Generative Neuro-Symbolic Models of Concept Learning

Reuben Feinman

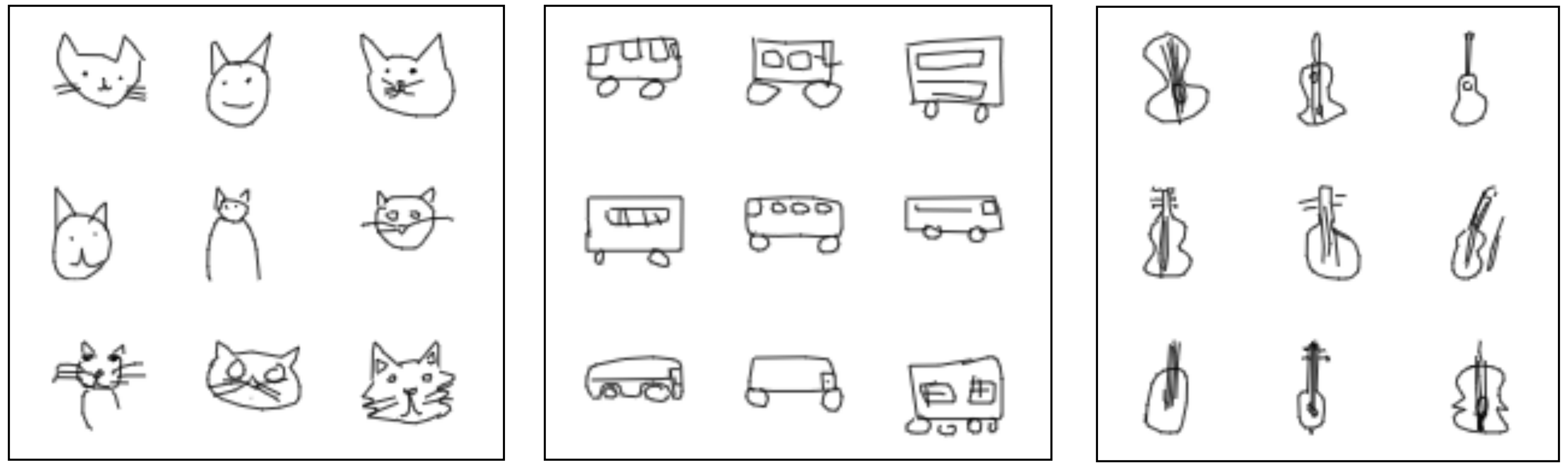
advised by
Brenden Lake

Human concepts are *task-general*

Recognition

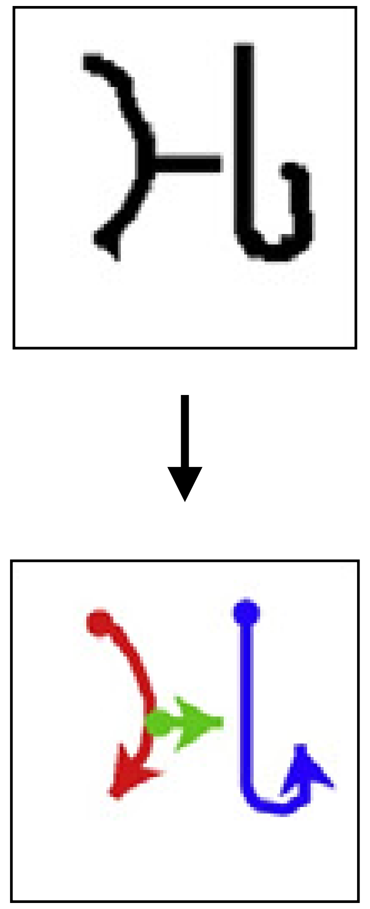
(Lake et al., 2015)

Generation

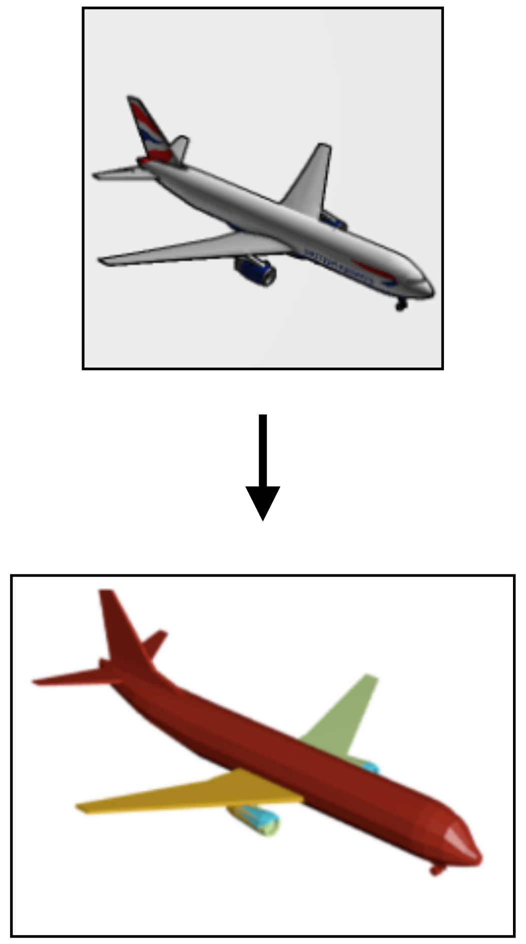


(Jongejan et al., 2016)

Parsing

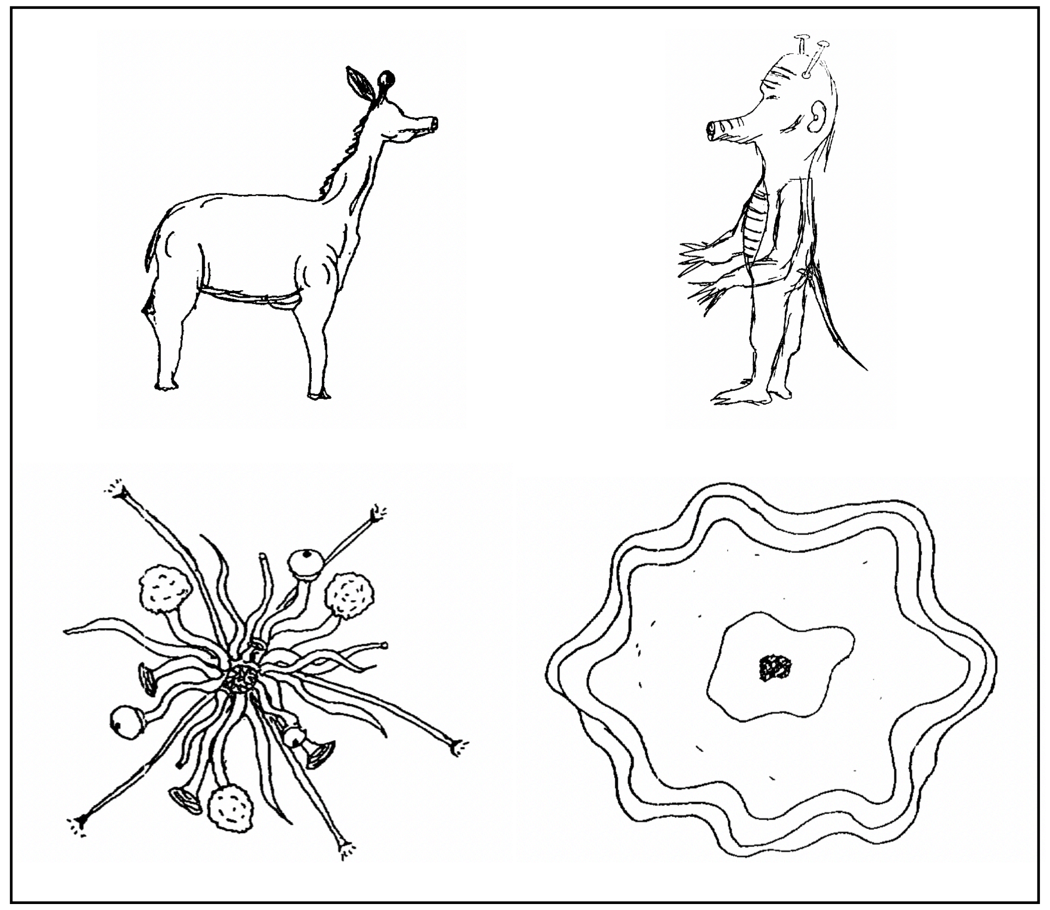


(Lake et al., 2015)



(Mo et al., 2018)

Imagination



(Ward, 1994)

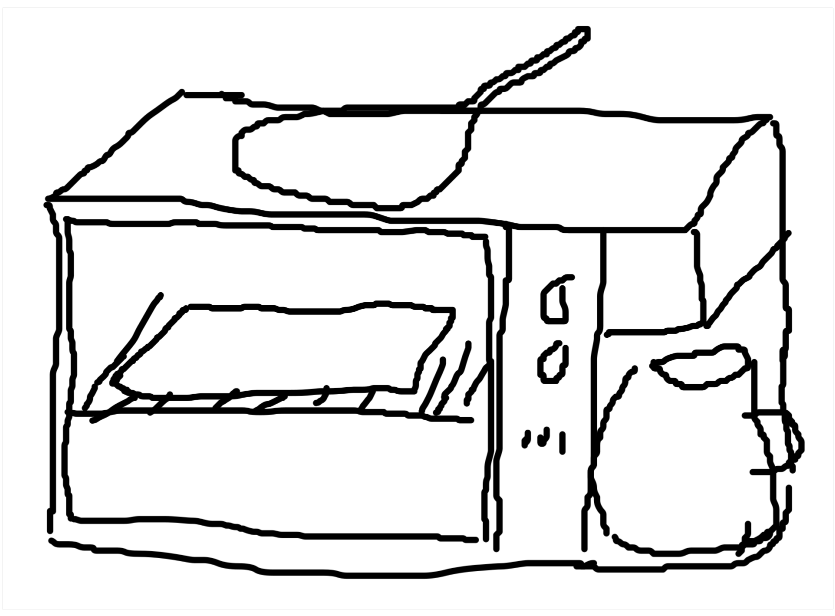
Human concept learning is *fast*

This is a
“breakfast machine.”

Which is another?

What are its parts?

Create a new one.



*example inspired by a blog post from Vicarious AI

Research questions

- What is the structure of human conceptual representations? How does this structure support a variety of discriminative and generative abilities?

- How do people acquire such rich representations from so little experience?

- How can we understand these abilities in computational terms?

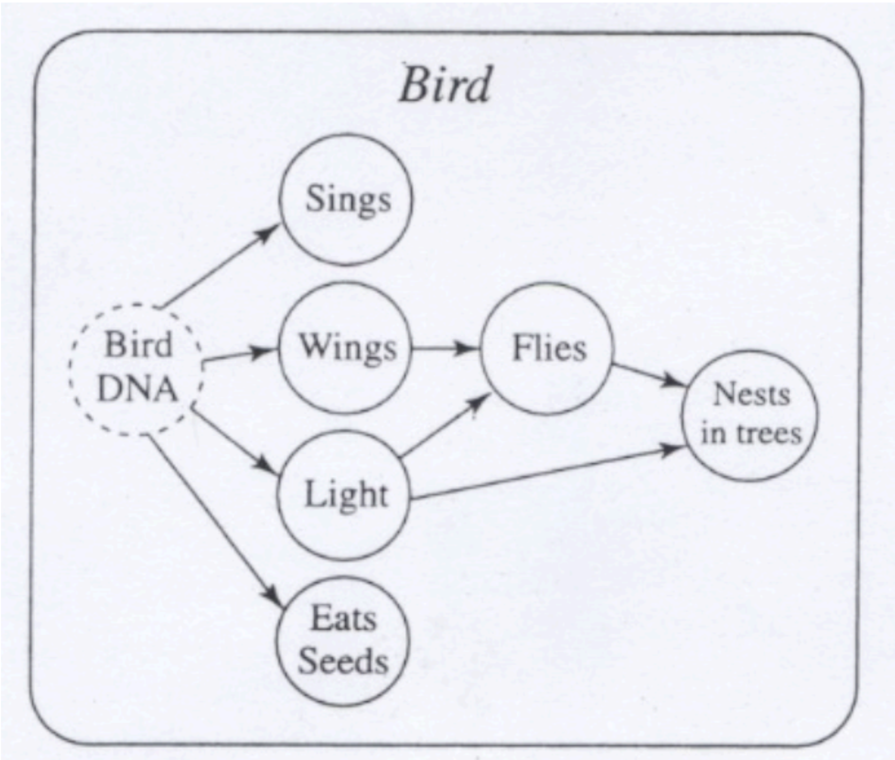
Modeling Traditions

Tradition 1: structured knowledge

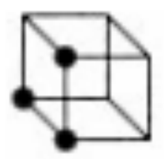
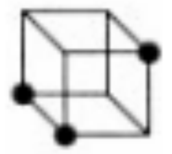
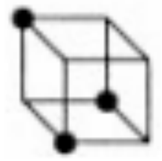
Intuitive theories
(Murphy & Medin, 1985)

The "language of thought"
(Fodor, 1975)

Causal-model theory
(Rehder, 2007)



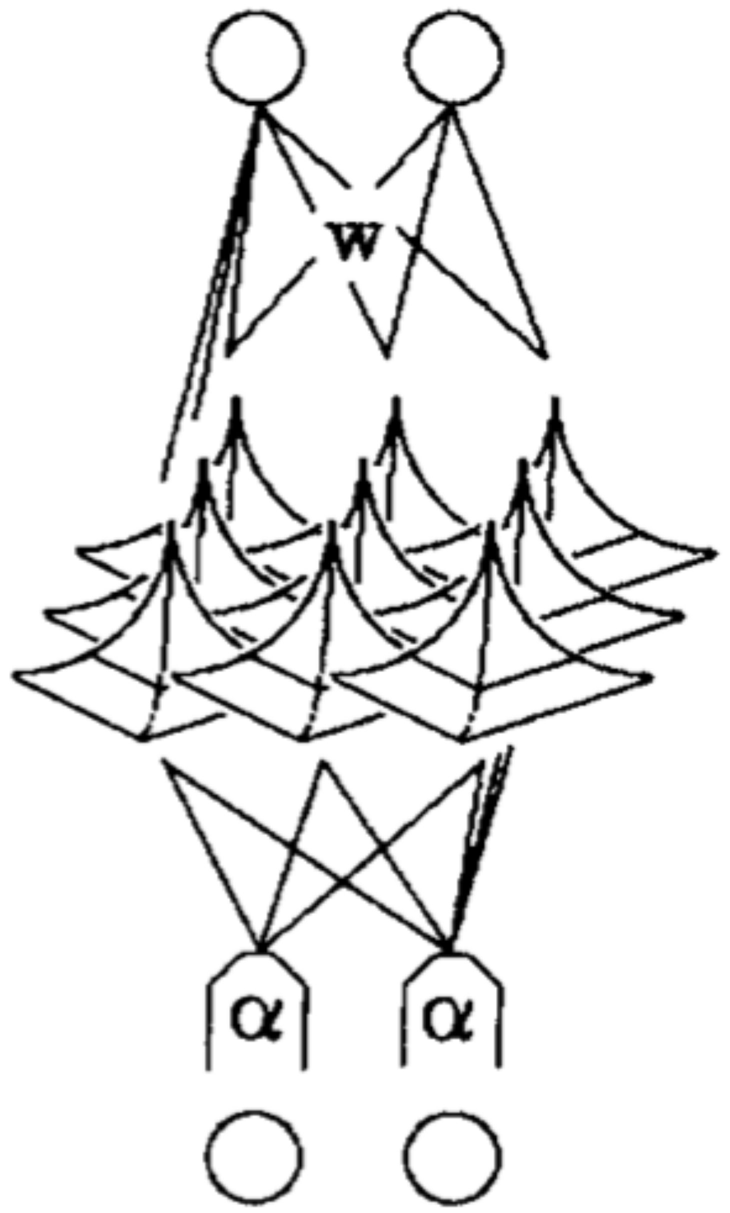
Boolean concepts
(Feldman, 2000)

DNF	Minimal formula	Complexity	Illustration
$a'b'c' + a'b'c + a'bc'$	$a'(bc)'$	3	
$a'b'c' + a'b'c + abc'$	$a'b' + abc'$	5	
$a'b'c' + a'bc + ab'c$	$a'(b'c' + bc) + ab'c$	8	

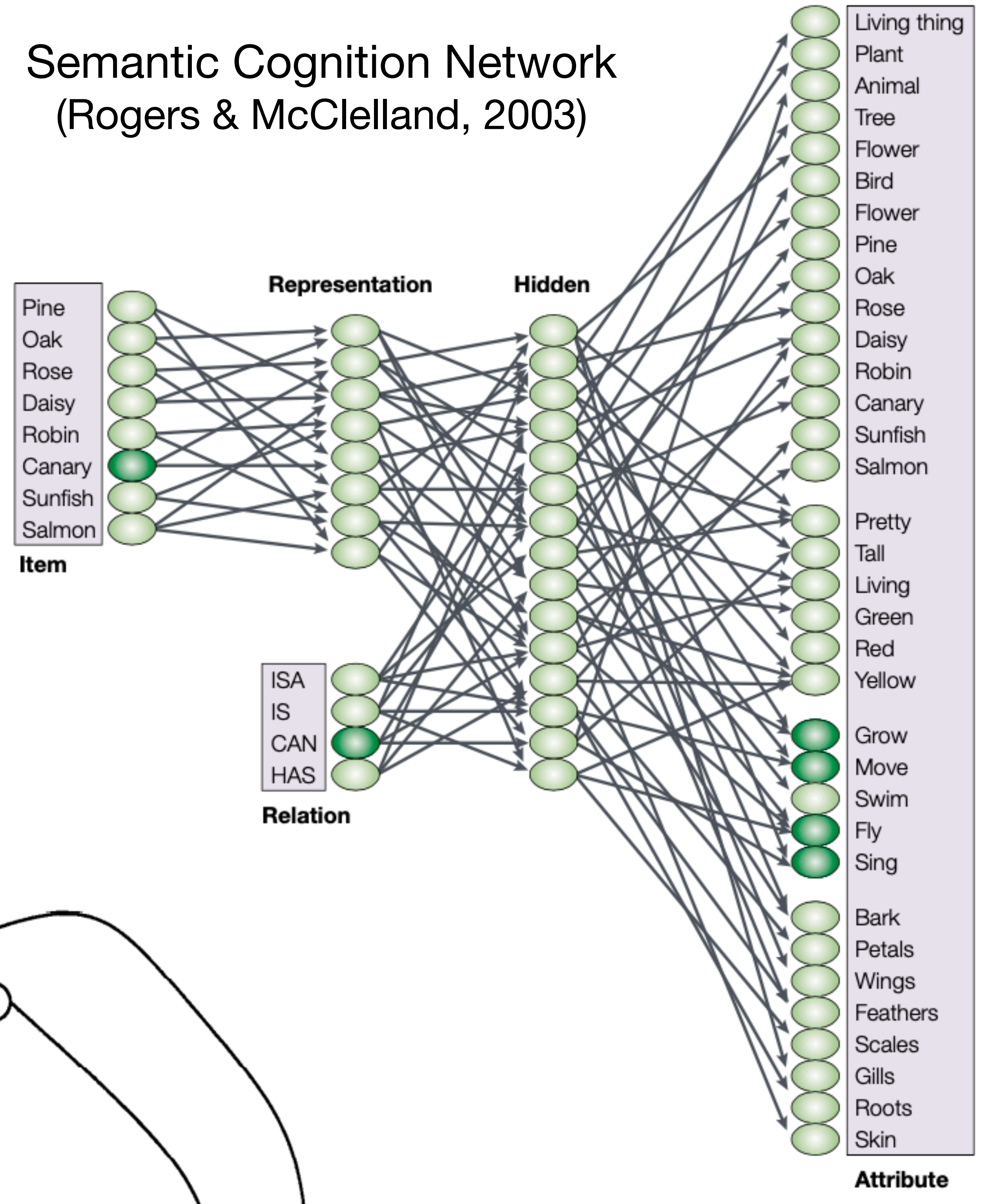
Synthesis?

