Barron space bound controls the estimation term as well, and places it at order $O((wd\log n)/n)$ for bounded weight two layer networks (Ma et al., 2022). An alternative Rademacher bound (Bartlett & Mendelson, 2002) gives, for $\|V\|_F \le B_V$, $\|W\|_F \le B_W$, and activations with Lipschitz constant one,

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \le O\big(B_V B_W \sqrt{w/n}\big).$$

The two estimates agree when $B_V B_W \asymp \sqrt{wd}$, and the main text uses only $n = \tilde{O}(wd/\varepsilon^2)$.

Linear floor. Let $\Delta = g_\star(Z) - A_\star Z$ with $\mathbb{E}\|\Delta\|^2 = \varepsilon_{\mathrm{lin}}$. Under $Y = g_\star(Z) + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 I)$ independent of Z, the mutual information and minimum mean square error relation (Guo et al., 2005) differentiates mutual information in the signal to noise ratio and identifies the derivative with the error. Replacing $g_\star(Z)$ by $A_\star Z$ increases that error by a term of order $\mathrm{Var}(\Delta \mid A_\star Z)$. When Δ is not determined by $A_\star Z$ and the signal to noise ratio is bounded below,

$$I(g_\star(Z); Y) - I(A_\star Z; Y) \ge c(\sigma)\,\varepsilon_{\mathrm{lin}}.$$

Linearising $(g_\star(Z), A_\star Z, Y)$ as jointly Gaussian gives the same gap as the CCA mass on the residual directions, which is Theorem 3 applied to Δ. The Gaussian $\tanh$ world in Figure 1(c) uses this linearisation, since Y is $\tanh(AZ)$ plus Gaussian noise.

When the target is linear. If $g_\star(z) = Wz$, the approximation term vanishes for a matrix and the variance term is $O(dd_v/n)$. A width w network then pays an extra $O(wd\log n/n)$, so the linear map has the lower risk when n is small. That regime is the left half of Figure 1(c), and it is the CLIP caption regime of Corollary 1.

Sample arithmetic. Balancing C_f^2/w against $wd \log n/n$ gives $w \asymp C_f \sqrt{n/(d \log n)}$ and total risk $\tilde{O}(C_f \sqrt{d/n})$. Fixing w and demanding risk ε forces $n = \tilde{O}(wd/\varepsilon^2)$. Whenever $w \gg 1$, the two layer class needs more samples for the same ε, which is why LLaVA-1.5 changes the map and the instruction set together (Liu et al., 2024). The six layer projector in Honeybee raises the effective width and inflates the same estimation term (Cha et al., 2024).

D.2 Token scale as a heuristic

The main text uses the scale $k_{\text{scale}} = (\beta c/(2\alpha))^{1/\gamma}$, with chosen weights α and β, and the calculation below reproduces the values 8.5 and 2.2. Assume $\lambda_i^2 = c\, i^{-\gamma}$ with $\gamma > 1$ and $ci^{-\gamma} < 1$ on the indices in use. Lemma 3 and an integral comparison then yield

$$\Delta I(k) = \Theta\big(k^{1-\gamma}\big)$$

as $m \to \infty$. A linear context price plus the tail, $J(k) = \alpha k + \beta \Delta I(k)$, has continuous derivative $J'(k) = \alpha + \beta \Delta I'(k)$ with $\Delta I'(k) \sim -\frac{c}{2} k^{-\gamma}$. The unique zero is $k_{\text{scale}} = (\beta c/(2\alpha))^{1/\gamma}$ after folding $1/(1-\lambda_1^2)$ into c, and convexity of the relaxation makes that critical point a minimum of J.

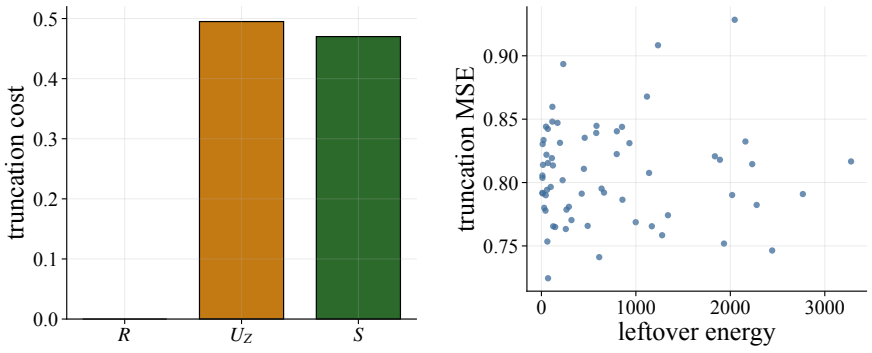
On the CLIP caption spectrum, $\gamma = 1.51$ and $c \approx 1$. The cheap token price $\alpha = 0.02$ used for LLaVA style grids yields $k_{\text{scale}} \approx 8.5$, and the dearer price $\alpha = 0.15$ used for query bottlenecks yields $k_{\text{scale}} \approx 2.2$. Both values lie well below the encoder grid, while the rank count on these spectra is $k_{\min} = 1$. Frozen connector negative log likelihood on a Qwen3-VL-2B reader has its minimum at $k = 16$, which matches the cheap scale in order.

