# MASK SCHEDULES ARE A PROGRAMMABLE SUPERVISION GRAPH

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Masked pretraining treats the mask distribution as a hyperparameter. But a mask draw also assigns regions to the context and to the target, and so decides which pairs of regions the objective makes predict one another. Averaged over draws this gives a co-exposure matrix over pairs of regions, a supervision graph that the mask distribution, or schedule, determines completely and that can be rewritten with the data, the network and the objective held fixed. Removing one edge takes predictive influence away from that pair alone (AUC $0.996$) and drives its relational readout to chance, while each region's own content remains readable. It does so even when both regions are predicted more often than before and still appear together in every input. The effect reverses when two schedules are exchanged and is graded in co-exposure. It holds on real images: with the regions of CelebA faces tiled onto noise, raising the co-exposure between one eye and the mouth lifts that eye's readout of mouth opening from $65.4\%$ to $71.1\%$, the level of the mouth region itself ($70.9\%$). The graphs a fixed geometry admits form a polytope with exact linear-programming certificates, and a specified graph can be compiled by constrained maximum-entropy projection, which gets the direction of the effect right but, at the same co-exposure of the edited pair, achieves less of it than placing target blocks by hand. The effect shrinks with the share of the objective the relation carries and vanishes when the image's own coherence already delivers the information, as on whole faces.

## 1 INTRODUCTION

Masked image modeling hides part of an image and trains a network to predict what was hidden, either in pixels (He et al., 2022; Xie et al., 2022), in discrete tokens (Bao et al., 2022), or in the representation of another view of the same image (Assran et al., 2023). What to hide is treated as a hyperparameter and tuned for downstream accuracy: the masked fraction and sampling strategy of He et al. (2022), the multi-block sampler of Assran et al. (2023), and the semantic, attention-guided and adaptive samplers that followed (Li et al., 2022; Kakogeorgiou et al., 2022; Bandara et al., 2023). What that choice does to the model, beyond its effect on accuracy, has not been measured directly.

A single mask draw does something more specific than hide a fraction of the image. It selects a context set $C$ and a target set $T$ of patches, and the objective then makes every patch in $C$ predict every patch in $T$. If we fix a few regions of the image, the draw decides which *pairs of regions* are put in a predictive relationship on that step. Averaged over the mask distribution, or schedule, this defines a co-exposure matrix over region pairs,

$$W(i \to j) \;=\; \mathbb{E}\big[\,|C \cap i| \cdot |T \cap j|\,\big],$$

which we call the *supervision graph* of the schedule: region $i$ supervises region $j$ in proportion to how often $i$ is seen while $j$ is predicted. The schedule determines this matrix completely, and the matrix can be rewritten while the data, the network and the objective all stay fixed.

If this graph is causally real, the mask schedule is a lever on what a self-supervised model learns. Rewriting which pairs of regions supervise one another would steer the model toward a chosen relation, with no new data and no labels. This paper tests whether it does, and under what conditions.

A concrete case runs through the paper. Take a face, cut out the two eyes and the two corners of the mouth, and paste them onto noise, so that nothing but the masking connects them. Raise the co-



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.

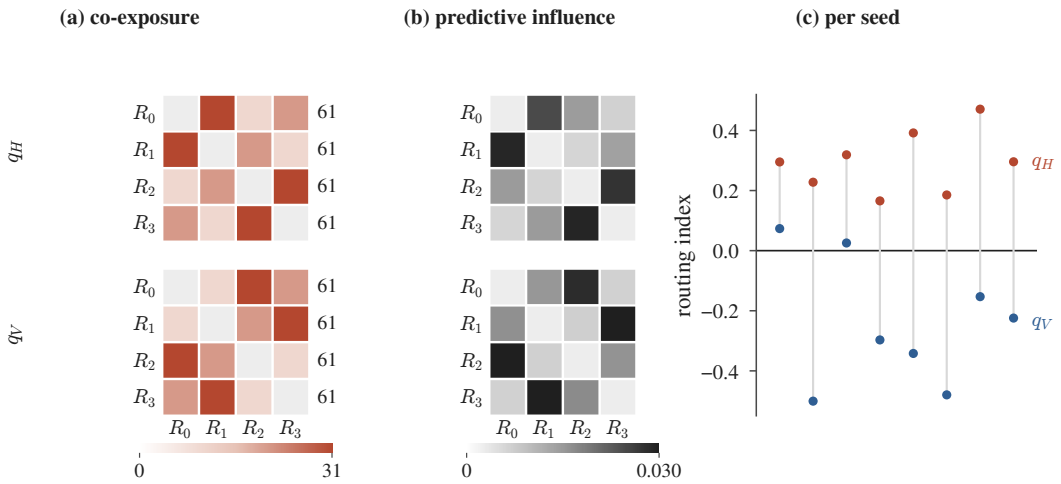


Figure 1: Two schedules with equal marginals and row sums, $q_H$ and $q_V$, co-expose the horizontal or the vertical pairs (a). The learned influence follows them (b), in every seed (c).

exposure of the left eye and the mouth corner below it while keeping how often each region is seen nearly fixed, retrain from scratch, and a linear probe on the left eye's representation reads whether the mouth is open at $71.1\%$ instead of $65.4\%$, about what the mouth region itself affords. The right eye does not move: the schedule has routed information about the mouth into one eye's representation and not the other's. Exchange the two schedules and the eyes exchange roles, although nothing about the images, the architecture or the loss differs between the runs. Figure 1 shows the same exchange on a benchmark in which every pair of regions carries the same information.

We rewrite the graph in two ways: by *editing* it, placing target blocks so that the sampler co-exposes the pairs we want, and by *compiling* it, solving for a mask distribution whose co-exposure matrix matches a specified graph. We find the following.

- Editing the graph changes what the model learns, and only where it was edited. Removing one edge drives the corresponding relational readout to chance, even when the two regions still appear together in every input, while each region's content remains readable. The loss of predictive influence falls only on the removed pair. The change reverses when two schedules are exchanged and grows with the size of the edge.
- The effect holds on rendered glyphs, natural photographs and real face regions, each set on noise, and its scope can be measured. It shrinks with the share of the objective the relation carries, and it vanishes when the image's own coherence has already brought the information to the source node, as it has on whole faces.
- Written in per-region target counts, the graph is a covariance at fixed marginals, and the graphs a fixed geometry admits form a polytope with exact certificates. A specified graph can be compiled by constrained maximum-entropy projection, which gets the direction of the effect right but not its size: at the same co-exposure of the edited pair, placing target blocks by hand achieves more.

## 2 Related Work

Work on masked pretraining has mostly asked *what to hide*. We ask what the manner of hiding builds between parts of an image, and whether that structure can be specified in advance.

**What masking induces.** Contrastive learning has been analyzed as spectral decomposition of an *augmentation graph* whose vertices are data points and whose edges join two views of the same example (HaoChen et al., 2021), and contrastive and non-contrastive objectives both recover spectral embeddings of such a graph (Balestriero & LeCun, 2022). For masked prediction, each random

mask singles out a set of latent variables linking the hidden and visible parts (Kong et al., 2023), masked autoencoding implicitly aligns mask-induced positive pairs (Zhang et al., 2022), and a fixed mask bounds what the prediction loss can detect, with low-visibility masks restoring the sensitivity that near-global context removes (Liu et al., 2022; Cai & Shi, 2026). Our graph lives inside a single image and takes its edges from the masking, so it can be changed without touching the data. We give the region-pair counterpart of these results and, rather than characterizing it, intervene on it.