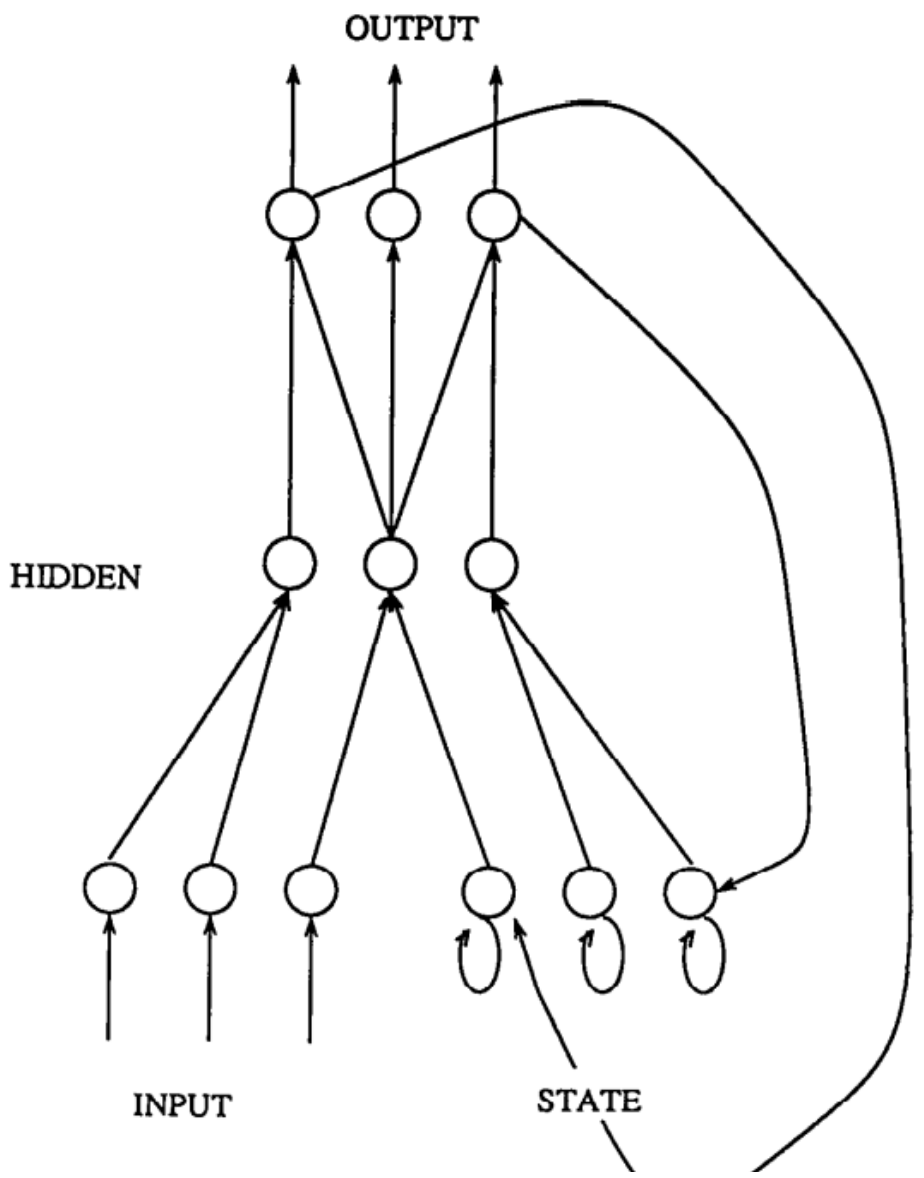
Tradition 2: statistical knowledge

ALCOVE
(Kruschke, 1992)



Semantic Cognition Network
(Rogers & McClelland, 2003)



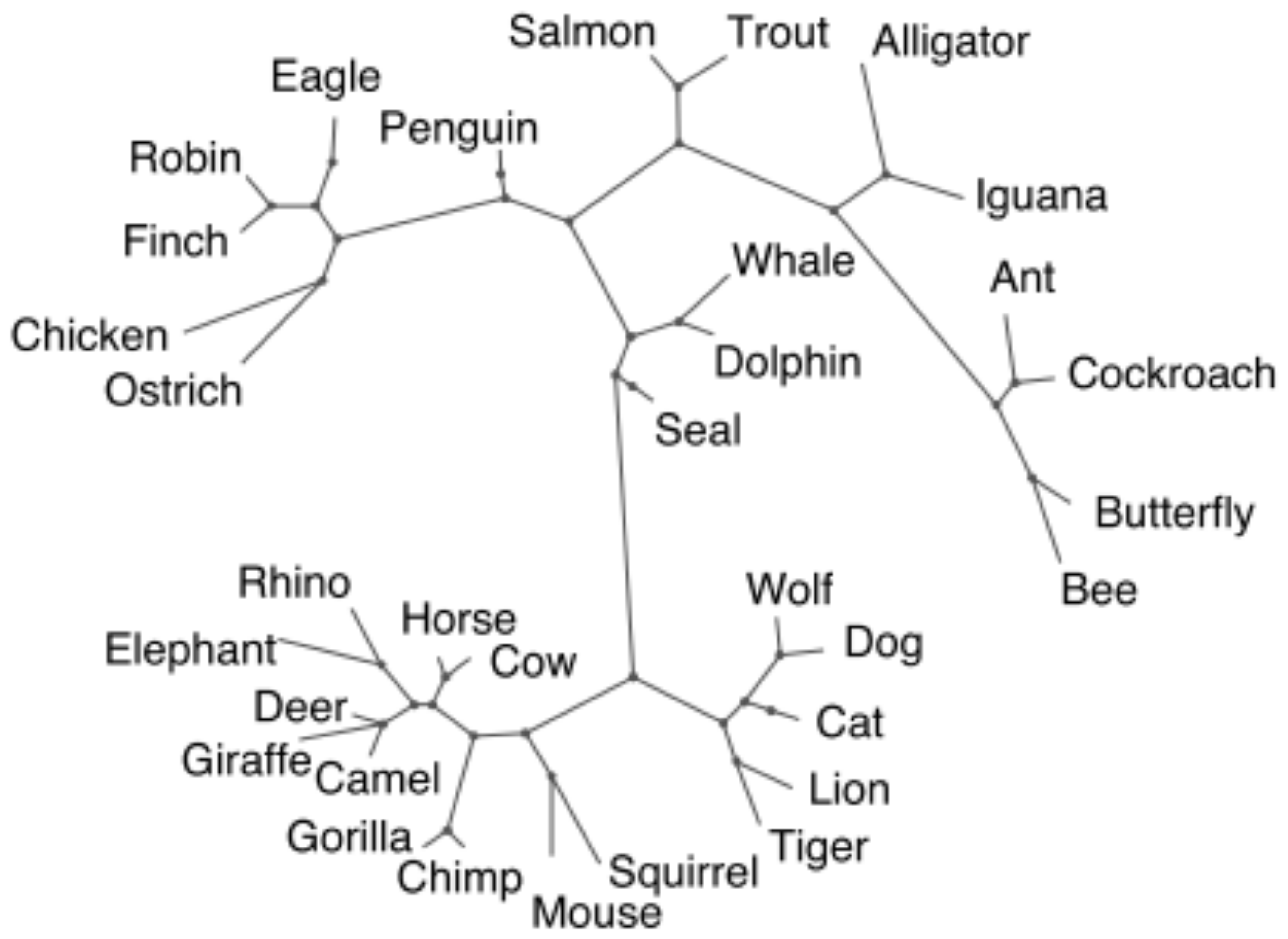


Finding Structure in Time
(Elman, 1990)

Prior work: Integrating structure and statistics

Bayes' rule: $P(\text{structure} \mid \text{data}) \propto P(\text{structure})P(\text{data} \mid \text{structure})$

Structural Forms
(Kemp & Tenenbaum, 2008)



Bayesian Program Learning
(Lake et al., 2015)

procedure GENERATETYPE
- $\kappa \leftarrow P(\kappa)$ ▷ Sample number of parts
- **for** i = 1 ... κ **do**
  - $n_i \leftarrow P(n_i|\kappa)$ ▷ Sample number of sub-parts
  - **for** j = 1 ... n_i **do**
    - $s_{ij} \leftarrow P(s_{ij}|s_{i(j-1)})$ ▷ Sample sub-part sequence
  - **end for**
  - $R_i \leftarrow P(R_i|S_1, ..., S_{i-1})$ ▷ Sample relation
- **end for**
- $\psi \leftarrow \{\kappa, R, S\}$
- **return** @GENERATETOKEN(ψ) ▷ Return program

Rational Rules
(Goodman et al., 2008)

S

$S1:$ $\forall x\ \ell(x) \Leftrightarrow (D)$

$D1:$ $\forall x\ \ell(x) \Leftrightarrow ((C) \vee D)$

$D2:$ $\forall x\ \ell(x) \Leftrightarrow ((C) \vee \text{False})$

$C1:$ $\forall x\ \ell(x) \Leftrightarrow ((P \wedge C) \vee \text{False})$

$C2:$ $\forall x\ \ell(x) \Leftrightarrow ((P \wedge \text{True}) \vee \text{False})$

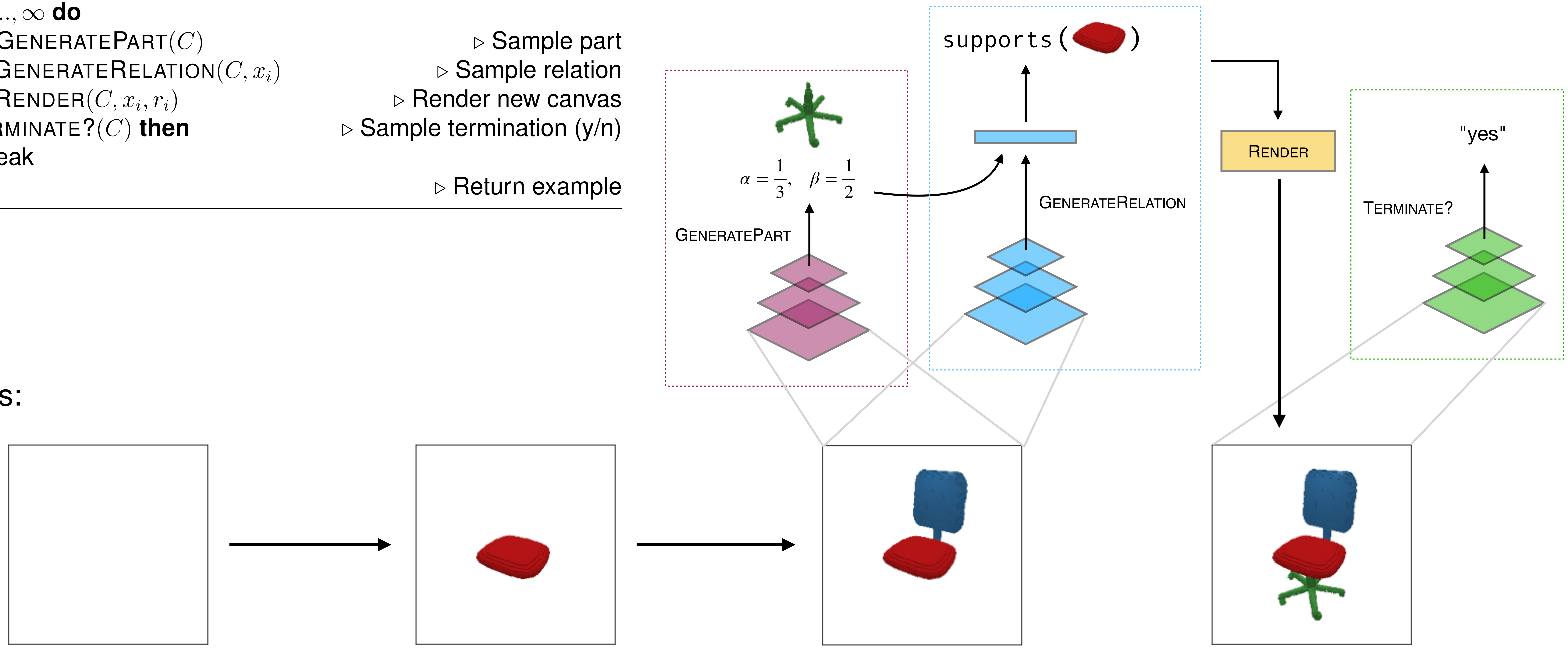
$P1:$ $\forall x\ \ell(x) \Leftrightarrow ((F_1 \wedge \text{True}) \vee \text{False})$

$F_1 2:$ $\forall x\ \ell(x) \Leftrightarrow (((f_1(x)=0) \wedge \text{True}) \vee \text{False})$

Proposal: Generative Neuro-Symbolic (GNS) modeling

procedure GENERATEEXAMPLE
  $C \leftarrow 0$ ▷ Initialize blank canvas
  for $i = 1..., \infty$ **do**
    $x_i \leftarrow$ GENERATEPART(C) ▷ Sample part
    $r_i \leftarrow$ GENERATERELATION(C, x_i) ▷ Sample relation
    $C \leftarrow$ RENDER(C, x_i, r_i) ▷ Render new canvas
    if TERMINATE?(C) **then** ▷ Sample termination (y/n)
      break
  return C ▷ Return example

Canvas:
C

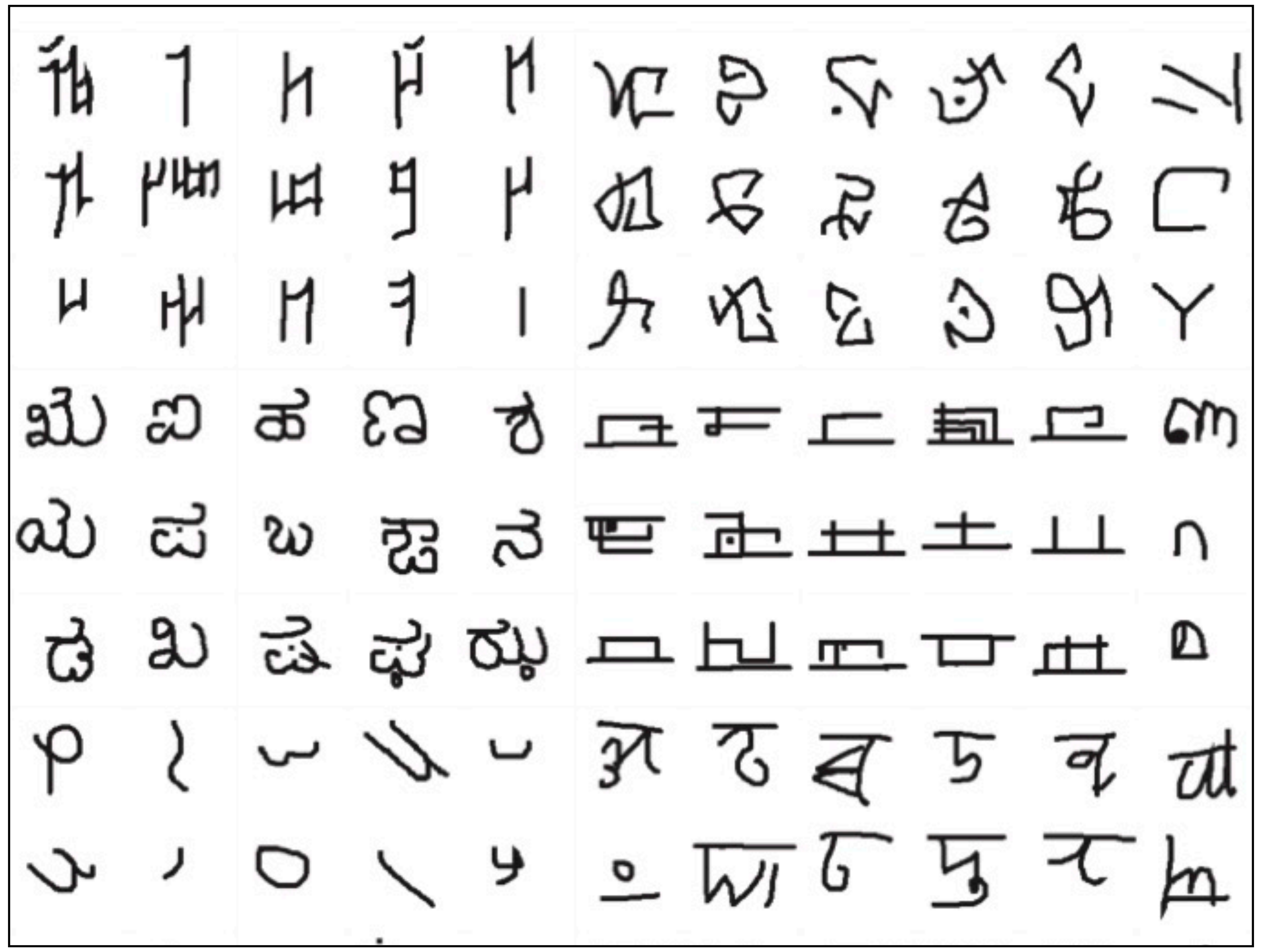


New example

Agenda

- Case study #1: handwritten characters
- Case study #2: structured visual concepts ("alien figures")
- Additional projects
- Summary & conclusions

Case study #1: handwritten characters

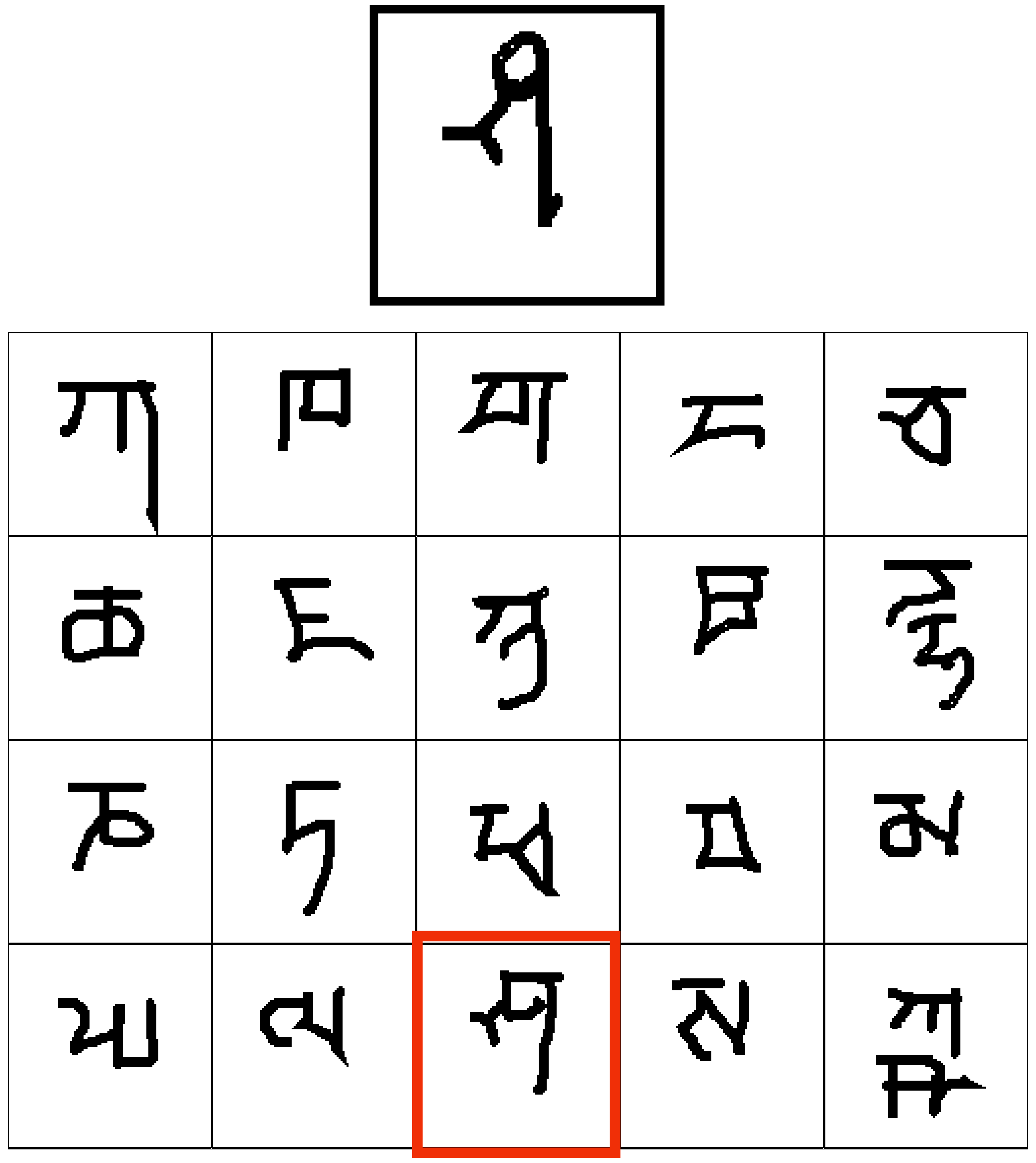


(Lake et al., 2015)

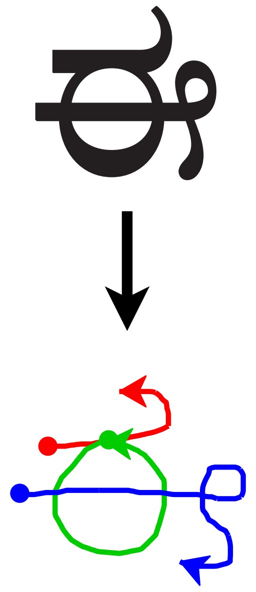
The Omniglot Challenge

(Lake et al., 2015)

