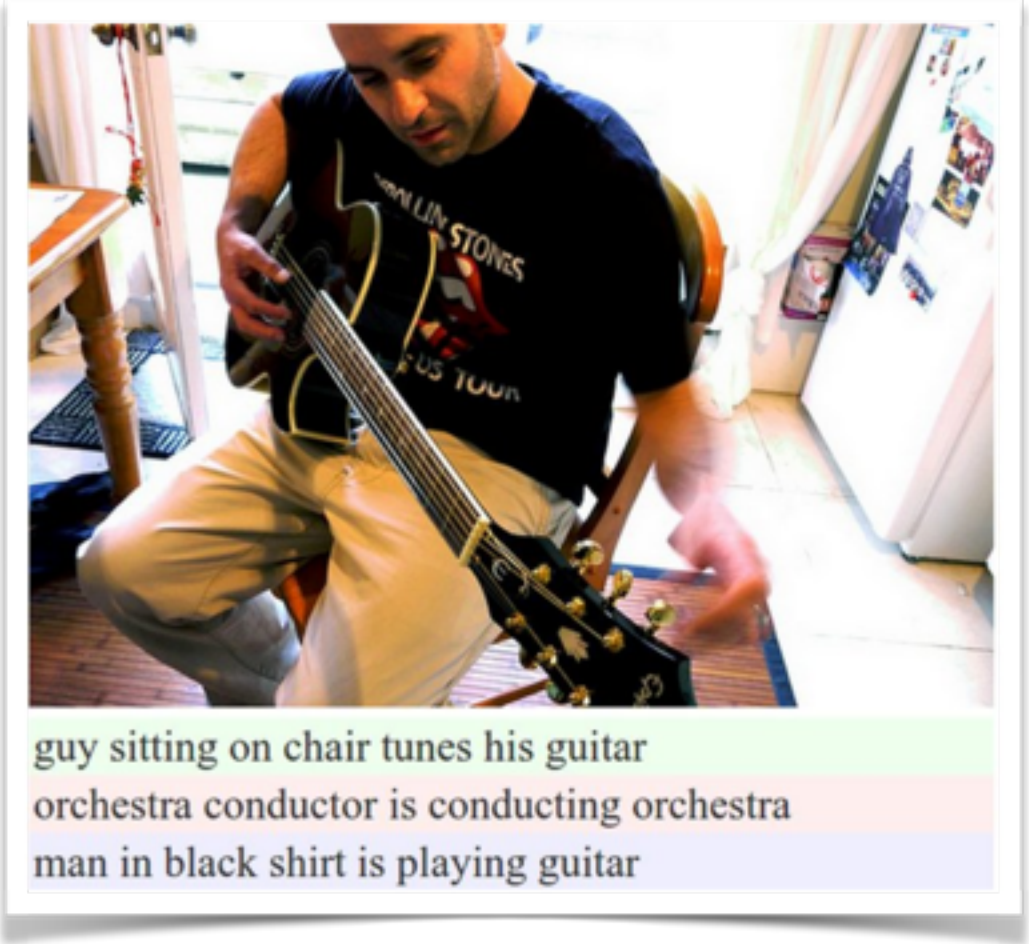
