A. Supplementary Material

A.1. Description of simulated RSV distributions

When evaluating the RSV on a synthetic distribution, we considered the following generative model that consists of a common component x_0 with additive noise:

$$\begin{aligned} x_a = x_0 + n_a, \quad x_b = x_0 + n_b, \\ z_i = w_i x_a + (1 - w_i) x_b, \\ w_i \sim \text{Beta}(\alpha, \beta), \quad x_0 \sim \mathcal{N}(0,\ 1), \quad n_a \sim \mathcal{N}(0,\ 1), \quad n_b \sim \mathcal{N}(0,\ 1). \end{aligned} \tag{6}$$

Depending on the values of α and β, the Beta distribution that the weights w_i are drawn from will take different shapes, changing how units in the representation z vary with inputs x_a and x_b. We find that the distribution of RSVs in Fig. 3 reflect the full spectrum of these various distributions, where the resulting RSVs can vary from an approximately Gaussian distribution where units vary equally with both modalities, to polarized representations where units vary uniquely with one modality

For this synthetic simulation, we can derive a closed form expression for the RSV. In particular (and dropping the subscript i for clarity),

$$z = x_0 + w n_a + (1 - w) n_b \tag{7}$$

and note that z will be distributed as a normal distribution. Then,

$$SV_i = Var(Z|X_a = x_a) \tag{8}$$
$$= \sigma_z^2 (1 - p^2) \tag{9}$$
$$= \sigma_z^2 (1 - \frac{Cov(z, x_a)^2}{\sigma_z^2 \sigma_{x_a}^2}) \tag{10}$$

We know that

$$\sigma_z^2 = \sigma_{x_0}^2 + w^2 \sigma_a^2 + (1 - w)^2 \sigma_b^2 \tag{11}$$

since x_0, n_a, and n_b are independent. Finally,

$$Cov(z, x_a) = \mathbb{E}[(Z - \mathbb{E}[Z])(X_a - \mathbb{E}[X_a]] \tag{12}$$
$$= \mathbb{E}[Z X_a] \tag{13}$$
$$= \mathbb{E}[(w X_a + (1 - w) X_b) X_a] \tag{14}$$
$$= \mathbb{E}[(w(X_0 + N_a) + (1 - w)(X_0 + N_b))(X_0 + N_a)] \tag{15}$$
$$= \mathbb{E}[(X_0 + w N_a + (1 - w) N_b)(X_0 + N_a)] \tag{16}$$
$$= \mathbb{E}[X_0^2] + w \mathbb{E}[N_a^2] \tag{17}$$
$$= \sigma_{x_0}^2 + w \sigma_a^2 \tag{18}$$

We also know that

$$\sigma_{x_a}^2 = \sigma_{x_0}^2 + \sigma_a^2. \tag{19}$$

We can then solve for SV_i by plugging Eq 9, 16, 17 into Eq 8 and obtain:

$$SV_i = \sigma_z^2 (1 - \frac{Cov(z, x_a)^2}{\sigma_z^2 \sigma_{x_a}^2}) \tag{20}$$
$$= (\sigma_{x_0}^2 + w^2 \sigma_a^2 + (1 - w)^2 \sigma_b^2)(1 - \frac{\sigma_{x_0}^2 + w \sigma_a^2}{(\sigma_{x_0}^2 + \sigma_a^2)(\sigma_{x_0}^2 + w^2 \sigma_a^2 + (1 - w)^2 \sigma_b^2)}) \tag{21}$$

We assumed that the representation z_i for half of the units were sampled from above generative model, while the other half the representation z_i were sampled from the reverse convex combination of inputs, i.e, $z_i = w_i x_b + (1 - w_i) x_a$.

For simulations 2-4, we set $\beta = 20$ and varied α in $[1, 20, 30]$ respectively. We considered a representation on $N = 20000$ units. For the first simulation we only considered the half of units in the generative model above, with $\alpha = 1$ and $\beta = 10$.

A.2. Generalization of RSV to arbitrary number of sensors

We can naturally generalize the RSV to an arbitrary number n of sources. To do so, define:

$$SV_i(X_j, x_1, ..., x_{j-1}, x_{j+1}, ..., x_n) = Var(f(\mathbf{X})_i | X_1 = x_1, ..., X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}, ..., X_n = x_n),$$

and then collect the individual source variances into a vector $\mathbf{SV}_i$ of size n. Then normalized sensor variance would be

$$RSV_i = \text{softmax}(\mathbf{SV_i}),$$

which provides a normalized quantification (between 0 and 1) of how much an individual unit varies with each sensor modality j.