**Mask design, and where the graph applies.** SemMAE (Li et al., 2022), AttMask (Kakogeorgiou et al., 2022) and AdaMAE (Bandara et al., 2023) choose what to mask, by parts, by attention or adaptively, to make the pretext task harder or more semantic, and are judged by downstream accuracy. We change the masking to change the predictive channel between two designated regions, and judge it by whether every other pair is left alone. The graph exists only where an objective predicts one part of an image from another. It applies to methods with a within-image masked prediction term, such as iBOT (Zhou et al., 2022) and data2vec (Baevski et al., 2022), and not to methods built only on two augmented views of a whole image (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021). Unlike curriculum learning (Bengio et al., 2009) and data pruning (Sorscher et al., 2022), which order or select examples, we change which regions within one example supervise one another.

**Intervention as evidence.** Asking whether a property is *used* rather than merely present is the move behind amnesic probing (Elazar et al., 2021), nullspace projection (Ravfogel et al., 2020), and causal mediation and abstraction analyses of trained networks (Vig et al., 2020; Geiger et al., 2021). Those methods edit a trained representation, which can push it off the distribution it was trained on (Grant et al., 2026). We intervene on the supervision and retrain, so every representation we compare lies within its own training distribution. The graph is in this sense a concrete, editable instance of the inductive bias that self-supervision is known to require (Saunshi et al., 2022; Locatello et al., 2019).

## 3 The Supervision Graph

This section fixes the objects the rest of the paper measures. Model configuration, optimizer settings, seed counts and the permutation test are given in Appendix A.

### 3.1 Task and data

Each $96 \times 96$ image is cut into 8-pixel patches and carries four non-overlapping regions of $3 \times 3$ patches, $R_0$ (top left), $R_1$ (top right), $R_2$ (bottom left) and $R_3$ (bottom right), which we also call *nodes*. Some pairs of nodes stand in a relation, from a *source* node to a *target* node. Figure 2a shows the three datasets, and Appendix E gives their construction.

RelForest is synthetic: each node carries a rendered glyph, and each row is a related pair. With probability $\kappa = 0.8$ the glyph at $R_1$ is the glyph at $R_0$ shifted by $s$ modulo ten, and otherwise it is drawn independently, and likewise for $R_3$ and $R_2$. The four pairs across rows carry no information. We take $s = 5$ throughout, except for the comparison with pixel reconstruction in Section 4.1 and the rank measurements in Section 5.4, which were run at $s = 0$. At $s = 5$ the two glyphs of a related pair differ and a probe that judges the relation from pixels alone sits at chance ($50.1\%$), so any change in the relational readout cannot be attributed to pixel similarity. At $s = 0$ that shortcut reaches $91.6\%$. A variant, RelGraph, makes all four glyphs noisy copies of one shared glyph, so that every pair carries the same information and only the schedule can make the learned influence asymmetric. We use it only for the exchange of Section 4.2.

RelCIFAR carries the same relations with natural photographs in the nodes, $E$ images per class (Krizhevsky, 2009). CelebA (Liu et al., 2015) carries a real relation: the source node is one eye and the target node is the corner of the mouth below it, with the crop chosen so that the eye centers land on the node centers.

Two layouts are used. In the *full-image* layout the nodes sit on a complete image and are connected by the image's own coherence as well as by the masking. In the *tiled* layout each node is pasted onto

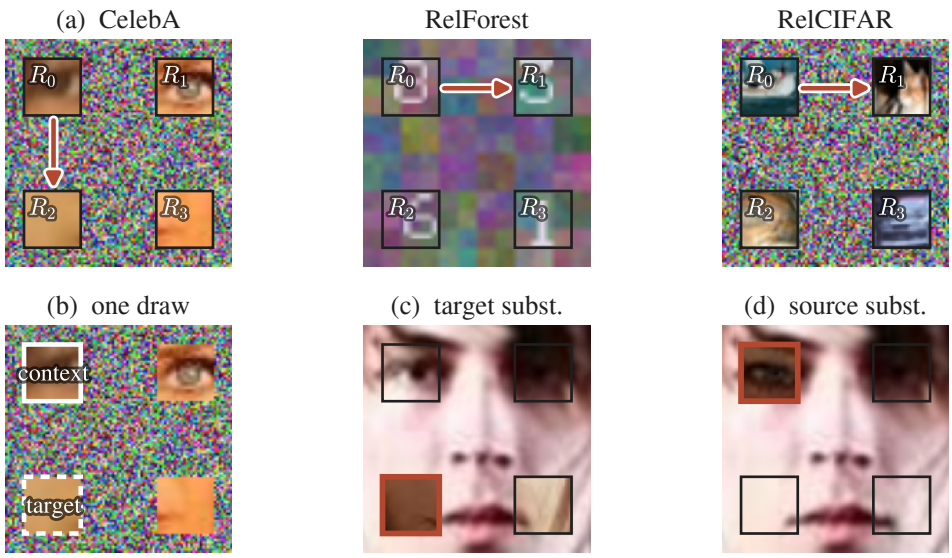


Figure 2: **(a)** The three datasets. **(b)** One mask draw. **(c, d)** Two ways to break an edge.

independent noise, so that the only path between two nodes is the one the masking provides. The distinction is decisive in Section 4.4.

The backbone is a Vision Transformer (ViT) (Dosovitskiy et al., 2021) of width 192 and depth 6, trained with the joint-embedding predictive objective of I-JEPA (Assran et al., 2023), to which we add the variance and covariance terms of VICReg (Bardes et al., 2022). Without them the effective rank of the representation is 4.7 out of 192 and no structure is measurable. With them it is 37.4. Section 4.1 repeats the central measurement with a pixel-reconstruction objective.

### 3.2 Co-exposure and marginals

One mask draw selects a context set $C$ and a target set $T$ of patches, and the network predicts the representation at $T$ from $C$. We define

$$W(i \to j) \;=\; \mathbb{E}\big[\,|C \cap i| \cdot |T \cap j|\,\big], \tag{1}$$

the expectation taken over the mask distribution, with rows indexed by the node acting as context and columns by the node acting as target. The schedule determines $W$ completely. The *marginals* of a node, the fractions of its patches that fall in the context and in the target, averaged over draws, can be held fixed while $W$ changes. Every intervention we report is a matched pair of *conditions*: two schedules that differ in the co-exposure of a designated pair of nodes. Either both hold the marginals fixed to within 0.04, or the condition that removes the co-exposure raises the target marginals, so that a change in readout cannot be explained by how often a region is predicted.

### 3.3 Readouts

After pretraining we freeze the encoder, run it on the whole input, mean-pool its output over each node, and fit linear probes with $k \in \{16, 64, 256, 1024, 4096\}$ labeled examples (from 64 on faces), reporting balanced accuracy. On RelForest and RelCIFAR the probes read the four node vectors concatenated: the *relational readout* judges whether a relation holds, so chance is 50%, and the *content readout* identifies one node's glyph or class. Since the probes are linear, a relational readout at chance means that the relation is not linearly decodable. On faces a probe reads a mouth attribute from a single eye node. The mouth is part of the input, so a gain means that the schedule has placed more mouth information in the eye node's representation. A *generic readout*, whether each face is male, read from the representation pooled over all patches, tracks damage that is not specific to the edited pair. Every result also reports the pair that was not intervened on, which controls for probe capacity (Hewitt & Liang, 2019; Belinkov, 2022).