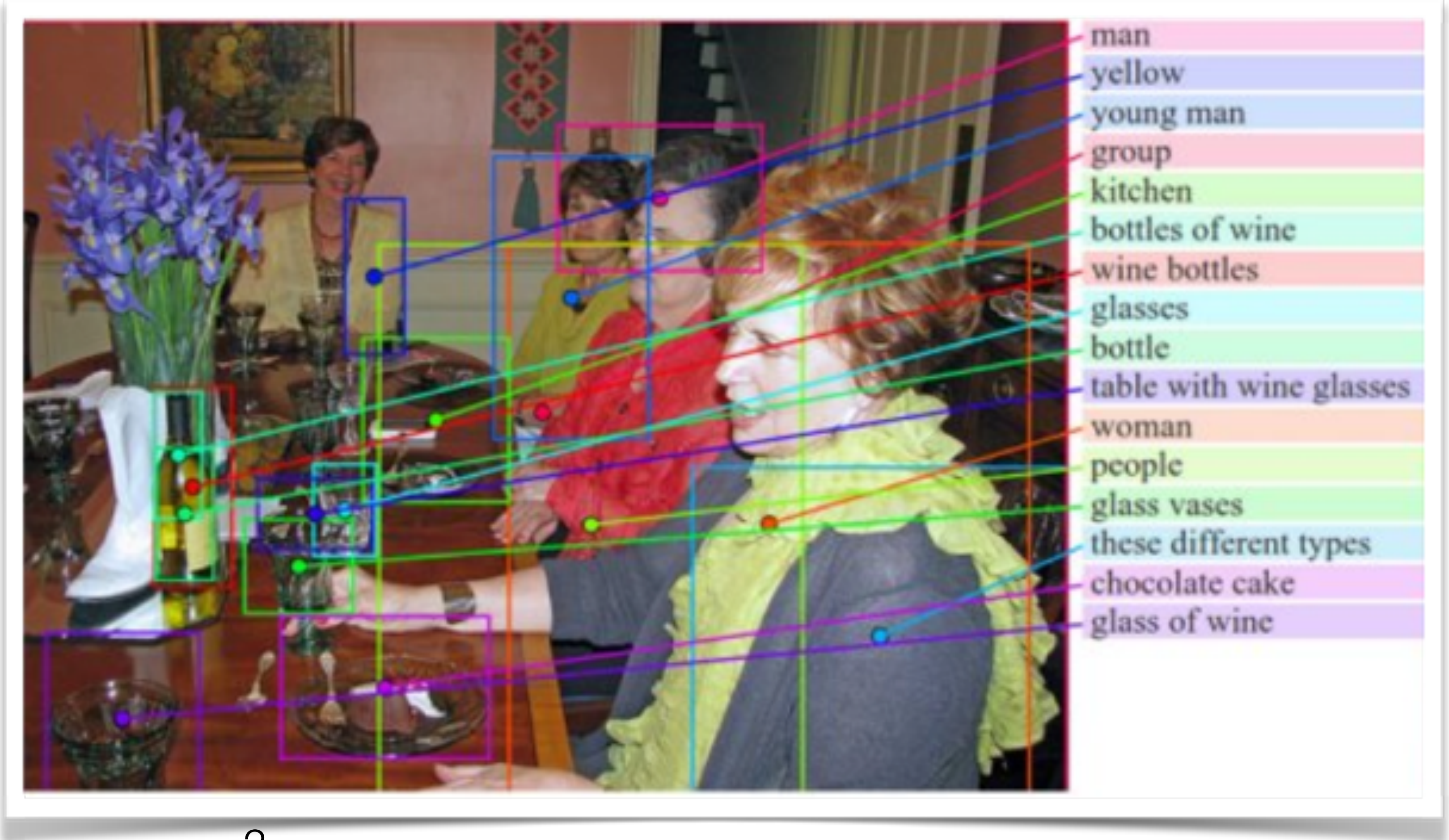
Deep Image Description

