CONTROLLING INFORMATION FLOW IN VISION TRANSFORMERS WITH VARIATIONAL BOTTLENECKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We make the information transmitted by attention in vision transformers a measurable and controllable quantity. By inserting variational information bottlenecks on attention-mediated writes to the residual stream, we train models under an explicit information cost, producing a spectrum from independent patch processing to fully expressive global attention. The same mechanism is compatible with both supervised classification and DINO-style self-supervised learning, allowing us to study how useful visual representations emerge as progressively more information is exchanged across patches. On ImageNet-100, we characterize how performance and representation quality vary with information rate, and how communication is allocated across depth, attention heads, and spatial locations. Under tight information constraints, only a small number of attention heads are active, exposing simple visual computations that precede richer global representations. These results provide a direct way to control internal information flow and yield vision transformers whose computations are more tractable for mechanistic analysis.

1 INTRODUCTION

Transformers have emerged as a dominant architectural paradigm across machine learning, distinguished by their ability to aggregate information globally across the constitutive elements of an input (Vaswani et al., 2017). In vision transformers (Dosovitskiy et al., 2021), an image is decomposed into patches, and computation proceeds through repeated alternation between two modules: attention blocks, which enable information sharing across patches, and multilayer perceptron (MLP) blocks, which refine patch representations independently. Because communication and local refinement occur in distinct computational steps, information flow becomes an explicit and manipulable axis of the architecture, rather than being entangled with processing as in prior paradigms. This naturally raises a mechanistic question: when communication is restricted, what circuits of attention-mediated interaction emerge first (Olah et al., 2020; Elhage et al., 2021), and what do they reveal about how transformers build global understanding from local evidence?

Most interpretability methods analyze the internal structure of models after training, revealing mechanisms that can be rich but difficult to disentangle or control (Rudin, 2019; Samek et al., 2021). Interpretable-by-design approaches instead impose particular forms of explanatory structure, such as prototypes or concept bottlenecks (Chen et al., 2019; Koh et al., 2020). We pursue a different form of constraint: rather than prescribing the content of internal representations, we restrict the amount of information that can be communicated between them. The information bottleneck framework provides a natural basis for this approach (Alemi et al., 2017), but applying it to internal information flow requires a representation in which transmitted information is operationally measurable. Deterministic, point-valued activations make such quantities difficult to define and control directly (Saxe et al., 2019). We therefore introduce stochastic representations on attention-mediated writes, making their information rate both measurable and optimizable.

Information-theoretic bottlenecks can be applied either post hoc or during training. Post-hoc methods (Schulz et al., 2020; Jiang et al., 2020; Hong et al., 2025) can probe pretrained models, but do so under perturbations that differ from the conditions under which those models were optimized. Training-time bottlenecks instead make information restriction part of the learning problem itself (Alemi et al., 2017; Achille & Soatto, 2018), biasing optimization toward solutions that operate under limited communication. We take this approach, using vision transformers as a setting in which communication between otherwise local computational streams can be isolated and controlled.

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

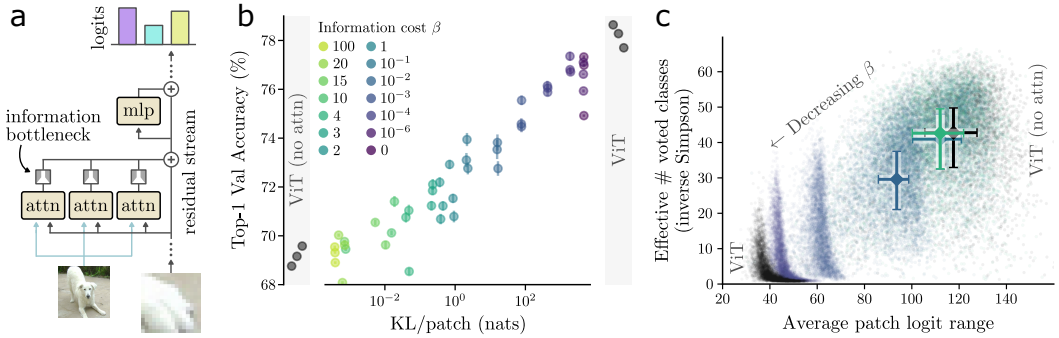

Figure 1: **Limiting information written by attention to the residual stream.** **(a)** We insert a variational information bottleneck after every attention head, just before its message is written to the patch residual stream. The resulting KL penalty upper bounds the information written by attention; scaling the summed penalty by β yields models ranging from independent patch processing to standard ViT attention. **(b)** For a ViT-tiny trained on ImageNet-100, this produces a continuous spectrum trading accuracy for information, where information is reported as total KL per patch. Models smoothly interpolate between a no-attention ViT and an unmodified ViT; stochastic IB error bars show standard deviation across 10 validation runs. **(c)** With global average pooling as the final readout, each patch contributes uniformly to the final prediction, allowing patch logits to be interpreted as votes. Patch voting behavior changes continuously along the information spectrum: with less attention-written information, patches show larger logit ranges and more diverse top-class votes, measured by the inverse Simpson effective count of voted classes. Each dot is one image, colored by β as in **b**; medians and interquartile ranges summarize overlapping point clouds.

Specifically, we place variational information bottlenecks (Alemi et al., 2017) on every attention-mediated write to the residual stream, without otherwise modifying the attention mechanism or transformer architecture. Varying the strength of this bottleneck produces a continuous family of models spanning standard global attention and a zero-communication limit in which patches are processed independently before permutation-invariant aggregation, closely related to DeepSets (Zaheer et al., 2017). This gives the communication axis a concrete endpoint and allows us to ask how transformer computation changes as global information exchange is progressively introduced.

Across this spectrum, we study both supervised classification and DINO self-supervised learning (Caron et al., 2021). We characterize how predictive performance and representation quality vary with information rate, how communication is distributed across depth and spatial locations, and which attention heads are first to carry nontrivial information. Under tight communication constraints, these heads expose simple visual computations that precede richer global representations, providing a tractable view of how attention begins to integrate information across patches.

Contributions. Concretely, we: (i) introduce a family of vision transformers in which attention-mediated information flow is explicitly measurable and continuously controllable; (ii) characterize the tradeoff between routed information and learned representations from independent patch processing to unrestricted global attention, under both supervised training and DINO self-supervision; and (iii) analyze how information is allocated across depth, patches, and attention heads, including the heads that remain active under the strongest constraints. Together, these results establish information restriction as a practical way to expose and control the emergence of global computation in vision transformers.

2 Related Work

Mechanistic interpretability and transformer circuits. This work aligns with mechanistic interpretability: understanding and ultimately controlling the internal computations of trained models. Prior work has traced attention-mediated information flow through transformer circuits (Olah et al., 2020; Elhage et al., 2021), decomposed representations into interpretable features using sparse dictionary learning and related tools (Bricken et al., 2023; Huben et al., 2024), and developed mechanistic

Table 1: Comparison with related approaches for constraining or simplifying internal computation. Our method imposes a continuous information-rate constraint during training directly on attention-mediated communication.

