Appendix A. Loss function derivation and pseudo code

A.1. Model M1

Here we derive the objective function in Eq. 4. The generative and recognition models are factorized as:

$$p_\theta(X_{\leq T}, y_{\leq T}, z_{\leq T}|\mathbf{x}_{\leq T}) = \prod_{t=1}^{T} p_\theta(X_t, y_t|z_{\leq t}, \mathbf{x}_{\leq t}) p_\theta(z_t)$$
$$q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}) = \prod_{t=1}^{T} q_\phi(z_t|\mathbf{x}_{\leq t})$$

The variational lower bound (ELBO) on the joint log-likelihood of the generated data, $\log p_\theta(X_{\leq T}, y_{\leq T}|\mathbf{x}_{\leq T})$, is derived as:

$$\begin{aligned}
&\mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\log p_\theta(X_{\leq T}, y_{\leq T}|\mathbf{x}_{\leq T}) \frac{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\log \frac{p_\theta(X_{\leq T}, y_{\leq T}, z_{\leq T}|\mathbf{x}_{\leq T})}{p_\theta(z_{\leq T}|\mathbf{x}_{\leq T})} \frac{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T} \log \frac{p_\theta(X_t, y_t|z_{\leq t}, \mathbf{x}_{\leq t}) p_\theta(z_t)}{p_\theta(z_t|\mathbf{x}_{\leq t})} \frac{q_\phi(z_t|\mathbf{x}_{\leq t})}{q_\phi(z_t|\mathbf{x}_{\leq t})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T} \Big[\log p_\theta(X_t, y_t|z_{\leq t}, \mathbf{x}_{\leq t}) - \log \frac{q_\phi(z_t|\mathbf{x}_{\leq t})}{p_\theta(z_t)} + \log \frac{q_\phi(z_t|\mathbf{x}_{\leq t})}{p_\theta(z_t|\mathbf{x}_{\leq t})}\Big]\Big] \\
&\geq \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T} \log p_\theta(X_t, y_t|z_{\leq t}, \mathbf{x}_{\leq t})\Big] - \sum_{t=1}^{T} D_{KL}\big(q_\phi(z_t|\mathbf{x}_{\leq t}), p_\theta(z_t)\big)
\end{aligned}$$

We assume, the modalities X_t and y_t are conditionally independent given the common latent variables (Wu and Goodman, 2018) and all observations till the current time. Therefore,

$$\begin{aligned}
\log p_\theta(X_{\leq T}, y_{\leq T}|\mathbf{x}_{\leq T}) \geq\ & \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T} \lambda_1 \log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t}) + \lambda_2 \log p_\theta(y_t|z_{\leq t}, \mathbf{x}_{\leq t})\Big] \\
& - \sum_{t=1}^{T} \beta D_{KL}\big(q_\phi(z_t|\mathbf{x}_{\leq t}), p_\theta(z_t)\big) \qquad \text{(A1)}
\end{aligned}$$

where λ_1, λ_2, β are the weights balancing the terms.

Algorithm 1: Learning the proposed network

Initialize parameters of the generative model θ, recognition model ϕ, sequence length T. Initialize optimizer parameters: $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\eta = 0.001$, $\epsilon = 10^{-10}$.
Initialize $x_1^{(1)} \leftarrow F(X_1, \ell_0)$, $x_1^{(2)} \leftarrow g_3(\ell_0)$, where ℓ_0 is the initial sampling location (ref. Experimental setup in Section 3), g_3 is an identity function (ref. Action selection in Section 2.4), and the function F extracts a sample $x^{(1)}$ (e.g., 5×5 patch) from the environment X (e.g., 28×28 image) at location ℓ (e.g., center of the image).