Table 1: Share of the objective carried by one relation, by source substitution unless marked.

| Dataset | Layout | $\Delta_{\mathrm{rel}}$ (%) | Positive seeds |
|---|---|---|---|
| RELFOREST | tiled | 2.22 | 8/8 |
| RELCIFAR | tiled, $E = 4$ | 5.12 | 15/15 |
| | tiled, $E \geq 16$ | 0.09 | 14/15 |
| CELEBA | tiled | 20.47 | 8/8 |
| | full-image | 2.30 | 8/8 |
| | *by target substitution* | 43.30 | 5/5 |
| | full-image, less context | 2.76 | 8/8 |

### 3.4 PREDICTIVE INFLUENCE

To read off what the trained predictor actually uses, we perturb the input at node $i$ and measure the relative change in the prediction at node $j$ made from every patch outside $j$, $S(i \to j) = \mathbb{E}\,\|\Delta\hat{z}_j\|/\|\hat{z}_j\|$, giving a $4 \times 4$ influence matrix. This query is the same for every schedule and gives the predictor more context than training does. Pairs that a specified graph asks to be connected are *lit*; the rest are *dark*. From $S$ we form four quantities, among them the graph routing score GRS, which ranks the measured influences against the specified graph.

$$\text{routing index} = \frac{\bar{S}_{\text{lit}} - \bar{S}_{\text{dark}}}{\bar{S}_{\text{lit}} + \bar{S}_{\text{dark}}} \in [-1, 1], \qquad \text{retention } r_p = S_p^{\text{int}} / S_p^{\text{base}},$$

$$\text{GRS} = \rho_{\text{Spearman}}\big(S,\ \text{specified graph}\big), \qquad \text{specificity} = r_{\text{removed}} - r_{\text{retained}}.$$

Here $\bar{S}$ averages $S$ over the lit or the dark pairs, and $S_p^{\text{int}}$ and $S_p^{\text{base}}$ are the influence on pair $p$ with and without the intervention. We also report two areas under the receiver operating characteristic curve (AUC): one separating lit from dark influence values, and, for an intervention that removes edges, one separating the retentions of removed pairs from those of kept pairs.

### 3.5 RELATIONAL SHARE OF THE LOSS, AND TWO WAYS TO BREAK AN EDGE

How much of the objective a relation accounts for is measured by

$$\Delta_{\mathrm{rel}} \;=\; \big(L_{\text{broken}} - L_{\text{intact}}\big)/L_{\text{intact}}, \tag{2}$$

the relative increase in prediction loss when the relation is destroyed. Figure 2c,d shows the two ways to destroy it. *Target substitution* replaces the target node with the corresponding region of another image. This also severs the target from its own surroundings. *Source substitution* replaces only the source node. On the full-image layout the two differ by an order of magnitude, as Table 1 shows, because target substitution measures the contribution of the entire surround rather than of the one edge. We therefore use source substitution on that layout. $\Delta_{\mathrm{rel}}$ measures a share of the objective, not whether the mechanism operates. A relation with a small share can still be routed, only less, as Section 4.4 shows.

## 4 EDITING THE GRAPH

This section rewrites the graph by hand, choosing where target blocks fall so that the sampler co-exposes the pairs we want. Appendix E specifies every schedule.

### 4.1 IS THE INTERVENTION SPECIFIC?

We construct two schedules, each of which removes the co-exposure of one of the two RELFOREST relations by always giving its two nodes the same role, and read the influence matrix of each. Data, model, loss, mask ratio and per-patch marginals are identical across the schedules, and only the co-exposure differs. Removing a relation costs the corresponding pair about four fifths of its predictive influence and costs the other pair nothing (Table 2). The specificity has the same sign in all eight seeds, and the retentions of removed pairs separate from those of retained pairs with AUC $0.996$.

Table 2: Retention of influence (rows: removed relation). Only the diagonal falls.

| | Measured pair | |
|---|---|---|
| Removed | A | B |
| A | **0.196** | 1.296 |
| B | 1.096 | **0.228** |

Table 3: Agreement of the influence matrix with the specified graph on RELGRAPH.

| | GRS ($\rho_{\max} = 0.819$) | AUC |
|---|---|---|
| Raw influence | 0.736 | 0.908 |
| Divided by baseline | 0.672 | 0.829 |

Removing both relations costs each what removing it alone costs, to within $0.08$, so neither collapse depends on the other relation surviving (Table 10).

A second experiment separates supervision from co-occurrence. Its schedule makes $R_2$ and $R_3$ targets together or leaves both in the context, so the two regions still appear together in every input, and each is predicted at least as often as under a schedule that strengthens their edge. Neither is ever predicted from the other, and $W(R_2 \to R_3) = W(R_3 \to R_2) = 0$. The relational readout stays at chance at every probe size, $49.5\%$ to $50.2\%$ (Table 9), where the I-JEPA block schedule it is derived from reaches $59.8\%$ at $k = 4096$. Both regions keep their content, read at over $94\%$, so a nonlinear comparison of the two could still recover the relation: what the schedule removes is its linear readability. The other relation, whose co-exposure the schedule leaves unchanged, falls from $62.5\%$ to $57.1\%$ and stays above chance. The supervision graph is therefore not a co-occurrence graph: what makes a relation readable is that one region is predicted from the other.

**A second objective.** Replacing the representation-prediction objective with pixel reconstruction (He et al., 2022), on RELFOREST at $s = 0$, leaves the result in place. Both specificities keep their sign in all 16 seeds, and removed and retained pairs still separate, with AUC $0.968$ against $1.000$ under representation prediction. Two things change, both detailed in Appendix B: a single cut removes nearly all influence, so additivity cannot be tested, and training is unstable across seeds.

### 4.2 DOES IT REVERSE?

Specificity alone is consistent with a schedule that merely damages one pathway. The stronger test is whether the influence matrix follows when two schedules are exchanged. We run it on RELGRAPH, where every pair carries the same information, with two compiled schedules (Section 5.3) of equal marginals. One, $q_H$, raises the co-exposure of both horizontal pairs to $1.5$ times its value under a uniformly random mask and halves that of both vertical pairs, and the other, $q_V$, does the reverse (Figure 1). The routing index differs between them by $0.53$, with the same sign in all eight seeds, and the influence matrix agrees with the specified graph at $90\%$ of the attainable rank correlation (Table 3).

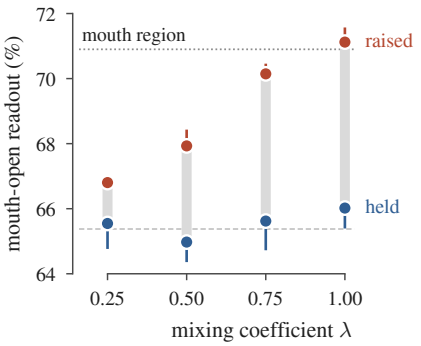


Figure 3: As $\lambda$ grows, the raised eye reaches the mouth region's level. Dashed: $\lambda = 0$. Whiskers: $\pm 1$ s.e.