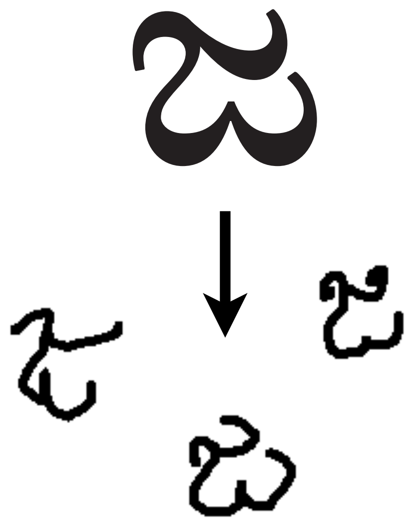
classification



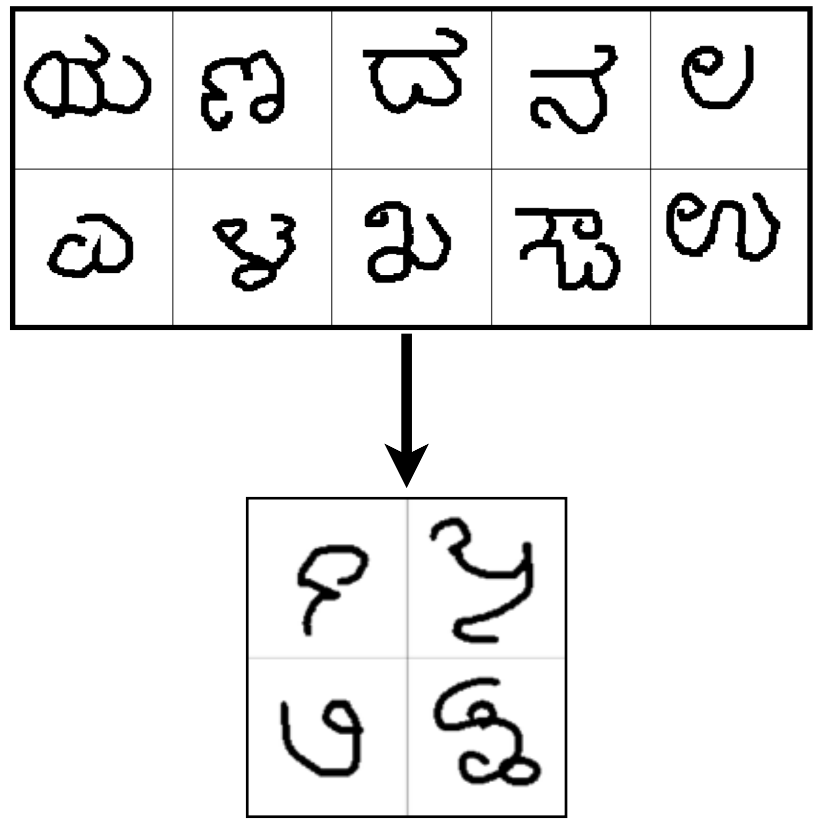
parsing



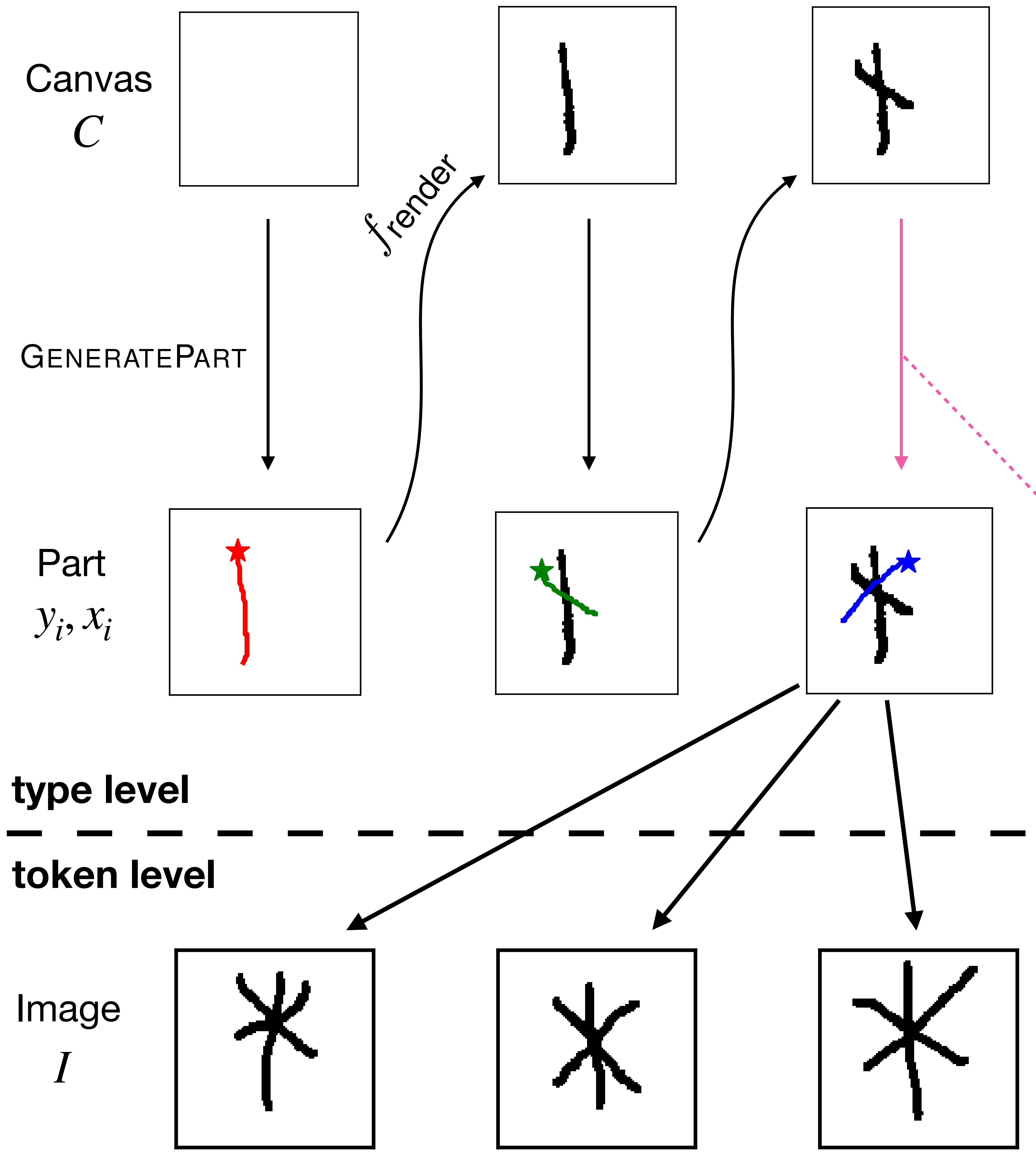
generating
new examples



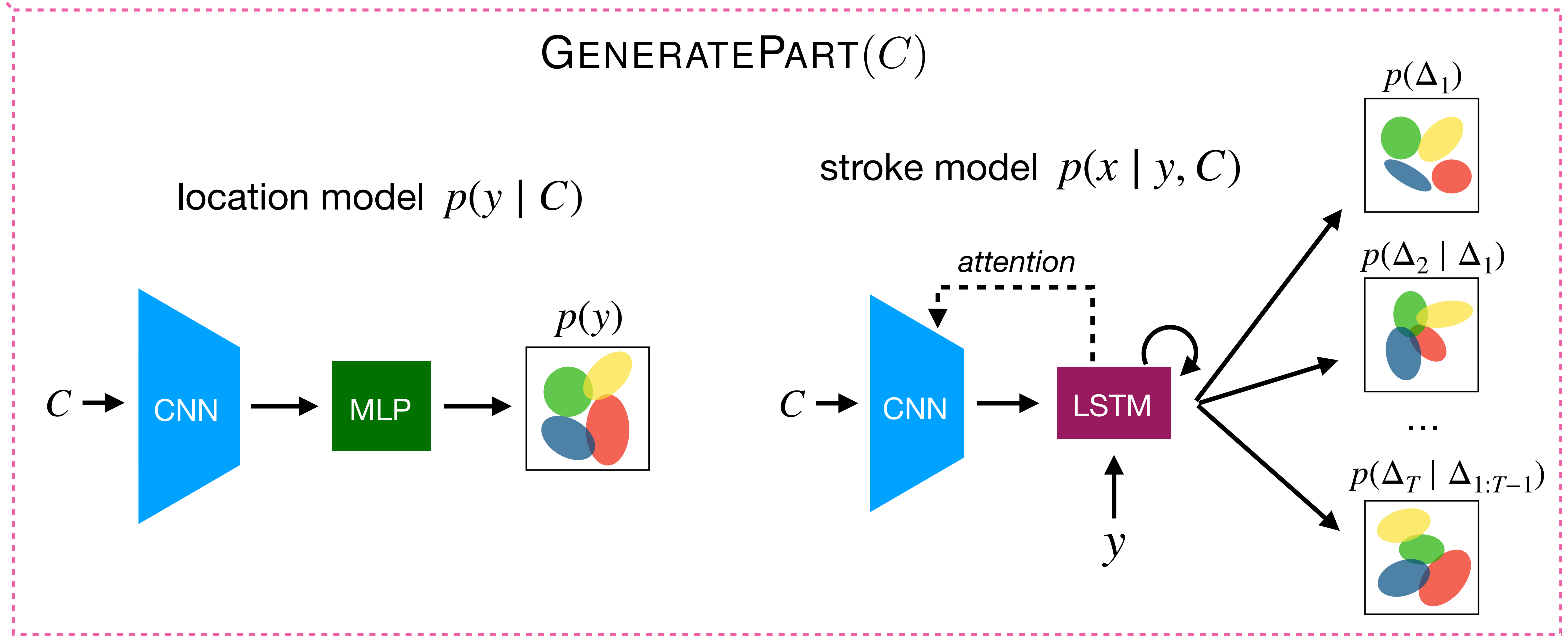
generating
new concepts



GNS model of character concepts

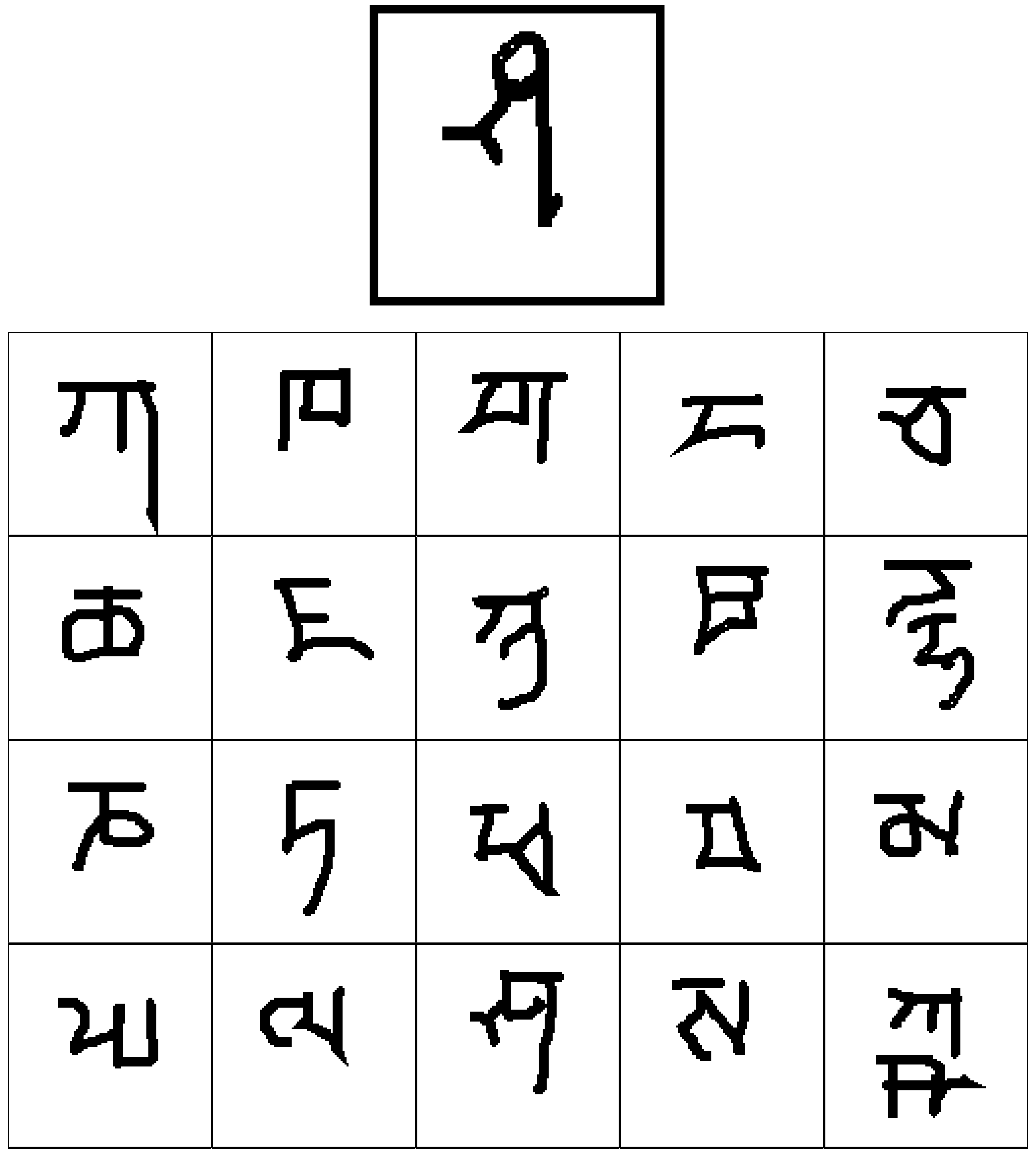


procedure GENERATETYPE
- $C \leftarrow 0$ ▷ Initialize blank image canvas
- **while** *true* **do**
  - $[y_i, x_i] \leftarrow$ GENERATEPART(C) ▷ Sample part location & parameters
  - $C \leftarrow f_{\text{render}}(y_i, x_i, C)$ ▷ Render part to image canvas
  - $v_i \sim p(v \mid C)$ ▷ Sample termination indicator
  - **if** v_i **then**
    - break ▷ Terminate sample
- $\psi \leftarrow \{\kappa, y_{1:\kappa}, x_{1:\kappa}\}$
- **return** ψ ▷ Return concept type

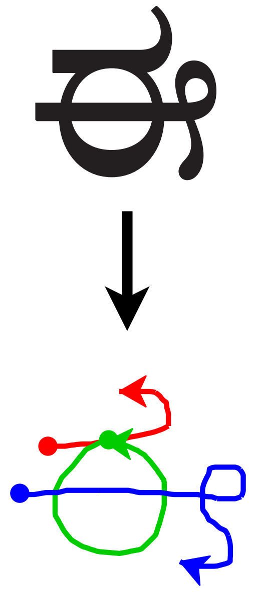


The Omniglot Challenge

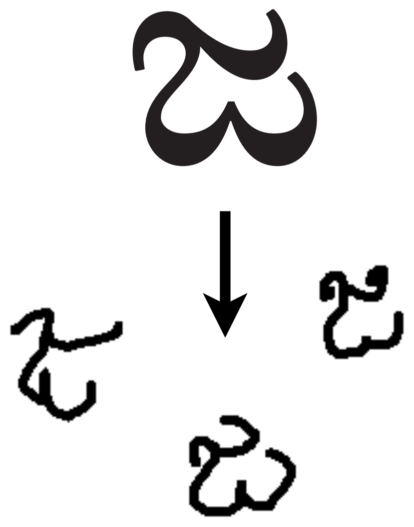
classification



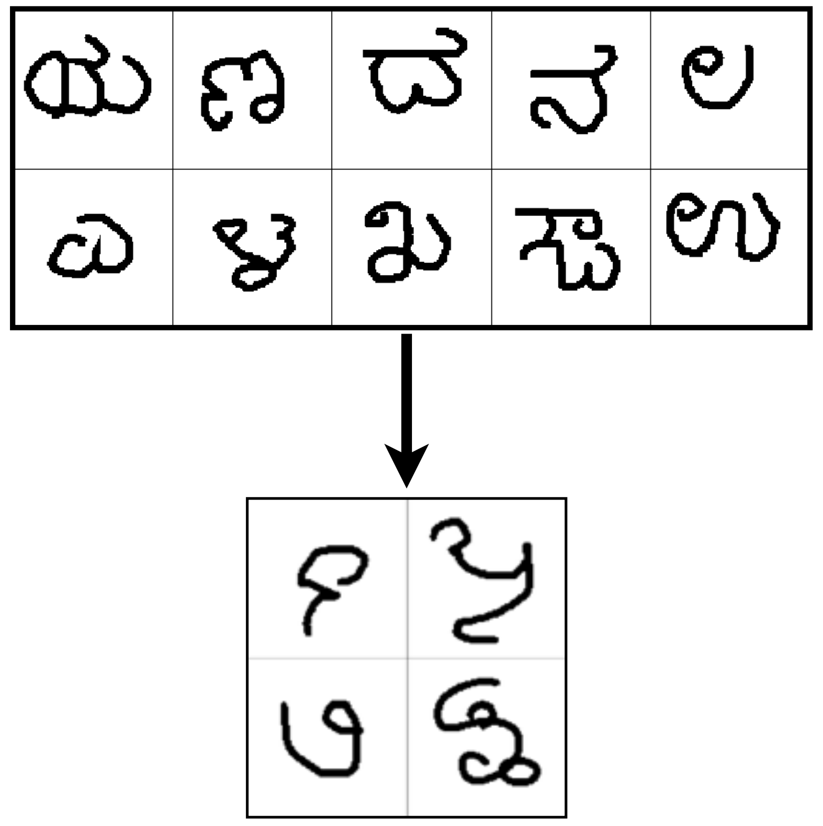
parsing



generating
new examples



generating
new concepts



procedure GENERATETYPE

Generating new concepts

1. Log-likelihoods (LL) of held-out concepts

Test loss per drawing trajectory

GNS	**-19.51**
H-LSTM	-20.16
LSTM	-19.66

Replicates across different train/test splits

Approximate test LL per pixel image

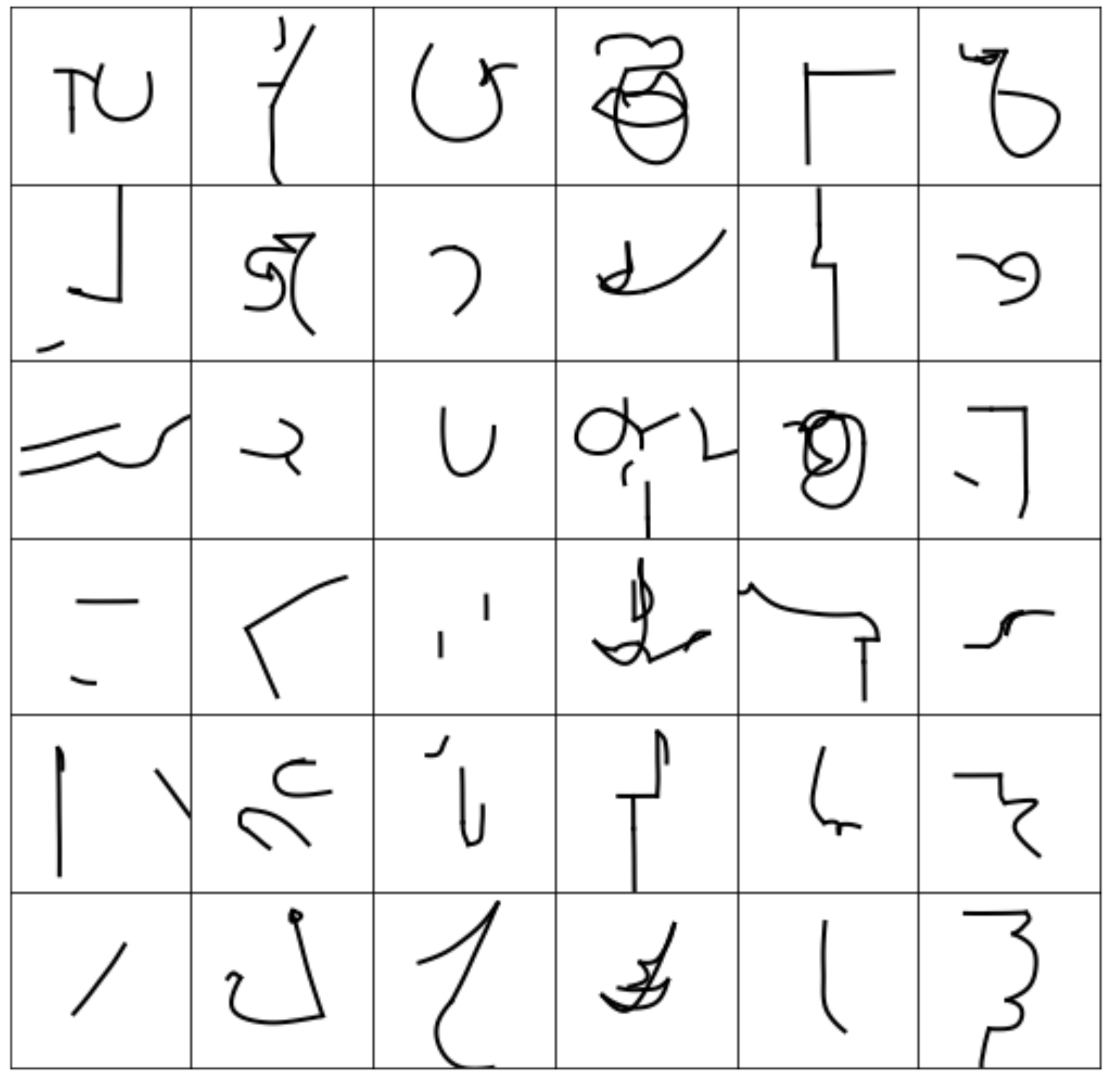
GNS	**-383.67**
VHE	-546.84
SG	-861.05

2. Model samples

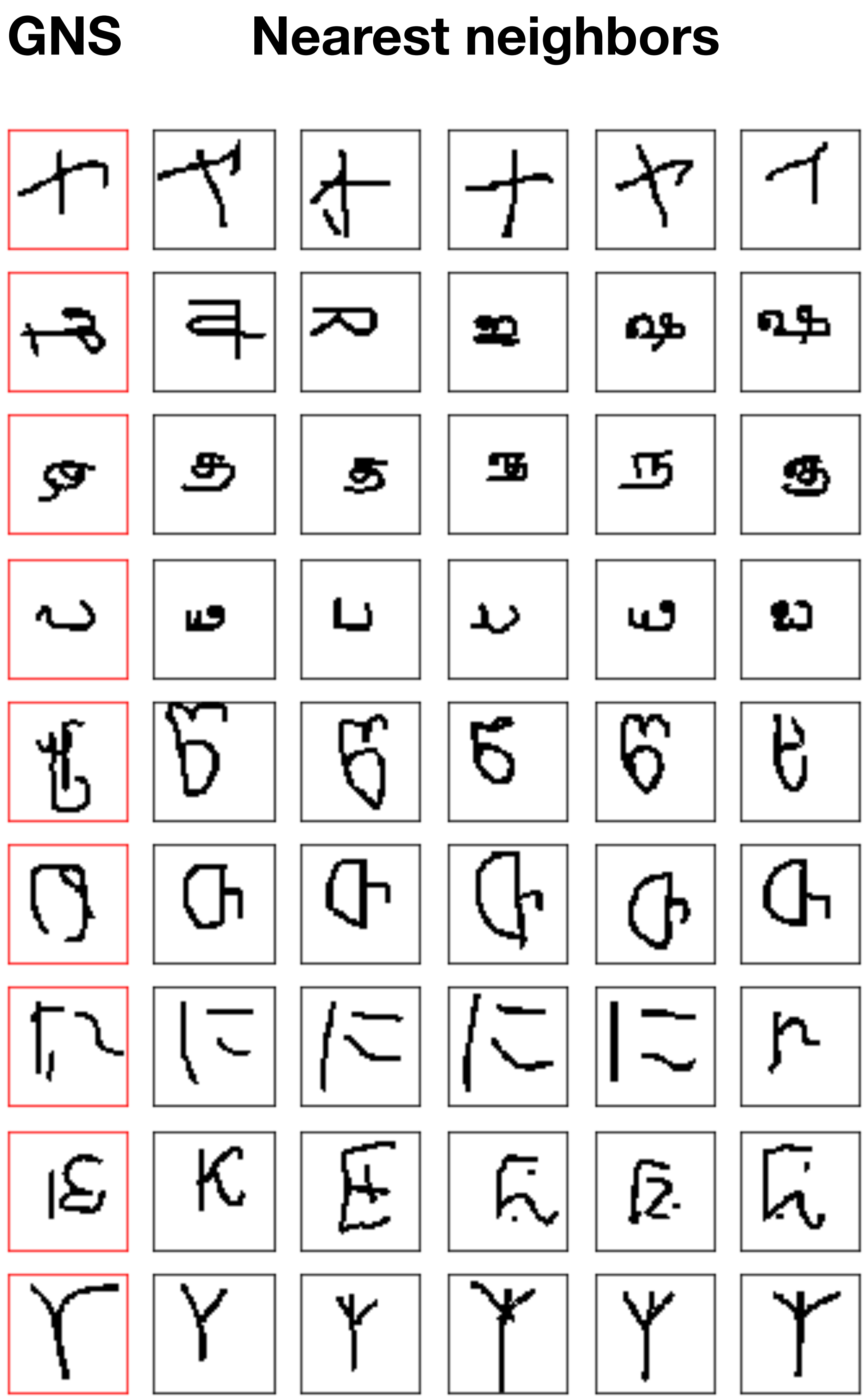
Omniglot

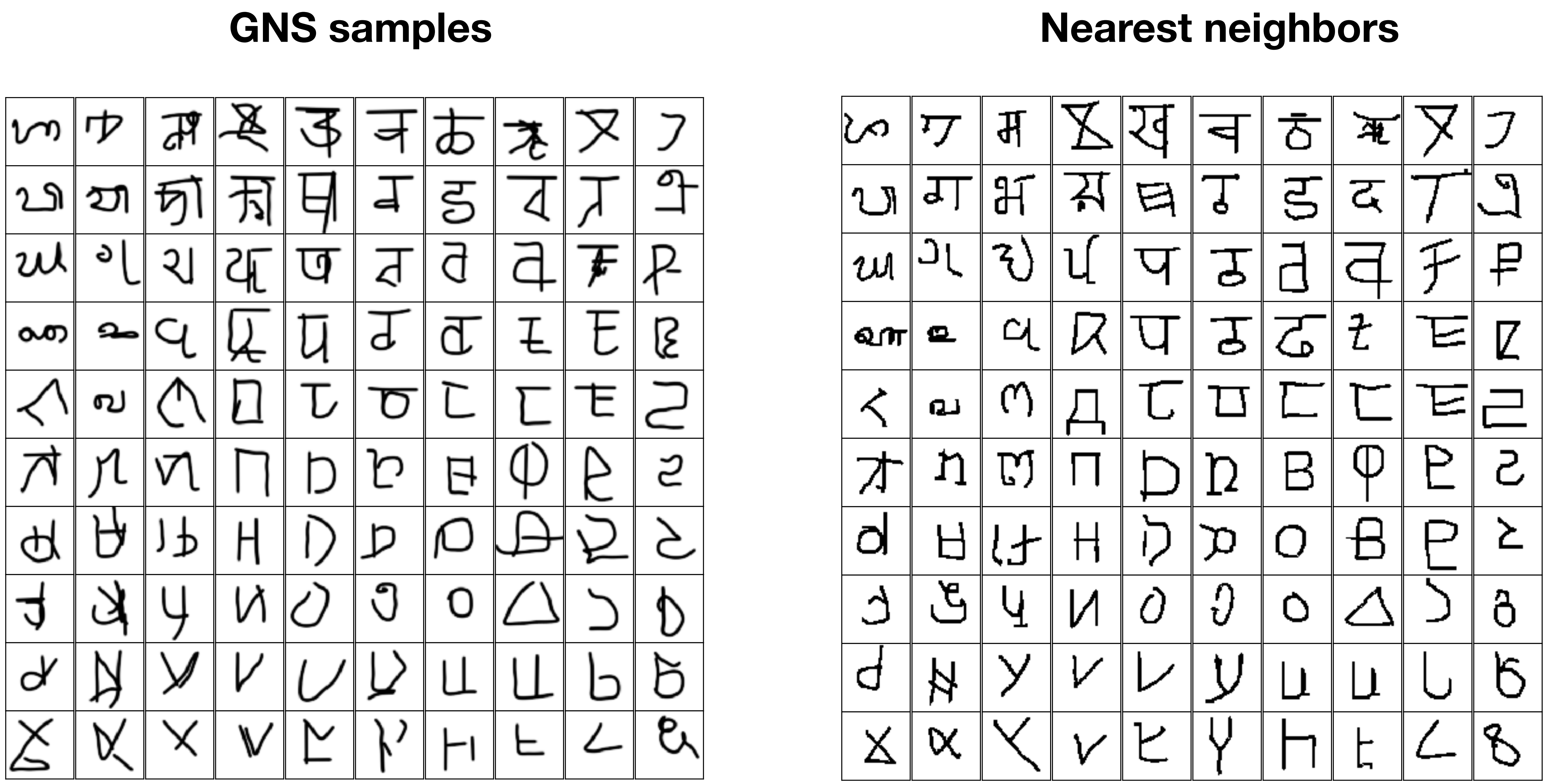
GNS model

fully-symbolic model (BPL)