Rui-Wei Zhao
rw.du.zhao@gmail.com

Outline

- Generating descriptions for the whole images{Vinyals2014, Karpathy2014}
- Generating descriptions for the regional images{Karpathy2014}


guy sitting on chair tunes his guitar
orchestra conductor is conducting orchestra
man in black shirt is playing guitar



Generating descriptions for the whole images

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

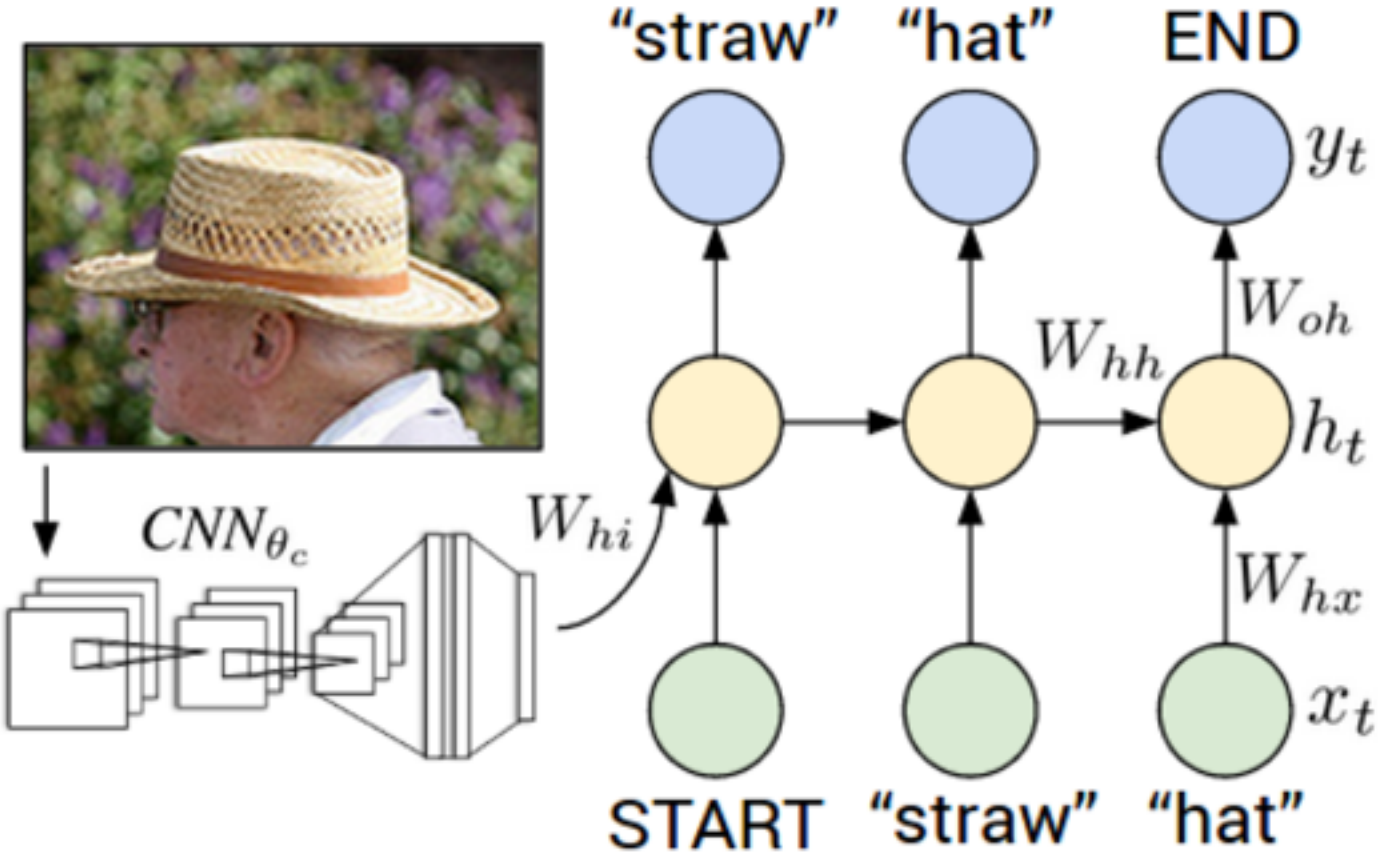
"'little girl is eating piece of cake."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."