### 4.3 IS IT GRADED IN CO-EXPOSURE?

If the graph is what carries the effect, then the size of the effect should follow the size of the edge. On tiled CELEBA we interpolate the schedule with a mixing coefficient $\lambda \in [0, 1]$ that raises the co-exposure of source and target, in both directions, from $7.3$ to $35.5$, while no node's marginals move by more than $0.04$. The eye whose co-exposure with the mouth corner below it rises we call the *raised* eye, and the other eye, whose pair the schedule does not target, the *held* eye (Figure 3). *Routing*, the raised eye's gain minus the held eye's, rises monotonically in $\lambda$, by about $5$ percentage points (pp) per unit $\lambda$ for mouth opening (positive in 7 of 8 seeds) and $8$ for smiling (8 of 8).

Table 4: One relation in glyphs and photos. The two $\Delta_{\text{rel}}$ are not comparable.

| | Glyphs | Photos |
|---|---|---|
| $\Delta_{\text{rel}}$ (%) | 2.22 | 5.12 |
| Relational gain (pp) | 32.2 (8/8) | 17.0 (16/16) |
| $\Delta$ content readout (pp) | $\geq -1.7$ | $> 0$ |
| Routing (pp) | — | 6.25 (15/16) |
| $\Delta_{\text{rel}}$ at $E \geq 16$ (%) | — | 0.09 |

Table 5: Eye to mouth in three settings. Last column: full image before intervention, tiles after routing.

| | $\Delta_{\text{rel}}$ (%) | Routing (pp) | Eye→mouth (%) |
|---|---|---|---|
| Full image | 2.30 | $\approx 0$ | 73.8 |
| Tiled | 20.47 | 4.99 (15/16) | 71.1 |
| Less context | 2.76 | $\approx 0$ | — |

### 4.4 Where does it hold?

**Content.** The relation in RelForest is carried by rendered glyphs and in RelCIFAR by natural photographs. The relation is the same, the content is not, and in both the pixel shortcut sits near chance. Table 4 shows that the intervention behaves the same way in both. On photographs, routing compares the gain of the raised relation with that of the other. The two values of $\Delta_{\text{rel}}$ are not comparable, since a glyph occupies a tenth of its noise tile and a photograph all of it.

Raising the number of images per class shrinks the effect along with the relation's share of the objective. At $E \geq 16$, $\Delta_{\text{rel}}$ falls to $+0.09\%$, a fiftieth of its value at $E = 4$: at this scale the objective barely uses class-level relations between natural images. Raising co-exposure then lifts the relational readout by $2.4$ pp at $E = 16$, in all eight seeds, and by less than one point at $E = 32$ and $64$, against $17.0$ at $E = 4$.

**A real relation.** On CelebA in the tiled layout, with the source node an eye and the target node the corner of the mouth below it, the effect appears at full strength. At $\lambda = 1$ the raised eye's readout of mouth opening rises from $65.4\%$ to $71.1\%$, the level the mouth region itself affords ($70.9\%$), while the held eye stays within a point of where it started. Routing is $5.0$ pp on mouth opening (15 of 16 seeds) and $7.8$ pp on smiling (16 of 16), and it reverses in all eight seeds when the schedules are exchanged. The generic readout does not fall. A larger encoder, of width $384$ and depth $8$, trained with the same budget nearly doubles routing on mouth opening, to $9.3$ pp in all eight seeds (Table 11).

**Layout, and why the mechanism has a scope.** On the full-image layout the same intervention routes nothing, and the reason can be measured. Before any intervention, the eye node's representation already reads mouth opening at $73.8\%$, above the $71.1\%$ that routing attains on tiles. The objective hardly needs the eye: any context block containing an eye also contains cheek, nose side, skin and lighting, all of which carry mouth state, and the relation's share of the objective on full images is about a ninth of its share on tiles (Table 5). There is nothing for the schedule to add.

A narrower context, which lets each block see less of the face, does not help: $\Delta_{\text{rel}}$ reaches only $+2.76\%$, and routing stays near zero. The schedule controls which regions are co-exposed, not whether the image already connects them.

## 5 Compiling a Specified Graph

Editing by hand requires knowing where to put the target blocks. The declarative version of the question is: given a specified graph, does a legal mask distribution realize it, and can we solve for one? Both questions become finite once the graph is written in the right variables.

### 5.1 The graph depends on the mask only through node counts

Let $t_i$ be the number of the $m = 9$ patches of node $i$ that a mask draw assigns to the target. The masks of this section take the context to be the complement of the target, so $|C \cap i| = m - t_i$, and Equation (1) becomes

$$W(i \to j) \;=\; \mathbb{E}_q\big[(m - t_i)\, t_j\big], \tag{3}$$

where $q$ is the distribution of the count vector $t \in \{0, \dots, m\}^4$. Under the default $q_0$ of this section the target patches are a uniformly random subset of fixed size. Every question about the graph is therefore a question about a distribution on $(m+1)^4 = 10^4$ states, few enough to enumerate exactly, although the count grows exponentially with the number of nodes.

**Proposition 1.** *Let $\mu_i = \mathbb{E}_q[t_i]$ and $|t| = \sum_i t_i$. Then for every ordered pair $i \neq j$*

$$W(i \to j) = m\mu_j - \mu_i\mu_j - \mathrm{Cov}_q(t_i, t_j), \qquad \sum_{i \neq j} W(i \to j) = K(\mu) + \sum_i \mathrm{Var}_q(t_i) - \mathrm{Var}_q(|t|),$$

*where $K(\mu)$, the sum of $m\mu_j - \mu_i\mu_j$ over ordered pairs, depends on the marginals alone.*

Both identities follow from expanding $\mathbb{E}[t_i t_j]$ (Appendix D). With the marginals fixed, editing the supervision graph is editing the covariance of the node counts, and the total supervision can change only through the variances of the counts and of their sum. At fixed variances, raising one edge must lower others. The first identity also gives $W(i \to j) - W(j \to i) = m(\mu_j - \mu_i)$, so at equal marginals, as in every compiled schedule here, the graph is symmetric. A directed edit needs unequal marginals or a context that is not the complement of the target, as the block schedules of Section 4 allow.

### 5.2 Which graphs a fixed geometry admits

Because Equation (3) is linear in $q$, so is feasibility. A specified graph $W^\star$ is realizable under marginals $\mu$ exactly when some distribution $q$ on the $10^4$ states satisfies $\mathbb{E}_q[(m - t_i)t_j] = W^\star(i \to j)$ for every ordered pair and $\mathbb{E}_q[t_i] = \mu_i$ for every node. A linear program decides this before any training, and returns either a distribution that realizes the graph or a certificate that none exists.

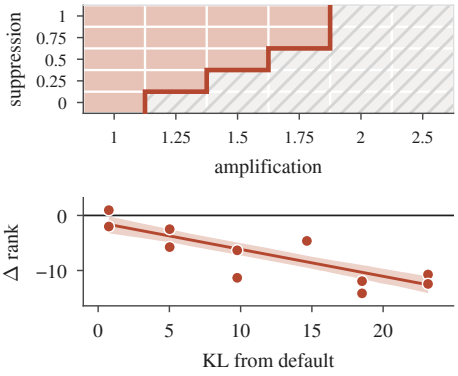


Figure 4: **(a)** Feasible multiplier pairs and their boundary $\alpha^*$. **(b)** Rank falls with KL from the default (80% bootstrap band).