Method	When constrained	Object	Constraint
ViT-IB (ours)	scratch	attention writes	information rate
Codebook features	scratch	features	quantized information
Concept bottlenecks	scratch	features	supervised concepts
Sparse attention	scratch / fine-tuned	attention edges	sparsity / count
Gated / pruned heads	fine-tuned / frozen	attention heads	count
Circuit discovery	frozen	edges	sparsity / count
IBA	frozen	activations / tokens	information rate

accounts in simplified settings such as small transformers or narrowly defined tasks (Akyürek et al., 2023; Nanda et al., 2023; Wang et al., 2023; Dutta et al., 2024). Extending these ideas to vision transformers remains challenging because image patches are continuous, attention is dense rather than causally masked, and semantic units are less discrete than language tokens. Recent work has therefore studied ViT representations across depth, training objectives, and architectures (Raghu et al., 2021; Caron et al., 2021; Park et al., 2023; Dorszewski et al., 2026), including feature-decomposition approaches applied to ViT activations (Stevens et al., 2025). Our work contributes a complementary simplification: rather than restricting model size, task complexity, or analyzing a fixed pretrained model, we restrict internal communication capacity during training, yielding a tunable spectrum of transformer computations.

Information bottleneck methods for interpretability. Our method builds on the information bottleneck principle (Tishby et al., 2000; Alemi et al., 2017), which formalizes a tradeoff between predictive sufficiency and compression. Beyond its use as a regularizer or robustness tool (Hu et al., 2024), the information bottleneck has also been used to identify which information is important to a model's computation. Post-hoc bottleneck methods, such as Information Bottleneck Attribution and its extensions (Schulz et al., 2020; Jiang et al., 2020; Hong et al., 2025), probe pretrained models by restricting information after training. Training-time bottleneck methods instead shape the solutions learned by optimization; for example, distributed information bottleneck approaches quantify feature importance by restricting communication about input dimensions (Murphy & Bassett, 2023; 2024). We extend this training-time perspective from input features to transformer internals by placing variational bottlenecks on attention-mediated residual writes.

Communication restriction and structured simplification in transformers. Several lines of work simplify transformer computation by pruning or restricting attention heads, sparsifying attention, or removing tokens (Michel et al., 2019; Voita et al., 2019; Pan et al., 2021; Rao et al., 2021; Liang et al., 2022; Bolya et al., 2023), primarily for efficiency. Gao et al. (2026) train transformers under weight sparsity for a capability-interpretability frontier and a complementary route to interpretable-by-design models through structural sparsity rather than constrained information flow. Table 1 situates our approach relative to codebook-based bottlenecks (Tamkin et al., 2023), concept bottlenecks (Koh et al., 2020), sparse attention (Correia et al., 2019; Peters et al., 2019), head gating and pruning (Voita et al., 2019; Michel et al., 2019), circuit discovery (Wang et al., 2023; Conmy et al., 2023), and post-hoc information bottlenecks (Schulz et al., 2020; Jiang et al., 2020; Hong et al., 2025).

3 Method

3.1 Vision transformers as residual stream communication

We adopt the residual stream framing of transformer computation developed in Elhage et al. (2021). A vision transformer (ViT) processes an image as a sequence of patch embeddings, $\mathcal{S} = \{x_i\}_{i=1}^{N}$, where each patch maintains its own residual stream representation throughout the network. Each transformer block alternates between two update mechanisms: (i) an attention block, which computes context-dependent interactions across patches, and (ii) an MLP block, which refines each patch representation independently.

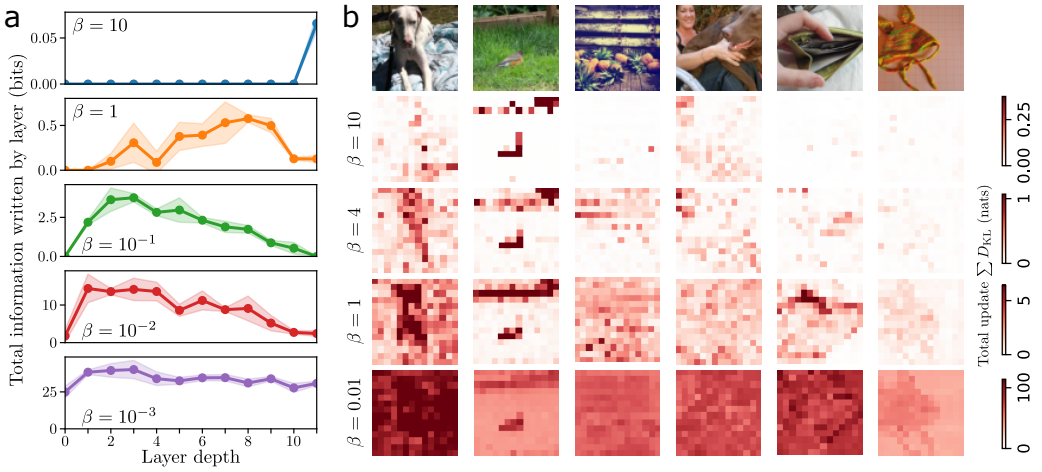

Figure 2: **Communication allocation across layers and across patches.** **(a)** Total transmitted information $I(\Delta; Z)$, estimated directly through density ratios, summed over heads in each layer and averaged across three seeds, shifts from late to early layers as β decreases. **(b)** For a random sample of validation images correctly classified by the unmodified ViT, we visualize total KL per patch, summed over all attention heads, for selected models near the Pareto front in Fig. 1b. Panel **a** quantifies layerwise information transmission, while panel **b** shows the spatial distribution of the corresponding KL-based communication cost within individual images. Colormaps are normalized separately for each model (row).

Crucially, only attention enables information exchange between sequence elements (Fig. 1a). MLP blocks operate pointwise on each residual stream and introduce no new cross-patch information. Thus, attention-mediated writes to the residual stream constitute the natural pathway through which global visual representations emerge from local patch features. Our approach directly monitors and restricts this communication channel.

3.2 Bottlenecking attention-mediated residual writes

Within an attention block, there are several possible points at which information flow could be restricted. We intervene at the final attention update, immediately before it is written back into the residual stream. This leaves the attention computation itself intact while limiting the information transmitted through the resulting update. To resolve communication at the level of individual heads, we instantiate a separate variational bottleneck on every attention head in every layer, yielding 36 independently regularized channels in ViT-Tiny.

Let $\Delta_i^{\ell,h} \in \mathbb{R}^{d_h}$ denote the attention update for patch i from head h in layer ℓ, before the output projection. In a standard transformer block, these head updates are concatenated, projected, and added deterministically to the residual stream. We instead route each head update through a head-specific variational bottleneck:

$$\Delta_i^{\ell,h} \;\to\; z_i^{\ell,h} \sim q_{\phi_{\ell,h}}(z \mid \Delta_i^{\ell,h}) \;\to\; \hat{\Delta}_i^{\ell,h} = g_{\theta_{\ell,h}}(z_i^{\ell,h}), \tag{1}$$

where each layer-head pair has its own encoder $q_{\phi_{\ell,h}}$ and decoder $g_{\theta_{\ell,h}}$. The decoded head updates are then concatenated, projected, and written to the residual stream as usual. Because attention-mediated residual writes are the mechanism by which patches exchange information, this intervention isolates the attention-mediated pathway through which inter-patch communication occurs.