If instead $\lambda_i^2 = e^{-i/\tau}$, a geometric sum produces $\Delta I(k) \sim \frac{\tau}{2} e^{-k/\tau}$ and a logarithmic scale $\tau \log(\beta/(2\alpha\tau))$, or a minimum at zero tokens when the tail is already cheaper than one token. Every caption spectrum in Section G follows a power law, and the exponential tail is the closed form of the same derivative.

E Truncation ranking and relaxations

E.1 Ranking and proxies

Williams and Beer decompose $I(Y; Z, L)$ on the redundancy lattice (Williams & Beer, 2010). When only the redundant atom is dropped and L remains, the Bayes risk is unchanged. When a unique visual atom of b bits is dropped, the Fano inequality gives a strictly positive excess risk (Cover et al., 1991). When a synergistic factor is removed, $I(Y; \widetilde{Z}, L) = I(Y; L)$, and on XOR that residual is a half error. The ranking $R \ll U_Z \sim S$ organises published preserve and compress comparisons. Fine grained and spatial tasks load U_Z and S, while coarse semantics load R (Lin et al., 2024; Cha et al., 2024; Li et al., 2025b).



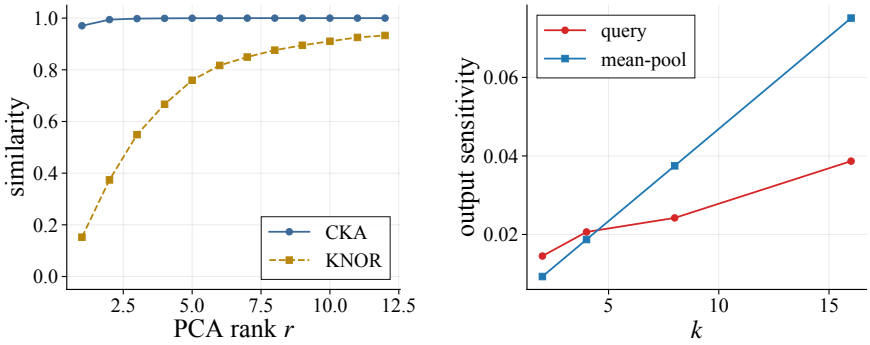
(a) PID truncation costs.

(b) ReLU leftover energy.

Figure 5: Truncation costs and a reconstruction proxy. (a) replaces redundant, unique, and synergistic coordinates by noise. (b) compares reconstruction error with leftover PCA energy on frozen ReLU encoders.

The CLIP bit construction in the main text treats a copy of the language as R, a visual unique coordinate as U_Z, and a pairing that needs both as S. Truncation costs are excess risks after the corresponding coordinate is replaced by noise. The stable quantity is the order of those costs, and the absolute values move across seeds on four hundred images. Figure 5 shows the CLIP bit truncation costs together with the leftover energy proxy.

Nearest neighbour overlap after a Lipschitz map of local Gram distances satisfies $\mathbb{E}[\text{overlap}] \geq 1 - C\sqrt{1 - \text{CKA}}$ (Kornblith et al., 2019). For jointly Gaussian $(Z, \widetilde{Z})$, optimal linear reconstruction error equals the rate distortion error at rate $I(Z; \widetilde{Z})$ (Cover et al., 1991). When $I_0 = I_{\mathcal{U}}(\widetilde{Z}; Y)$ is packed into effective dimension q, a unit perturbation of $\widetilde{Z}$ has mean quadratic excess risk at least $cI_0\varepsilon^2/q$. Softmax Jacobians concentrate on a selectable support, whereas spatial mean pooling keeps a uniform Jacobian (Zhang et al., 2025c). A sharp attention map therefore sends a perturbation of the attended token into the whole output, while pooling N tokens into k bins keeps the Jacobian spread across bins. On sixty four frozen ReLU worlds, leftover PCA energy tracks reconstruction error with Spearman correlation 0.9998 and tracks task error with correlation 0.55, so the non Gaussian truncation bound is nearly tight for reconstruction and remains a monotone proxy for the task. Figure 6 shows the neighbourhood overlap and the query versus pool sensitivity.



(a) CKA and neighbourhood overlap. (b) Query versus mean pool sensitivity.

Figure 6: Geometry proxies under concentration. (a) tracks neighbourhood overlap against CKA along a PCA rank sweep. (b) compares a query map with a spatial pool at matched token count.

On the same CLIP bits, dropping a copy of the language costs 0, dropping a visual unique bit costs 0.495, and dropping a synergistic bit costs 0.47, which matches the order in Proposition 1. Inverse eigenvalues are a weaker reconstruction proxy than leftover energy, as shown in Figure 7.