We fix the geometry and the marginals and ask for a pair of multipliers: an amplification of the lit horizontal edges and a suppression of the dark vertical ones, with the diagonals held at their value under $q_0$. On a $6 \times 5$ grid, 14 of the 30 pairs are feasible, and the largest feasible amplification at each level of suppression traces the boundary of the feasible region, which we write $\alpha^*$ (Figure 4a). Two limits shape it, and they bind independently. At the total supervision of $q_0$, the contrast between lit and dark edges can reach exactly the co-exposure of one edge under $q_0$. At that contrast, the total can reach 1.27 times that of $q_0$, and no schedule with these marginals exceeds 1.30 times. The positions of the nodes never enter $q$, so moving them leaves every certificate unchanged: $\alpha^*$ describes what a fixed geometry admits and cannot rank one geometry against another.

### 5.3 The compiler

Given a specified graph, the compiler returns the count distribution that realizes it and lies closest to the default $q_0$ in Kullback–Leibler (KL) divergence,

$$q^\star = \arg\min_q \ \mathrm{KL}(q \,\|\, q_0) \quad \text{subject to} \quad \mathbb{E}_q[(m - t_i)t_j] = W^\star(i \to j), \quad \mathbb{E}_q[t_i] = \mu_i, \qquad (4)$$

a constrained I-projection (Csiszár, 1975). The constraints are linear in $q$, so the solution is an exponential tilt of the default, $q^\star(t) \propto q_0(t) \exp(\theta^\top f(t))$, where $f(t)$ collects the constrained statistics and $\theta$ holds one multiplier per constraint, found by minimizing the convex dual. A mask is drawn by sampling $t$ from $q^\star$ and placing $t_i$ target patches uniformly at random within each node, so every draw is a fresh mask.

Minimizing KL to the default maximizes entropy relative to the default, not in absolute terms. We use two *operating points*, which raise one pair by a factor of 1.3 or 1.5, halve its mirror and keep every other entry of $W$, with 24% or 48% of patches as targets. Table 7 shows that both keep most of the entropy of $q_0$, whereas graphs on the contrast boundary of Figure 4a force $q$ onto few states.

On RELFOREST the compiled schedule routes the relation it is asked to route. Two compiled schedules that mirror each other, and are therefore identical in KL divergence, entropy, marginals and total supervision, raise one relation and suppress the other. The same relation is read 23.6 pp better at $k = 4096$ under the schedule that raises its edge than under the one that suppresses it, in all eight seeds (Table 9). On faces, at both operating points, the compiled schedule routes in the specified direction and reverses when the specified graph is reflected.

### 5.4 WHAT THE CHANGE COSTS THE REPRESENTATION

Moving a schedule away from $q_0$ costs representational capacity. Within a fixed geometry on RELFOREST at $s = 0$, the further a schedule's count distribution lies from $q_0$ in KL divergence, the more the effective rank (Roy & Vetterli, 2007) of the representation falls ($\rho = -0.79$, $p = 0.002$, Figure 4b).

Measured on the pair that is actually being read, rather than pooled, the relationship runs the other way: effective rank and relational readout of the raised pair move in opposite directions ($\rho = -0.47$ to $-0.57$, $n = 48$). Raising the co-exposure of a pair increases the alignment of its two representations at the cost of the uniformity of that subspace (Wang & Isola, 2020), so a rank drop on that pair is part of the effect rather than a price paid for it. Appendix C gives the derivation, the full data and the same drop on faces.

### 5.5 DIRECTION, NOT MAGNITUDE

Compiling and placing can be set to the same co-exposure of the raised pair and compared, and they do not give the same result (Figure 5). On faces, placement lifts the raised eye's readout of mouth opening to 70.1% with less co-exposure than compilation needs to reach 68.9%. Compilation also costs 2.5 to 3.9 pp of the generic readout, which placement never does.

Placing and compiling agree in that one entry of $W$ and differ in the others, in the marginals and in the geometry of the masks, all listed in Table 8. The co-exposure of the edited pair therefore does not by itself set the size of the effect. In every schedule family we tested it does set the sign: the effect follows the edited pairs, and every reflected schedule reverses it.

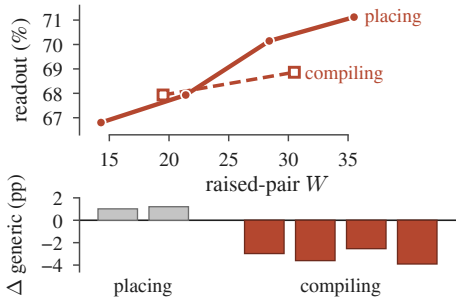


Figure 5: At equal co-exposure of the raised pair, compiling attains less than placing and costs generic readout.

## 6 CONCLUSION

The schedule of a masked pretraining objective determines a graph over pairs of image regions, and that graph can be rewritten with the data, the network and the objective held fixed. Rewriting it changes what the model learns to predict from what: the loss of predictive influence falls only on the pair whose edge was removed, exchanging two schedules exchanges the learned structure, and the effect follows the size of the edge. What carries it is supervision, not co-occurrence: when two regions always appear together but neither is predicted from the other, their relation does not become linearly readable. The effect holds on glyphs, photographs and face regions set on noise, and is larger in a larger encoder. In per-region target counts the graph is a covariance at fixed marginals, its feasible set is a polytope with exact certificates, and maximum-entropy projection compiles a specified graph into a mask distribution.

The scope is measurable. The effect shrinks with the share of the objective the relation carries, and it vanishes when the image's own coherence has already brought the information to the source node, as on whole faces, where the eye reads the mouth better than any schedule can make it. In every schedule family tested, the edited pairs set the direction of the effect, but their co-exposure alone does not set its magnitude. These results rest on four-node layouts, small encoders and one representation-prediction objective with variance–covariance regularization, and pixel reconstruction was tested only at $s = 0$.

### AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, designing experiments and giving feedback on them, cleaning and reformatting datasets, proposing and refining hypotheses, and interpreting results. We have not used generative AI tools for generating synthetic data sets, developing conceptual frameworks or translation, and formulating or proving mathematical claims, writing proofs, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for writing code, creating figures, suggesting experimental parameters, searching for and summarizing literature, formatting references, suggesting the paper's structure, drafting the paper and editing it for 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.

### ETHICS STATEMENT

The CELEBA experiments use face images and the attribute annotations distributed with that dataset. Those annotations are known to encode annotator judgments about appearance and to be unbalanced across demographic groups, and we use them only as readout targets for measuring where predictive information sits in a representation, not to build or evaluate any attribute classifier for deployment. We report two attributes (mouth opening, smiling) because they are localized in a region the schedule can target, not because predicting them is a task worth doing, and we read the annotated gender attribute only as a generic control for damage to the representation. No new data were collected and no human subjects were involved.

### REPRODUCIBILITY STATEMENT

Section 3 specifies the datasets, the relations, the layouts, the probes and the influence measurement. Appendix A gives the model configuration, the optimizer settings, the seed counts and the sign-flip permutation test, Appendix D the compiler and its certificates, and Appendix E the construction of each dataset, schedule and measurement. Code, schedules and per-seed outputs will be released.

## REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning (ICML)*, 2022.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M. Patel. AdaMAE: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, 2009.

Yichao Cai and Javen Qinfeng Shi. An identifiability theory of masked prediction: Mode blindness and mask schedules. *arXiv preprint arXiv:2608.01383*, 2026.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

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 (ICLR)*, 2021.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning (ICML)*, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations (ICLR)*, 2026.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision (ECCV)*, 2022.

Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. SemMAE: Semantic-guided masking for learning masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bingbin Liu, Daniel Hsu, Pradeep Ravikumar, and Andrej Risteski. Masked prediction: A parameter identifiability view. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, 2007.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning (ICML)*, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

# A SETUP DETAILS

**Model and objective.** The encoder is a ViT (Dosovitskiy et al., 2021) of width 192 and depth 6 on $96 \times 96$ images, trained with the I-JEPA objective (Assran et al., 2023) augmented with the variance and covariance terms of VICReg (Bardes et al., 2022) at weights $0.1$ and $0.004$. The larger encoder of Section 4.4 and Appendix C has width 384 and depth 8. The pixel-reconstruction comparison in Section 4.1 swaps the objective for the pixel reconstruction of the masked autoencoder (MAE) (He et al., 2022) and changes nothing else.

The regularization terms are not optional. Without them the trained encoder reaches an effective rank of $4.7$ out of 192, at which point every influence and readout we report is measuring a collapsed representation. At the weights above it settles at $37.4$. Both are measured on otherwise identical runs, and the rank keeps rising with the weights: $17.4$ at a fifth of them and $42.8$ at three times.

**Optimization.** AdamW (Loshchilov & Hutter, 2019), learning rate $3 \times 10^{-4}$, weight decay $0.05$, linear warmup over the first $10\%$ of steps, target-encoder exponential moving average (EMA) with momentum starting at $0.996$. Every run in the paper uses $8000$ steps at batch size $128$ over $50{,}000$ samples, with these settings.

**Probes.** Linear probes on frozen features, trained for $40$ epochs on $k$ examples drawn from a pool of $8192$ and evaluated on $4096$ held-out examples, with $k \in \{16, 64, 256, 1024, 4096\}$ (from $64$ on faces). We report balanced accuracy. The relational probe is binary. Effective rank is the exponential of the entropy of the normalized covariance spectrum of centered features (Roy & Vetterli, 2007).

**Seeds and the permutation test.** Routing and readout results use 8 seeds per condition, or 16 where a count out of 16 is given. $\Delta_{\text{rel}}$ uses 8 seeds, except 15 for the two RELCIFAR rows of Table 1 and 5 for target substitution on full images. The permutation test enumerates all $2^n$ sign assignments of the per-seed differences and reports the two-sided proportion at least as extreme as the observed mean. Its smallest attainable value is $2/2^n$: $0.0625$ at $n = 5$, $0.0078$ at $n = 8$. At that floor the effect size and the sign count carry the evidence, not $p$.

# B PIXEL RECONSTRUCTION: FULL NUMBERS

Section 4.1 reports that removal stays specific under a pixel-reconstruction objective, with two differences, both in Table 6. Both columns are on RELFOREST at $s = 0$, not the $s = 5$ configuration of Table 2.

The first concerns additivity: removing both relations should cost each what removing it alone costs. Under representation prediction the single-cut retentions are $0.084$ and $0.088$ against $0.139$ for the double cut, far enough from zero to compare. Under pixel reconstruction they are $0.0036$ and $0.0059$ against $0.0006$, all within rounding of zero, so additivity cannot be assessed.

The second difference is stability. When relation A is removed, the retention of the kept pair is $1.937 \pm 0.233$ under representation prediction and $1.289 \pm 0.878$ under pixel reconstruction, the latter ranging from $0.008$ to $4.193$ across 16 seeds. Two of the 32 single-cut runs collapse outright, in opposite directions: one falls to an effective rank of $1.8$ and one rises to $87.3$, and both drive every cross-region influence to zero. Effective rank across all pixel-reconstruction runs spans $1.1$ to $99.9$, against $33.2$ to $44.0$ under representation prediction.

# C EFFECTIVE RANK AND RELATIONAL READOUT

Section 5.4 reports two correlations of opposite sign, computed on different things.

**Pooled rank against the size of the edit.** The two analyses on RELFOREST are at $s = 0$. The calibration sweeps six operating points, each with a schedule and its reflection, and correlates the KL divergence of the schedule from the default against the change in the pooled effective rank of the representation. Over those twelve schedules, $\rho = -0.792$, $p = 0.002$. Adding three further operating points, at KL $8.78$, $13.18$ and $15.66$, gives $\rho = -0.726$, $p = 0.0007$, $n = 18$.

Table 6: Specificity and retention under the two objectives, 16 seeds each.

| | Representation | Pixels |
|---|---|---|
| Specificity, removing A | $-1.850 \pm 0.236$ | $-1.286 \pm 0.877$ |
| Specificity, removing B | $-1.303 \pm 0.192$ | $-1.126 \pm 0.485$ |
| Graph AUC | 1.000 | 0.968 |
| Retention, removed pair | 0.084, 0.088 | 0.0036, 0.0059 |
| Retention, both removed | 0.139 | 0.0006 |
| Retention, retained pair | $1.937 \pm 0.233$ | $1.289 \pm 0.878$ |
| Effective rank, range | 33.2–44.0 | 1.1–99.9 |

**Pair-local rank against that pair's readout.** Measured instead on the pair being read, and on the features the probe actually fits, the correlation between the change in effective rank and the change in relational readout is negative at every probe size: $\rho = -0.474$, $-0.502$, $-0.478$ and $-0.570$ at $k = 16$, 256, 1024 and 4096, each over $n = 48$ paired runs (four operating points, twelve seeds). Removing the between-point variation by centering within each operating point strengthens it to $\rho = -0.730$ at $k = 4096$, so the relationship holds inside a point and is not an artifact of comparing points with different KL.

The two are consistent. Pooled rank falls as the schedule moves further from the default, which is a cost. Rank on the designated pair also falls, but that pair's readout rises: co-exposure buys alignment between its two representations at the expense of the uniformity of that subspace (Wang & Isola, 2020). Pricing the readout by pooled rank therefore has the sign backwards, and results relating global rank to downstream performance (Garrido et al., 2023) do not transfer to one pair.

**On faces.** The pair-local drop needs capacity to show on faces. In the encoder of width 192 it is not detectable, whereas in the encoder of width 384 and depth 8 the rank of the raised pair falls from 130.8 to 74.8 in all eight seeds.

## D THE COMPILER

**The default and the count distribution.** In Section 5 a mask draw assigns a fixed number of the 144 patches to the target, 35 or 69 for the two target fractions, and the context is the rest. Under the default $q_0$ these patches are a uniformly random subset, so the node counts $t$ follow a multivariate hypergeometric distribution, which we compute exactly on all $10^4$ states. A compiled schedule changes only the distribution of $t$. To draw a mask we sample $t$ from $q^\star$, choose $t_i$ of the nine patches of each node $i$ uniformly at random, and fill the remaining target patches uniformly from outside the nodes. Patches inside a node are exchangeable under this construction, so matching the node means $\mathbb{E}[t_i]$ matches the target marginal of every patch.