A.3. Description of deep linear network experiment

We considered the original input-output correlation (before dropping a sensor) to be

$$\mathbf{\Sigma}^{yx}_{pre} = \begin{bmatrix} 1 & 0 & 3 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{22}$$

Our perturbation involved dropping a sensor, in this case the third column, leading to

$$\mathbf{\Sigma}^{yx}_{post} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{23}$$

Using the analytical equations for the learning dynamics given by [26] for the shallow and deep network, we investigated how learning the task (row 5) was affected (Fig. 2), finding that such a perturbation had a significant on the dynamics of sensor learning in the deep, but not shallow, network.

A.4. Description of architectures and training

Most of our experiments are based on the ResNet-18 architecture [15]. We modified the architecture to process multi-sensor input with what we call a SResNet-18. We separately process two initial pathways which we combine in an additive manner. In particular, the initial pathway followed the architecture of [15] directly up to (and including) `conv3_x` (See Table 1 of [15]). After combining the pathways, the remaining layers followed the ResNet-18 architecture directly.

To examine the effect of depth, we modified the All-CNN architecture [28], following [1]. In particular we processed each pathway with the following architecture:

$$\texttt{conv 96 - [conv } 96 \cdot 2^{i-1} \texttt{ - conv } 96 \cdot 2^{i} \texttt{ s2]}_{i=1}^{n} \texttt{ - conv } 96 \cdot 2^{n} \texttt{ - conv1 } 96 \cdot 2^{n} \texttt{ - conv1 10}$$

where s refers to the stride. We then merged the final representation from each pathway in an additive manner. We examined the setting when $n = 1, 2, 3$. We used a fixed learning rate of 0.001 in these experiments.

A.5. Description of Blurring Experiments (Fig. 4)

We attempted to simulate a cataract-like deficit by blurring the image to one pathway. We reduced the resolution of the image being passed to one pathway by first resizing the Cifar images to 8×8, and then resizing to its original size (32×32 pixels, decreasing the available information.

While training, we applied standard data augmentation on the uncorrupted pathway (random translation of up to 4 pixels, and random horizontal flipping. We then retained a width w of the leftmost and rightmost pixels from uncorrupted and corrupted pathway respectively, setting $w = 16$ unless otherwise stated. At inference time, no data augmentation was applied and the leftmost w pixels and rightmost w pixels was supplied to each pathway respectively. We used an initial learning rate of 0.075, decaying smoothly at each epoch with a scale factor of 0.97. We also found that using a fixed learning rate of

0.0005 (Fig. 15) and different initial learning rates (Fig. 16, right) had similar RSV and performance changes as a result of the initial deficits.

To quantify the information contained in the representation, we randomly masked out each pathway with $p = 0.1$ during training, and computed the usable information I_u contained in the representation Z abbout the task Y following [19, 35] by computing $I_u(Z;Y) = H(Y) - L_{CE}$, with $H(Y)$ being known and equal to $\log_2 10$ since the distribution of targets is uniform, and L_{CE} being the cross-entropy loss on the test set. We reported the corresponding RSV plots, and network performance in Appendix Fig. 9, which reveal similar performance trends and polarization of units, when pre-training with the random masking as in Fig. 4.

A.6. Description of Independent Pathways Experiment (Fig. 6)

We followed the same setup as above, but instead randomly permuted the images fed to the 'right' pathway across the batch, breaking the correlation between the views. We trained using an initial learning rate of 0.05, decaying smoothly with a scale factor of 0.97. When training with the deficit we randomly sampled the target from the different views with $p = 0.5$. We also modified the architecture to produce multiple classification outputs, corresponding to a classification based on both views, or each pathway respectively. This modification was helpful for interpreting the polarization plots. While training, the loss function was applied on the head that contained the proper input-target correspondence. After the deficit, and during inference, only the head corresponding to both views was used.

A.7. Description of Masking + Supervised MultiViT training

These experiments were based on the MultiMAE architechture [4], using their implementation and closely following their default settings. We adapted their implementation to process two separate RGB views coming from Kinetics-400 dataset [7]. We used a patch size of 16 in all experiments, and the AdamW optimizer [24]. All inputs were first resized to 224×224 pixels. Our learning rate followed the linear scaling rule [13].