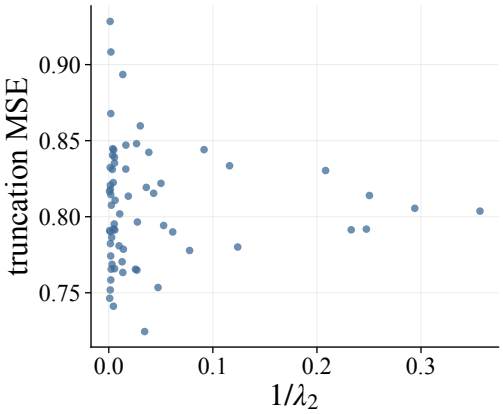


Figure 7: Inverse eigenvalues of frozen ReLU features, compared with leftover PCA energy as a reconstruction proxy.

E.2 RELAXATIONS

AlignVLM constrains visual tokens to the convex hull of word embeddings (Masry et al., 2025). Let $\mathcal{U}_{\mathrm{conv}}$ denote that compact convex set. Replacing u by the logarithm of its covering number yields a capacity bound for a nonlinear reachable set, and the linear statements continue to apply on the linear span of $\mathcal{U}_{\mathrm{conv}}$. The construction extends $\mathcal{U}$ beyond a subspace. Hidden state principal components and prefix gradients remain the operational constructions in the experiments.

Training the encoder can lengthen $\{\lambda_i\}$ and raise $I(Z;Y)$, and training the language model can enlarge $\mathcal{U}$ toward $\mathbb{R}^d$. Every frozen upper bound remains a bound on the connector alone, and the estimates of $\mathcal{U}$ and of the spectrum are then repeated on the new pair. The present paper studies the frozen slice that LLaVA-1.5 and BLIP-2 already train (Liu et al., 2024; Li et al., 2023). Flamingo style gated cross attention inside the language model is a further processing of the same k tokens (Alayrac et al., 2022), so the mixing bound still applies to the visual prefix and the context price includes the extra blocks.

F IDENTIFICATION OF THE READABLE SET

Embedding PCA of a released checkpoint is the convenient construction of $\mathcal{U}$. On Vicuna-7B the leading 256 principal components hold 24% of the row centred variance, against 6.3% under isotropy, and the corresponding shares are 20% against 10% on Qwen3-VL-4B and 19% against 6.3% on Qwen3-VL-8B. Ninety percent of that variance still requires $u_{90} = 3015$ on Vicuna, $u_{90} = 2116$ on Qwen3-VL-4B, $u_{90} = 1692$ on Qwen3-VL-2B, and $u_{90} = 2594$ on Qwen3-VL-8B, as shown in Table 2. A rank count $k_{\min} = \lceil r_\star/u \rceil$ is then 1 for any caption spectrum whose CCA rank lies below a few thousand. The count describes the word table, and token count is spent on spatial detail and on mixing.

Table 2: Readable dimension on released checkpoints.

source	d	u_{90}	top 32 overlap with embed PCA
Vicuna-7B embed PCA	4096	3015	none
Qwen3-VL-4B embed PCA	2560	2116	none
Qwen3-VL-2B embed PCA	2048	1692	1
Qwen3-VL-8B embed PCA	4096	2594	none
Qwen3-VL-2B hidden PCA	2048	104	0.040
Qwen3-VL-8B hidden PCA	4096	111	0.009
Qwen3-VL-2B prefix Jacobian	2048	1844	0.016
Qwen3-VL-8B prefix Jacobian	4096	3687	0.008
random baseline, $d = 2048$	2048	none	0.015
random baseline, $d = 4096$	4096	none	0.008

Figure 8 shows the same anisotropy on the released embedding tables. The curves rise slowly, in line with the large embedding values of u_{90}.

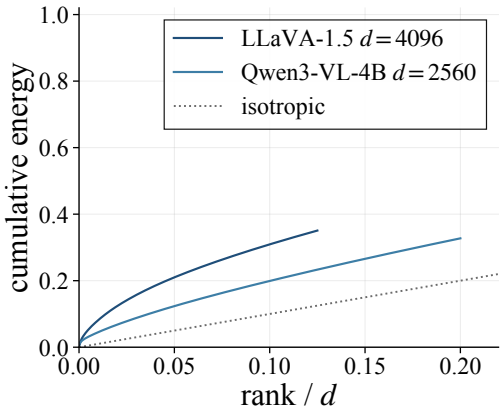


Figure 8: Cumulative variance of released embedding tables. Each curve is the row centred energy in the leading principal components.

Last layer hidden states on teacher forced COCO captions occupy a much thinner set. On Qwen3-VL-2B, $u_{90} = 104$, and the leading 64 hidden principal components already hold 84% of that variance, against 9.2% for the embedding table. On Qwen3-VL-8B, $u_{90} = 111$, and the leading 64 components hold 84%, against 7.7% for the embedding table. Table 3 gives the cumulative shares at several ranks.

