Appendices

A. Neuron-wise Analysis

We present additional examples illustrating how the infant model perform classification using visual concept neurons. Furthermore, we provide complete results of Age of Acquisition (AoA) to quantify the cognitive level of visual concepts for both in-vocabulary and out-of-vocabulary cases.

A.1. Neuron-based Classification Examples

For this analysis, we conducted 4-way classification experiments on the Konkle object dataset to evaluate out-of-vocabulary classification. The examples are derived from neurons selected randomly under the specified experimental settings, with classification trial images and neuron activations presented, details in Figure 11 and Figure 12.

A.2. Age-of-Acquisition Ratings

Age-of-Acquisition (AoA) ratings, defined by Kuperman et al. [26], estimate the age at which a person learns a word. These ratings were obtained via crowdsourcing using 30,121 English content words, organized into frequency-matched lists based on the SUBTLEX-US corpus [4]. Each list included calibrator and control words for validation. It strongly correlated with prior norms ($r = 0.93$ with Cortese and Khanna [9], $r = 0.86$ with the Bristol norms [36]), confirming their reliability for studying vocabulary development.

Participants on Amazon Mechanical Turk rated the age they first understood each word on a numerical scale (in years). Words unfamiliar to participants could be marked with "x" to exclude outliers. Data cleaning removed non-numeric responses, ratings exceeding participant age, low-correlating responses ($r < 0.4$), and extremely high AoA ratings (> 25 years). This yielded 696,048 valid ratings.

A.2.1. Detailed AoA Results

Here we present visual concepts that from Konkle object dataset[23] class, by applying neuron labeling, we found many visual concept neurons with corresponding class inside vision encoder's hidden representation. For founded classes, we investigate their AoA values to prove the alignment between computational model and infant cognition.

A.3. Further Clarification on n-Way Classification Results

In Figure 8, the infant model CVCL (●) using "ours" has limited class coverage. As shown in the Venn diagram (Figure 5) and the coverage results (Figure 6), the "in-vocab" class in this setting is restricted. Consequently, the results include at most 31 classes. Therefore, in Figure 8, the rightmost point for "CVCL-ResNeXt50 (Ours)" corresponds to $n = 31$ instead of $n = 32$.

Table 2. **In-vocabulary Classes and Corresponding AoA Values.** The table lists the identified in-vocabulary classes along with their Age-of-Acquisition (AoA) values. For some classes, closely related words (shown in parentheses) were used to derive AoA values.

Vocab (Col 1)	AoA (Col 1)	Vocab (Col 2)	AoA (Col 2)
bike	2.9	abagel	4.79
stamp	2.94	umbrell	4.79
microwave	3.23	desk	5.00
pen	3.33	hat	5.11
knife	3.37	cookie	5.50
broom	3.43	stool	5.56
scissors	4.05	necklace	5.61
button	4.15	sofa	5.63
hairbrush	4.15	fan	5.68
pizza	4.26	chair	6.00
kayak	4.42	ball	6.21
bucket	4.5	sandwich	6.33
clock	4.5	pants	7.67
apple	4.67	socks (sock)	8.80
tricycle	4.7	bowl	8.90
camera	4.78		

Table 3. **Out-of-Vocabulary Classes and Corresponding AoA Values.** The table lists the identified out-of-vocabulary classes along with their Age-of-Acquisition (AoA) values. For some classes, closely related words (shown in parentheses) were used to derive AoA values.

Vocab (Col 1)	AoA (Col 1)	Vocab (Col 2)	AoA (Col 2)
sippycup (cup)	3.57	collar	6.56
toyrabbit (rabbit)	3.94	yarn	6.61
toyhorse (horse)	4.15	necktie	6.63
dresser	4.28	hanger	6.78
roadsign (sign)	4.32	binoculars	6.79
rug	4.61	telescope	6.95
doorknob	4.70	seashell	7.06
mask	4.80	golfball (golf)	7.16
dollhouse	4.86	dumbbell	7.56
muffins (muffin)	5.11	bathsuit (bathrobe)	7.90
tent	5.16	bowtie	7.94
hammer	5.42	rosary	8.21
frisbee	5.50	calculator	8.22
cushion	5.53	suitcase	8.22
watergun (gun)	5.58	trunk	8.30
ceilingfan (fan)	5.63	chessboard	8.37
helmet	5.71	compass	8.44
stapler	5.83	cupsaucer (saucer)	8.44
axe	6.11	lantern	8.55
speakers (speaker)	6.11	licenseplate (license)	8.70
lawnmower	6.11	pokercard (poker)	9.10
domino	6.17	keyboard	9.32
recordplayer	6.37	ringbinder (binder)	10.42
pitcher	6.42	powerstrip	12.01
grill	6.53		