For the masking sensitivity experiments in Fig. 8, we used a fixed delay of 1.33 seconds (4 frames) between frames, and trained with an initial base learning rate of 0.0001, with 40 epochs of warmup for the learning rate. We trained for 800 epochs, with a 200 epoch deficit of independent frames during the pre-training starting at different epochs during training. We used a masking ratio of 0.75. We pre-trained with a batch size of 256 per GPU on 8 GPUs. After the pre-training, we fine-tuned for 20 epochs with all the tokens and the corresponding action classification label. We fine-tuned on 8 GPUs with a batch size of 32. We fine-tuned with a learning rate of 0.0005, with 5 epochs of warmup.

For the supervised experiments, we trained our networks with an initial base learning rate of 0.01 for 120 epochs using all the tokens, with 20 epochs of warmup. We applied a temporary deficit of independent frames for 20 epochs, starting at various epochs during the training. We used in cutmix (1.0) and mixup (0.8) applied to each view) while training and we used a random baseline between frames. For the supervised experiments, we used a batch size of 64 per GPU.

In both the masking and supervised experiments in Fig. 8, we reported the difference of networks trained with a deficit starting at different epochs of training against a corresponding model trained without any deficit. In Fig. 7, we show example reconstructions from our Multi-View transformer pre-trained without a deficit for 800 epochs with a random baseline between frames.

B. Additional Plots

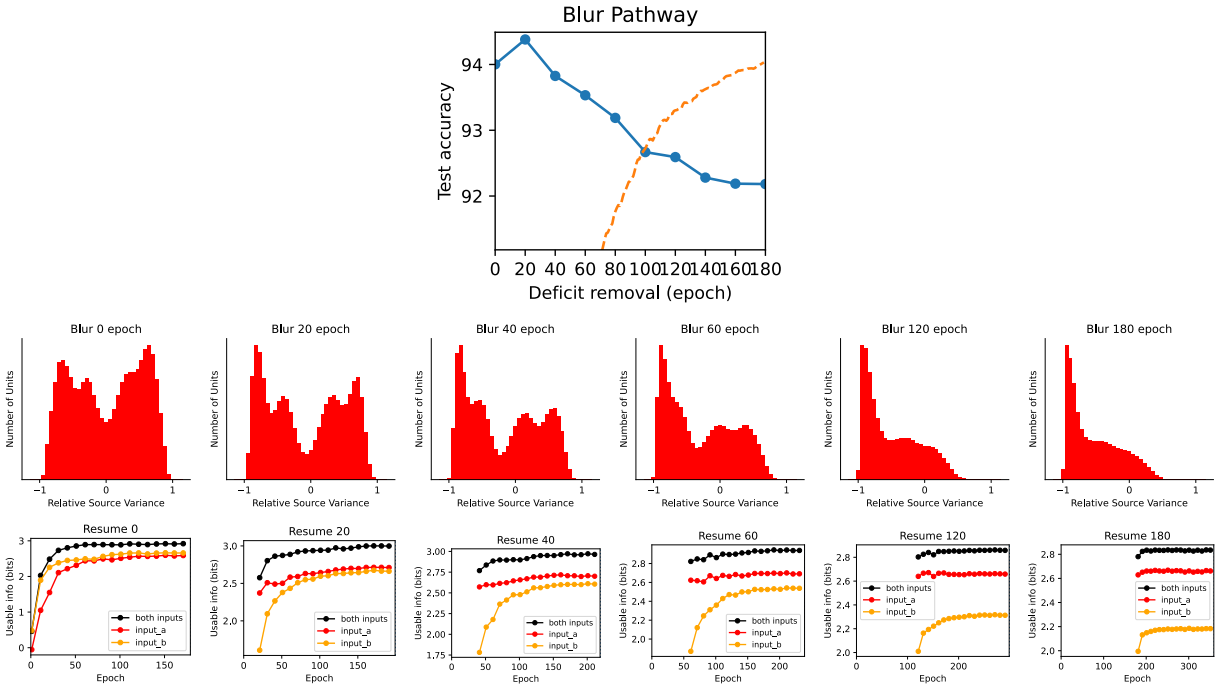

Figure 9. Same blurring experiment as Fig. 5 with corresponding Relative Source Sensitivity, Fig. 4, but with the addition of random masking on each view with $p = 0.1$, allowing the decoding of the usable information [19] (bottom row). Note that the polarization (second row) is similar to Fig. 4, which is also reflected by the inability to decode the inhibited pathway, after exposure to a sufficiently long deficit (orange trace in bottom row).