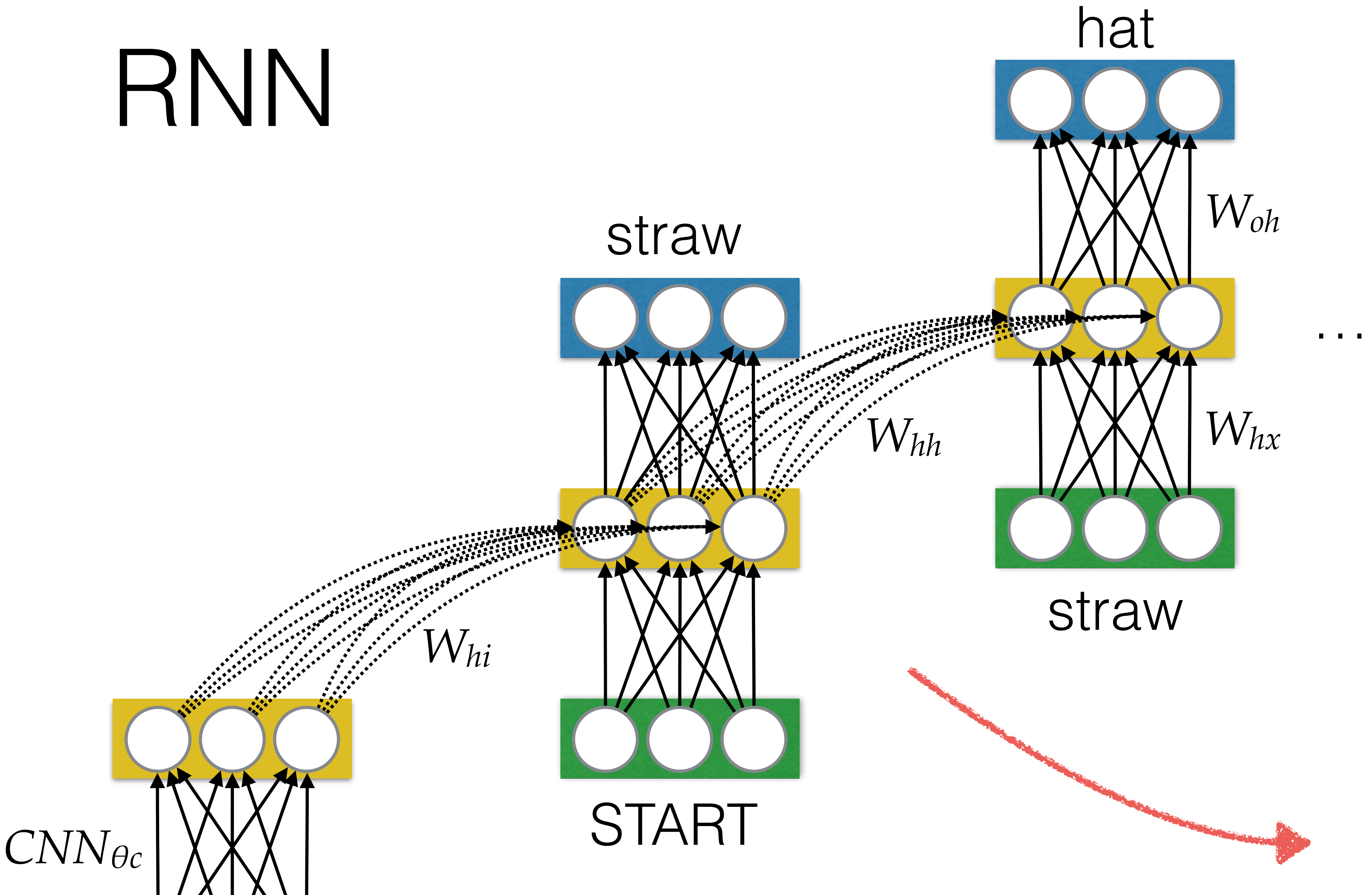
"black cat is sitting on top of suitcase."

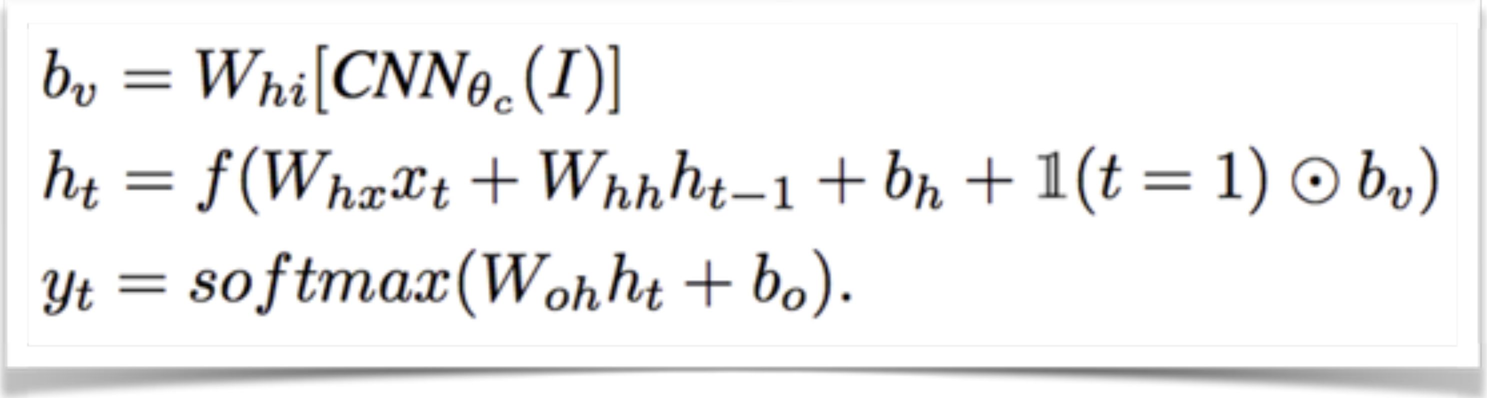
Predictive Model



$$f(s \mid v; \Theta) = p(s_1 \mid v, s_0)\, p(s_2 \mid v, s_0, s_1) \cdots p(s_T \mid v, s_0, \ldots, s_{T-1})$$