3.3 Training objective and information restriction

The communication bottlenecks are implemented using the variational information bottleneck (VIB) principle (Tishby et al., 2000; Alemi et al., 2017). Each layer-head bottleneck uses a diagonal Gaussian encoder,

$$q_{\phi_{\ell,h}}(z \mid \Delta) = \mathcal{N}\big(\mu_{\phi_{\ell,h}}(\Delta), \mathrm{diag}(\sigma^2_{\phi_{\ell,h}}(\Delta))\big), \tag{2}$$

and is regularized toward an isotropic unit Gaussian prior on its latent space, denoted $p(z) = \mathcal{N}(0, I)$. The resulting KL divergence upper bounds the mutual information between Δ and z, yielding a tractable proxy for transmitted information. For analyses requiring transmitted information rather than its variational upper bound, we estimate $I(\Delta; Z)$ directly via Monte Carlo likelihood-ratio estimation.

For a general task loss $\mathcal{L}_{\text{task}}$, we optimize

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \beta \sum_{\ell,h,i} D_{\text{KL}}\big(q_{\phi_{\ell,h}}(z_i^{\ell,h} \mid \Delta_i^{\ell,h}) \,\|\, p(z_i^{\ell,h})\big), \tag{3}$$

where $\mathcal{L}_{\text{task}}$ can be any differentiable objective; in this work, we use cross-entropy for supervised classification and the DINO loss for self-supervised training (Caron et al., 2021). The coefficient β sets the cost of transmitting attention-mediated information, thereby tuning the capacity of each head-level communication channel.

This is not a single-bottleneck compression objective in the classical IB sense. Rather, it defines a distributed information budget over the attention-mediated messages that carry information between patches. Each bottleneck penalizes the information retained in a local write, while the task loss selects messages that remain useful to downstream computation. The summed KL penalty is therefore a decomposable communication cost, allowing information flow to be measured by head and by patch.

Comparing probabilistic head updates via normalized mutual information. To compare probabilistic head updates across seeds and β values, we use the probabilistic representation-space normalized mutual information (NMI) estimator of Murphy et al. (2025), treating each head's bottleneck posterior as a soft clustering over patch instances. Details are in Appx. C.

3.4 Implementation details

All experiments use the ViT-Tiny backbone from `timm` (12 blocks, embedding dimension 192, 3 attention heads per block). We adopt this lightweight backbone to enable extensive sweeps over information budgets and detailed mechanistic analysis, while preserving the canonical transformer computation pattern. We insert variational information bottlenecks on the attention-mediated residual writes in every block, instantiating a separate encoder–decoder bottleneck for each head in each layer (36 bottlenecks total). All bottlenecks share the same architecture and latent dimensionality, differing only in their learned parameters. For ViT-tiny, the unmodified architecture contains approximately 5.5M parameters, while the bottlenecked variant contains approximately 7.6M parameters, corresponding to a roughly 40% parameter overhead from the bottleneck modules.

For supervised classification, we use global average pooling over patch representations followed by a linear classifier head (Zhai et al., 2022), ensuring that the zero-communication limit remains well-defined. We train on ImageNet-100, a 100-class subset of ImageNet (Deng et al., 2009) introduced by Tian et al. (2020), using AdamW with learning rate 6×10^{-4} and weight decay 0.05 for 1000 epochs. Training uses a cosine learning-rate schedule with 10 epochs of warmup, and the information penalty coefficient β is fixed over the course of training. We apply RandAugment for standard appearance-level augmentation, but omit patch-mixing strategies such as Mixup and CutMix since these distort patch-level statistics and complicate interpretation.

For self-supervised experiments, we train the same bottlenecked ViT-Tiny models with the DINO objective (Caron et al., 2021) on the same ImageNet-100 subset. To preserve the communication-limited readout, we remove the CLS token and replace DINO's projection head with global average pooling followed by a single linear projection to a 256-dimensional representation before the prototype layer. We use 2^{16} prototypes and train for 800 epochs with batch size 256, learning rate 5×10^{-4}, teacher-temperature warmup over 30 epochs, and weight decay ramped from 0.04 to 0.4. Additional optimization and augmentation details are provided in Appendix B.

4 Results

We present results in two stages. Our primary analysis uses supervised ImageNet-100 models, where the classification objective and GAP readout allow us to study the full information-accuracy spectrum, patch voting behavior, and the individual attention heads that first transmit nontrivial messages. We then ask whether the same communication-restriction framework extends beyond supervised

classification by applying it to DINO self-supervised training. Together, these experiments test both the mechanistic value of the low-information regime and the broader applicability of attention-written information as a tunable constraint.

Within the supervised models, we analyze the system at two complementary scales. First, we characterize the continuum induced by the information cost β, demonstrating systematic changes in attention-mediated communication, predictive performance, and patch voting behavior. Second, we exploit the sparsity of the low-information regime to perform fine-grained mechanistic analyses of individual attention heads.

4.1 Properties of the spectrum of models

The entire spectrum from independent patch processing (attention blocks are removed from the ViT) to freely flowing information as in an unmodified ViT, spans $69.2\% \pm 0.3$ to $78.2\% \pm 0.3$ top-1 validation accuracy (mean and standard deviation over three random seeds) (Fig. 1b). That the attention-less baseline remains competitive underscores that substantial visual recognition can be achieved from simple pooling of local patch evidence, consistent with the strong performance of local-feature-based image representations prior to deep networks (Lowe, 2004; Sánchez et al., 2013).

Intuitively, as information communicated by attention increases, so too does classification performance. With only around one nat of KL for every ten patches, accuracy increases by $1 - 2\%$. When around one nat of KL is communicated per patch, it further increases by $\sim 2\%$. Empirically, accuracy increases approximately linearly with $\log$ KL over much of the spectrum. The least restricted models ($\beta = 0$) approach, though do not fully reach, the accuracy of an unmodified ViT (within $\sim$1% top-1), suggesting a small residual effect of the IB parameterization even in the weakly constrained regime.

Without attention to pass information between patches, there can clearly be no consensus between patch representations. How does patch independence give way to consensus as attention becomes less restricted? We show in Fig. 1c characteristics of patch voting behaviors for several models close to the Pareto front. These patch-level measures provide a complementary view of global integration: rather than only tracking top-1 accuracy, they quantify how communication induces consensus among local predictions. The spectrum therefore captures not just performance gains, but a qualitative transition from diverse independent patch hypotheses to coordinated global representations.

We summarize patch consensus using two quantities: the inverse Simpson effective count of top-class votes across patches (Simpson, 1949; Hill, 1973), $1/\sum_c p_c^2$, and the range of patch logit contributions. Independent patch processing yields many distinct patch-level class votes (effective counts often exceeding 50 different classes per image) and large logit ranges (total logit range $\gtrsim 100$), while increasing communication progressively drives consensus. Importantly, we are able to smoothly and controllably interpolate between these extremes by tuning the information cost coefficient β.

We visualize aspects of the processing of a handful of images by different models in Fig. 2 and in Appx. D. For the IB models, we display a heatmap of the information penalty (in terms of KL divergence) summed across all attention heads for each patch in the image. Because each bottleneck yields a per-patch KL cost, we obtain a direct, quantitative map of where attention allocates its limited communication budget. For large β, update KL tends to be insignificant for most of the image and then concentrated on patches in the foreground. As β decreases and communication is more plentiful, significant KL is spent on all patches. In Appx. D, we show samples of images randomly selected based on other criteria (e.g., large difference between classification assignments by the two ViT baselines) and additionally visualize examples of patch voting behavior.