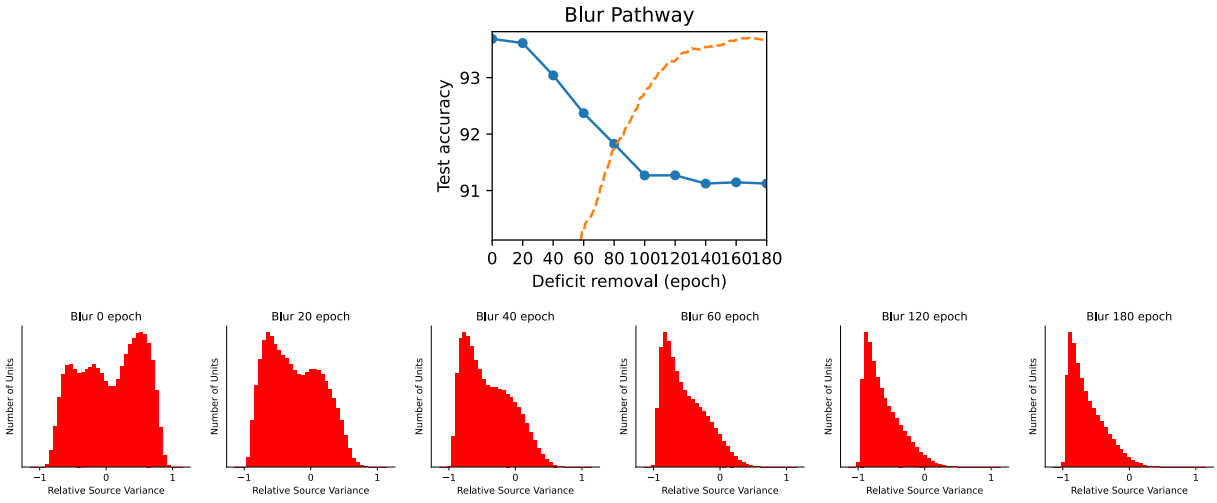

Figure 10. Same blurring experiment as Fig. 5 with corresponding Relative Source Sensitivity, Fig. 4 for crop width of 16 (used in the main text) for easier comparison against different crop widths in Fig. 11 and Fig. 12.

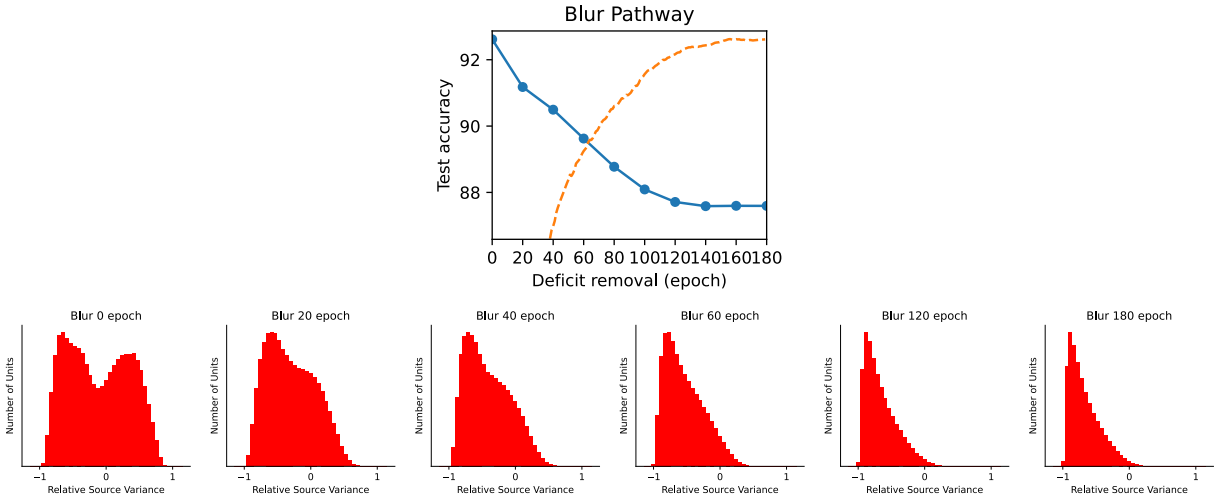

Figure 11. Same blurring experiment as Fig. 5 with corresponding Relative Source Sensitivity, Fig. 4 for crop width of 14.