while *true* **do**
  for $\tau \leftarrow 1$ *to* T **do**
    Model M1
    $\hat{X}_\tau, \hat{y}_\tau \leftarrow PatComClassM1(x_{1:\tau}^{(1:2)})$
    Model M2
    $\hat{X}_\tau, \hat{y}_\tau \leftarrow PatComClassM2(x_{1:\tau}^{(1:2)})$
    Model M3
    $\hat{X}_\tau \leftarrow PatComClassM1(x_{1:\tau}^{(1:2)})$
    $\hat{y}_\tau \leftarrow Classifier(\hat{X}_\tau)$
    Saliency Computation
    $S_\tau \leftarrow g_1(X_{\tau+1}, \hat{X}_\tau)$ [ref. Eq. 2]
    $\ell_\tau \leftarrow g_2(S_\tau)$ [ref. Eq. 3]
    $x_{\tau+1}^{(2)} \leftarrow g_3(\ell_\tau)$
    $x_{\tau+1}^{(1)} \leftarrow F(X_{\tau+1}, \ell_\tau)$
    Learning
    Update $\{\theta, \phi\}$ or $\{\theta, \phi, \pi\}$ by maximizing Eq. 4, 5 or 6.
  end
end

A.2. Model M2

Here we derive the objective function in Eq. 5. The generative and recognition models are factorized as:

$$p_\theta(X_{\leq T}, y_{\leq T}, z_{\leq T} | \mathbf{x}_{\leq T}) = \prod_{t=1}^{T} p_\theta(X_t, y_t | z_{\leq t}, \mathbf{x}_{\leq t}) p_\theta(z_t)$$

$$q_\phi(z_{\leq T} | \mathbf{x}_{\leq T}, y_{\leq T}) = \prod_{t=1}^{T} q_\phi(z_t | \mathbf{x}_{\leq t}, y_t)$$

The variational lower bound (ELBO) on the log-likelihood of the generated data, $\log p_\theta(X_{\leq T}, y_{\leq T}|\mathbf{x}_{\leq T})$, when the true label is given is derived as:

$$
\begin{aligned}
&\mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big[\log p_\theta(X_{\leq T}, y_{\leq T}|\mathbf{x}_{\leq T})\frac{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big[\log \frac{p_\theta(X_{\leq T}, z_{\leq T}, y_{\leq T}|\mathbf{x}_{\leq T})}{p_\theta(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\frac{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big[\sum_{t=1}^{T}\log \frac{p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t})p_\theta(z_t)p_\theta(y_t)}{p_\theta(z_t|\mathbf{x}_{\leq t}, y_t)}\frac{q_\phi(z_t|\mathbf{x}_{\leq t}, y_t)}{q_\phi(z_t|\mathbf{x}_{\leq t}, y_t)}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big[\sum_{t=1}^{T}\Big[\log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t}) + \log p_\theta(y_t) - \log \frac{q_\phi(z_t|\mathbf{x}_{\leq t}, y_t)}{p_\theta(z_t)} \\
&\qquad\qquad + \log \frac{q_\phi(z_t|\mathbf{x}_{\leq t}, y_t)}{p_\theta(z_t|\mathbf{x}_{\leq t}, y_t)}\Big]\Big] \\
&\geq \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big[\sum_{t=1}^{T}\log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t}) + \log p_\theta(y_t)\Big] - \sum_{t=1}^{T} D_{KL}\big(q_\phi(z_t|\mathbf{x}_{\leq t}, y_t), p_\theta(z_t)\big) \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big[\sum_{t=1}^{T}(\log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t}) + \log p_\theta(y_t))\Big] - \sum_{t=1}^{T} D_{KL}\big(q_\phi(z_t|\mathbf{x}_{\leq t}, y_t), p_\theta(z_t)\big)
\end{aligned}
$$

After adding the classification loss, the final objective function can be written as:

$$
\begin{aligned}
&\mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}, y_{\leq T})}\Big[\sum_{t=1}^{T}\log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t}) + \log p_\theta(y_t)\Big] \\
&- \sum_{t=1}^{T} D_{KL}\big(q_\phi(z_t|\mathbf{x}_{\leq t}, y_t), p_\theta(z_t)\big) + \sum_{t=1}^{T}\alpha \log q_\phi(y_t|\mathbf{x}_{\leq t})
\end{aligned} \tag{A2}
$$

where α controls the relative weight between generative and purely discriminative learning.