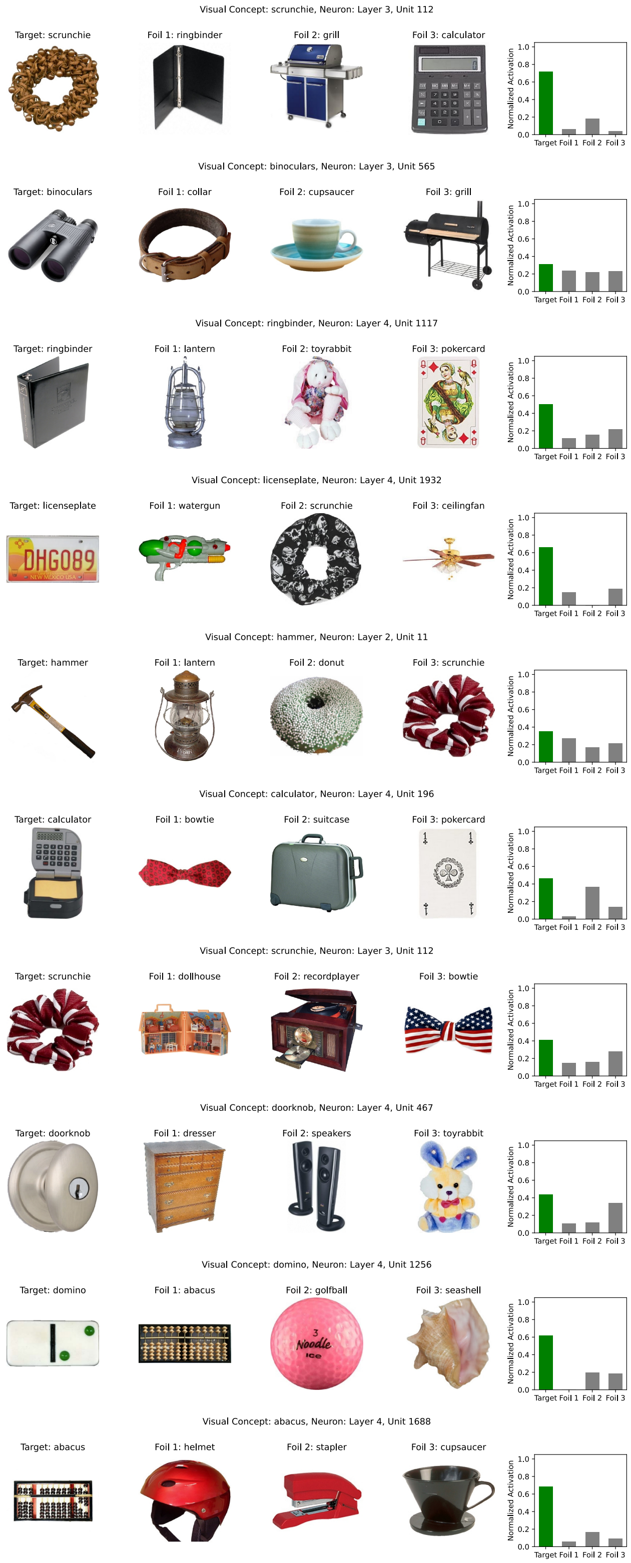

Figure 11. **Correctly Classified Examples.** Green bars indicate the highest normalized activation values, corresponding to the target image for correct classifications. Subtitles display information about visual concept neurons. These examples represent out-of-vocabulary classes from the Konkle object dataset [23].