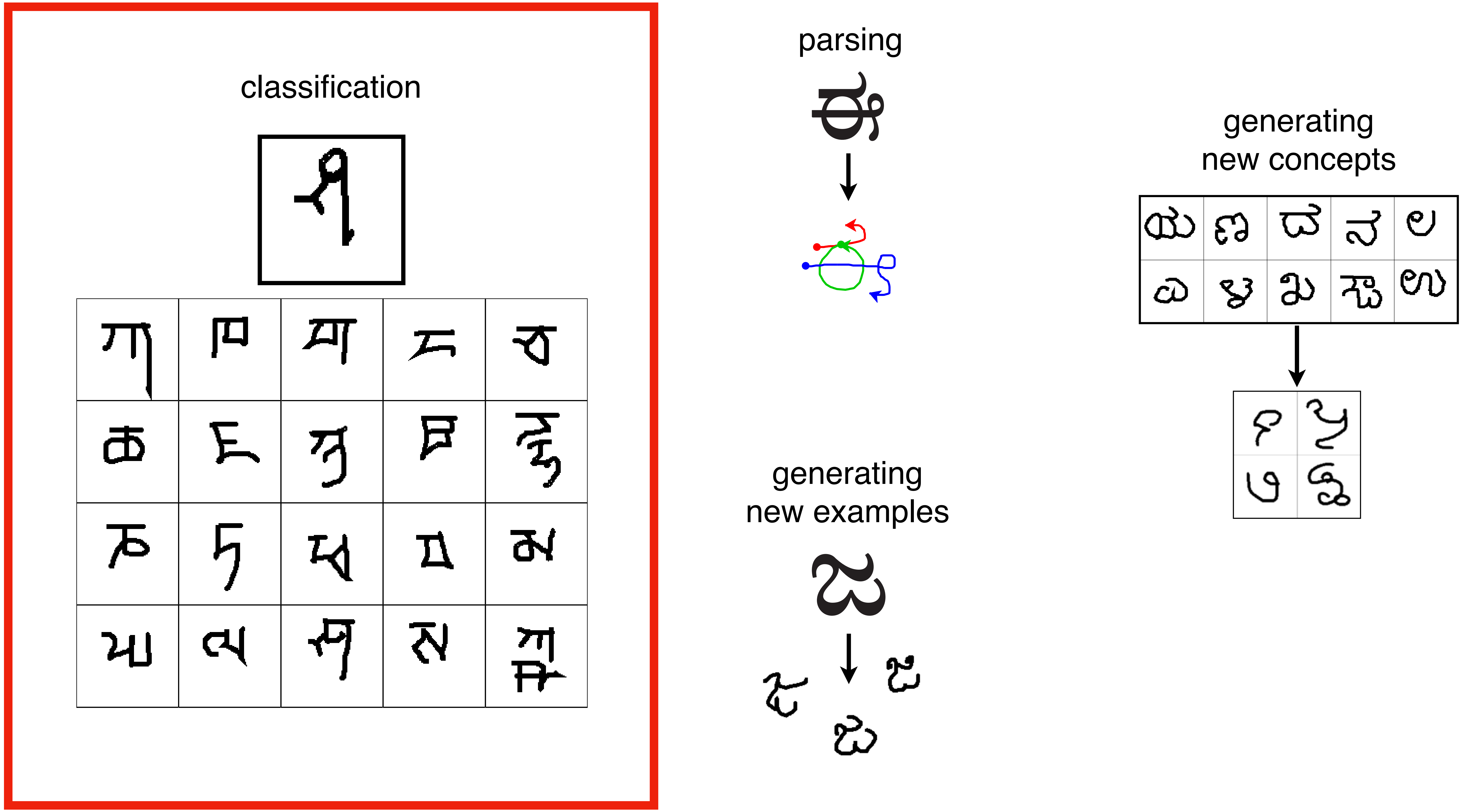
Generating new concepts



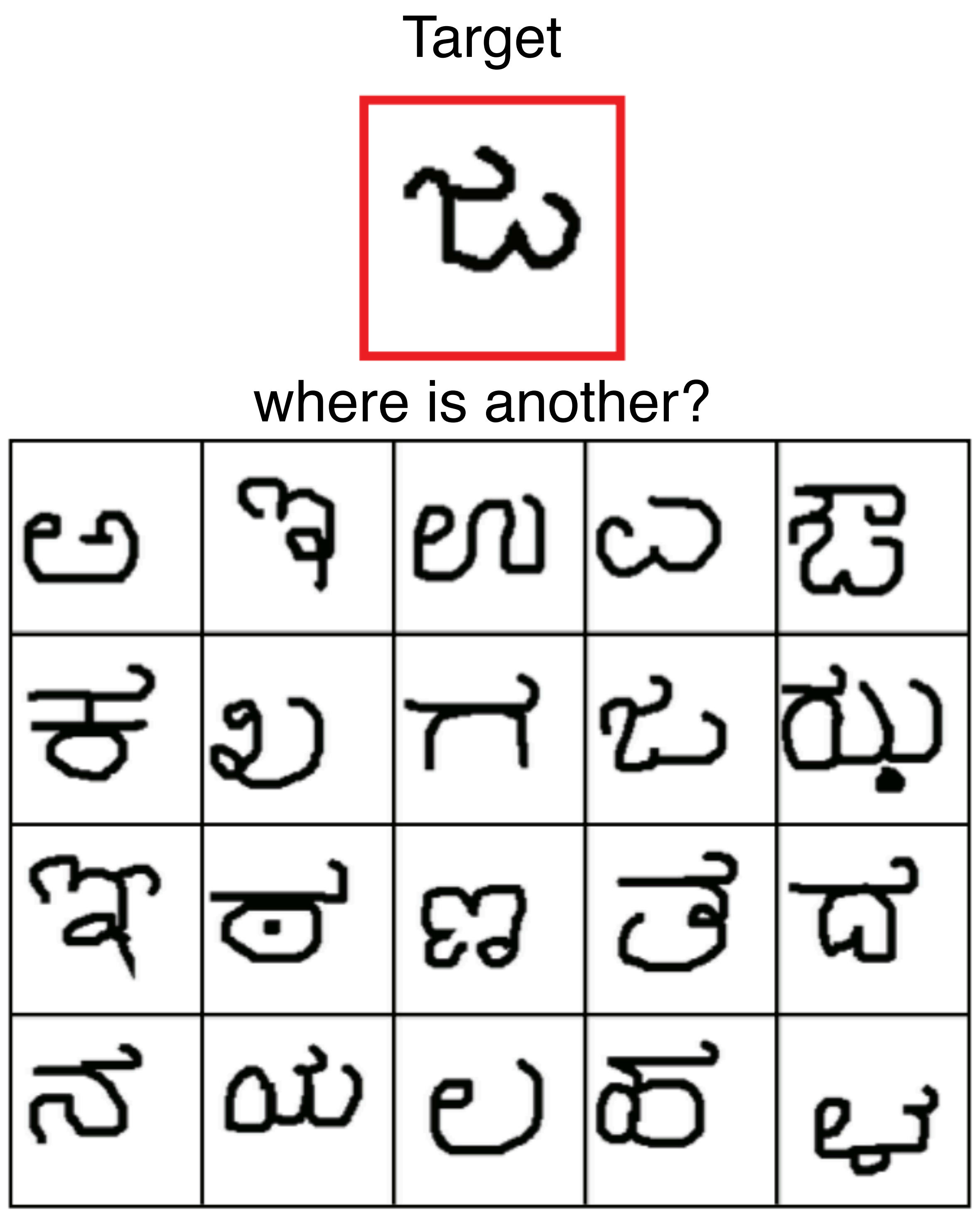


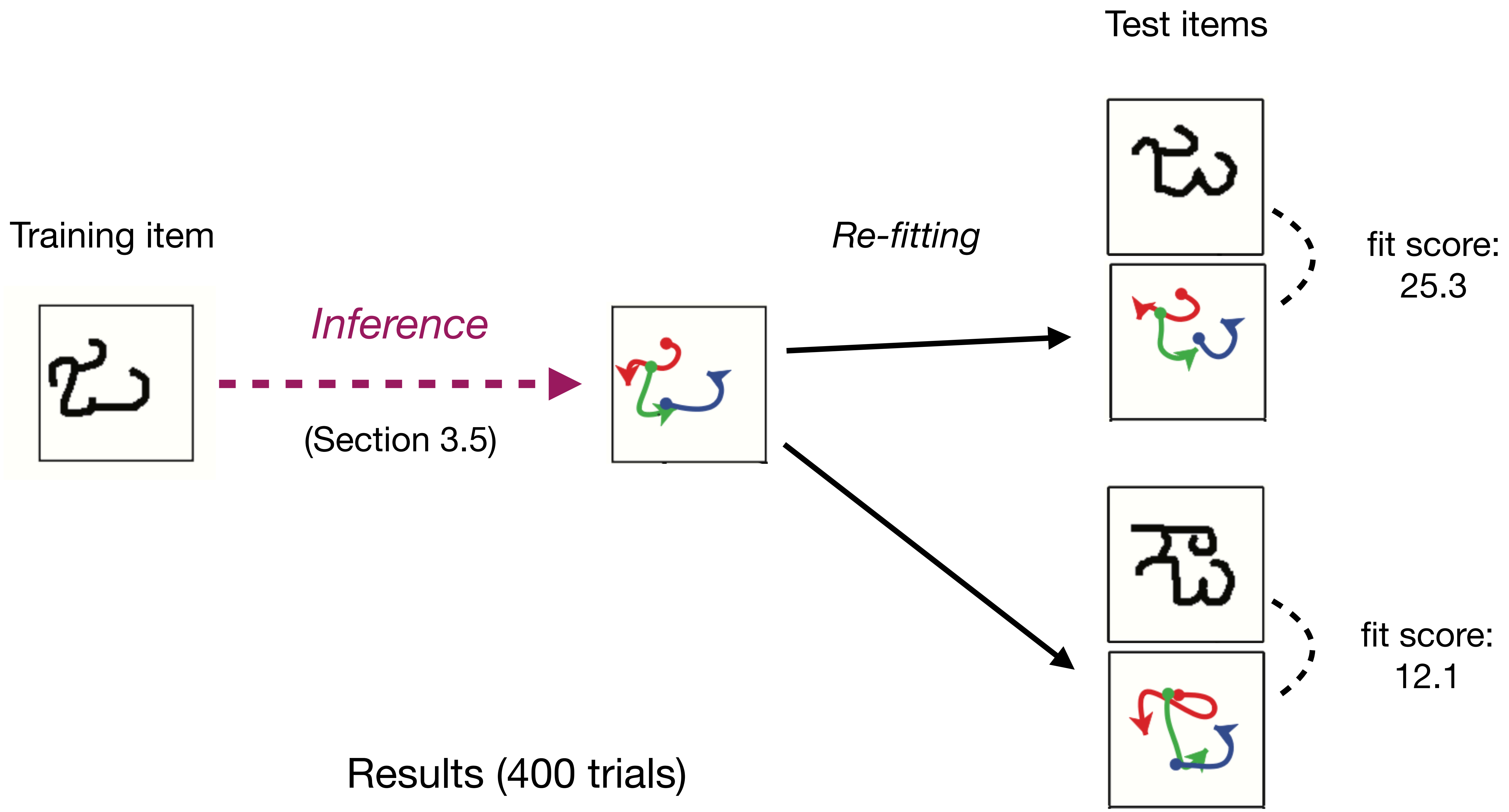
Nearest neighbors are located using the embedding of a convolutional neural network (CNN)

The Omniglot Challenge



One-Shot Classification

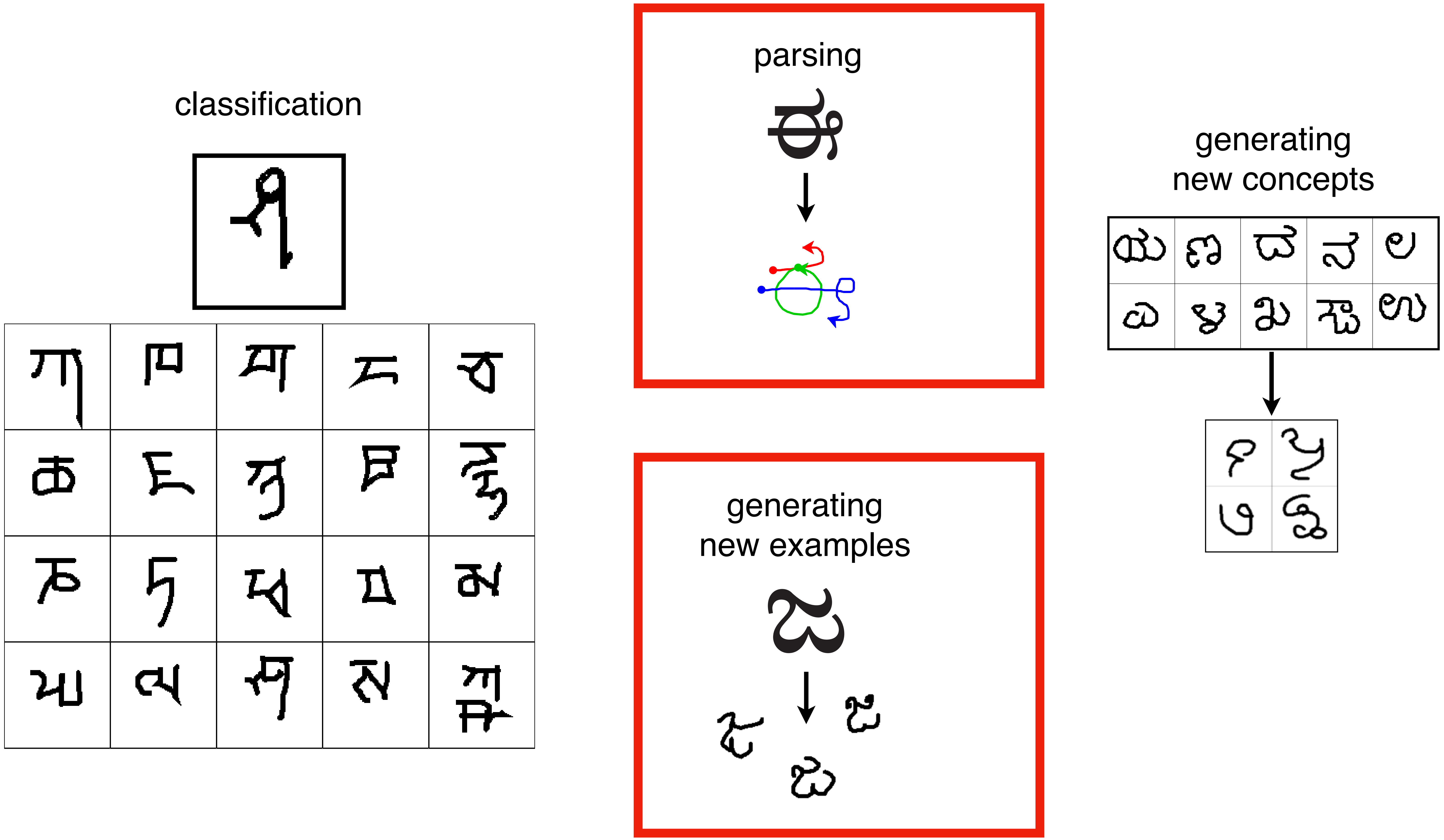




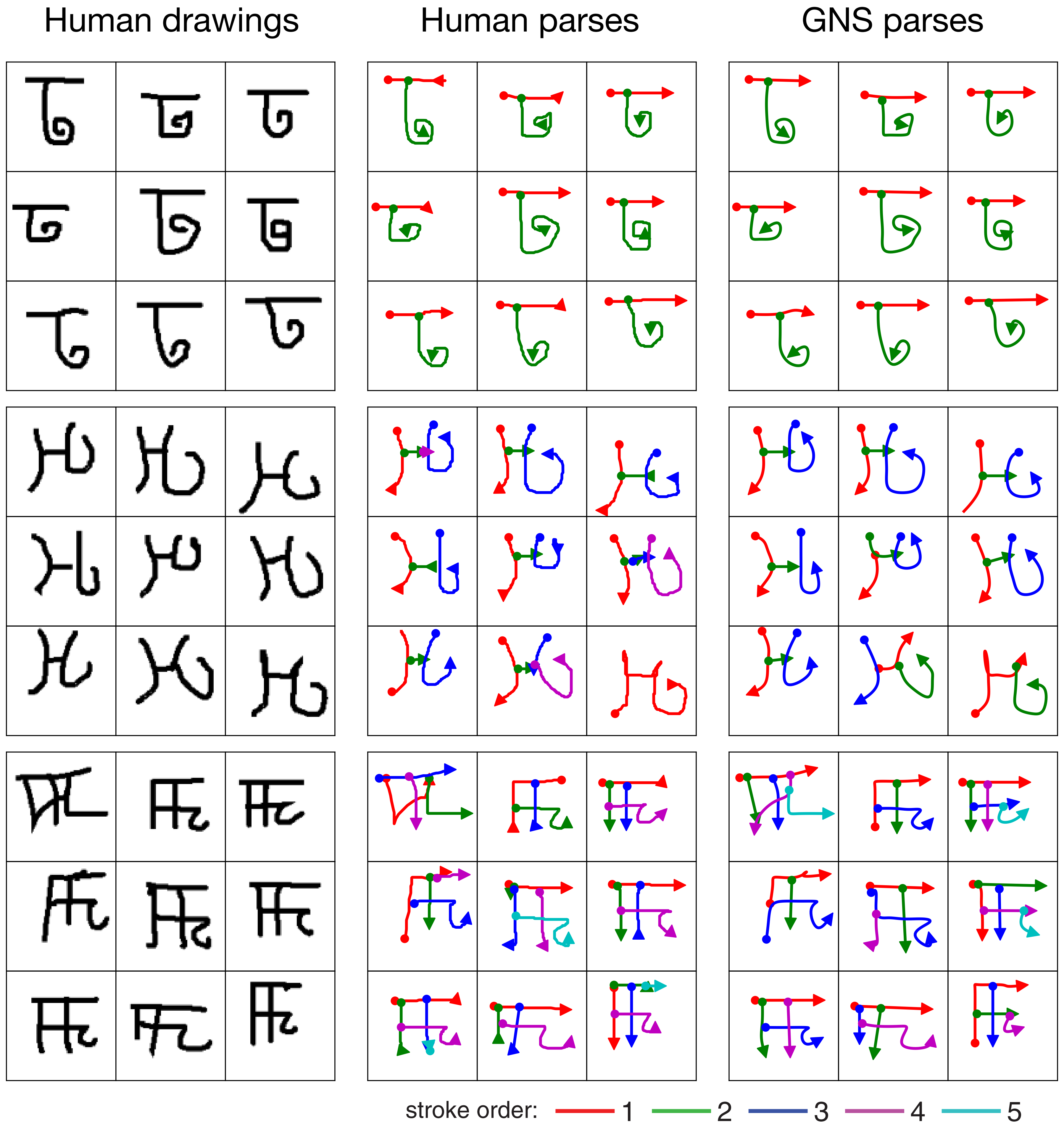
Results (400 trials)