A.3. Model M3

Here we derive the objective function in Eq. 6. The generative and recognition models are factorized as:

$$
\begin{aligned}
p_\theta(X_{\leq T}, z_{\leq T}|\mathbf{x}_{\leq T}) &= \prod_{t=1}^{T} p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t})p_\theta(z_t) \\
q_\phi(z_{\leq T}|\mathbf{x}_{\leq T}) &= \prod_{t=1}^{T} q_\phi(z_t|\mathbf{x}_{\leq t})
\end{aligned}
$$

Algorithm 2: $PatComClassM1(x_{1:\tau}^{(1:2)})$

Recognition Model

for $i \leftarrow 1$ *to* 2 **do**

  $h_\tau^{enc_i} \leftarrow RNN_\phi^{enc}(x_{1:\tau}^{(i)}, h_{\tau-1}^{enc_i})$

  $[\mu_\tau^{(i)}\ ; \Sigma_\tau^{(i)}] \leftarrow \varphi^{enc}(h_\tau^{enc_i})$

end

Product of Experts

$z_\tau \sim \mathcal{N}(\mu_\tau, \Sigma_\tau)$, where $\Sigma_\tau \leftarrow \Big(\sum_{i=1}^{2} {\Sigma_\tau^{(i)}}^{-2}\Big)^{-1},\ \mu_\tau \leftarrow \Big(\sum_{i=1}^{2} \mu_\tau^{(i)} {\Sigma_\tau^{(i)}}^{-2}\Big)\Sigma_\tau$

Generative Model

Pattern completion

$h_\tau^{dec_1} \leftarrow RNN_\theta^{dec}(z_\tau, h_{\tau-1}^{dec_1})$

$\hat{X}_\tau \leftarrow f_\sigma(h_\tau^{dec_1}, \hat{X}_{\tau-1})$

Classification Model

$h_\tau^{dec_2} \leftarrow RNN_\theta^{dec}(z_\tau, h_{\tau-1}^{dec_2})$

$\hat{y}_\tau \leftarrow softmax(h_\tau^{dec_2})$

The variational lower bound (ELBO) on the log-likelihood of the generated data, $\log p_\theta(X_{\leq T}|\mathbf{x}_{\leq T})$, is derived as:

$$
\begin{aligned}
&\mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\log p_\theta(X_{\leq T}|\mathbf{x}_{\leq T})\frac{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\log \frac{p_\theta(X_{\leq T}, z_{\leq T}|\mathbf{x}_{\leq T})}{p_\theta(z_{\leq T}|\mathbf{x}_{\leq T})}\frac{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T}\log \frac{p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t})p_\theta(z_t)}{p_\theta(z_t|\mathbf{x}_{\leq t})}\frac{q_\phi(z_t|\mathbf{x}_{\leq t})}{q_\phi(z_t|\mathbf{x}_{\leq t})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T}\Big[\log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t}) - \log \frac{q_\phi(z_t|\mathbf{x}_{\leq t})}{p_\theta(z_t)} + \log \frac{q_\phi(z_t|\mathbf{x}_{\leq t})}{p_\theta(z_t|\mathbf{x}_{\leq t})}\Big]\Big] \\
&\geq \mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T}\log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t})\Big] - \sum_{t=1}^{T} D_{KL}\big(q_\phi(z_t|\mathbf{x}_{\leq t}), p_\theta(z_t)\big)
\end{aligned}
$$

After adding the classification loss, the final objective function can be written as:

$$
\mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T}\log p_\theta(X_t|z_{\leq t}, \mathbf{x}_{\leq t})\Big] - \sum_{t=1}^{T} D_{KL}\big(q_\phi(z_t|\mathbf{x}_{\leq t}), p_\theta(z_t)\big) + \log q_\pi(y|X) \quad \text{(A3)}
$$

where $q_\pi(y|X)$ is the classification model whose input is the entire image (completed pattern) and not a sequence of observations. So the subscript t is dropped.