4.2 Update behavior of individual attention heads

The low-information regime provides a natural entry point for mechanistic analysis because the KL penalty causes most headwise channels to collapse to the prior. At $\beta = 10$, only a single head transmits nontrivial information; at $\beta = 4$, four heads do so. Thus, instead of analyzing dense attention across all layers and heads, we can inspect the first communication channels that optimization finds worth paying for.

Active heads in this regime often use only one or two latent dimensions, making their messages directly visualizable. When a head is effectively one-dimensional, its update reduces to a signed scalar message, allowing patches to be grouped by whether the head transmits a positive or negative signal. In Fig. 3, we show high-update patches and the corresponding attention maps for the single

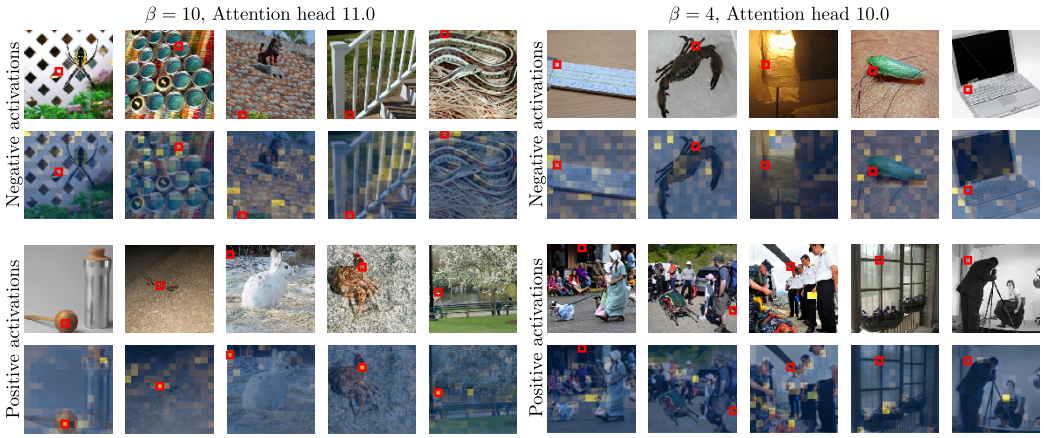

Figure 3: **Head update behavior.** For the single active attention head in the $\beta = 10$ model (head 11.0), and for one of the four active heads of the $\beta = 4$ model (head 10.0), we randomly sample positive and negative activations from patches in the top 0.1% of KL cost. The patch is indicated by a red square in both the original image (top of each pair of rows) and the head's attention map (bottom of each pair of rows).

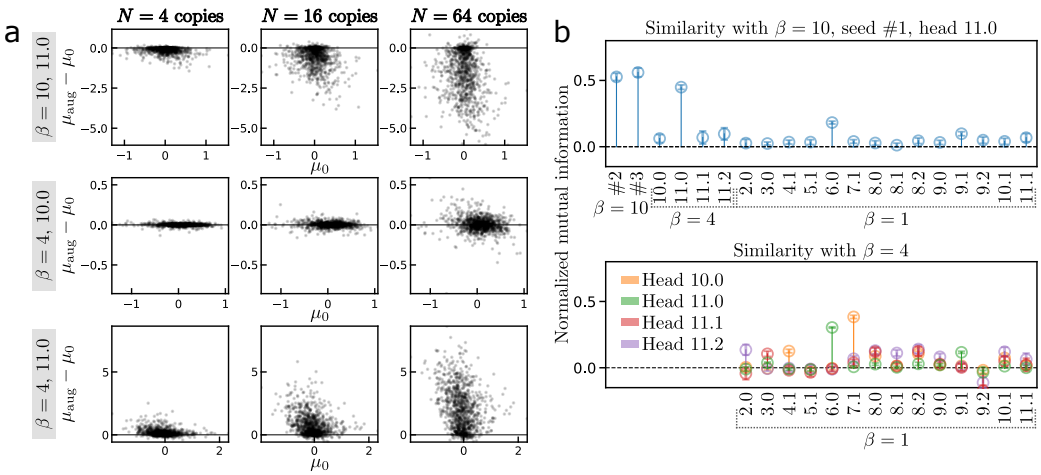

Figure 4: **Controlled perturbations and cross-model similarity reveal repeated-pattern messaging. (a)** We test whether selected heads encode within-image redundancy by copy-pasting a queried patch to N random locations before passing the augmented image through the model. For 1024 randomly selected patches, we compare the bottleneck posterior mean for the original image, μ_0, with the signed displacement induced by augmentation, $\mu_{\text{aug}} - \mu_0$. Repetition-sensitive heads show systematic displacement as the number of copies increases, while other active heads do not. **(b)** Pairwise similarity between active attention-head messaging channels, measured by NMI between probabilistic representation spaces. For $\beta = 10$ head 11.0 (*top*), we compare against the active heads from two other $\beta = 10$ seeds, the four active heads at $\beta = 4$, and the 14 active heads at $\beta = 1$. We also compare each active $\beta = 4$ head against all active $\beta = 1$ heads (*bottom*). Error bars show propagated uncertainty from Monte Carlo estimates of mutual information using 2×10^7 samples.

active head in the $\beta = 10$ model, as well as a representative active head from the $\beta = 4$ model. The earliest active head appears to encode a coarse contextual statistic: whether the current patch is part of a repeated or redundant pattern elsewhere in the image. Across three random seeds, each $\beta = 10$ model contains an active head with qualitatively similar repetition-sensitive behavior (Appendix E).

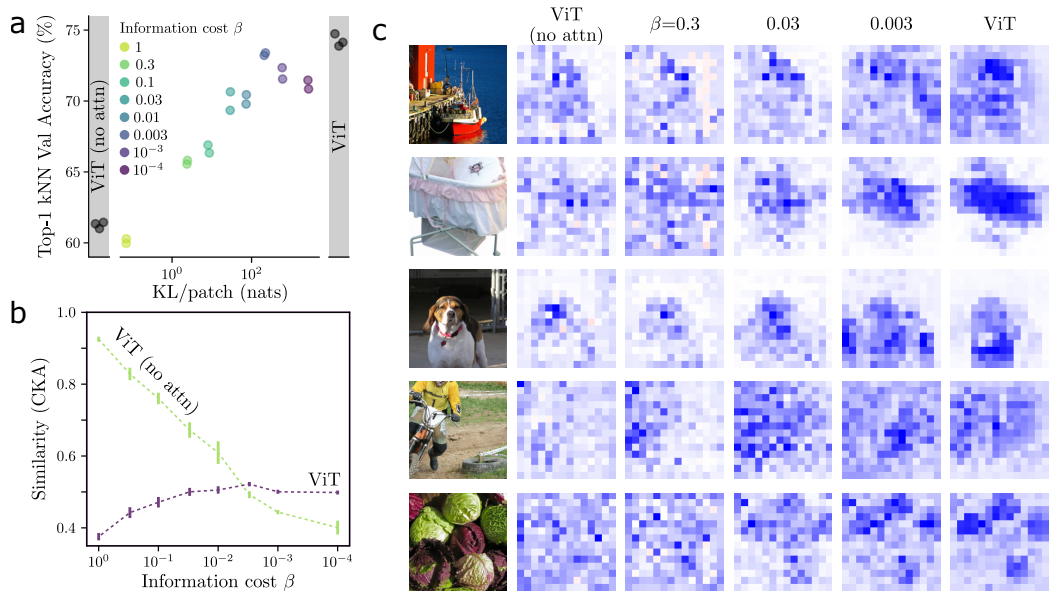

Figure 5: **Restricted communication in SSL. (a)** DINO kNN accuracy interpolates between the attention-free and unmodified ViT endpoints, peaking at $\beta = 0.003$. **(b)** CKA similarity (Kornblith et al., 2019) compares image representations from bottlenecked models to the unmodified ViT (purple) and the independent-patch-processing ViT (green). Error bars show the standard deviation across cross-seed comparisons. **(c)** Patch-image alignment maps for random validation images, computed as the dot product between each patch representation and the final global image representation.