	Accuracy
Humans	95.5%
GNS	94.3%
BPL	96.7%

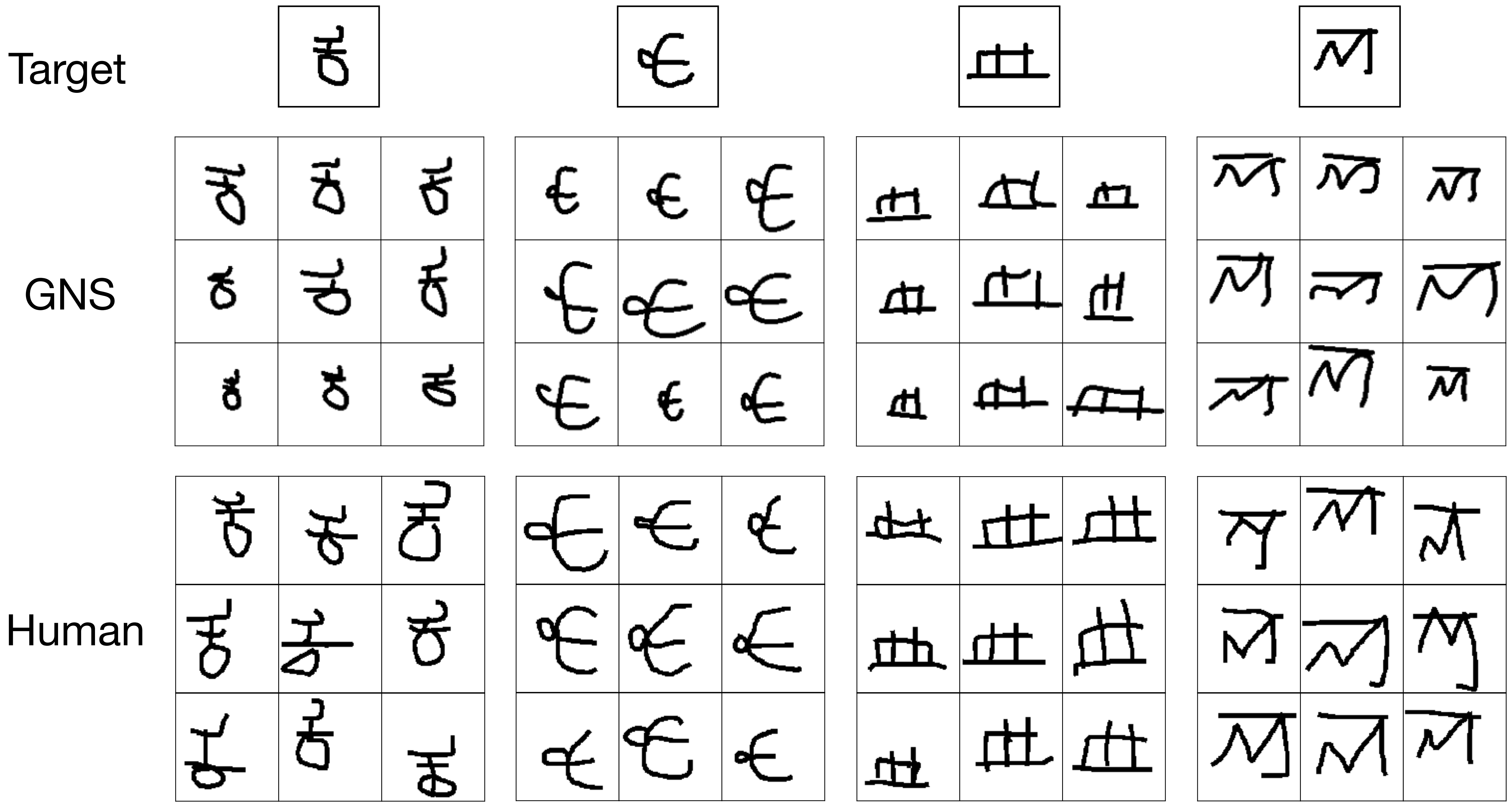
The Omniglot Challenge



Parsing



Generating new exemplars



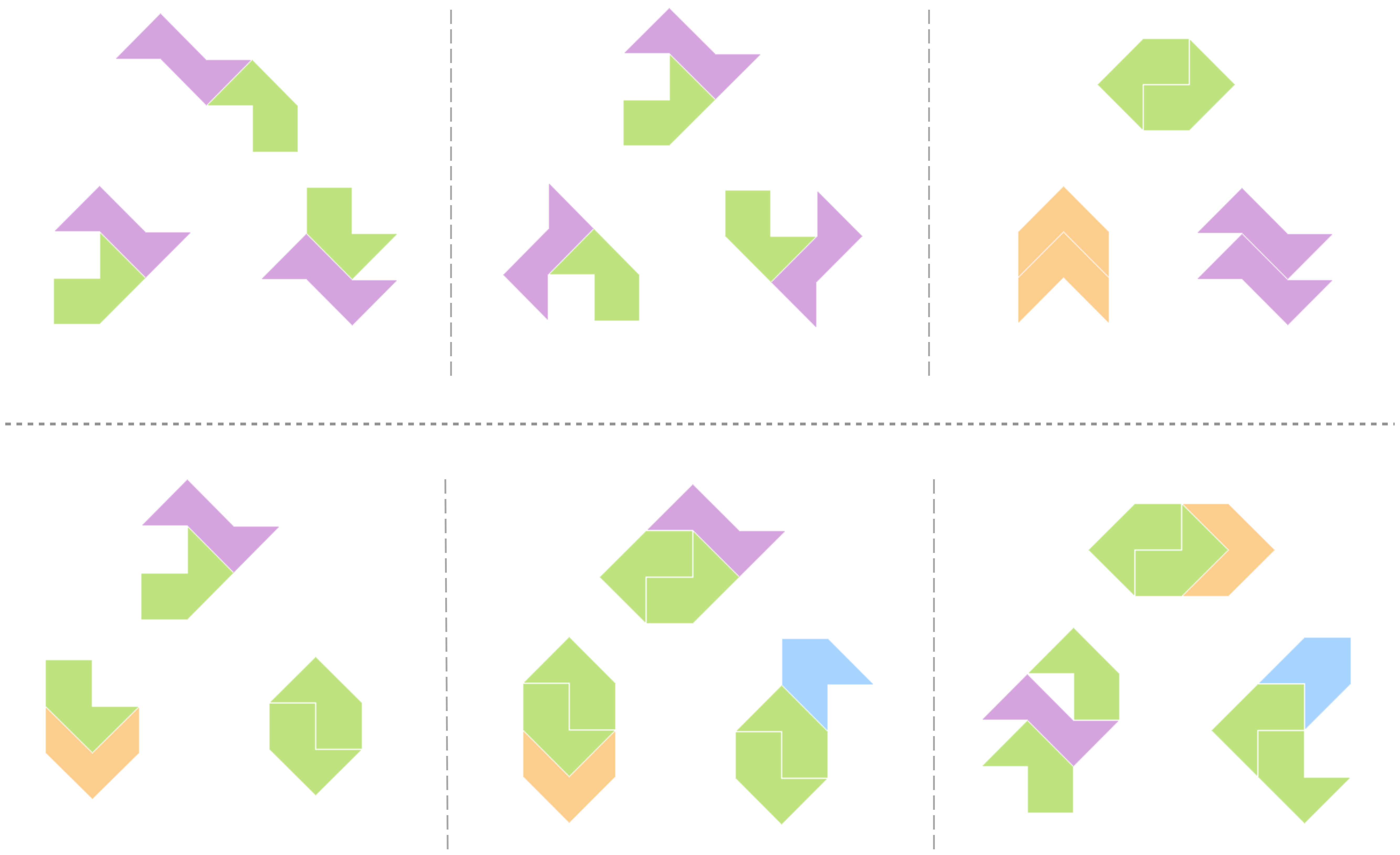
Many more examples are provided in Section 3.6 and Appendix A.5

Conclusions: Case study #1

- Humans quickly grasp new concepts and use them in a variety of ways
- Generative Neuro-Symbolic (GNS) models capture the dual structural and statistical components of character concepts and generalize to novel alphabets and a range of tasks
- GNS models offer an account for how previous experience can support the rapid acquisition of new concepts via priors

Case study #2:
structured visual concepts
("alien figures")

Alien figures



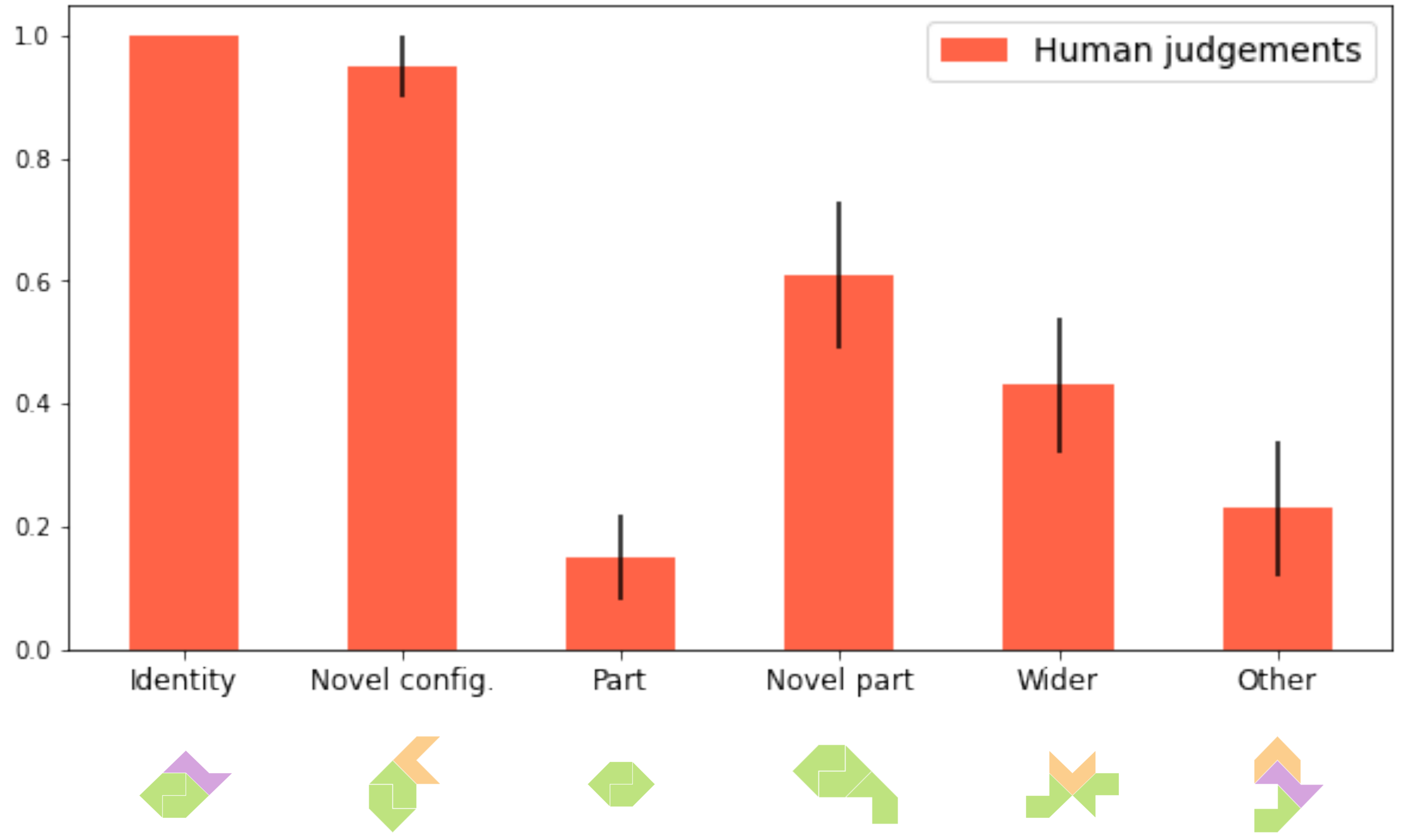
Yanli Zhou

Human experiments

Categorization

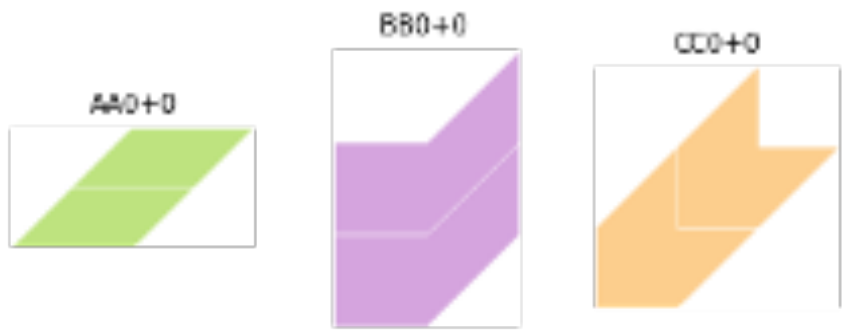
Here are 3 examples of a "wif":

Is this also a "wif?"

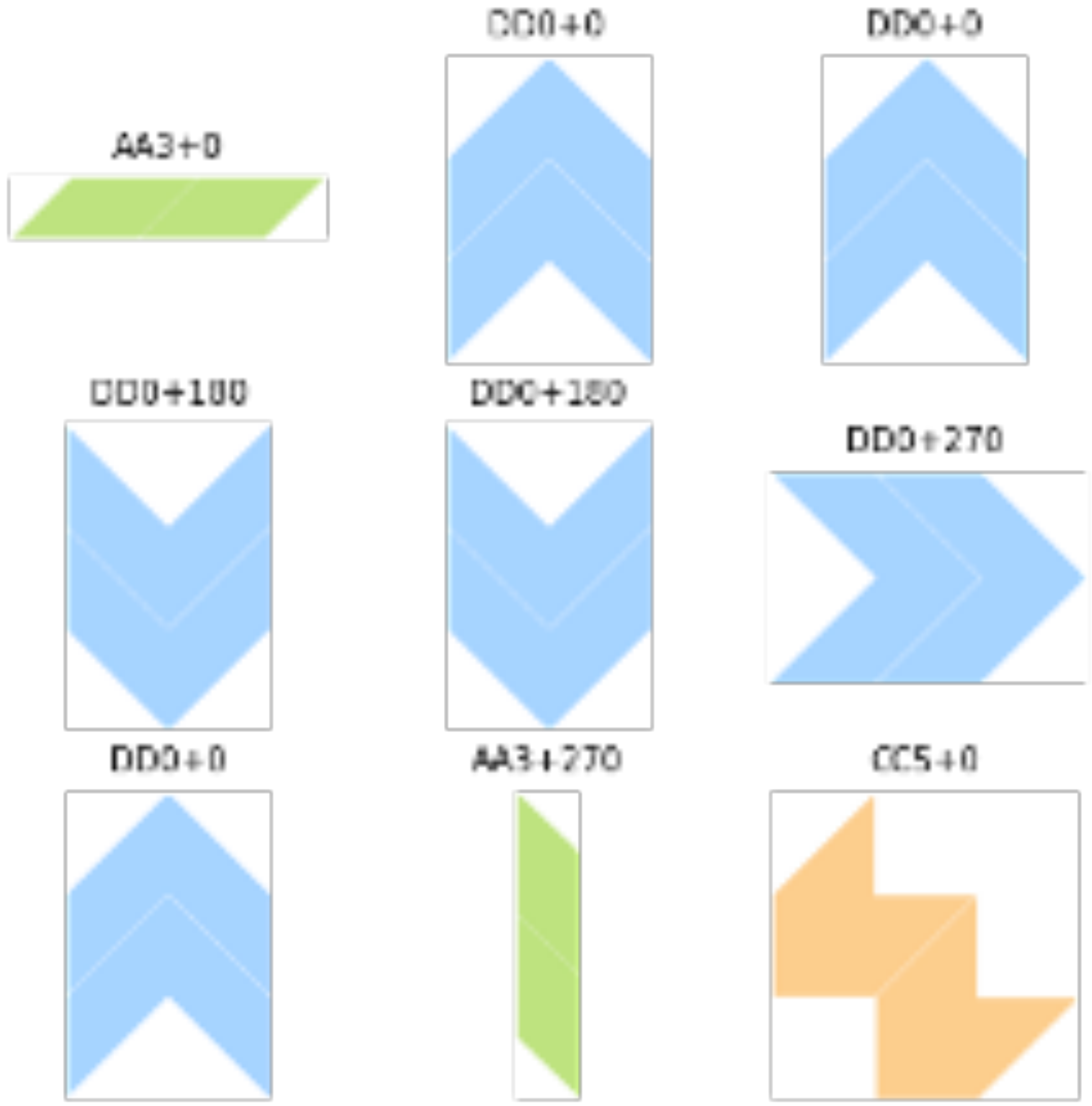


Generation

Here are 3 examples of a "dax":

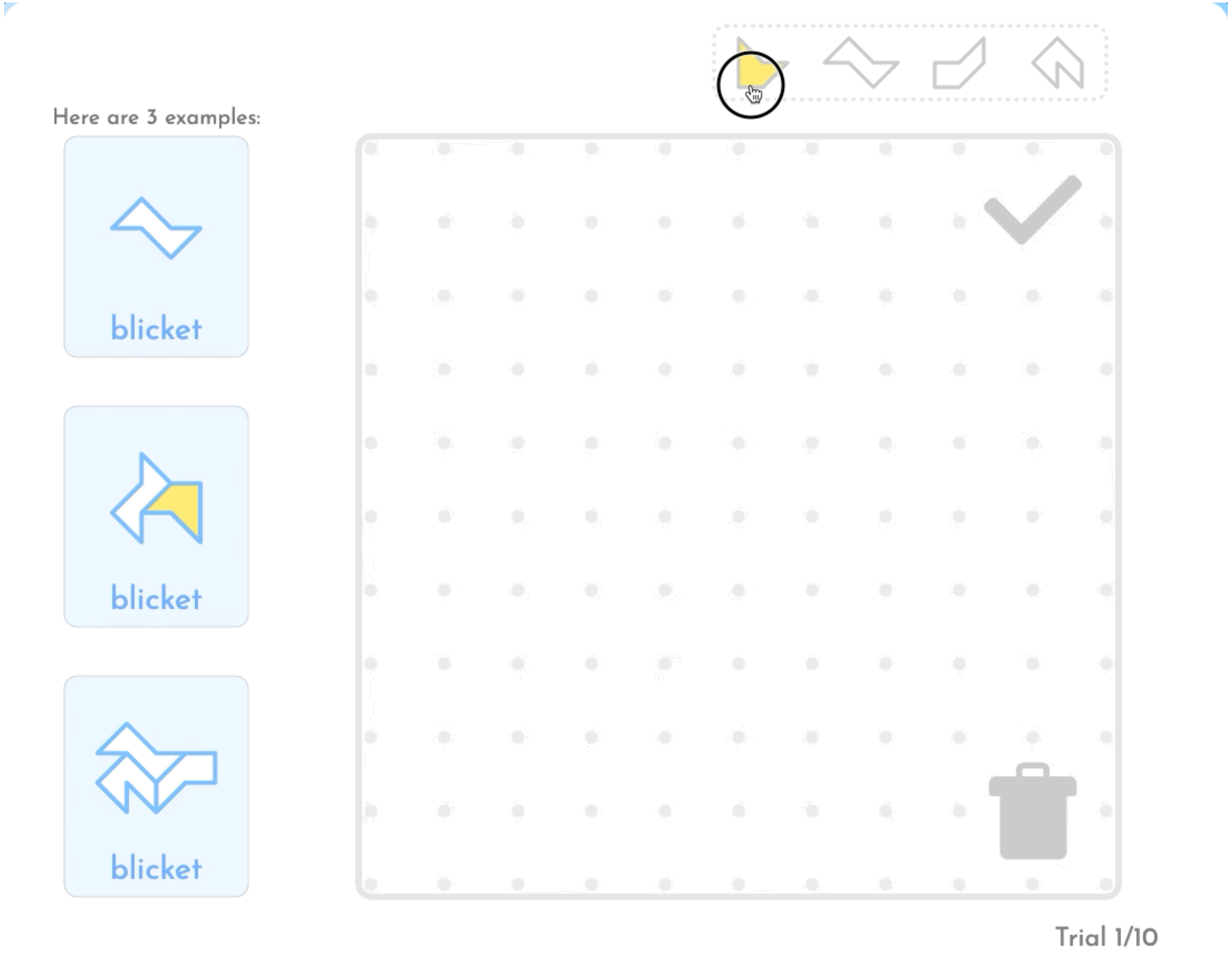


Can you make another "dax?"