Table 3: Cumulative variance shares. Isotropy is r/d.

source	$r = 64$	$r = 128$	$r = 256$	$r = 512$
Vicuna embed (isotropy)	0.107 (0.016)	0.161 (0.031)	0.238 (0.063)	0.351 (0.125)
Qwen3-VL-4B embed	0.078 (0.025)	0.124 (0.050)	0.199 (0.100)	0.327 (0.200)
Qwen3-VL-2B embed	0.092 (0.031)	0.149 (0.063)	0.241 (0.125)	0.392 (0.250)
Qwen3-VL-8B embed	0.077 (0.016)	0.121 (0.031)	0.191 (0.063)	0.309 (0.125)
Qwen3-VL-2B hidden	0.843 (0.031)	0.922 (0.063)	0.980 (0.125)	1.000 (0.250)
Qwen3-VL-8B hidden	0.841 (0.016)	0.915 (0.031)	0.976 (0.063)	1.000 (0.125)

The overlap between two r dimensional orthonormal bases Φ and Ψ is $r^{-1}\|\Phi^\top \Psi\|_F^2$, which equals one when the subspaces coincide and equals r/d in expectation for a Haar random Ψ in $\mathbb{R}^d$. On the 2B reader the overlap between embedding PCA and hidden PCA is 0.040, and on the 8B reader it is 0.009, with $r = 32$. The overlap between embedding PCA and the prefix gradient span is 0.016 on 2B and 0.008 on 8B, which matches the random baselines 0.015 and 0.008. The overlap between hidden PCA and the gradient span is 0.019 on 2B and 0.005 on 8B. The leading embedding principal components are therefore nearly orthogonal to the set the language model reads. Prefix gradients use four learnable visual tokens prepended to the caption, and the Jacobian of caption negative log likelihood is taken in embedding space and reduced by a left singular decomposition. The Jacobian u_{90} is 1844 on eighty captions in 2048 dimensions on 2B, and 3687 in 4096 dimensions on 8B, so the span supplies an overlap check and hidden PCA supplies the dimension. At $r = 32$ the chance values are 0.016 and 0.008, matching Table 4.

Table 4: Top 32 subspace overlap. Chance is r/d.

pair	2B ($d = 2048$)	8B ($d = 4096$)
embed PCA / hidden PCA	0.040	0.009
embed PCA / prefix Jacobian	0.016	0.008
hidden PCA / prefix Jacobian	0.019	0.005
embed PCA / random	0.015	0.008
chance r/d	0.016	0.008

Figure 9 shows the same 2B overlaps against the chance line. Hidden PCA sits above that line, and the gradient span sits on it.

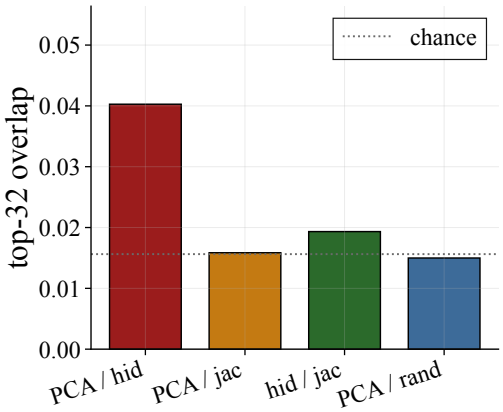


Figure 9: Top 32 subspace overlap on the 2B reader. The dotted line is the chance overlap r/d.

A hidden overlap of 0.040 on 2B is about two and a half times chance, and the 8B overlap 0.009 sits at chance, so their leading sets remain nearly orthogonal. Embedding u_{90} makes $k_{\min} = 1$. Hidden u_{90} is 104 on 2B and 111 on 8B, so a rank count would bind only for a task whose CCA rank exceeds one hundred, and none of the caption spectra in Section G reach that rank. The operational content of $\mathcal{U}_{\text{hid}}$ is that the unread gap equation 3 can be nonzero once the reader is this thinner set. Hidden PCA is therefore the operational $\mathcal{U}$, and embedding PCA remains the control.

The released LLaVA projector places $1.74\times$ more energy in the leading 64 embedding principal components than a matched random network, with ratios 1.70, 1.51, and 1.31 at widths 128, 256, and 512, as shown in Table 5. That energy describes the training of the projector. The reader T is the hidden state construction above.

Table 5: LLaVA projector energy in leading embedding principal components versus a matched random network.

leading components	projector	random	ratio
64	0.027	0.016	1.74
128	0.053	0.031	1.70
256	0.095	0.063	1.51
512	0.164	0.125	1.31

TinyAlign effective mutual information is the closest published object to $I_{\mathcal{U}}$ (Hu et al., 2026). The present paper identifies $\mathcal{U}$ from hidden states and prefix gradients of a released checkpoint, and it prices mixing and token count on that set.