We test this hypothesis with a controlled copy-paste augmentation (Fig. 4a). For each of 1024 randomly sampled images, we select a patch and paste copies of it at random locations elsewhere in the same image, increasing within-image redundancy while preserving the local content of the queried patch. For the repetition-sensitive heads, the latent message shifts systematically as the number of copies increases from 4 to 16 to 64. This effect is pronounced for $\beta = 10$ head 11.0 and $\beta = 4$ head 11.0, but absent for $\beta = 4$ head 10.0, supporting the interpretation that these heads transmit information about repeated structure rather than merely responding to local patch content.

To assess whether this message is reproducible across seeds and persists as the communication budget increases, we compare head update distributions using normalized mutual information for probabilistic representation spaces (Murphy et al., 2025). We treat each head's bottleneck posterior as a soft clustering of patch instances conditioned on image context, allowing comparison across heads even when their latent dimensionalities differ. As shown in Fig. 4b, the active heads of the three $\beta = 10$ models have high mutual agreement, indicating that the same repetition-sensitive message emerges reliably across seeds. Related information also appears in less constrained models, most clearly in $\beta = 4$ head 11.0 and $\beta = 1$ head 6.0. Moreover, the NMI comparisons mirror the copy-paste perturbation results: heads that respond similarly to repeated patches have high NMI, while visually distinct heads, such as $\beta = 4$ head 10.0, have low NMI.

These results suggest that the first attention-mediated messages to emerge under strong information pressure encode simple, high-value global context. Coarse redundancy signals are cheap to communicate, recur across seeds, and remain partially identifiable as the communication budget increases, providing a concrete example of how restricted transformers can expose attention circuits.

4.3 Communication-restricted self-supervised learning

Finally, we test whether attention-written information can be restricted in a training objective that does not directly optimize class labels. We train the same bottlenecked ViT-Tiny architecture with the DINO self-supervised objective on ImageNet-100, using GAP followed by a single linear projection so that the final representation remains communication-limited rather than being aggregated by a

large projection head. Communication restriction again induces a structured performance spectrum, but with dynamics that differ from supervised classification. As measured by kNN accuracy, the gap between the attention-less and unmodified ViTs is larger under DINO than in the supervised setting (Fig. 5a), indicating that self-supervised representation learning places greater demand on global attention-mediated communication. The IB models interpolate across this range, with peak performance at $\beta = 0.003$ before degrading at still lower values of β. This non-monotonicity suggests that, in the weakly constrained regime, the IB parameterization can affect optimization or representation geometry even when the explicit information penalty is small.

To compare the resulting representation spaces, we compute linear CKA (Kornblith et al., 2019) between the communication-restricted models and the two endpoint ViTs (Fig. 5b). High-β models closely align with the attention-less baseline. As communication increases, representations initially move toward the unmodified ViT, with similarity peaking near the same β that maximizes kNN accuracy before declining in the weakest-constraint regime. Thus, the best-performing DINO models are not simply those closest to the unmodified ViT endpoint; instead, the information budget traces a distinct path through representation space.

We also visualize how semantic localization emerges along the DINO spectrum. For each image, we compute the dot product between each patch representation and the global image representation obtained by averaging over patches. In the attention-less regime, these patch–image alignment maps remain largely local and diffuse. As communication becomes less restricted, object-aligned structure emerges, with clearer semantic localization appearing below approximately $\beta = 0.03$ (Fig. 5c). Thus, the same intervention that controls supervised patch voting also modulates the emergence of globally organized self-supervised representations.

Together, the DINO experiments show that communication bottlenecks are not tied to supervised classification losses. Across both supervised and self-supervised training, restricting attention-mediated residual writes yields an interpretable axis along which performance, representation geometry, and spatial organization change systematically. The DINO results also separate two roles of the bottlenecked spectrum. In the supervised setting, GAP and patch logits make the transition from local voting to global consensus directly measurable. In the self-supervised setting, the same communication constraint instead organizes representation geometry and patch-image alignment, suggesting that the information budget controls a broader property of transformer computation rather than a peculiarity of supervised classification.

5 Discussion

Our results suggest that strong communication pressure exposes simple, high-value forms of global integration. In the lowest-information regimes, most attention heads collapse to the prior, leaving only a handful of active channels that can be inspected directly. The first active heads are sparse, low-dimensional, and reproducibly sensitive to patch repetition or redundancy across random seeds (Figs. 3,4). This indicates that coarse global texture statistics may be among the most information-efficient messages available to the model: cheap to transmit, but useful for coordinating otherwise independent patch representations.

Several limitations remain. We study ViT-Tiny models on ImageNet-100, and the most interpretable circuits occur far from the dense communication patterns of models approaching the performance of an unrestricted ViT. Because the bottleneck modules add parameters and stochasticity, weakly constrained models need not exactly recover the unmodified ViT endpoint, even when the explicit KL penalty is small. Global average pooling provides a clean zero-communication endpoint but also constrains the class of readouts we study. In addition, each information budget requires training a separate model, making dense sweeps computationally expensive. Extending the framework to larger backbones, richer datasets, alternative aggregation structures, and more efficient ways of traversing the information spectrum will be important for demonstrating generality of observed circuits.

More broadly, this work treats attention as a measurable communication channel rather than only as an architectural primitive. By restricting attention-mediated residual writes during training, we obtain models whose internal communication is intrinsically constrained, sparse, and tunable. This provides a complementary route to mechanistic interpretability: instead of only analyzing fully expressive models after training, we can bias optimization toward lower-bandwidth solutions and study how global computation emerges as that constraint is relaxed.

AI USE STATEMENT

LLM-based programming assistance was used during implementation, including for code snippets related to architectural modifications for inserting variational information bottlenecks into the ViT attention blocks. The overall method design, integration into the codebase, experimental design, training, analysis, and interpretation were performed by the authors. The LLM was not used as a component of the proposed method, did not autonomously generate experimental results, and did not determine the scientific claims or conclusions. All implementation details, numerical results, figures, references, and manuscript content were reviewed and verified by the authors.

REFERENCES

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 (12):2897–2905, 2018.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2174–2184, 2019.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18613–18624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Teresa Dorszewski, Lenka Tětková, Robert Jenssen, Lars Kai Hansen, and Kristoffer Knutsen Wickstrøm. From colors to classes: Emergence of concepts in vision transformers. In Riccardo Guidotti, Ute Schmid, and Luca Longo (eds.), *Explainable Artificial Intelligence*, pp. 28–47. Springer Nature Switzerland, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=uHLDkQVtyC.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V. Govande, Bowen Baker, and Daniel P Mossing. Weight-sparse transformers have interpretable circuits. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Vw10TpuFCp.

Mark O Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2): 427–432, 1973.

Jung-Ho Hong, Ho-Joong Kim, Kyu-Sung Jeon, and Seong-Whan Lee. Comprehensive information bottleneck for unveiling universal attribution to interpret vision transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25166–25175, 2025. doi: 10.1109/CVPR52734.2025.02343.

Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5325–5344, 2024.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. Inserting Information Bottlenecks for Attribution in Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3850–3857, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.343.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

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.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Kieran A Murphy and Dani S. Bassett. Interpretability with full complexity by constraining feature information. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=R_OL5mLhsv.

Kieran A Murphy and Dani S Bassett. Information decomposition in complex systems via machine learning. *Proceedings of the National Academy of Sciences*, 121(13):e2312988121, 2024.

Kieran A. Murphy, Sam Dillavou, and Dani S. Bassett. Comparing the information content of probabilistic representation spaces. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=adhsMqURI1.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. Ia-red^2: Interpretability-aware redundancy reduction for vision transformers. *Advances in neural information processing systems*, 34:24898–24911, 2021.

Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=azCKuYyS74.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, 2019.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D. Tracey, and David D. Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, dec 2019. doi: 10.1088/1742-5468/ab3985. URL https://doi.org/10.1088%2F1742-5468%2Fab3985.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xWh1rYwB.

Edward H Simpson. Measurement of diversity. *nature*, 163(4148):688–688, 1949.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

Alex Tamkin, Mohammad Taufeeque, and Noah D. Goodman. Codebook features: Sparse and discrete interpretability for neural networks, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 776–794, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58621-8.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

A DECLARATIONS

Broader impacts. This work is foundational research on controlling and measuring information flow in vision transformers. Its potential positive impacts include improved interpretability, more transparent analysis of transformer computation, and better tools for understanding how models use attention-mediated communication. By making internal communication measurable and controllable, the method may help researchers identify simpler or more inspectable model behaviors and develop models whose internal mechanisms are easier to audit.

The potential negative impacts are indirect. Improved understanding of model internals could contribute to stronger models in domains where the models are later used harmfully. In addition, mechanistic analyses can be overinterpreted: sparse information flow or interpretable head behavior does not by itself guarantee that a model is safe, fair, or reliable under deployment. We therefore present the method as an analysis and control tool rather than as a guarantee of trustworthy behavior, and we support our claims with controlled experiments, quantitative measurements, and explicit limitations.

Compute resources. Experiments were conducted on an institutional high-performance computing cluster using NVIDIA A100 GPUs. As reported in Methods, the base ViT-tiny architecture has approximately 5.5M parameters, while the bottlenecked variant has approximately 7.6M parameters. For the supervised ImageNet-100 experiments, each bottlenecked ViT training run required approximately 3 days on a single A100 for 1000 epochs; the corresponding non-bottlenecked ViT baselines required roughly half this time. For the self-supervised DINO experiments, each bottlenecked ViT run required approximately 6 days on a single A100 for 800 epochs, with non-bottlenecked baselines again requiring roughly half this time. The reported experimental results include approximately 50 trained models spanning information-cost settings, random seeds, supervised baselines, and self-supervised runs. Beyond the experiments reported in the paper, the broader project involved approximately 50–100 additional preliminary exploratory runs used to tune hyperparameters, evaluate training schedules such as β-ramping, validate implementation choices, and select the final experimental protocol.

B IMPLEMENTATION DETAILS

We include training code with this submission.

All experiments are based on Vision Transformer (ViT) models implemented using the `timm` library. Unless otherwise stated, we use standard ViT architectures (e.g., patch size 16, image resolution 224) with modifications described below.

To study information-restricted attention, we wrap the attention module to bottleneck the per-head residual updates. For every head, the message passes through a variational information bottleneck (IB) module before being written to the residual stream. The IB is applied independently per attention head and per token, and returns both the transformed messages and a KL-divergence term used for regularization during training. For classification, we use global average pooling over the token representations and then apply the final projection to class logits. The remainder of the ViT architecture (patch embedding, positional embeddings, MLP blocks, normalization layers) follows the standard ViT design. All IB encoders and decoders had a single hidden layer with 256 dimensions, and the IB latent space was 32 dimensional.

DINO details. For DINO training, we follow the standard multi-crop augmentation protocol with global crop scale $(0.4, 1.0)$, local crop scale $(0.05, 0.4)$, and local crop size 96. We use warmup teacher temperature 0.04, final teacher temperature 0.07, student temperature 0.1, center momentum 0.9, teacher momentum 0.996, freeze the last layer for 1 epoch, use a 256-dimensional bottleneck projection space, and normalize the last layer.

Optimization. Models are trained using the AdamW optimizer with a base learning rate of 6×10^{-4} and a weight decay of 0.05, for 1000 epochs. Bias parameters, parameters with dimensionality one (e.g., LayerNorm scale parameters), and all parameters in the IB encoder and decoder are excluded from weight decay. All other parameters, including patch embeddings and positional embeddings, are subject to weight decay. Automatic mixed precision (AMP) is used with gradient scaling.

Training uses a cosine learning-rate schedule with linear warmup over the first 10 epochs. The coefficient β is fixed over the course of training.

Data and Augmentation. Experiments are conducted on the 100-class subset of Imagenet used in Tian et al. (2020) and specified in the accompanying github repository. Training images are augmented using random resized cropping, horizontal flipping, and RandAugment (Cubuk et al., 2020). We use a random resized crop with scale range [0.08, 1.0] and RandAugment with two augmentation operations of magnitude 9. Validation images are resized and center-cropped without stochastic augmentation. All images are normalized using standard ImageNet mean and standard deviation values.

C SIMILARITY ASSESSMENTS

Normalized mutual information. To compare the information content of two attention-head bottlenecks, we treat each head as inducing a soft clustering assignment over patches (Murphy et al., 2025). Let X denote a discrete index over all patch instances in the validation set, and let $Z^{\ell,h}$ denote the bottleneck latent variable for head h in layer ℓ. For a fixed head, the variational posterior defines a channel $q_{\ell,h}(z \mid x) \equiv q_{\phi_{\ell,h}}(z \mid \Delta_x^{\ell,h})$.

To quantify correspondence between two heads, let Z_1 and Z_2 denote their latent message variables. We form the joint channel by concatenating the two latent messages, equivalently using the product posterior $q(z_1, z_2 \mid x) = q_1(z_1 \mid x)q_2(z_2 \mid x)$. Using the identity

$$I(Z_1; Z_2) = I(X; Z_1) + I(X; Z_2) - I(X; Z_1, Z_2), \tag{4}$$

we estimate the information shared by the two headwise messages. Finally, we report the normalized mutual information (Strehl & Ghosh, 2002),

$$\mathrm{NMI}(Z_1, Z_2) = \frac{I(Z_1; Z_2)}{\sqrt{I(Z_1; Z_1')\, I(Z_2; Z_2')}}, \tag{5}$$

where Z_1' and Z_2' denote independent samples from the same corresponding headwise channels. Uncertainty is propagated from the standard errors of the Monte Carlo estimates of the constituent mutual informations.

JSD similarity of classification outputs. We compared the similarity of classification probability vectors for models across the spectrum, using the Jensen-Shannon divergence. The divergence values, grouped into five bins, are shown in Fig. 6.