**Proof of Proposition 1.** For an ordered pair $(i, j)$, and for the variance of the total,

$$W(i \to j) = m\mu_j - \mathbb{E}[t_i t_j] = m\mu_j - \mu_i\mu_j - \mathrm{Cov}(t_i, t_j),$$
$$\mathrm{Var}(|t|) = \sum_i \mathrm{Var}(t_i) + \sum_{i \neq j} \mathrm{Cov}(t_i, t_j).$$

The first line is the first identity of the proposition. Summing it over the twelve ordered pairs and substituting the second line gives the second. Numerically, both identities hold to $10^{-13}$ on the two defaults and on every distribution in Table 7. For the geometry of Figure 4a, $K(\mu) = 242.578$ and the default has $\sum_{i \neq j} W(i \to j) = 244.274$.

**Feasibility and the two limits.** The linear program of Section 5.2 has one variable per count state and 17 equality constraints: twelve ordered pairs, four node means and normalization. We solve it with HiGHS. On the 30 cells of Figure 4a it agrees with the projection in every cell, finding a realizing distribution for 14 and proving infeasibility for the other 16. The two limits are linear programs in $(q, \alpha)$ with a free scale $\alpha$. At the total of $q_0$, the largest contrast between lit and dark

edges is exactly the co-exposure of one edge under $q_0$, reached at amplification $1.5$ and suppression $0.5$. At unit contrast, the largest amplification is $1.907$, a total of $1.271$ times that of $q_0$. Under the four node means alone, the total can reach $1.299$ times that of $q_0$.

**Solving the projection.** The dual of Equation (4) is

$$\min_{\theta} \ \log \sum_{t} q_0(t) \exp\left(\theta^\top f(t)\right) - \theta^\top c,$$

where $f$ stacks the constrained statistics and $c$ their targets. We standardize each statistic, minimize with L-BFGS-B, and accumulate the partition function in log space so that it does not underflow when the tilt moves mass into states where $q_0$ is tiny. A target counts as solved when every constraint is met to within $10^{-4}$. Table 7 lists the compiled distributions used in the paper.

Table 7: Compiled count distributions. Effective states: $1/\sum_t q(t)^2$.

| Specified graph | Target fraction | KL (nats) | Entropy (nats) | Effective states |
|---|---|---|---|---|
| One pair $\times 1.3$, its mirror $\times 0.5$ | $0.24$ | 6.77 | $6.48 \to 5.70$ | 130 |
| One pair $\times 1.5$, its mirror $\times 0.5$ | $0.48$ | 8.78 | $7.16 \to 7.15$ | 714 |
| Lit $\times 1.25$, dark $\times 1.00$ (interior) | $0.48$ | 3.03 | $7.16 \to 7.79$ | 1814 |
| Lit $\times 1.50$, dark $\times 0.50$ (boundary) | $0.48$ | 24.53 | $7.16 \to 1.73$ | 5 |

The first row is used on RELFOREST and on faces, the second on faces, and the last, the boundary cell at amplification $1.5$ and suppression $0.5$ in Figure 4a, for the exchange of Section 4.2. The third is an interior cell, listed for comparison. The face version of the first row raises a vertical pair instead of a horizontal one, which is the same problem with the layout transposed, and gives the same KL, entropy and achieved graph.

**What Figure 5 matches.** Figure 5 places each schedule at the co-exposure of its raised pair, and Table 8 gives the rest of the graph. Placing and compiling agree only in that entry. Placing puts the two nodes of the raised pair in the target far more often than those of the held pair, so that the latter, nearly always in the context, supervise the former with a co-exposure near $40$. Compiling keeps all four nodes at one marginal and every entry outside the raised and the held pair at its value under $q_0$.

Table 8: Co-exposure and target marginals of the schedules in Figure 5, tiled CELEBA. A pair's entry averages its two directions, which differ by less than $0.4$. Placing: $10^5$ draws per schedule, standard error below $0.12$. Compiling: exact.

| | | Co-exposure | | | Target marginal | |
|---|---|---|---|---|---|---|
| | Schedule | Raised pair | Held pair | Other eight | Raised | Held |
| Placing | $\lambda = 0$ | 7.3 | 5.5 | 5.5–44.4 | 0.554 | 0.075 |
| | $\lambda = 0.25$ | 14.3 | 6.0 | 5.1–42.8 | 0.544 | 0.079 |
| | $\lambda = 0.5$ | 21.4 | 6.3 | 4.6–41.6 | 0.537 | 0.081 |
| | $\lambda = 0.75$ | 28.4 | 6.7 | 4.1–40.5 | 0.531 | 0.085 |
| | $\lambda = 1$ | 35.5 | 7.2 | 3.7–39.1 | 0.523 | 0.088 |
| Compiling, $0.24$ | $q_0$ | 15.0 | 15.0 | 15.0 | 0.243 | 0.243 |
| | $q^\star$ | 19.5 | 7.5 | 15.0 | 0.243 | 0.243 |
| Compiling, $0.48$ | $q_0$ | 20.4 | 20.4 | 20.4 | 0.479 | 0.479 |
| | $q^\star$ | 30.5 | 10.2 | 20.4 | 0.479 | 0.479 |

## E DATASETS, SCHEDULES AND MEASUREMENT

**Geometry.** Images are $96 \times 96$ with 8-pixel patches, a $12 \times 12$ grid. The four nodes are $24 \times 24$-pixel blocks with top-left corners at $(8, 8)$, $(8, 64)$, $(64, 8)$ and $(64, 64)$, so horizontally and vertically adjacent nodes are $56$ pixels apart, center to center.

**RELFOREST.** The glyphs follow a forest. Each root is drawn uniformly from ten classes, and each child takes its parent's class shifted by $s$ modulo ten with probability $\kappa = 0.8$ and an independent class otherwise. Children are conditionally independent given their parent, so an edge of the forest is the only path between the subtrees it joins. We use two edges, $R_0 \to R_1$ and $R_2 \to R_3$, rather than a star with one center. In the two-edge forest $I(R_0; R_1) = 2.071$ bits, unchanged when conditioning on $R_2$, and every pair across rows carries zero. In a star with the same $\kappa$, conditioning on a second leaf drops a spoke from $2.071$ to $0.685$ bits, so the other leaves could stand in for a cut spoke and a cut would not remove its information. Glyphs are rendered on independent noise. A classifier that judges the relation from pixel similarity alone reaches $91.6\%$ on held-out pairs at $s = 0$ and $49.6\%$ to $50.1\%$ at $s = 1, 3, 5$.

**RELGRAPH.** Each image draws one glyph class, and each of the four nodes shows it with probability $\kappa = 0.8$ and an independent class otherwise. The data are then symmetric under exchanging the horizontal and vertical pairs, so any asymmetry in the learned influence comes from the schedule.

**RELCIFAR.** The same forest over CIFAR-10 images, with the class propagating along the edges, so the relation is semantic identity between two natural images. Each class contributes $E$ images, and the nodes are pasted on noise so that the background cannot predict them.