Algorithm 3: $PatComClassM2(x_{1:\tau}^{(1:2)}, y_{1:\tau})$

Classification Model

$h_\tau^{cls} = RNN_\alpha^{cls}(h_{\tau-1}^{cls}, \mathbf{x}_{1:\tau})$

$\hat{y}_\tau = softmax(h_\tau^{cls})$

Recognition Model

for $i \leftarrow 1$ *to* 2 **do**

  $h_\tau^{enc_i} \leftarrow RNN_\phi^{enc}(x_{1:\tau}^{(i)}, h_{\tau-1}^{enc_i})$

  $[\mu_\tau^{(i)}\ ; \Sigma_\tau^{(i)}] \leftarrow \varphi^{enc}(h_\tau^{enc_i})$

end

if *labels are present* **then**

  $h_\tau^{enc_3} \leftarrow tanh(y_\tau)$

else

  $h_\tau^{enc_3} \leftarrow tanh(\hat{y}_\tau)$

$[\mu_\tau^{(3)}\ ; \Sigma_\tau^{(3)}] \leftarrow \varphi^{enc}(h_\tau^{enc_3})$

Product of Experts

$z_\tau \sim \mathcal{N}(\mu_\tau, \Sigma_\tau)$, where $\Sigma_\tau \leftarrow \Big(\sum_{i=1}^{3} {\Sigma_\tau^{(i)}}^{-2}\Big)^{-1}$, $\mu_\tau \leftarrow \Big(\sum_{i=1}^{3} \mu_\tau^{(i)} {\Sigma_\tau^{(i)}}^{-2}\Big)\Sigma_\tau$