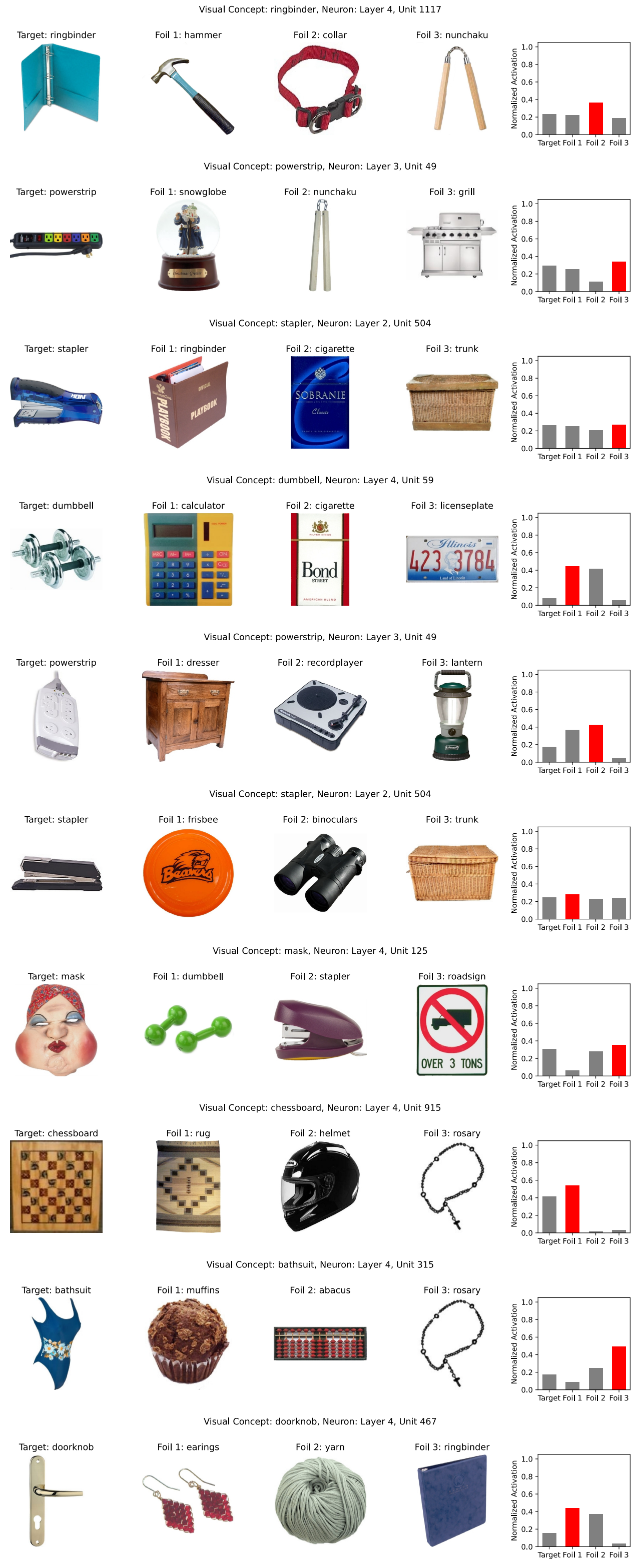

Figure 12. **Incorrectly Classified Examples.** Red bars indicate the highest normalized activation values, corresponding to incorrect classifications. Subtitles display information about visual concept neurons. These examples represent out-of-vocabulary classes from the Konkle object dataset [23].

B. Centered Kernel Alignment (CKA)

For two sets of activations, $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$, from corresponding layers of two models, where n is the number of examples and p and q are the feature dimensions (i.e., the number of neurons in each layer), the linear CKA is defined as:

$$\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{HSIC}(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathrm{HSIC}(\mathbf{X}, \mathbf{X})\,\mathrm{HSIC}(\mathbf{Y}, \mathbf{Y})}}, \tag{8}$$

where $\mathrm{HSIC}(\cdot, \cdot)$ is the Hilbert-Schmidt Independence Criterion [18], which measures the dependence between two datasets. A higher CKA score indicates more similar representations between two models at the given layer.