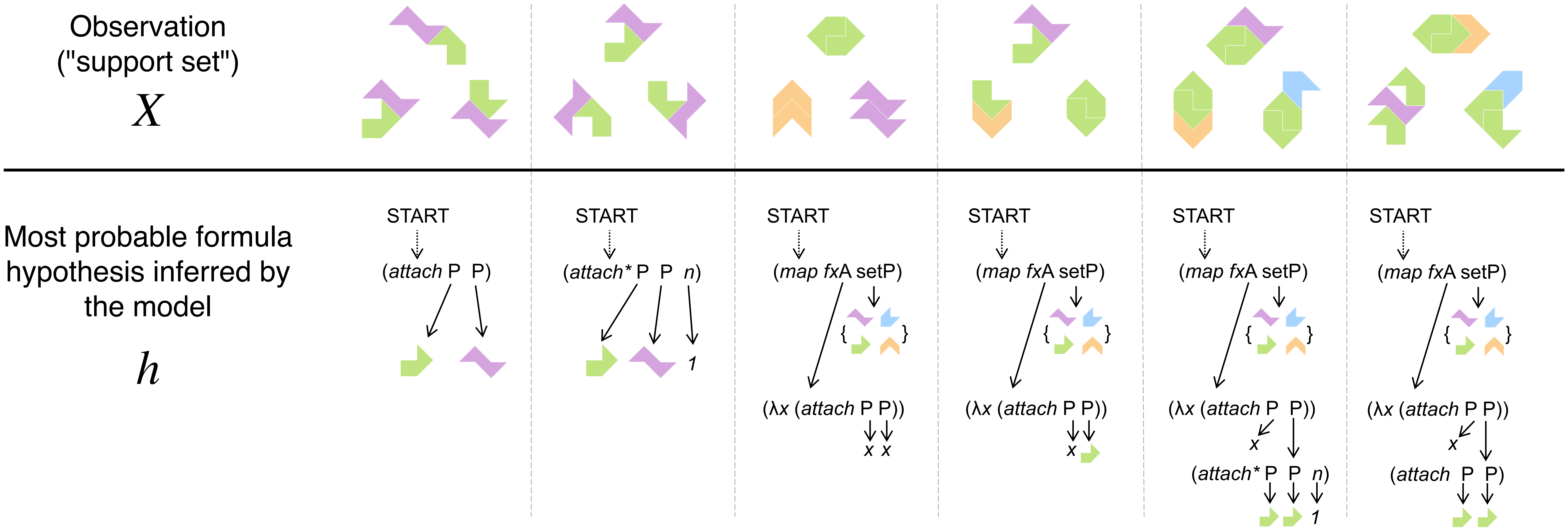
Human generations

Generation task MTurk interface



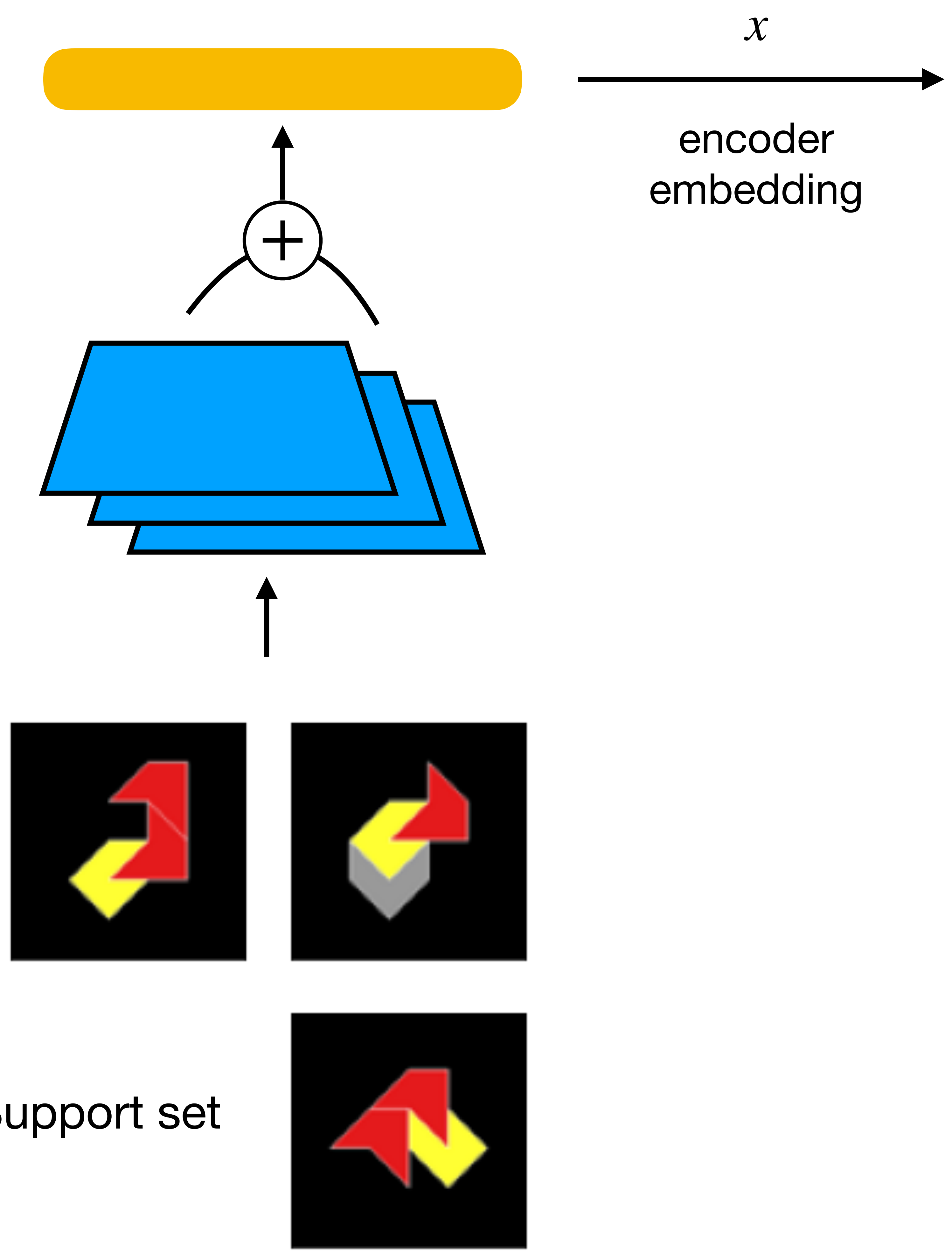
Symbolic Bayesian model

$$p(h\,|\,X) \propto p(h)p(X\,|\,h)$$



Generative neuro-symbolic (GNS) model

1. Neural encoder

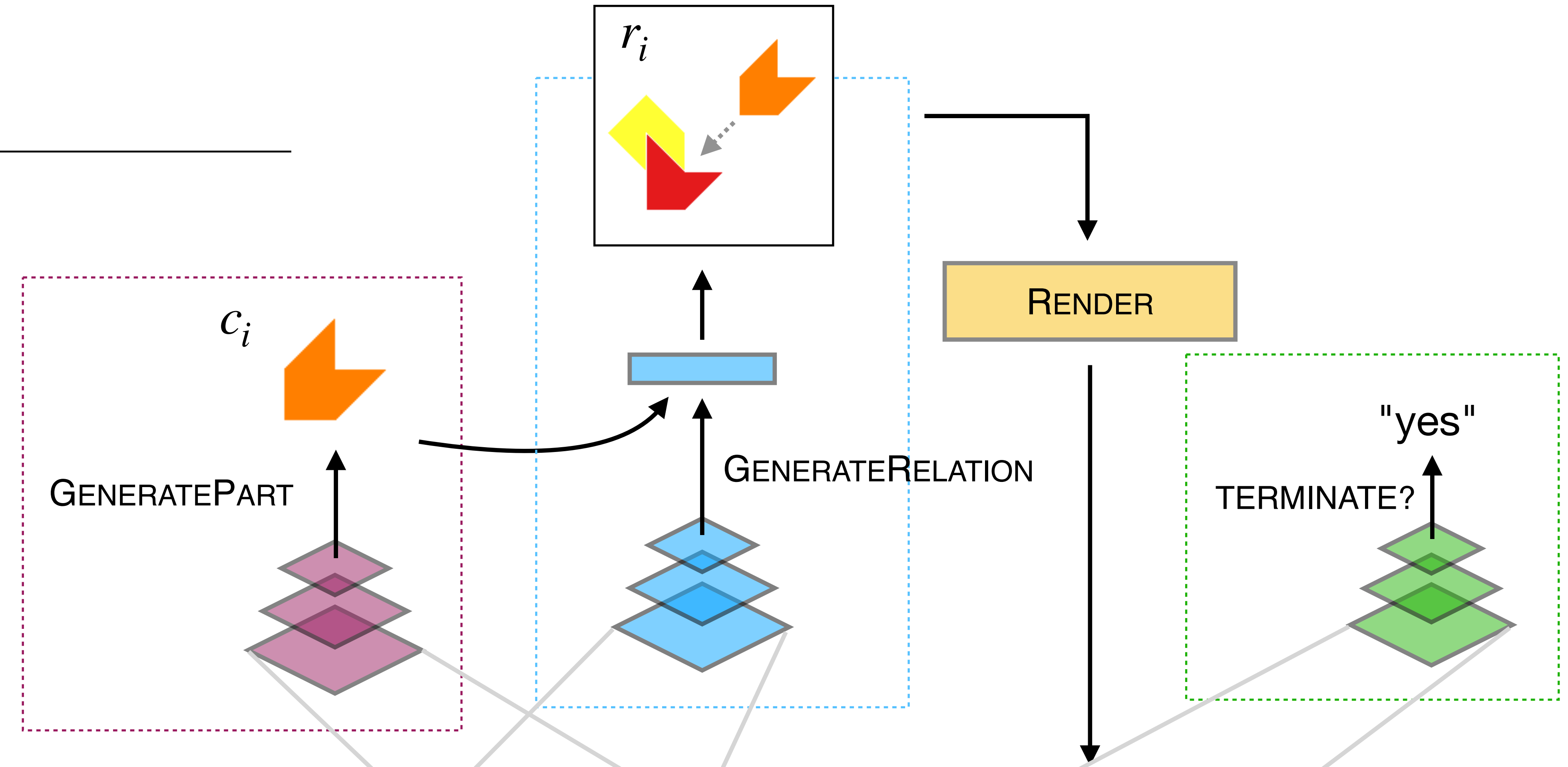


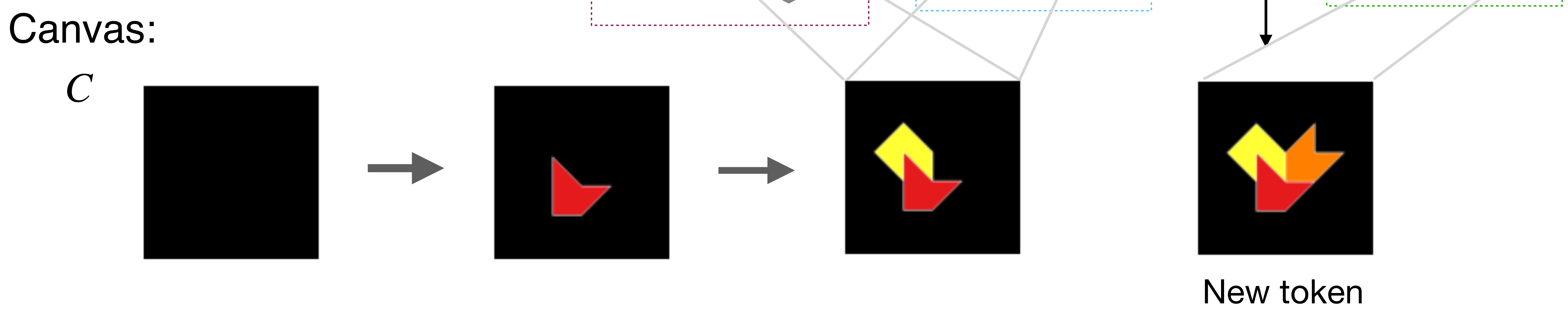
2. Generative neuro-symbolic decoder

procedure GENERATETOKEN(x)
  $C \leftarrow 0$
  while True **do**
    $c_i \leftarrow$ GENERATEPART(x, C)
    $r_i \leftarrow$ GENERATERELATION(x, C, c_i)
    $C \leftarrow$ RENDER(C, c_i, r_i)
    if TERMINATE(x, C) **then**
      break
  return C

key

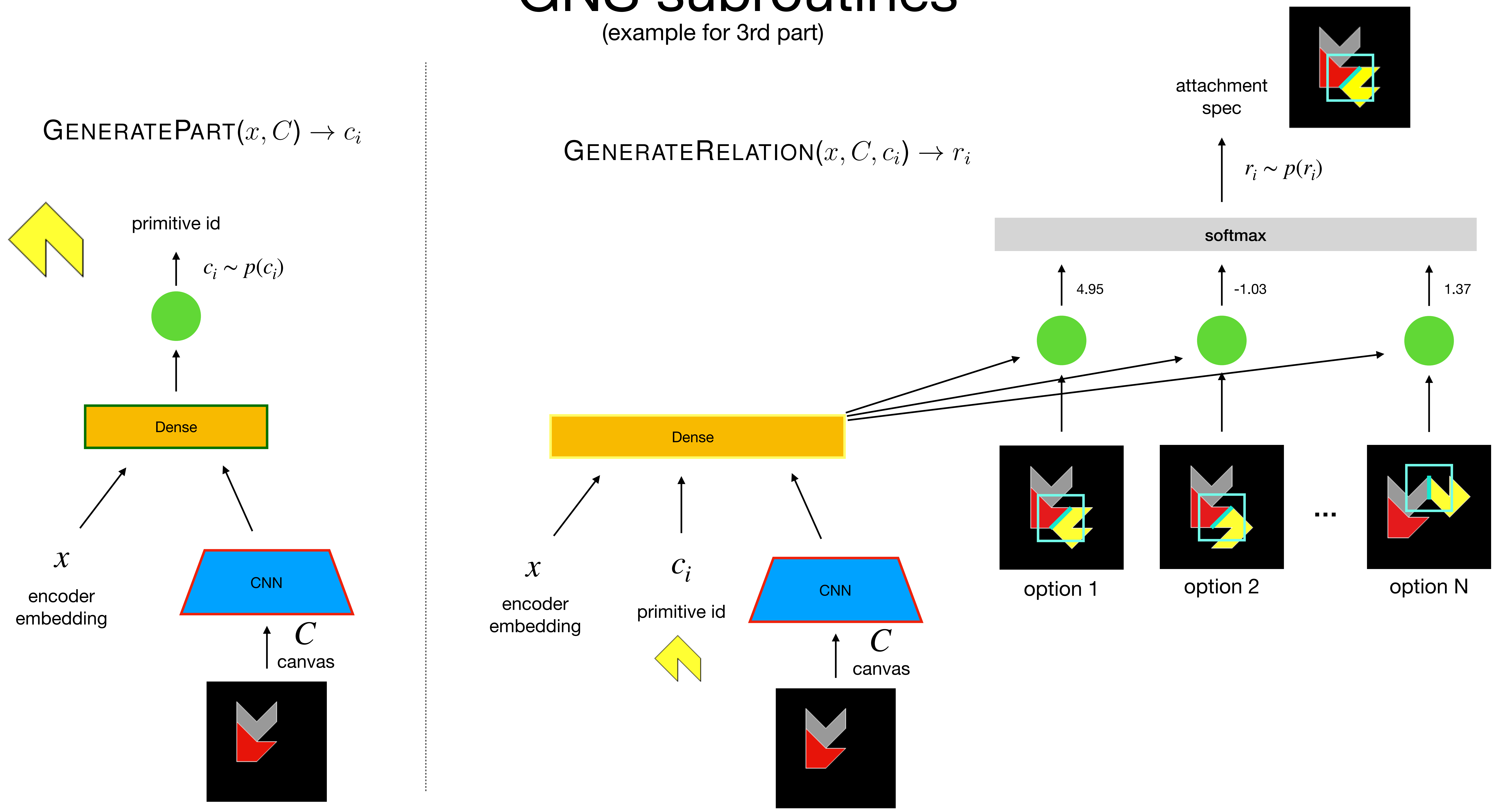
- x: encoder embedding
- C: canvas (partial object)
- c_i: primitive ID for part i
- r_i: symbolic relation for part i



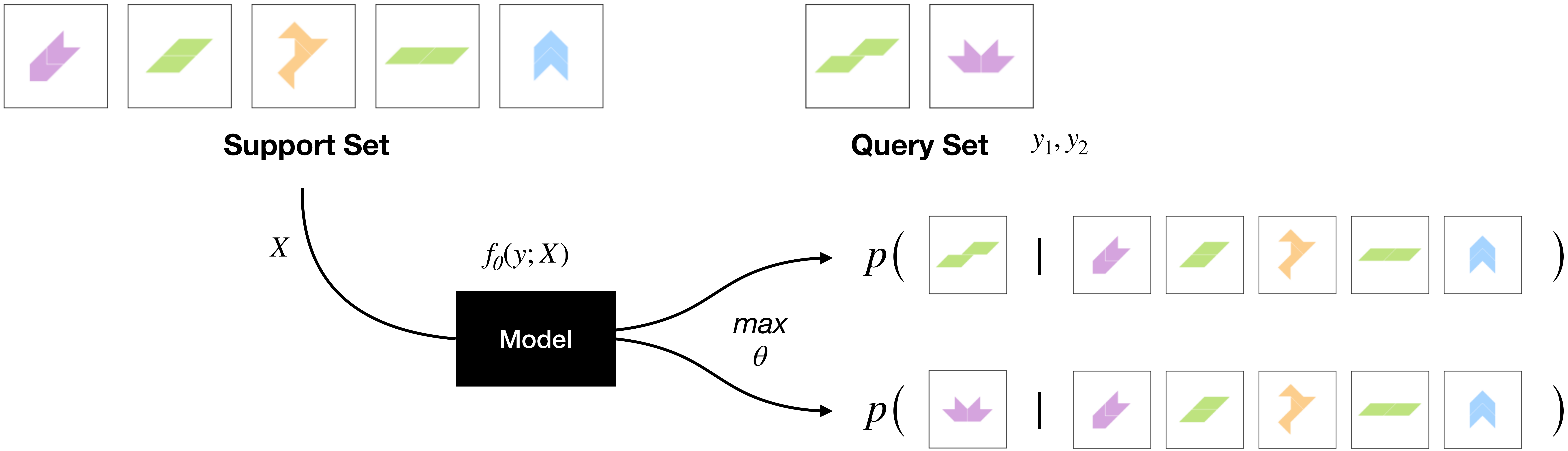


GNS subroutines

(example for 3rd part)



Meta-learning



Objective: maximize log-likelihood of query tokens conditioned on the support

Meta-learning training data

bootstrapping the symbolic Bayesian model

P
Synthetic data distribution

procedure P
$h \sim p(h)$ ▷ Sample formula hypothesis from prior
$S = x_1, ..., x_n \sim p(x \mid h)$ ▷ Sample support set from formula
$Q = x'_1, ..., x'_n \sim p(x \mid h)$ ▷ Sample query set from formula
return S, Q

R
Resampled synthetic data distribution

procedure R
$S \sim$ Uniform(Φ) ▷ Sample support set from human trials
$h \sim p(h \mid S)$ ▷ Sample formula hypothesis from posterior
$Q = x'_1, ..., x'_n \sim p(x \mid h)$ ▷ Sample query set from formula
return S, Q

H
Human distribution

procedure H
$S, Q \sim$ Uniform(Φ) ▷ Sample support & query sets from human trials
return S, Q

C
Bias training distribution

(see Section 4.4 and Appendix B.2)

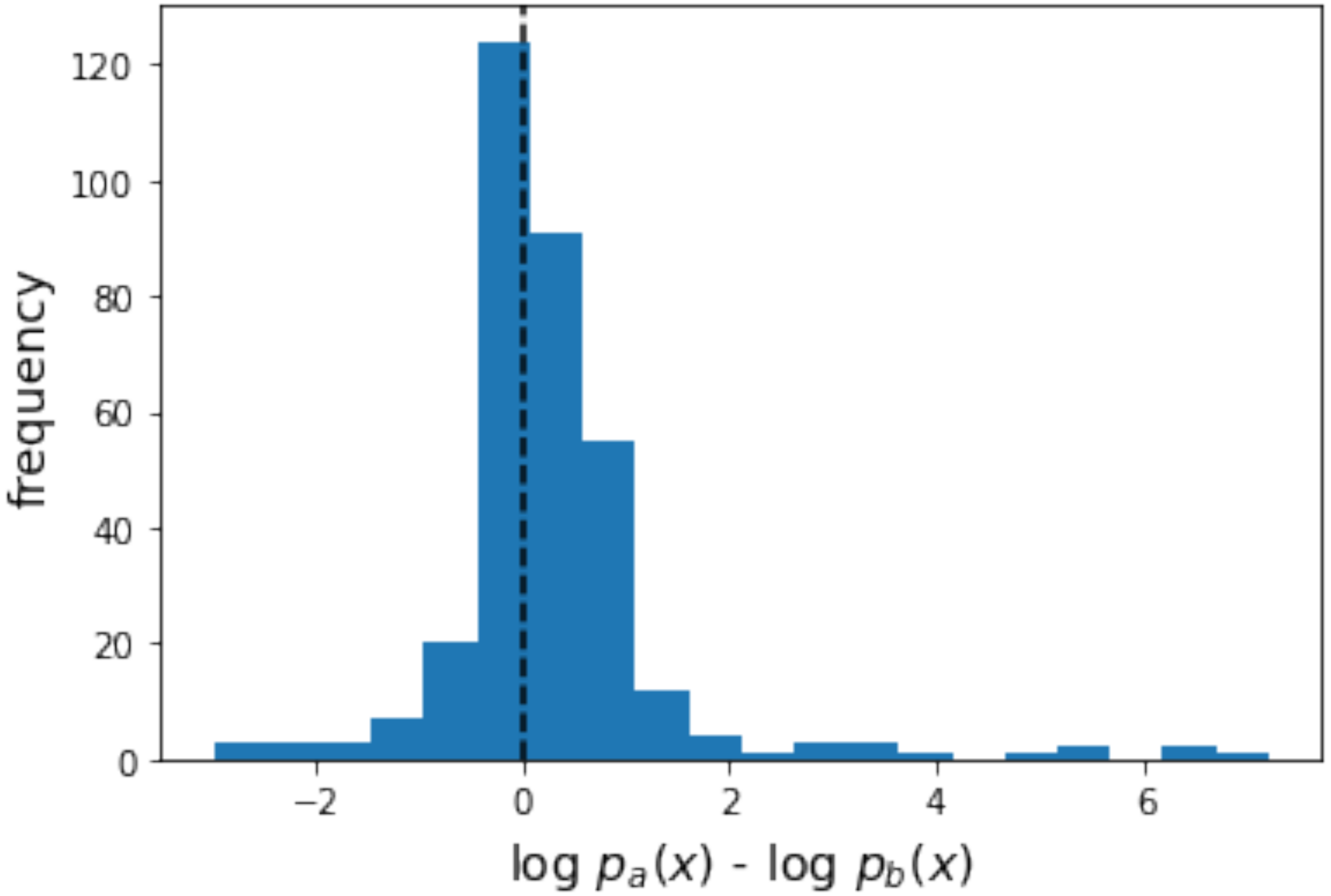
Log-likelihood evaluations

	Test log-likelihood
Bayesian	-4.741
GNS (P/R/H/C)	**-4.444**
GNS (P/R/H)	-4.535
GNS (P/R)	-4.645
GNS (P)	-4.930

Likelihood of held-out human generations. For each model, the total log-likelihood averaged over the holdout set is reported.

Paired t-test comparing per-example log-likelihood of GNS (P/R/H/C) vs. Bayesian

$$t(336) = 6.197, \quad p < 0.001$$



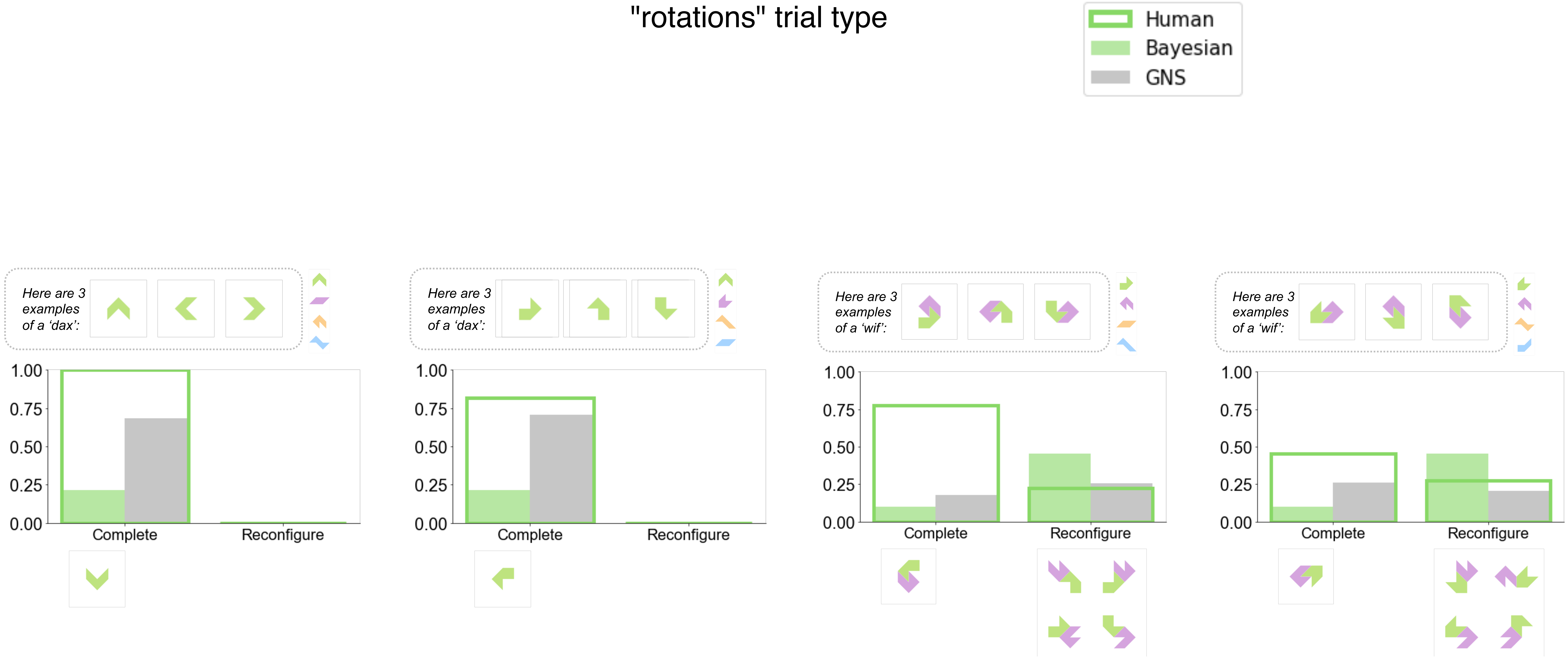
Log-likelihood evaluations

	support set	new token	freq.	delta
1			(2)	4.70
2			(1)	3.22
3			(1)	3.19
4			(5)	2.18
5			(1)	1.90
6			(1)	1.69
7			(3)	1.62
8			(1)	1.47

Accounting for human inductive biases

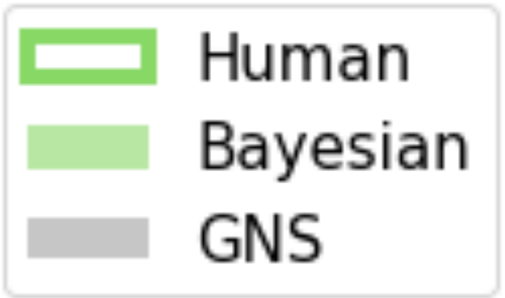
Accounting for human inductive biases

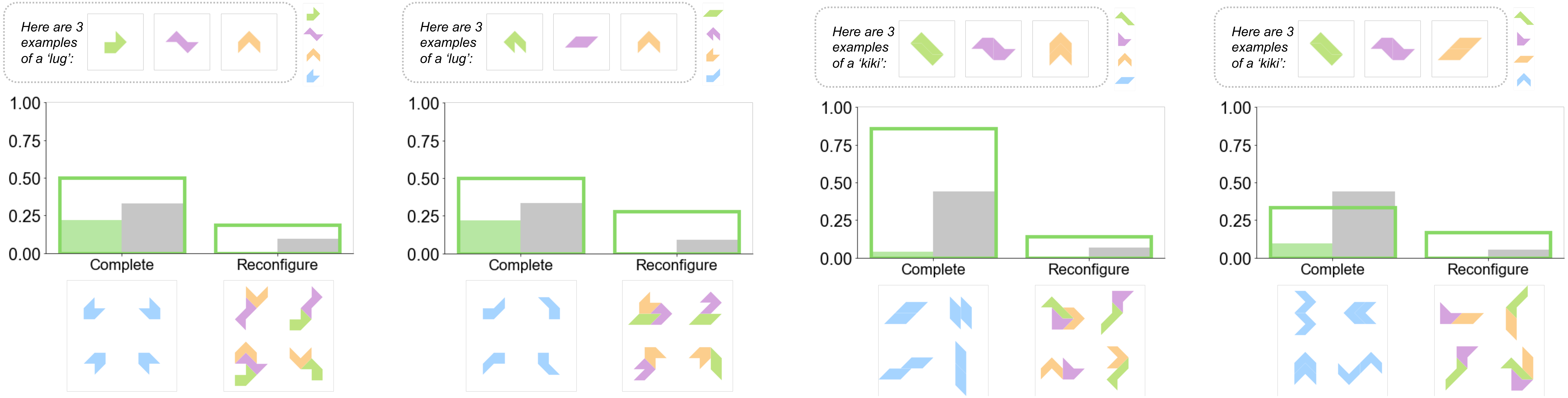
"rotations" trial type



Accounting for human inductive biases

"primitives" trial type





Conclusions: Case study #2

- GNS models are an effective way to understand and simulate human few-shot learning of structured visual concepts

- Compared to a strong symbolic baseline model, GNS provides an improved likelihood account of human few-shot generation