G Task spectra and the choice of Y

Canonical spectra are computed after a joint PCA to a rank that is a fraction of n, which keeps the sample covariance invertible. Reported I_{cca} uses equation 5 on that rank, and in sample ridge formulae that interpolate when $d > n$ are set aside. On Qwen merger tokens of width 2048 with $n = 400$, the reported spectra therefore use the reduced features. Table 6 collects the pairs used in the main text.

Table 6: Task spectra on four hundred COCO images.

pair	I_{cca}	$r_\star$	γ
CLIP glob / CLIP text	16.26	25	1.51
Qwen Z / language model caption embed	12.44	23	1.55
Qwen Z / teacher forcing hidden	14.36	24	1.49
Qwen Z / CLIP caption text	17.90	26	1.26
Qwen Z / COCO supercategory	2.47	6	4.53
Qwen3-VL-8B Z / teacher forcing hidden	12.29	23	1.65
Qwen3-VL-8B Z / CLIP caption text	15.00	26	1.34

Teacher forcing hidden states on Qwen3-VL-2B give $I = 14.36$ nats and $r_\star = 24$, while language model caption embeddings of the same strings give $I = 12.44$ nats, $r_\star = 23$, and tail exponent $\gamma = 1.55$. CLIP text of those captions gives $I = 17.90$ nats, $r_\star = 26$, and a flatter tail $\gamma = 1.26$, and a CLIP ViT-B/32 global feature against CLIP text gives $I = 16.26$ nats, $r_\star = 25$, and $\gamma = 1.51$. A twelve class COCO supercategory indicator falls to $I = 2.47$ nats and $r_\star = 6$. On Qwen3-VL-8B the teacher forcing value is $I = 12.29$ nats with $\gamma = 1.65$, and CLIP text of the same captions gives $I = 15.00$ nats with $\gamma = 1.34$, so the gap between generation and retrieval remains after the change of tower.

Teacher forcing and CLIP text differ by about one fifth in mutual information, and the two retrieval constructions sit on either side of the generation spectrum. The ordering of the families is stable under that swap, which shows that the reading is coarse, and the task variable for generation remains the teacher forcing state. The supercategory collapse shows, in addition, that the spectrum follows the choice of Y. Estimating $\{\lambda_i\}$ from CLIP text and then training on teacher forced negative log likelihood therefore compares two teachers. Section I is the loss of the frozen reader, and Section H together with the CLIP ceiling in Figure 3(b) is the retrieval teacher. Corollary 1 allows these two scores to separate the families differently, and the measurements do separate them. Figure 10 shows the caption embedding spectrum used for that comparison.

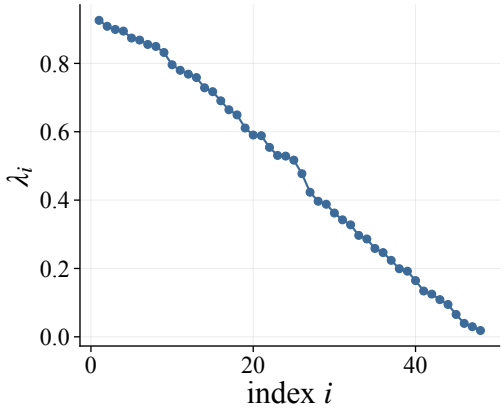
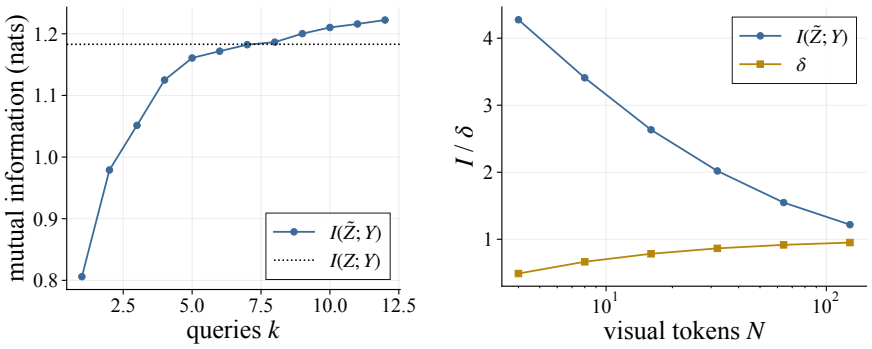


Figure 10: Canonical correlations of Qwen merger tokens against language model caption embeddings.

H Query mixing after concentration

Random queries on real CLIP patches are already nearly uniform, with $\delta \approx 0.96$. Figure 11 shows this diffuse control. Attention entropy is high and η sits near chance, so the lengthening cut has no concentrated signal set to dilute, and the figure is the control for Theorem 2.



(a) Query capacity in a Gaussian world.

(b) Random queries as N grows.

Figure 11: Controlled query worlds. (a) shows usable information against token count under a prescribed spectrum. (b) shows random queries as encoder length grows, with attention remaining diffuse.