RNN

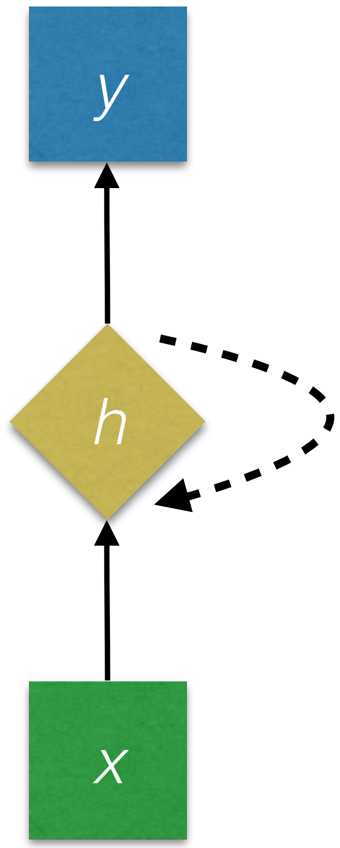




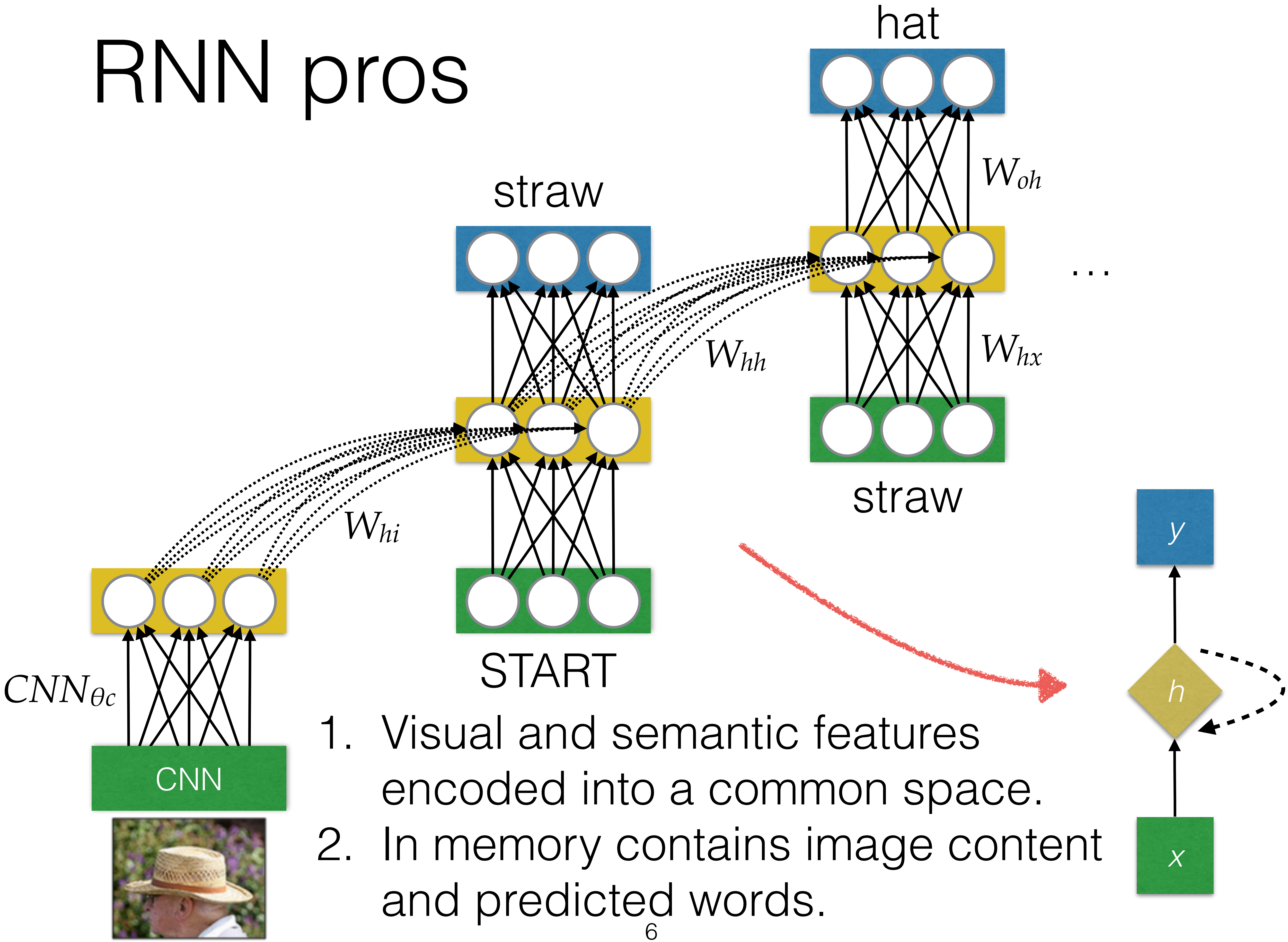
$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$