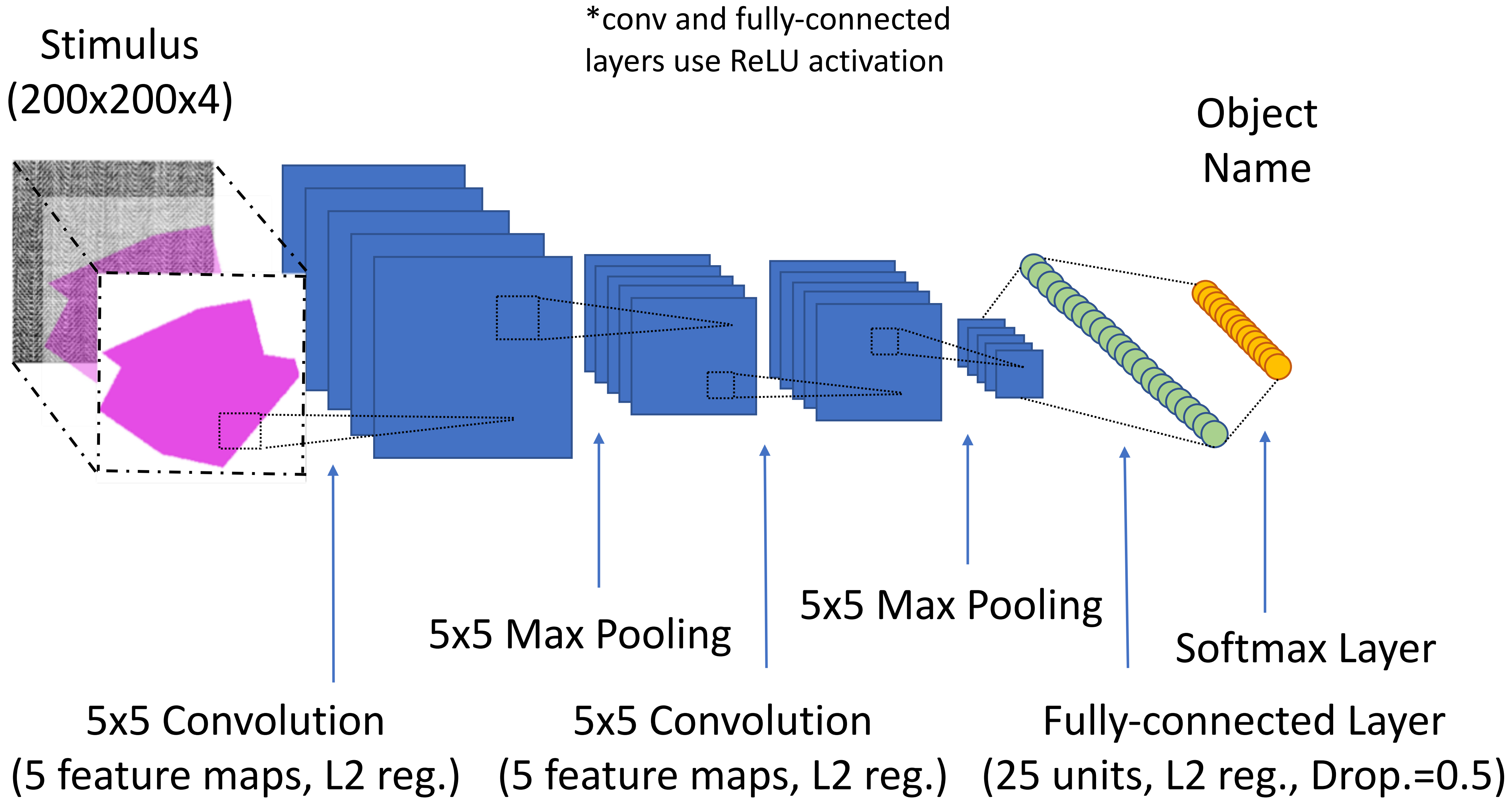
- GNS can account for human inductive biases that are not well-explained by alternatives

Additional projects

Learning inductive biases with simple neural networks

(Feinman & Lake, 2018)

Convolutional neural network (CNN) architecture

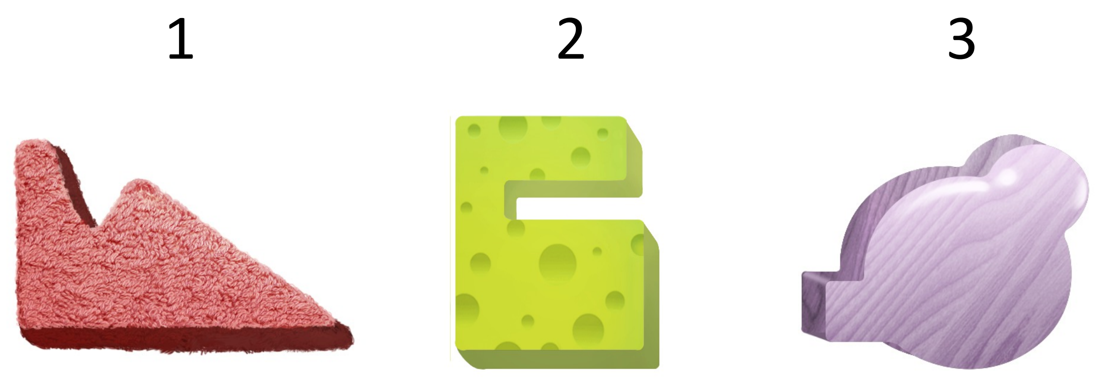


Shape bias test

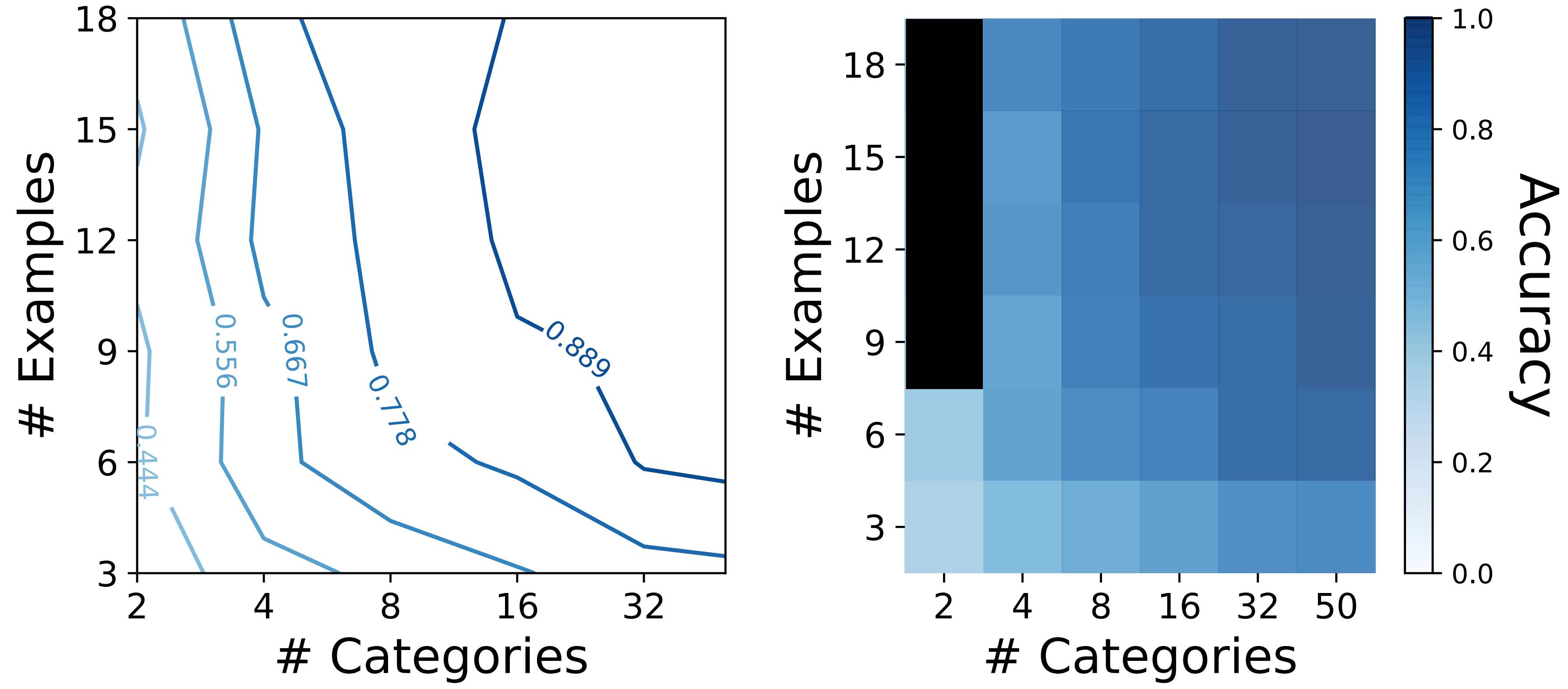
This is a "dax."

(Smith et al., 2002)

Where is the other "dax?"

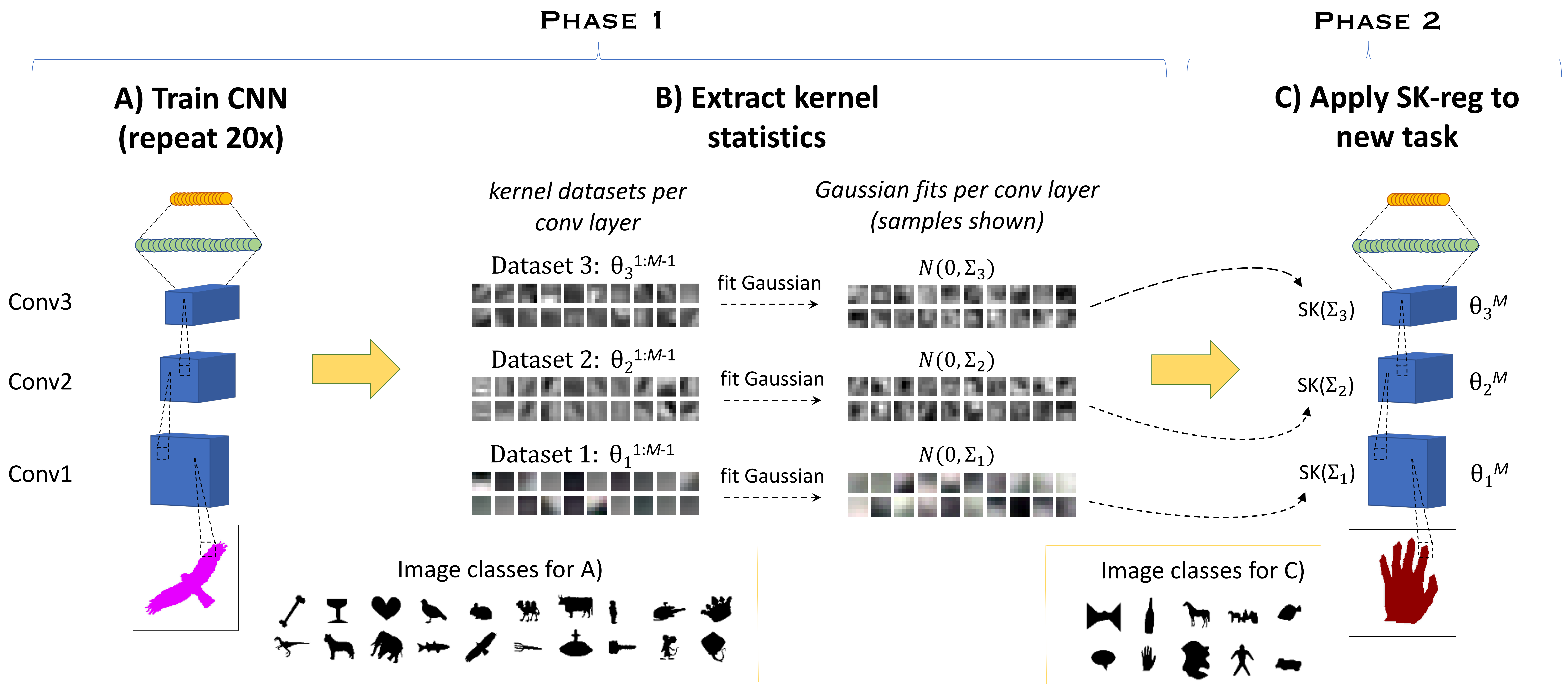


CNN shape bias strength vs. dataset size



Learning a smooth kernel regularizer for convolutional neural networks

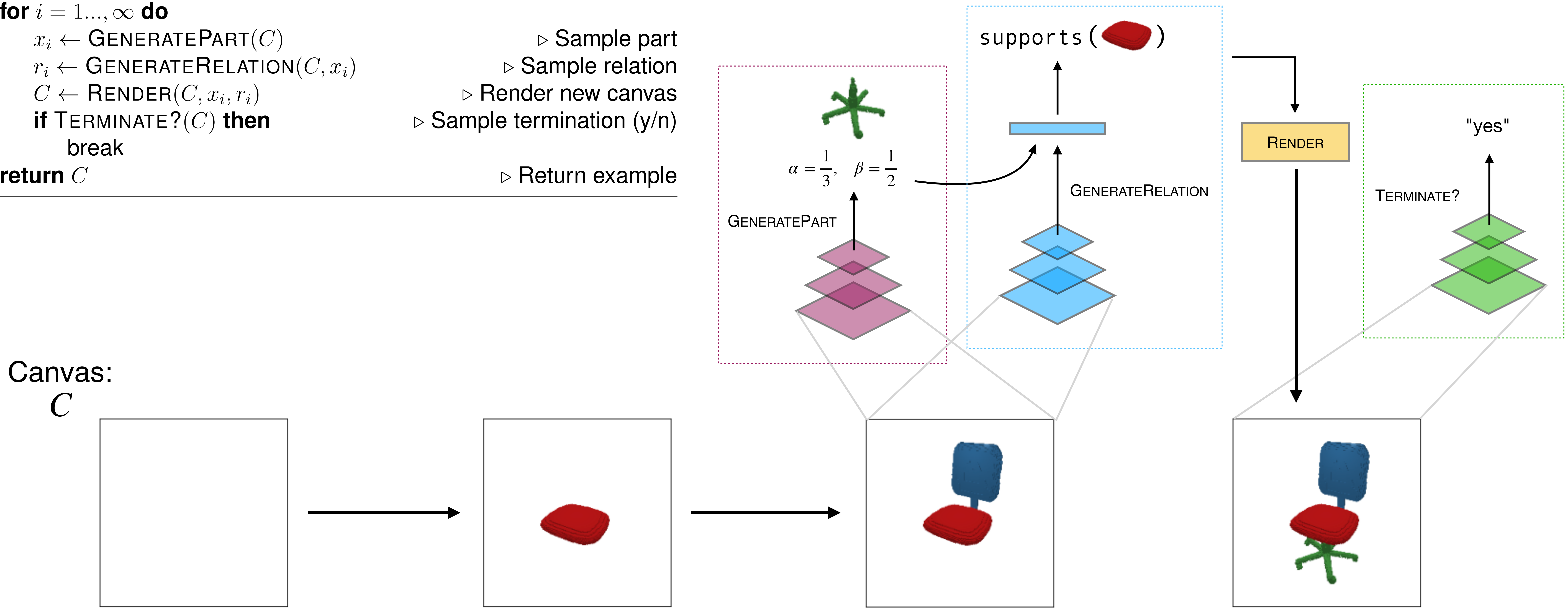
(Feinman & Lake, 2019)



Summary & Conclusions

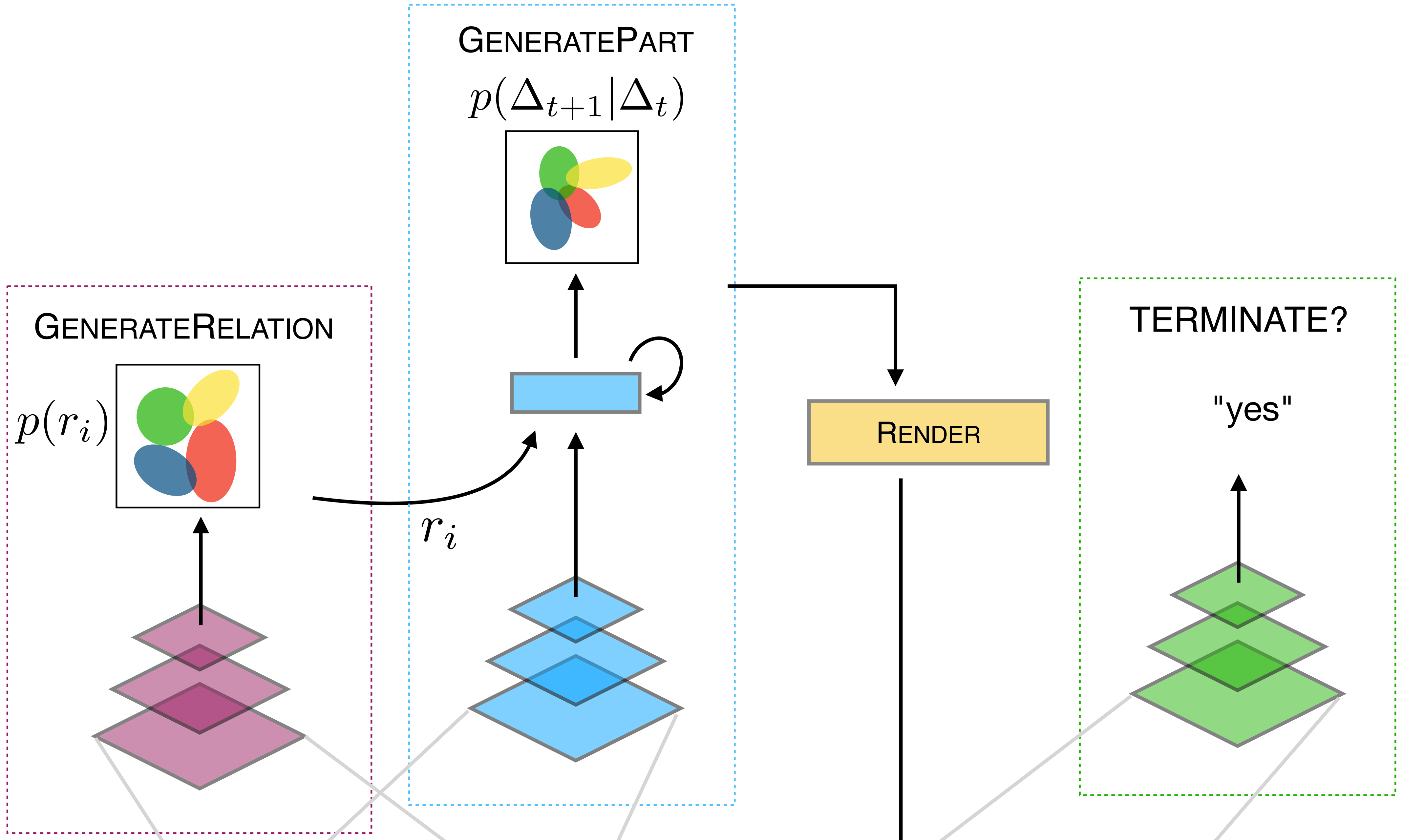
introduced Generative Neuro-Symbolic (GNS) modeling

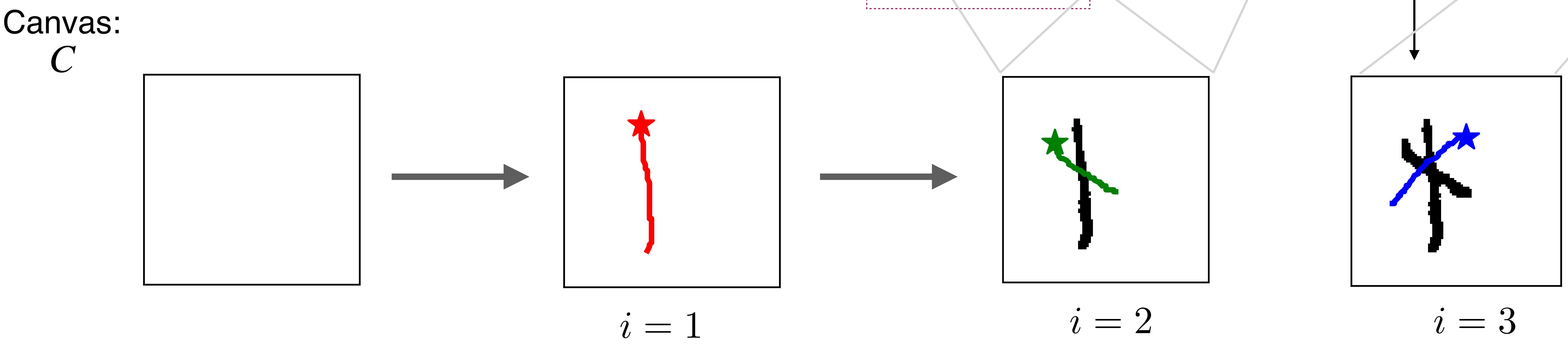
procedure GENERATEEXAMPLE
  $C \leftarrow 0$ ▷ Initialize blank canvas
  for $i = 1 ..., \infty$ **do**
    $x_i \leftarrow$ GENERATEPART(C) ▷ Sample part
    $r_i \leftarrow$ GENERATERELATION(C, x_i) ▷ Sample relation
    $C \leftarrow$ RENDER(C, x_i, r_i) ▷ Render new canvas
    if TERMINATE?(C) **then** ▷ Sample termination (y/n)
      break
  return C ▷ Return example



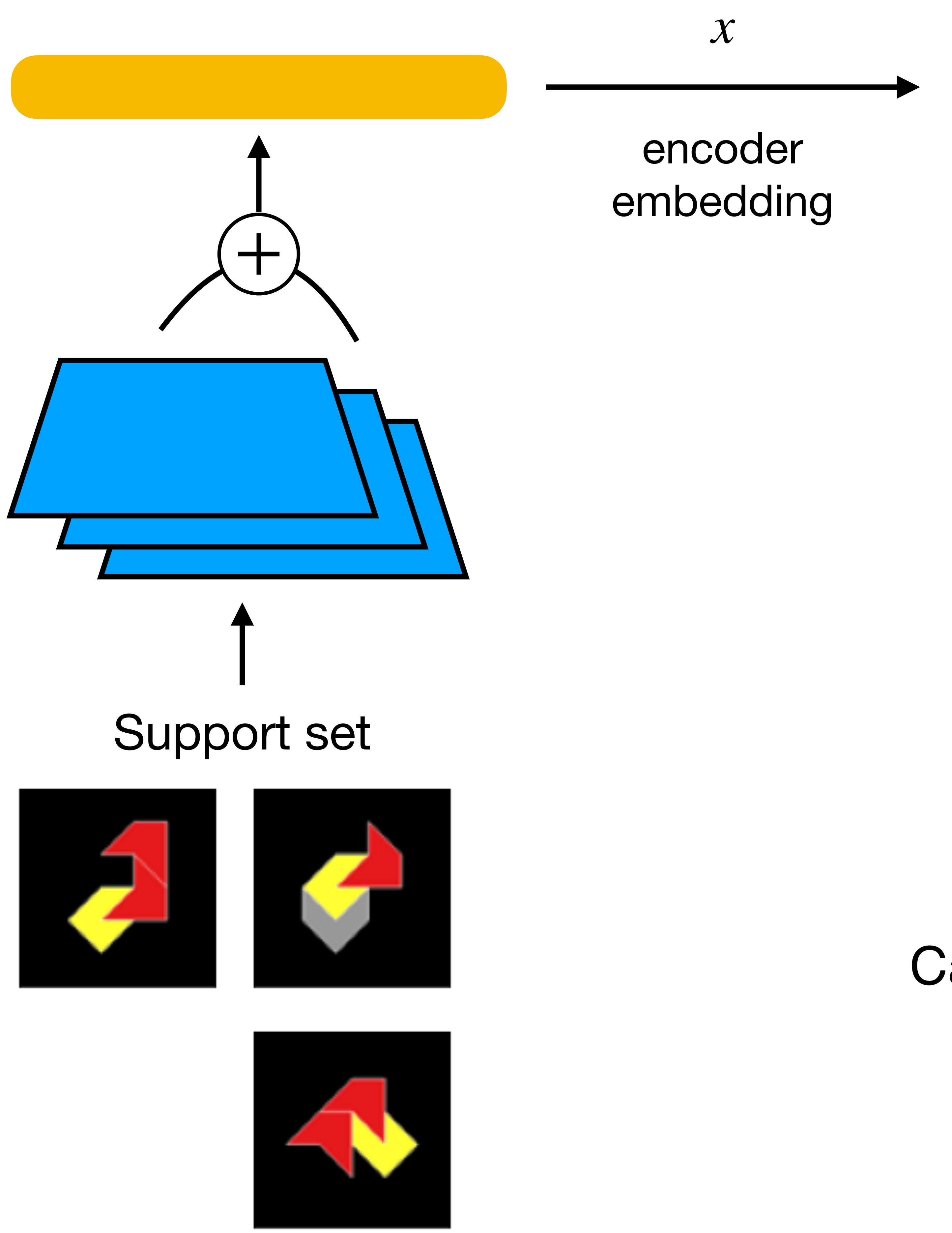
GNS model of handwritten character concepts

procedure GenerateCharacter
  $C \leftarrow 0$ ▷ Initialize blank canvas
  for $i = 1...,\infty$ **do**
    $r_i \leftarrow$ GenerateRelation(C) ▷ Sample relation
    $x_i \leftarrow$ GeneratePart(C, r_i) ▷ Sample part
    $C \leftarrow$ Render(C, x_i, r_i) ▷ Render to canvas
    $v_i \leftarrow$ Terminate?(C) ▷ Sample termination indicator
    if v_i **then**
      break ▷ Terminate sample
  return C