D ADDITIONAL COMPARISONS BETWEEN MODEL UPDATES AND PATCH VOTING

We show in Fig. 7 the survival function $\Pr(\mathrm{KL} \geq x)$ of per-patch information for all patches in the validation set and for all heads. We deem a head active if its KL exceeds 10^{-2} nats for at least one patch in the validation set. Only one attention head is 'active" at $\beta = 10$, and four heads are active for the $\beta = 4$ model. The heads that are active spend KL sparingly: 99% of patches receive an update costing less than 0.5 nats in the $\beta = 10$ case. For the low-information models ($\beta = 10, 4, 1$), the behavior is the same across heads, where the majority of patches get the same null update and the representation nearly mirrors the prior of the channel. As the information restriction loosens, $\beta = 0.01$ utilizes all 36 attention heads for non-trivial updates, and the distribution of KLs spent per patch shifts significantly, to where the majority deviate from the prior.

In Figs. 8, 9, and 10, we show how different models along the spectrum "vote" for different images, and for the information restricted models, how the KL budget is spent.

E ADDITIONAL ATTENTION HEAD HIGHLY ACTIVATING PATCHES

In Figs. 11&12, we show top activating patches of both polarities for other seeds at $\beta = 10$ and for the four active heads of the $\beta = 4$ model.

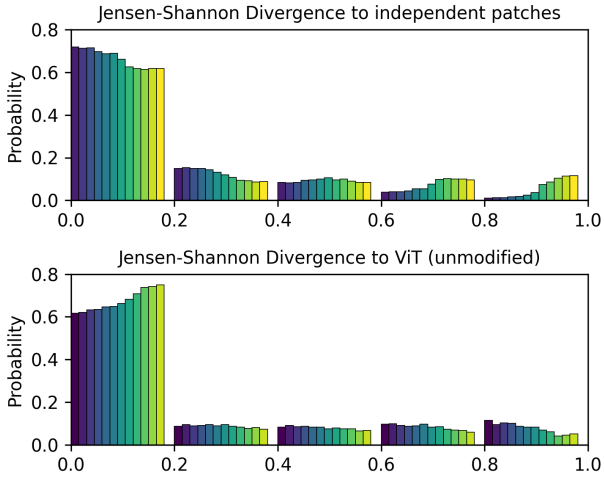

Figure 6: **Jensen-Shannon divergence between classification outputs across the spectrum.** For all images in the validation set, we measure the similarity of model classification outputs across the spectrum to that of the independent patch processing model (ViT w/o attention, dark blue, top row) and to that of the unmodified ViT (yellow, bottom row).

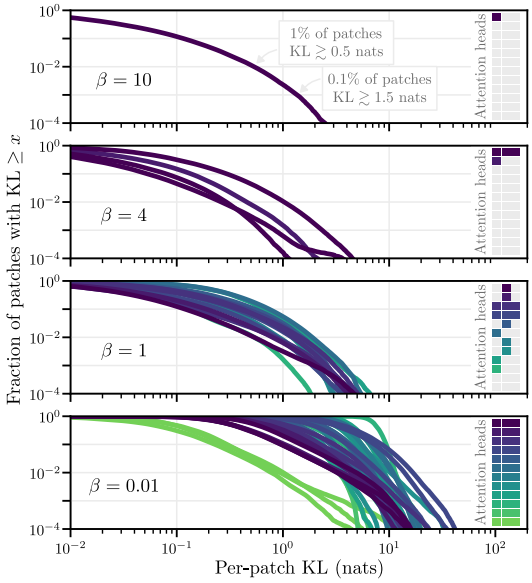

Figure 7: **Attention head firing patterns.** Survival function $P(\mathrm{KL} \geq x)$ of per-patch information across attention heads, showing that only a small fraction of patches carry significant routed information in the high-β regime. The grids on the right show which heads are active, starting with block 0 at the bottom.

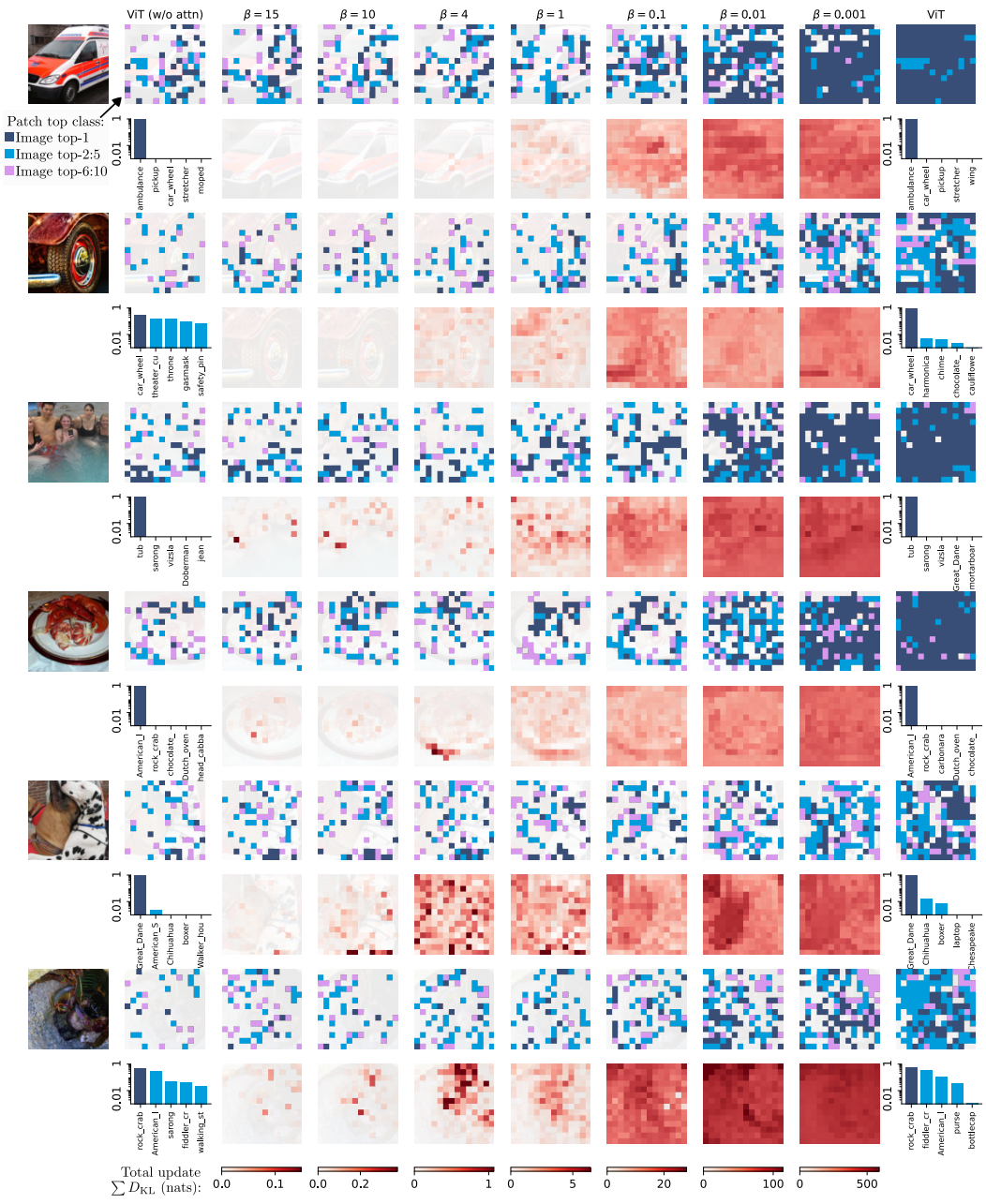

Figure 8: **Patch voting and attention head update behavior for a random selection of images correctly classified by ViT (seed #1).** Patches are colored according to the particular model's class assignments: dark blue for the patches whose top class matches that of the full image prediction top class, light blue for ranks 2 through 5, and pink for ranks 6 through 10.