The reported cut plants the task in $m = 2$ of $N = 16$ tokens, trains two queries until signal mass exceeds twice chance, with $\eta = 0.27$ against 0.125, and only then appends noise keys that are independent of Y. Table 7 gives that lengthening sweep. As N moves from 16 to 80, η falls from 0.273 to 0.168 and $I(\widetilde{Z};Y)$ falls from 4.25 to 3.19 nats, with rank correlation 0.40, which is the cut named by the theorem.

Table 7: Signal mass and usable I after a trained peak, then encoder lengthening without new signal.

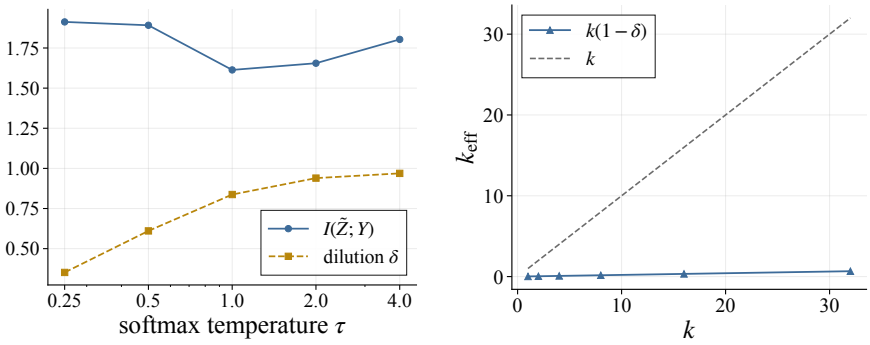
N	η	$I(\widetilde{Z};Y)$	MSE
16	0.273	4.25	0.815
32	0.207	3.11	0.980
48	0.178	3.37	0.956
80	0.168	3.19	1.090

Table 8 gives temperature at fixed N after the same peak. Raising τ from 0.25 to 4 lowers η from 0.273 to 0.176 and raises a Gaussian plug in estimate of I from 4.25 to 7.40 nats. Once the joint is poorly conditioned, that estimate separates from the population monotone of Theorem 2, and the main text uses the lengthening cut.

Table 8: Temperature after the trained peak. The columns are signal mass, a Gaussian plug in estimate of mutual information, and attention peakedness.

τ	η	plug in I	δ
0.25	0.273	4.25	0.260
0.70	0.276	5.08	0.452
1.00	0.271	5.50	0.547
2.00	0.234	6.41	0.724
4.00	0.176	7.40	0.839

Figure 12(a) is a temperature sweep on random queries, with no planted signal. Figure 12(b) is the effective token count on CLIP patches once attention is already diffuse.



(a) Temperature without a planted peak.

(b) Effective token count on CLIP patches.

Figure 12: Temperature and effective token count. (a) sweeps softmax temperature on random queries. (b) shows $k(1-\delta)$ on CLIP patches under diffuse attention.

On CLIP patches a hard spatial pool at $k = 32$ reaches the held out mutual information $I(Z;Y) \approx 6.53$ nats, as shown in Table 9. The remaining columns compare a linear map, a two layer map, and a query map at the same token counts.

Table 9: CLIP patch ceilings. Held out $I(Z;Y) = 6.53$ nats.

k	linear	two layer	query	$k(1-\delta)$
1	5.80	3.77	4.23	0.02
4	5.72	4.12	4.20	0.08
8	6.04	3.41	4.32	0.17
16	6.39	3.97	5.49	0.33
32	6.56	3.56	5.82	0.67

A two layer readout of the same tokens stays below that pool, and a trained query map remains 0.71 nats below it. The effective count $k(1-\delta)$ stays well below k whenever attention is diffuse. On Qwen merger tokens the held out mutual information is 4.79 nats, and the three families reach 3.98, 3.04, and 2.20, as shown in Table 10.

Table 10: Qwen merger token ceilings. Held out $I(Z;Y) = 4.79$ nats.

k	linear	two layer	query
1	5.13	2.62	3.46
4	2.58	2.84	2.53
8	2.62	3.03	2.23
16	3.53	2.64	2.29
32	3.98	3.04	2.20

In both tables the query map is the weakest reader of a caption vector at matched k. That gap is the mixing cost of a concentrated query, evaluated at a token count shared with the linear map and the two layer map.

I FROZEN CONNECTOR TRAINING

Both the CLIP ViT-B/32 encoder and the Qwen3-VL-2B language model are frozen, and only the connector is trained. The objective is teacher forced caption loss on four hundred COCO validation images, with a held out set of eighty images. The three families are a linear map, a two layer GELU map, and a query map with k learnable queries. The grid uses $k \in \{1, 2, 4, 8, 16, 32\}$ and $n \in \{128, 320\}$, two seeds where recorded, and twenty seven completed runs, with four hundred steps, batch size four, learning rate 10^{-3}, and a caption cap of thirty two tokens. Query temperature is swept in $\{0.5, 1, 2, 4\}$ at $k = 8$. Table 11 gives the main losses. At $k = 8$ the linear map moves

from 6.329 at $n = 128$ to 6.074 at $n = 320$, and the two layer network moves from 6.324 to 6.030, a gap of 0.044 nats at the larger sample. The query map reaches the lowest loss on the grid, 5.344 at $k = 16$.