Generative Model

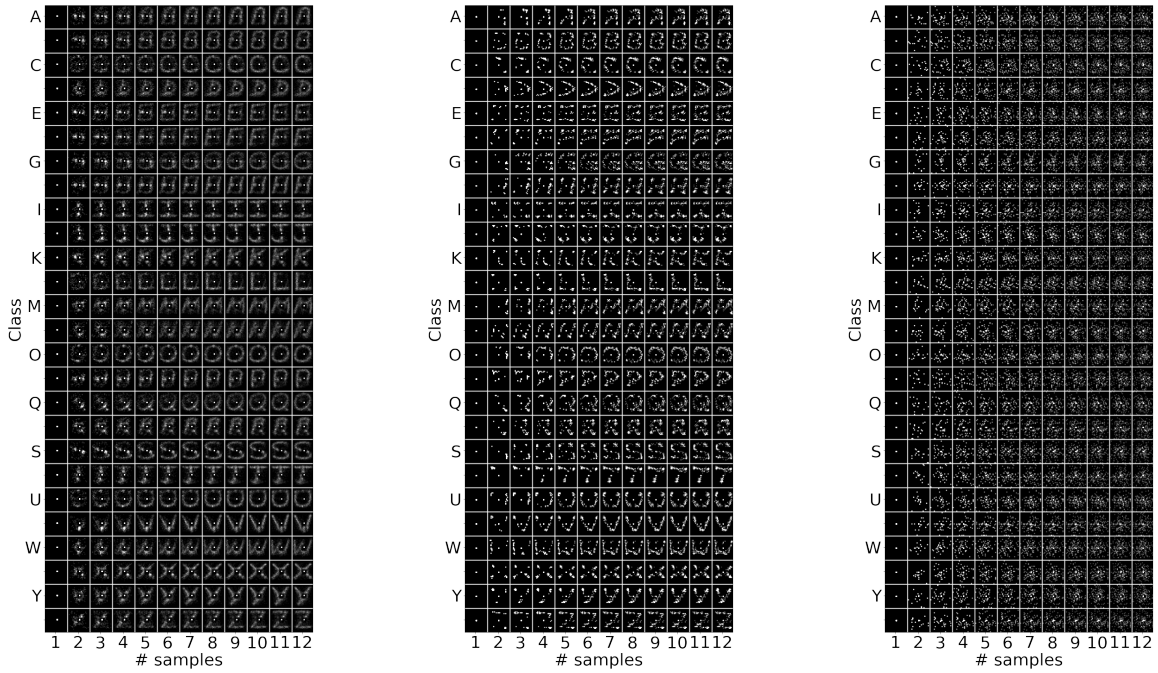
Pattern Completion

$h_\tau^{dec(1)} \leftarrow RNN_\theta^{dec}(z_\tau, h_{\tau-1}^{dec_1})$

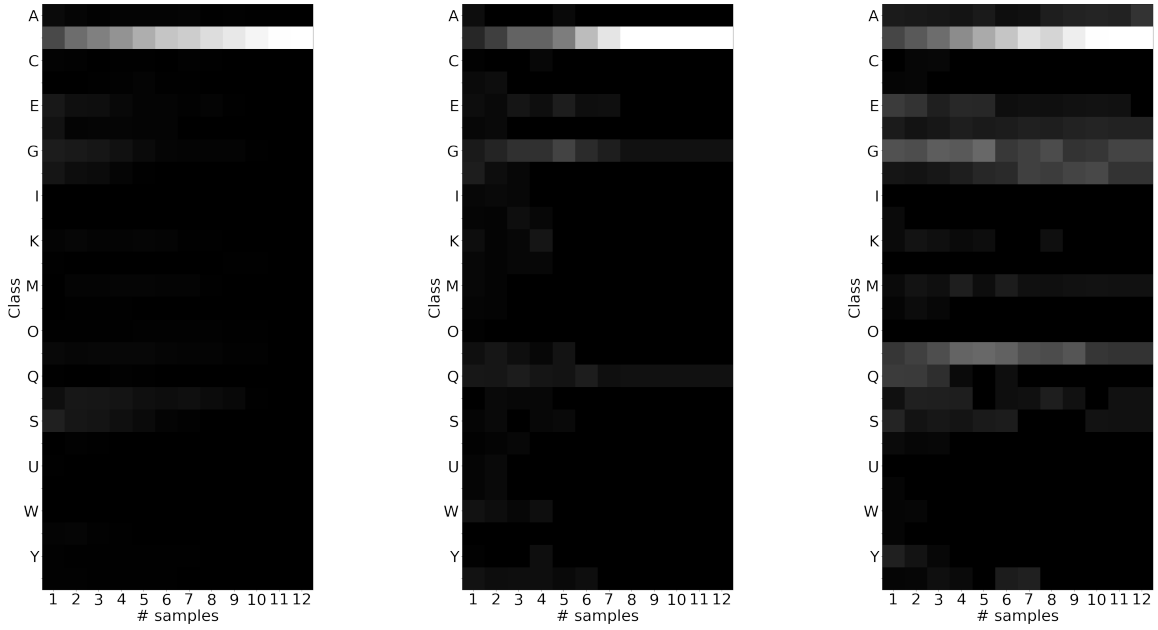
$\hat{X}_\tau \leftarrow f_\sigma(h_\tau^{dec_1}, \hat{X}_{\tau-1})$

Appendix B. Visualization of fixation maps

Visualization of the fixation maps obtained from our model (M1), RAM (Mnih et al., 2014), and the participants in (Baruah et al., 2023b), on uppercase and lowercase alphabets are shown in Figs. A1 and A2 respectively.