Figure 9: **Patch voting and attention head update behavior for a random selection of images correctly classified by ViT (seed #1) AND in the top 10% of images in terms of JSD between the ViT without attention and the unmodified ViT.**



Figure 10: **Patch voting and attention head update behavior for a random selection of images incorrectly classified by ViT (seed #1)**

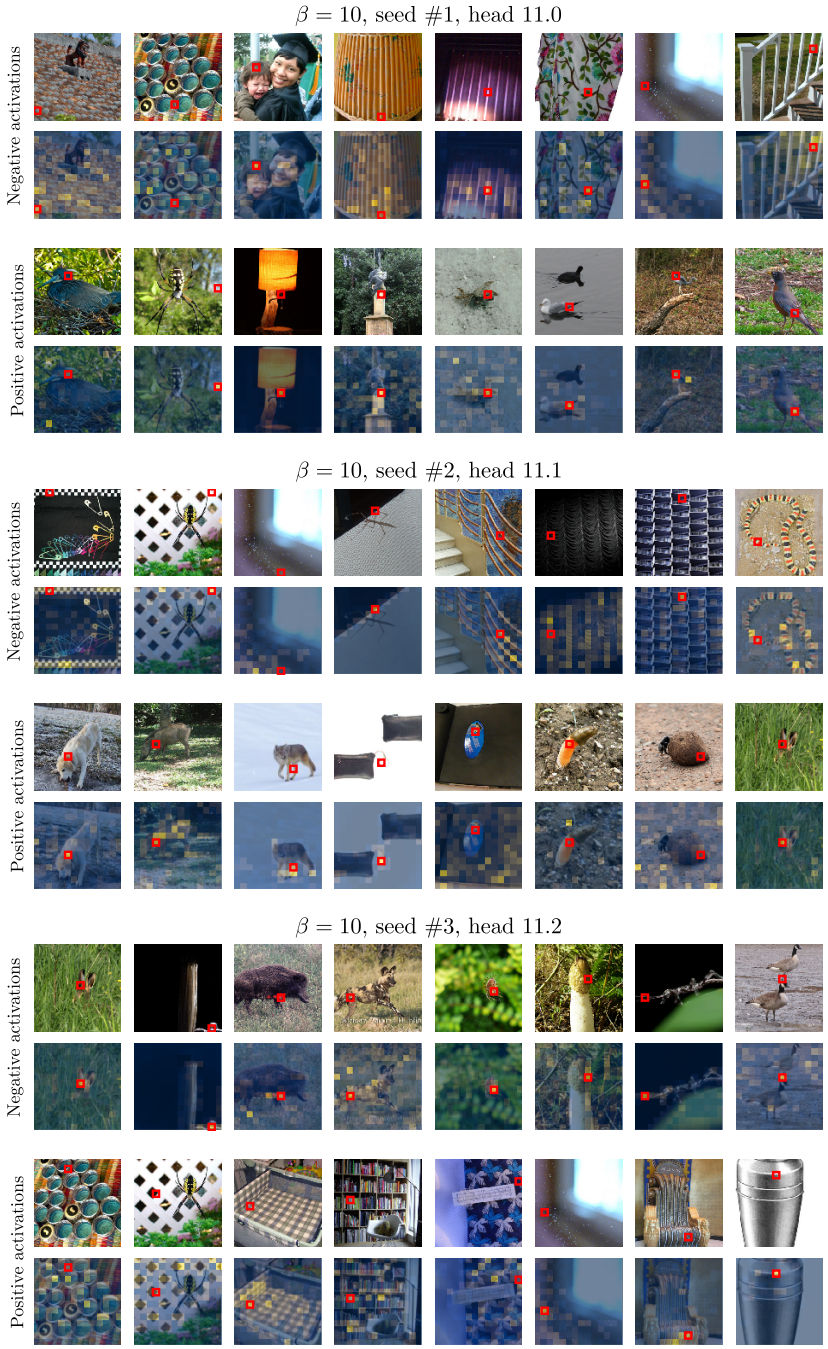

Figure 11: **Additional top activation examples for heads with one active latent dimension: different $\beta = 10$ seeds.**

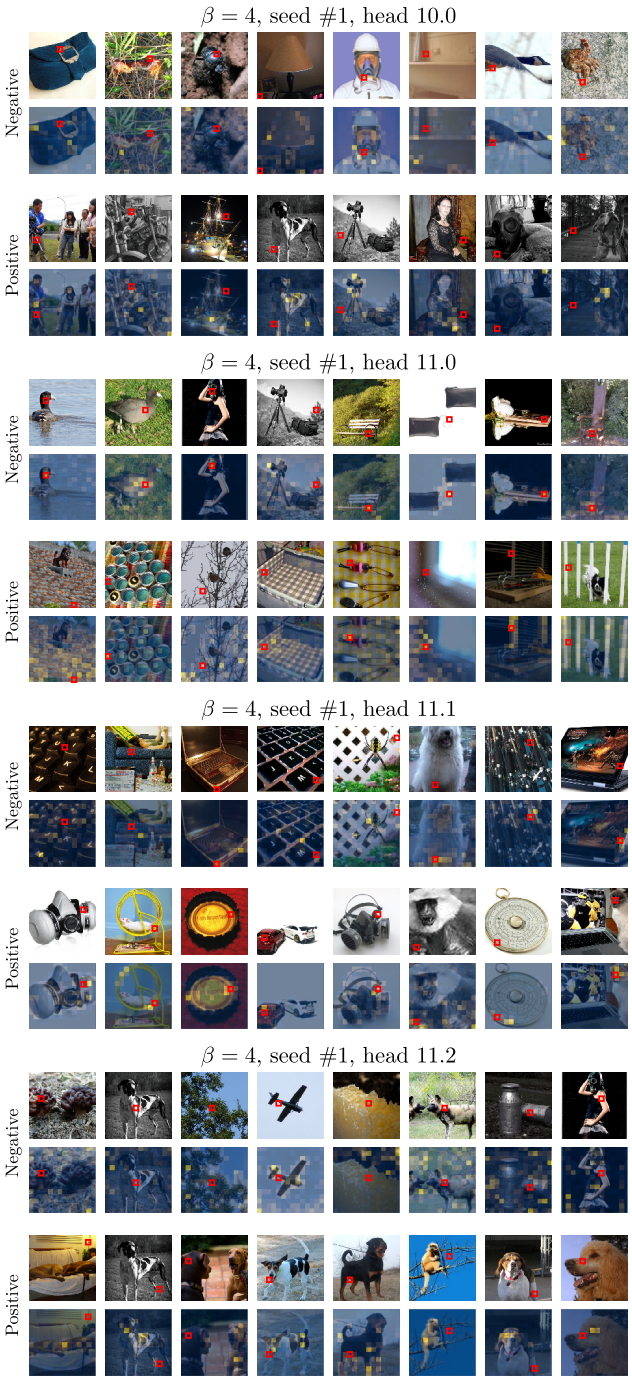

Figure 12: **Additional top activation examples for heads with one active latent dimension: different heads in the $\beta = 4$ model.**