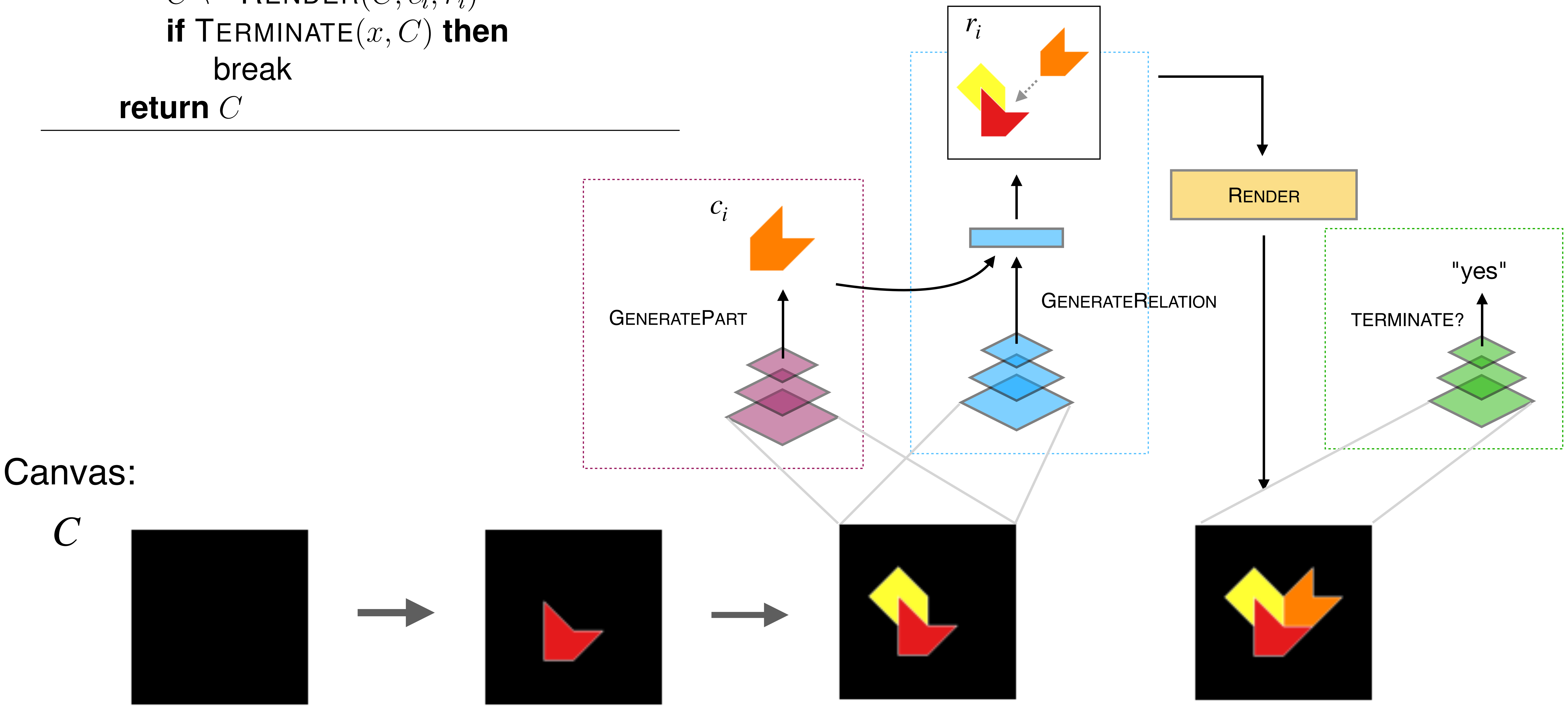




GNS model of synthetic part-based concepts ("alien figures")



procedure GENERATETOKEN(x)
    $C \leftarrow 0$
    while True **do**
        $c_i \leftarrow$ GENERATEPART(x, C)
        $r_i \leftarrow$ GENERATERELATION(x, C, c_i)
        $C \leftarrow$ RENDER(C, c_i, r_i)
        if TERMINATE(x, C) **then**
            break
    return C



General conclusions

- Generative neuro-symbolic (GNS) modeling provides a novel synthesis of ideas from the structured and statistical modeling traditions
- By combining these ingredients in a computational model, we can account for human concept learning in ways that purely- symbolic and neural models fall short
- GNS models can help us understand the dual structural and statistical natures of human knowledge and direct us toward a more accurate representation of concepts

Thank You

Brenden Lake

Yanli Zhou

Guy Davidson
Emin Orhan
Wai Keen Vong
+ HMLL lab

Tuan-Anh Le
Maxwell Nye
Joshua Tenenbaum
Lucas Tian
+ CoCoSci lab

Nikhil Parthasarathy

NYU Neuroscience cohort

Andy Feinman
Mary Van Hoomissen
Nick Feinman

Charlotte Walmsley

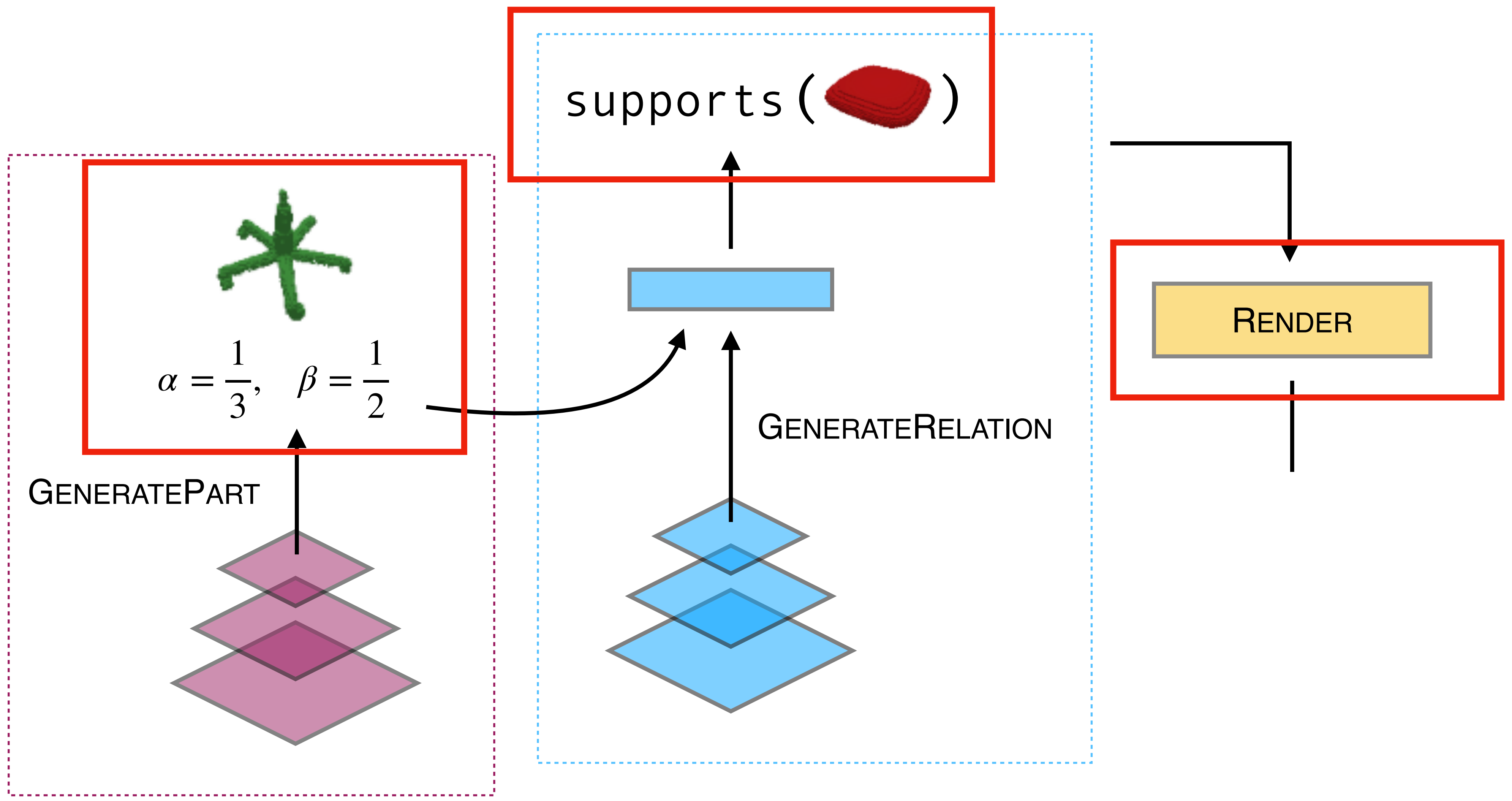
Questions?

"*What I cannot create, I do not understand.*"

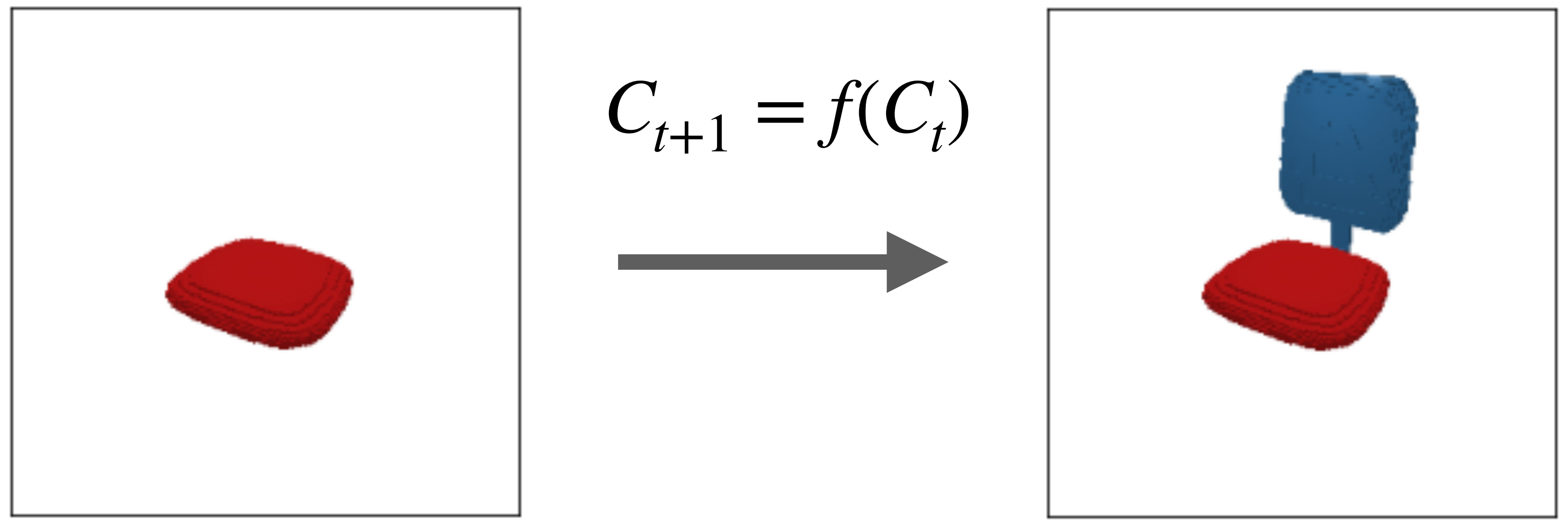
—Richard Feynman

Inductive biases of GNS architecture

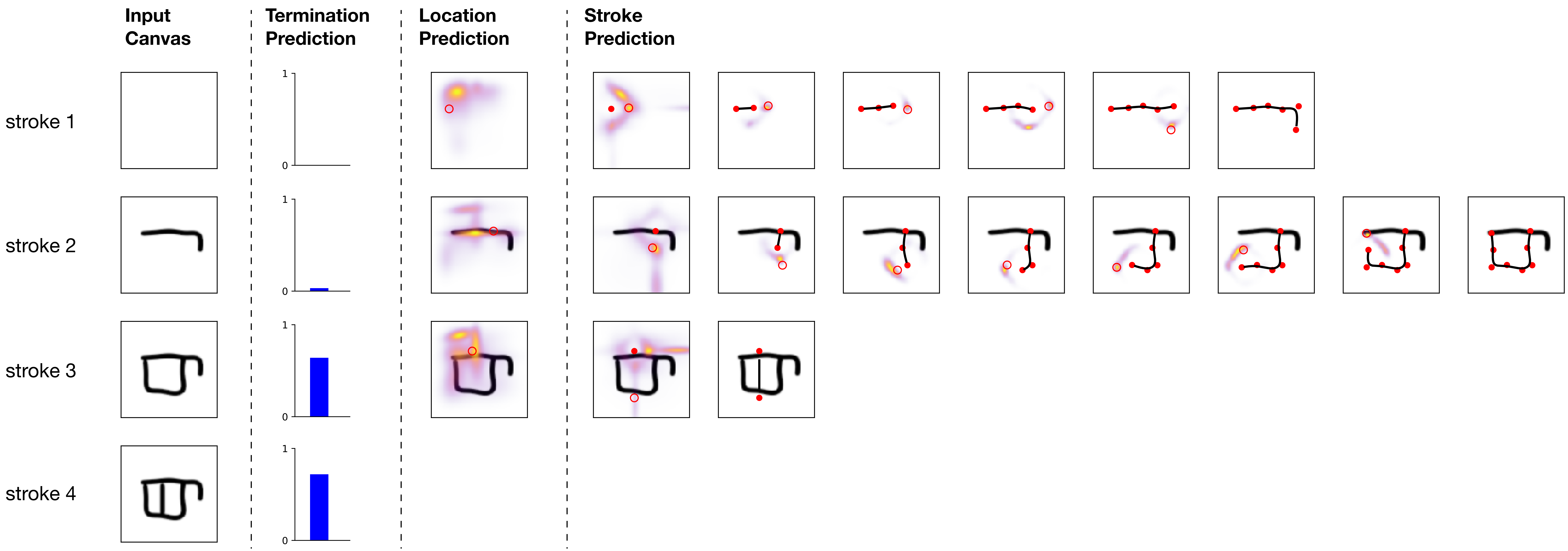
Explicit notion of *causality* via symbolic primitives and renderer



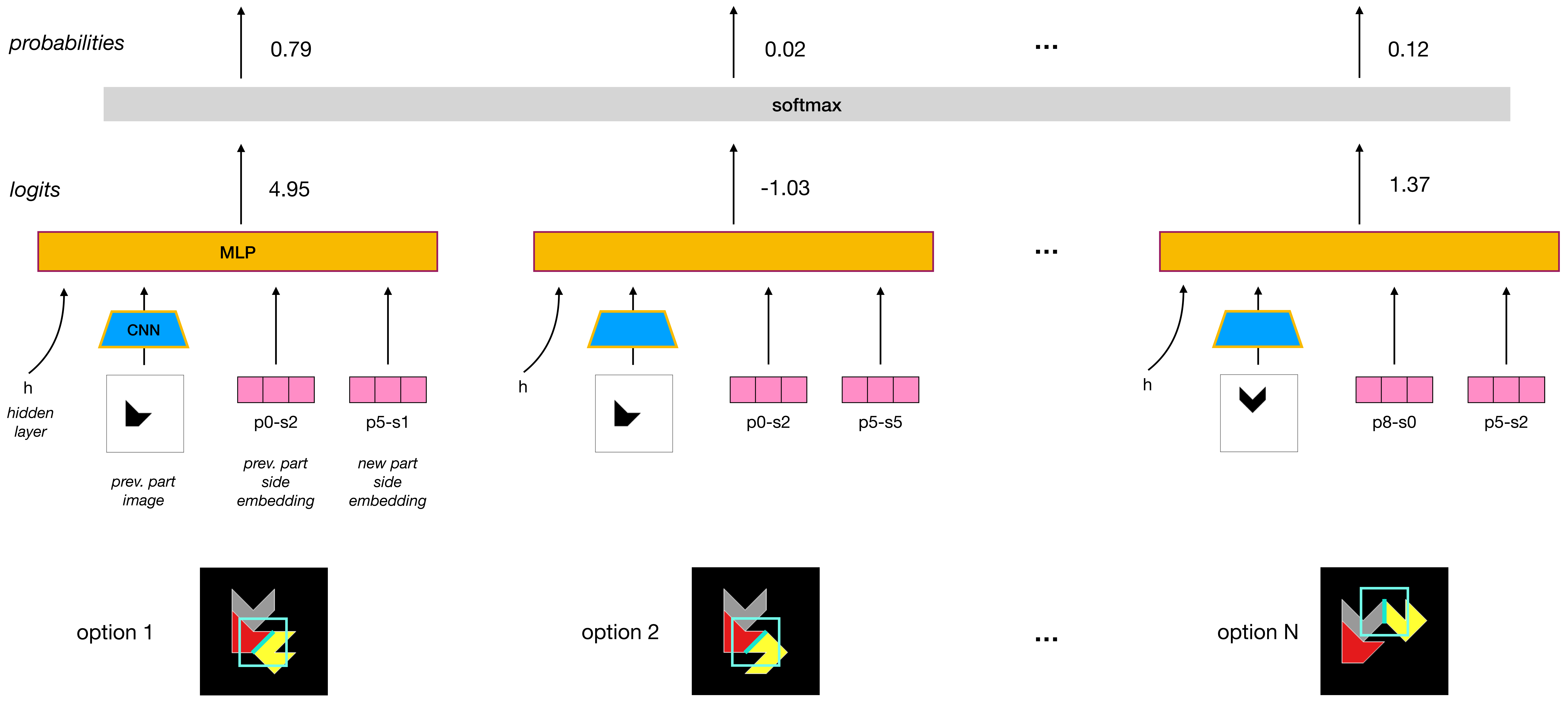
Compositional representation via modular subroutines and controlled memory state



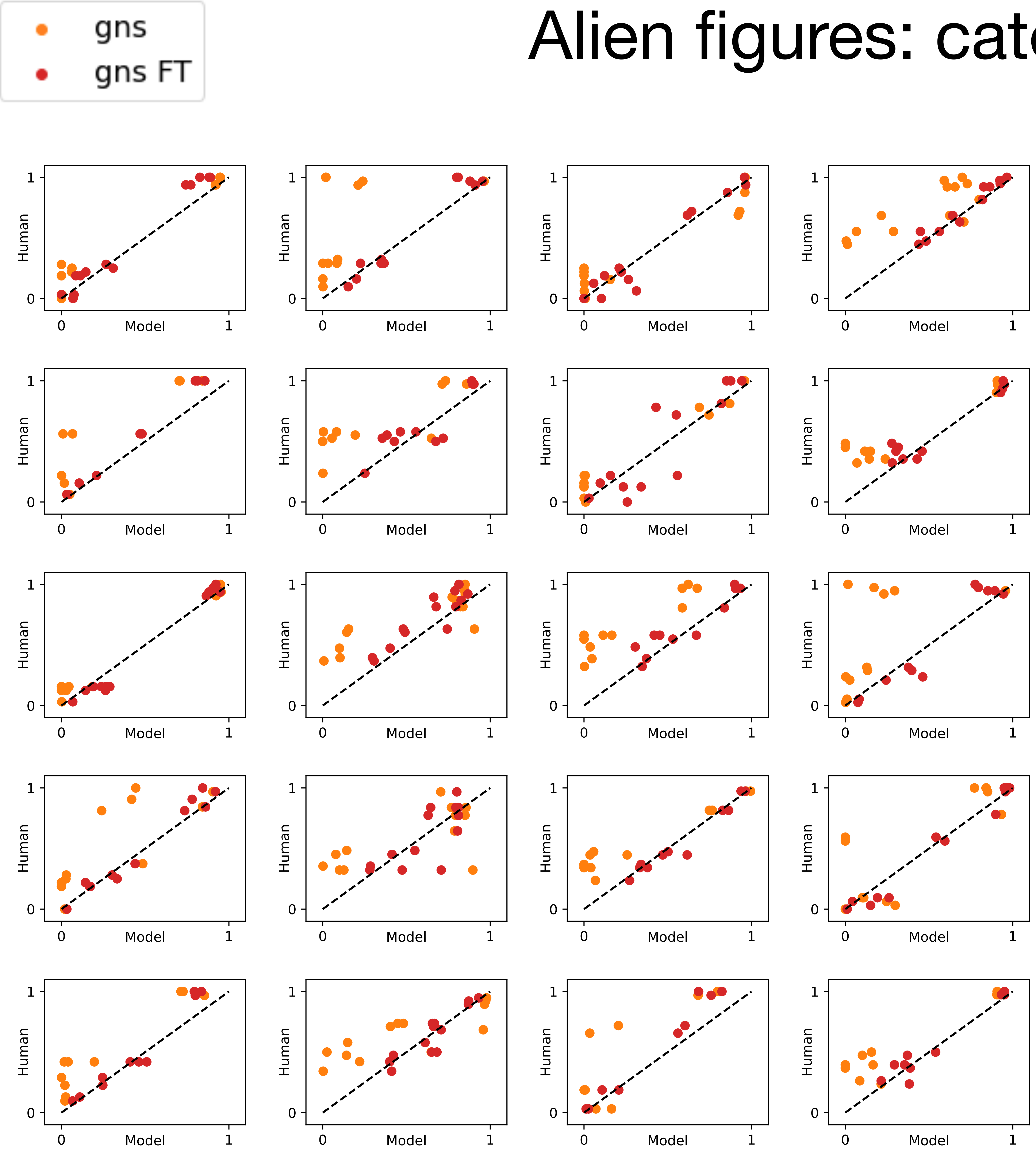
Forward model in action



$\textsc{GenerateRelation}(x, C, c_i) \rightarrow r_i$



Alien figures: categorization task



GNS: Best-performing GNS model from the generation task (experiment 1), evaluated without any modification

GNS FT: A *finetuned* variant of the GNS model from generation. The model is initialized with the generation parameters and further optimized using (a subset of) human categorization data

	Pearson r	Spearman r
GNS	0.761	0.637
GNS FT	0.953	0.881

Correlation with human judgements. Correlation coefficients are computed for each concept type, and the average coefficient across types is reported.