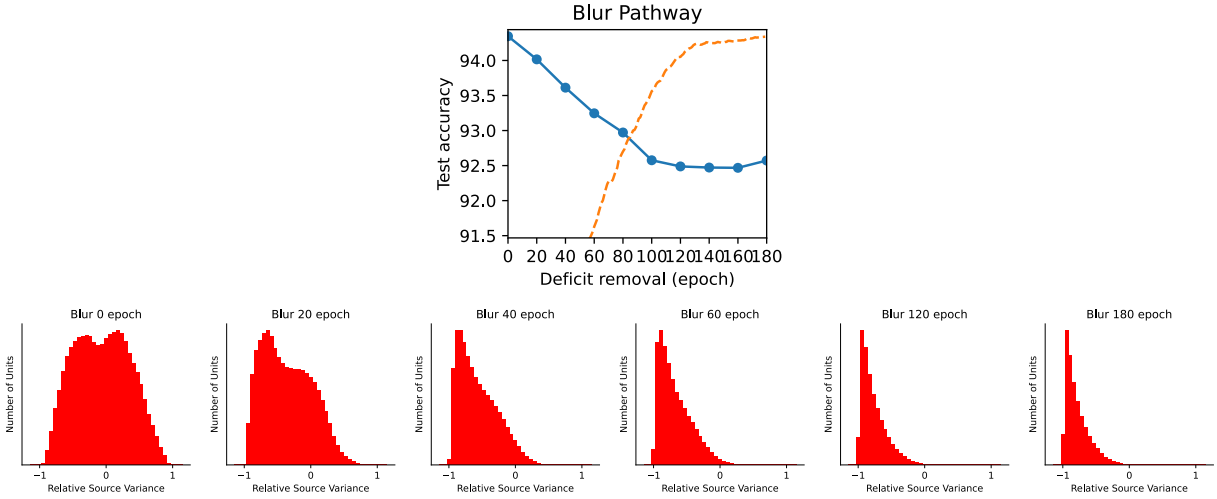

Figure 12. Same blurring experiment as Fig. 5 with corresponding Relative Source Sensitivity, Fig. 4 for crop width of 18.

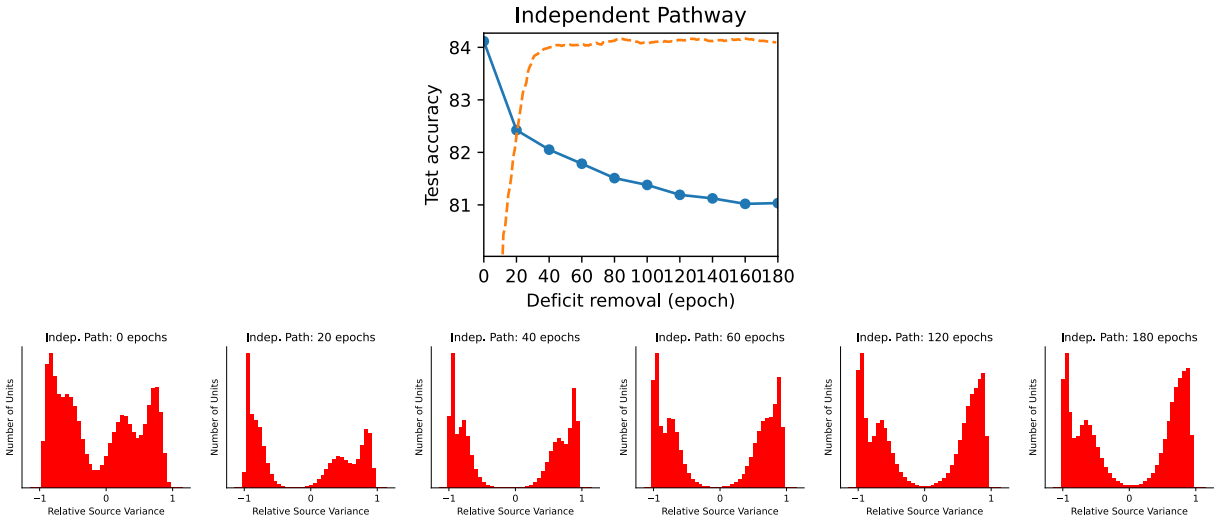

Figure 13. Strabismus-Like Deficit for ablation of no weight decay (wd $= 0$), no data augmentation and initial lr $= 0.05$. We also observe a polarized representation. Note the performance is reduced in comparison to Fig. 6, due to the lack of data augmentation and weight decay.