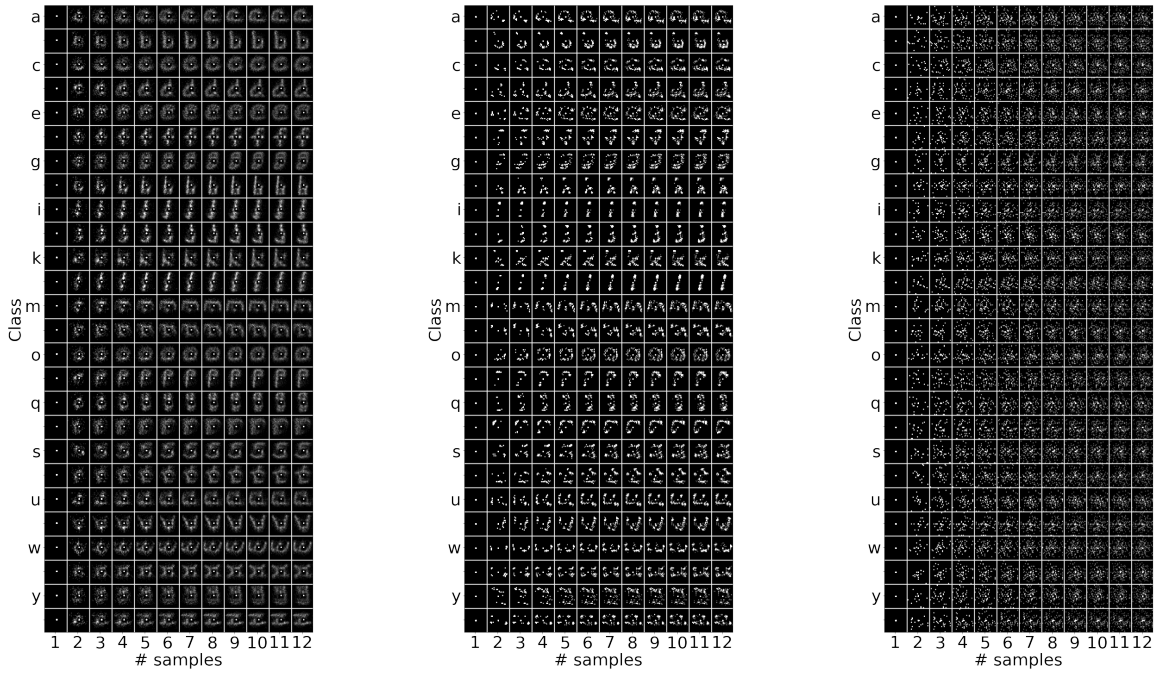


(a) Participants (b) Our model (M1) (c) RAM

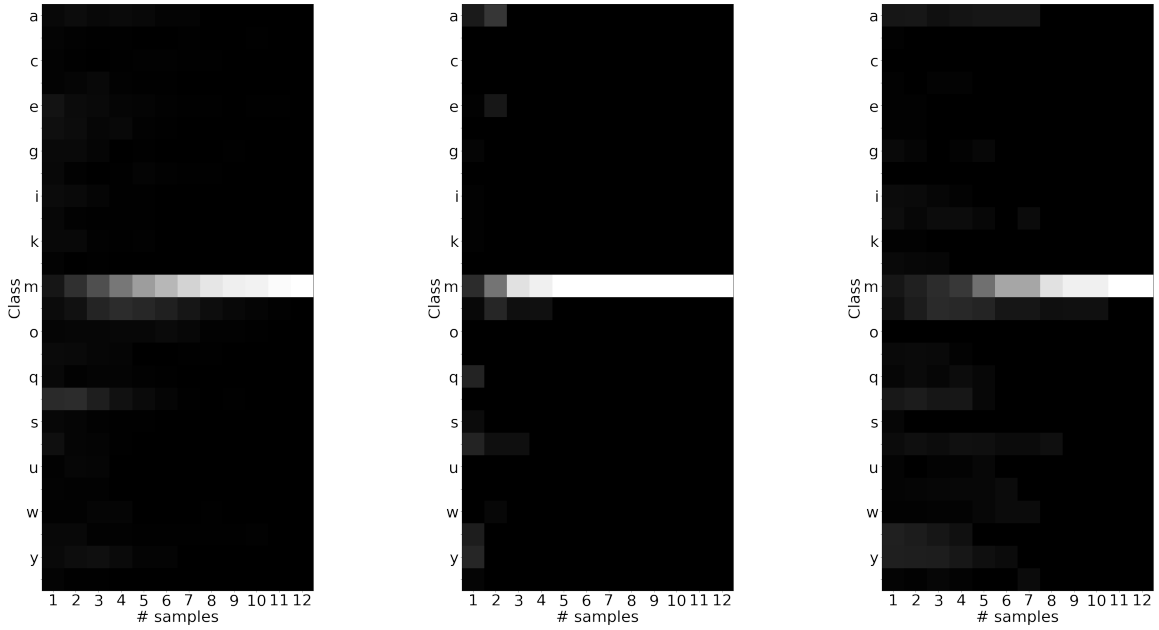


(d) Participants (e) Our model (M1) (f) RAM

Figure A1: (a)–(c) Distribution of sampling locations (or fixation maps) for each uppercase alphabet and each sampling instant. Qualitatively, the participants' fixation maps are more similar to our model's than RAM's. (d)–(f) Class distribution for class 'B'. The distributions are obtained by averaging the responses over all stimuli presented from each class. Each row corresponds to a class, and each column corresponds to a sampling instant which increases from left to right.



(a) Participants (b) Our model (M1) (c) RAM



(d) Participants (e) Our model (M1) (f) RAM

Figure A2: (a)–(c) Distribution of sampling locations (or fixation maps) for each lowercase alphabet and each sampling instant. Qualitatively, the participants' fixation maps are more similar to our model's than RAM's. (d)–(f) Class distribution for class 'm'. The distributions are obtained by averaging the responses over all stimuli presented from each class. Each row corresponds to a class, and each column corresponds to a sampling instant which increases from left to right.