Table 11: Frozen connector training. Caption NLL of a frozen Qwen3-VL-2B reader.

family	$k = 8, n = 128$	$k = 8, n = 320$	best k	best NLL
linear	6.329	6.074	4	5.916
two layer	6.324	6.030	16	5.575
query	untrained	5.743	16	5.344

Table 12 gives the loss against k at $n = 320$. The linear map is lowest at $k = 4$, and the two layer and query maps are lowest at $k = 16$. Retrieval at rank one stays between 0.01 and 0.04, so the caption loss remains the training score.

Table 12: Caption NLL against k at $n = 320$.

family	1	2	4	8	16	32
linear	6.113	5.985	5.916	6.046	6.058	6.217
two layer	5.971	6.130	6.347	6.030	5.575	6.115
query	6.006	5.794	5.454	5.743	5.344	5.796

Table 13 gives query loss against temperature at $k = 8$ and $n = 320$. Effective token count falls as τ rises from 0.5 to 4, and the loss on this slice is lowest at $\tau = 2$. Theorem 2 ties η to usable information after a peak exists, and the loss minimum in τ is a property of the frozen reader.

Table 13: Query connector at $k = 8$, $n = 320$. Effective count falls with temperature.

τ	caption NLL	$k(1 - \delta)$	δ
0.5	5.717	0.48	0.940
1.0	5.743	0.32	0.960
2.0	5.363	0.33	0.959
4.0	5.948	0.18	0.978

Caption loss of a frozen reader and mutual information against a CLIP text vector use different coordinates. A query map can sit below the CLIP ceiling and still lower the teacher forced loss, because the reader is T. On CLIP caption vectors the linear map has the lower loss at every sample size in the retrieval sweep, which is the linear teacher of Corollary 1. Figure 13 shows this comparison.

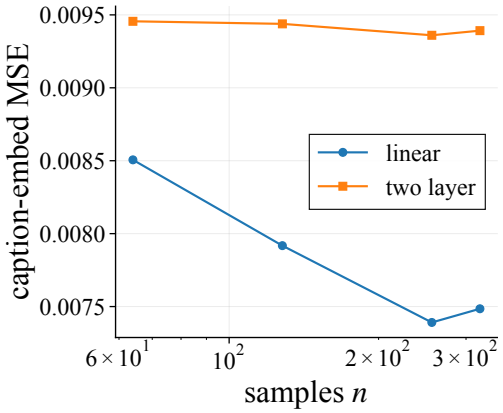


Figure 13: Caption loss on CLIP text vectors for a frozen reader. The linear map stays below the two layer network across the sample sweep.

A $\tanh$ teacher in a Gaussian world crosses at $n = 80$, as shown in Table 14. The columns compare ridge with a width 32 network on that nonlinear teacher.

Table 14: Ridge versus a width 32 $\tanh$ network on a nonlinear Gaussian teacher.

n	MSE linear	MSE two layer	difference
40	0.256	0.272	$+0.016$
80	0.220	0.176	-0.044
160	0.204	0.135	-0.069
320	0.196	0.105	-0.091
640	0.194	0.078	-0.116
1280	0.190	0.101	-0.089
2560	0.190	0.086	-0.103

A high gain $\tanh$ teacher keeps a small two layer advantage out to $n = 2560$. A quadratic readout of trailing CLIP principal components keeps the linear map ahead at $n = 320$. The comparison among families therefore tracks the teacher, and it tracks whether n covers the estimation term $wd \log n/n$.

J Gaussian checks and finite sample information

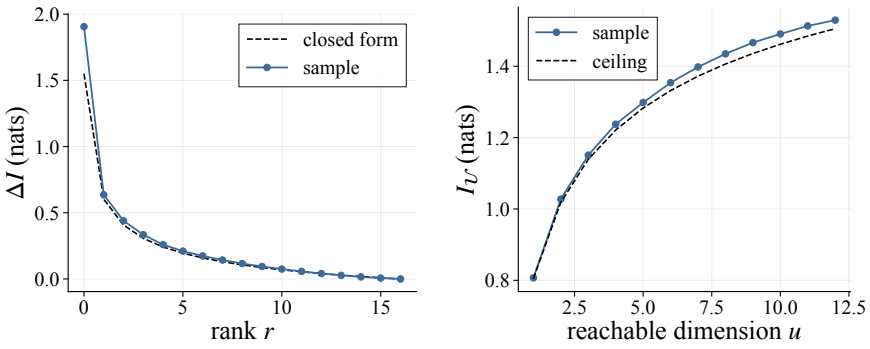
J.1 Gaussian world checks

Jointly Gaussian worlds check the closed forms, and the released checkpoints remain the empirical object. A prescribed CCA spectrum is sampled, a linear map of rank r is fitted, and the empirical residual is compared with $-\frac{1}{2}\sum_{i>r}\log(1-\lambda_i^2)$. Table 15 gives that sweep. From $r \geq 1$ the mean absolute gap is 0.032 nats, or 2.1% of the full $I(Z;Y) = 1.550$. The gap at $r = 0$ is the plug in bias of a $24 + 16$ dimensional covariance on $n = 8000$, and the comparison uses the ranks above zero.