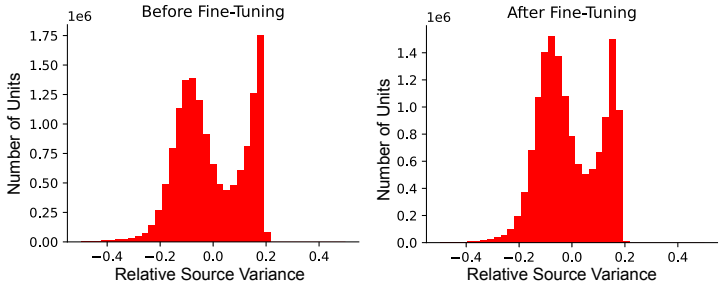

Figure 14. **Relative Source Variance for Multi-View Transformer. (Left)** We show the distribution of RSV evaluated on the units at output of the encoder before fine-tuning, revealing a bimodal distribution. Here, training was performed without any deficits. **(Right)** During fine-tuning, the representations appear to adapt to become slightly more balanced, depending more evenly on each view, while retaining the initial bimodal structure learned during pre-training.

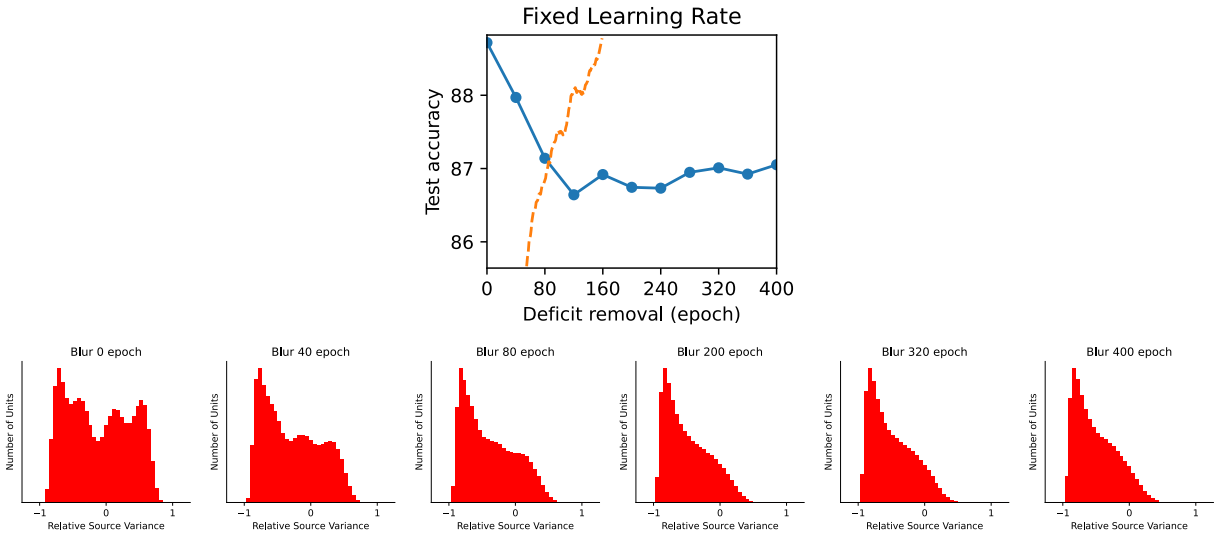

Figure 15. Fixed learning rate of 0.0005 during training have similarly shaped critical periods to those in paper, and similar RSV distributions as a result of the deficit.

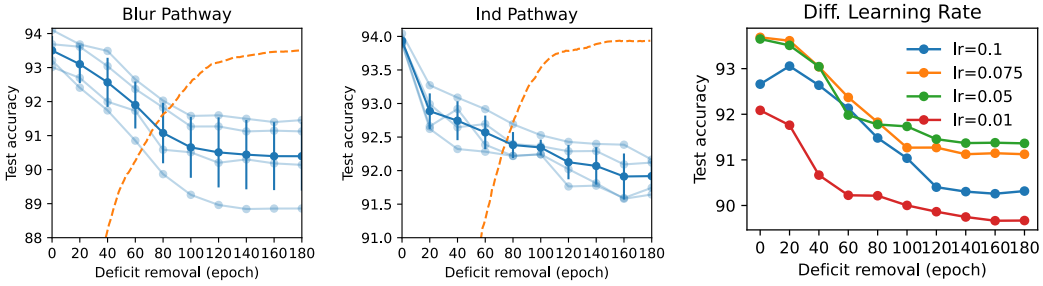

Figure 16. Results of multiple runs (light blue), their average (dark blue), and std (bars) for **(Left)** blurring and **(Center)** dissociation deficit. **(Right)** Different initial learning rates (for blur deficit) have have similarly shaped critical periods to those in paper.