**CELEBA.** In the aligned images the eye centers sit at $(69.4, 111.2)$ and $(107.6, 111.2)$ with a standard deviation of $1.7$ pixels. A $65.6$-pixel window rescaled to $96$ pixels places the eye centers on the centers of $R_0$ and $R_1$ and the mouth corners inside $R_2$ and $R_3$. The relations are the two vertical pairs, eye to the mouth corner below it. The full-image layout keeps the whole face. The tiled layout cuts out the four node boxes and pastes them on noise, taking $R_0$ and $R_2$ from one face and $R_1$ and $R_3$ from another, so each edge is the only path between its two nodes. The readouts are the attributes *mouth slightly open* and *smiling*, read from the eye node. In the tiled layout each image carries two faces, and the generic readout averages two probes on the pooled representation, one for the label of each face.

**Schedules.** Block sizes are fractions of the $144$ patches. Target blocks have aspect ratios in $[0.75, 1.5]$, the context block is square, and the context is the context block minus the target. A block is placed uniformly at random unless it is centered on a node. *Default block schedule*, the reference of the zero-co-exposure control: one target block of $10$–$15\%$ and a context block of $85$–$100\%$. *Placement*, used in Sections 4.3 and 4.4 and Figure 5 except for the less-context rows: two disjoint target blocks of $5$–$8\%$ and a context block of $85$–$100\%$. Under the joint rule, half the draws center one target block on each node of the pair and the other half place both blocks uniformly. Under the alternating rule, every draw centers one block on a node of the pair, each with probability $1/2$, and places the other uniformly, so that one node can be predicted while the other is in the context. A block that would overlap one already placed is redrawn. The mixing coefficient $\lambda$ is the probability of the alternating rule, and $\lambda = 0$ is the baseline. Both rules center the same expected number of blocks on each node, and Table 8 gives the resulting co-exposure and marginals. *Zero co-exposure* (Section 4.1): the default block schedule with the two nodes of the pair removed from the target, after which both become targets together with probability $0.35$. *Removal* (Tables 2, 6 and 10): the nodes are partitioned into groups, and a relation is cut by putting its two nodes in one group. Each group becomes target with probability $0.48$ and context otherwise, the rest of a fixed budget of $69$ target patches is drawn from outside the nodes, and the context is the complement of the target. Every node is then a target with probability $0.48$, and two nodes in one group are never co-exposed. *Less context* (Table 5): in half the draws one target block of $5$–$8\%$ is centered on a node drawn from the four, and a second is placed uniformly. The context block covers $25$–$35\%$. The edited schedule centers it on the other node of the pair whenever a target block is centered on one of them, and its control centers it on the pair as often, but only when no target block is.

**Influence and its summaries.** To measure influence on node $j$, the predictor sees every patch outside $j$ as context and predicts the patches of $j$. This query is the same for every schedule, and its $135$ context patches are more than the $75$ that the schedules on which we measure influence provide in training. We render pairs of inputs $x, x'$ that differ only in the content of node $i$ and set $S(i \to j) = \mathbb{E}\,\|\hat{z}_j(x) - \hat{z}_j(x')\| / \|\hat{z}_j(x)\|$ over held-out pairs. The second row of Table 3 divides each entry by the same entry of the same seed's baseline model. This removes a geometric floor that

holds whatever the relation: influence is about $1.4$ times higher between horizontally adjacent nodes than between vertically adjacent ones. GRS is the Spearman correlation, over the twelve ordered pairs of one seed, between the influence matrix and the specified graph written as $1$ on lit pairs and $0$ elsewhere, averaged over seeds. Because the specified graph is binary, ties cap the correlation at $\rho_{\max} = 0.819$ with four lit pairs and $0.648$ with two. Of the two AUCs, the one in Table 3 compares lit with dark influence values, and the one in Section 4.1 compares the retentions of removed pairs with those of kept pairs, pooled over seeds and conditions.

## F ADDITIONAL RESULTS

**Removing co-exposure without removing co-occurrence.** Table 9 gives the zero-co-exposure control of Section 4.1 at every probe size. Its schedule raises the target marginal of both nodes to $0.35$, above the $0.33$ and $0.09$ of the schedule it controls for, measured over $10^5$ draws. It leaves the other relation's co-exposure at $9.5$, raises that from $R_0$ and $R_1$ to $R_2$ and $R_3$ from $9.5$ to $25.0$, and lowers the reverse to $6.2$, all on pairs that carry no information. At $k = 4096$ the content readouts of the two nodes fall from $97.6\%$ and $97.2\%$ to $94.1\%$ and $94.0\%$, those of the other two nodes by less than one point, and the other relation from $62.5\%$ to $57.1\%$.

**Compiled routing at every probe size.** The right half of Table 9 gives the compiled swap of Section 5.3: the readout of one relation under the schedule that raises its edge from $15.0$ to $19.5$, minus its readout under the mirror schedule that lowers it to $7.5$.

Table 9: Relational readout on RELFOREST at $s = 5$ by probe size, eight seeds.

| | Co-exposure removed (%) | | | | Compiled swap (pp) | |
|---|---|---|---|---|---|---|
| $k$ | Default | Control | Difference | Lower | Raised − lowered | Higher |
| 16 | 50.89 | 49.81 | −1.08 | 6/8 | +3.05 | 7/8 |
| 64 | 50.46 | 49.73 | −0.73 | 5/8 | +7.97 | 8/8 |
| 256 | 52.96 | 50.15 | −2.81 | 7/8 | +13.07 | 8/8 |
| 1024 | 55.29 | 50.14 | −5.15 | 8/8 | +18.30 | 8/8 |
| 4096 | 59.83 | 49.53 | −10.30 | 8/8 | +23.59 | 8/8 |

**The full factorial of cuts.** Table 10 adds the condition that removes both relations. Each relation then keeps within $0.08$ of its retention when removed alone, so neither collapse depends on the other. Effective rank differs by less than three across the four conditions, ruling out a general collapse.

Table 10: Retention of influence under the four schedules, RELFOREST at $s = 5$, eight seeds.

| Removed | Relation A | Relation B | Effective rank |
|---|---|---|---|
| Neither | 1 | 1 | 45.1 |
| A | 0.196 | 1.296 | 43.9 |
| B | 1.096 | 0.228 | 43.5 |
| Both | 0.228 | 0.300 | 42.5 |

Table 11: Routing on tiled CELEBA at $\lambda = 1$ by encoder size, eight seeds.

| | Mouth opening (pp) | | | Smiling (pp) |
|---|---|---|---|---|
| Encoder | Raised eye | Held eye | Routing | Routing |
| Width 192, depth 6 | +5.71 | +0.90 | +4.81 (7/8) | +7.08 (8/8) |
| Width 384, depth 8 | +8.22 | −1.04 | +9.25 (8/8) | +10.49 (8/8) |

**Model size.** Table 11 compares the face experiment at $\lambda = 1$ for the two encoders on the same eight seeds. The larger one ends training at a higher loss, $0.358$ against $0.300$, so it is not yet converged at this budget, and it routes more on both attributes.

**The relation, not the permutation.** At the largest mixing coefficient, raising co-exposure lifts the relational readout at $k = 4096$ by $40.0$ pp at $s = 5$ and by $34.4$ pp at $s = 3$, in all eight seeds at both. The effect belongs to the relation rather than to one particular shift.