Table 15: Closed form CCA residual versus sample, prescribed power law spectrum, $n = 8000$.

r	ΔI closed	ΔI sample	I kept
0	1.550	1.906	0.000
1	0.602	0.636	0.949
2	0.407	0.439	1.143
4	0.242	0.259	1.309
8	0.107	0.116	1.444
12	0.041	0.041	1.509
16	0.000	0.000	1.550

Figure 14(a) repeats the residual comparison at a larger size than the main text row, and Figure 14(b) shows saturation in the reachable dimension. The closed form and the sample residual are the two curves in Figure 14(a).



(a) CCA residual, closed versus sample.

(b) Saturation in reachable dimension.

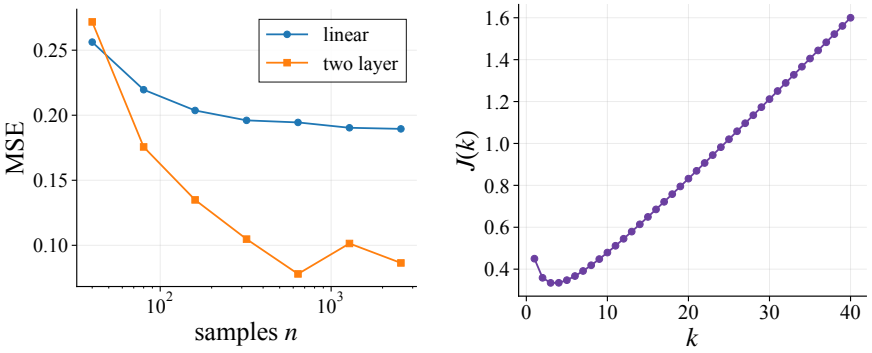
Figure 14: Gaussian identities at a larger size than the main text. (a) compares the linear residual with the CCA tail. (b) shows usable information as the reachable dimension grows.

A reachable dimension sweep at fixed k saturates within 0.025 nats of the $u = 3$ ceiling once u meets the CCA rank, as shown in Table 16. The right hand columns of the same table add tokens at $u = 3$.

Table 16: Reachable ceiling versus u, and extra tokens at $u = 3$.

u	ceiling	sample I	k at $u = 3$	sample I
1	0.805	0.807	1	1.143
3	1.139	1.151	2	1.150
6	1.332	1.354	4	1.157
9	1.436	1.466	8	1.168
12	1.505	1.529		

Extra tokens at fixed $u = 3$ change I by 0.025 nats, which is the expected scale when the bottleneck is u. A two layer network on a $\tanh$ teacher first falls below the linear map at $n = 80$, as already shown in Table 14. A linear context price plus a power law tail produces an empirical minimum within 0.27 tokens of the stationary point, averaged over four (α, β) pairs. Figure 15 shows the sample split and the token scale. These checks are Figure 1, and they support the Gaussian formulae as the linear reading of a measured spectrum.



(a) Nonlinear teacher, sample split.

(b) Token scale heuristic.

Figure 15: Sample split and token scale. (a) shows a two layer network and the linear map on a $\tanh$ teacher. (b) shows a linear context price plus a power law tail, with weights α and β chosen.

J.2 Finite sample mutual information

The Gaussian plug in $\hat{I} = \frac{1}{2} \log(\det \hat{\Sigma}_X \det \hat{\Sigma}_Y / \det \hat{\Sigma}_{XY})$ is biased upward when the joint dimension is not small compared with n. A consistent alternative evaluates I on the empirical canonical correlations and then applies equation 5, after a joint PCA that keeps n several times the retained rank. All I_{cca} numbers in Table 6 use that route, and in sample ridge formulae that interpolate when $d > n$ are set aside.

Table 8 is the case in which the plug in estimate and the population monotone separate. As τ rises, attention flattens, η falls, and the plug in estimate of I rises. Under the correlation comparison of Section C, the population mutual information moves with η, so the main text uses the lengthening cut.

Assumption 3 is the joint law for which mutual information has a closed form in a spectrum, the optimal linear connector is explicit, and rate distortion, the information bottleneck, and CCA coincide. Data processing, Barron approximation, the per token entropy cap, the mixing monotone after an affine reparameterisation, and the redundancy lattice on finite alphabets stand without that law. The closed forms that use it are equation 5, the equality cases in Theorem 3, and reconstruction error equal to reverse water filling.