$$y_t = softmax(W_{oh}h_t + b_o).$$

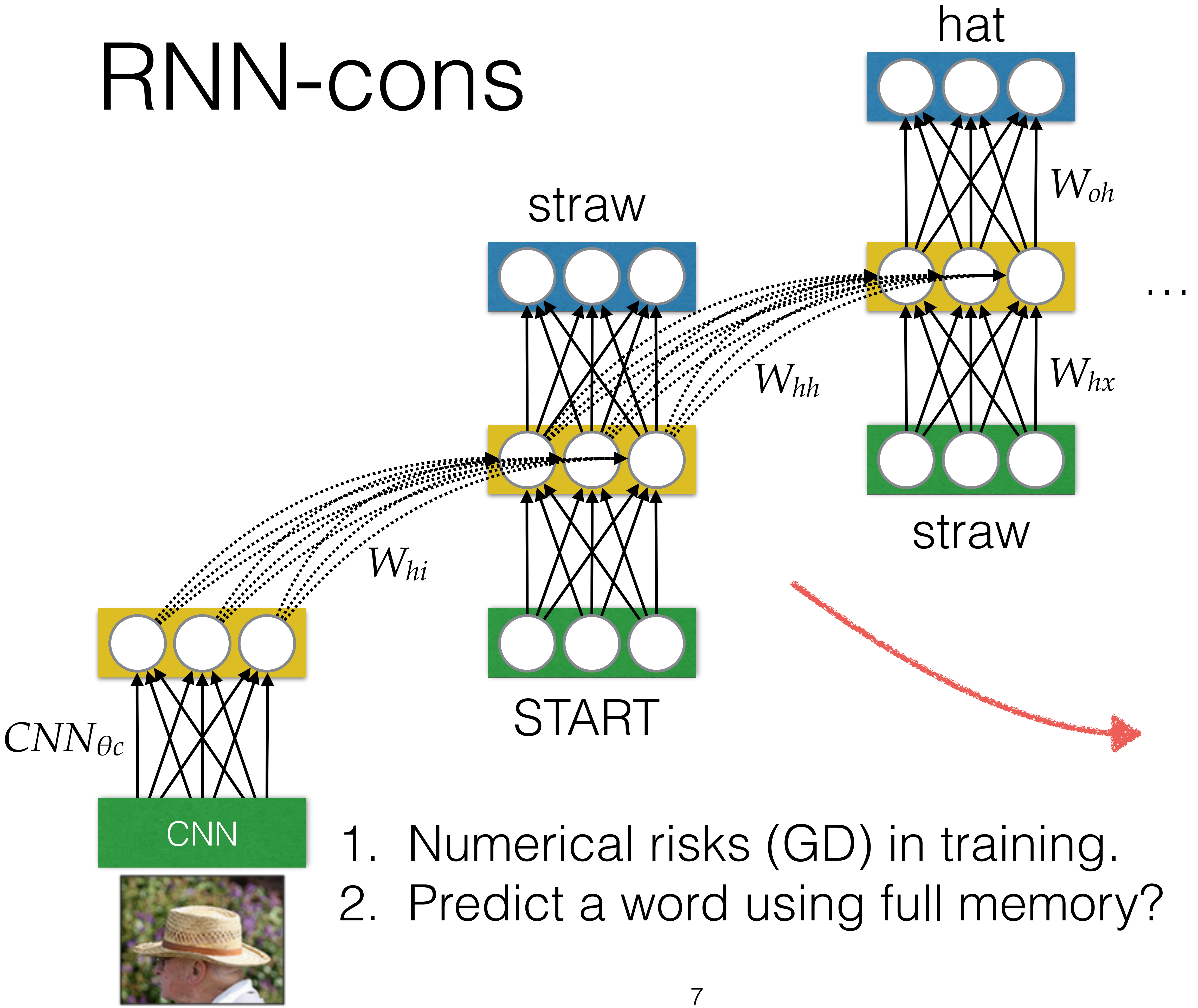


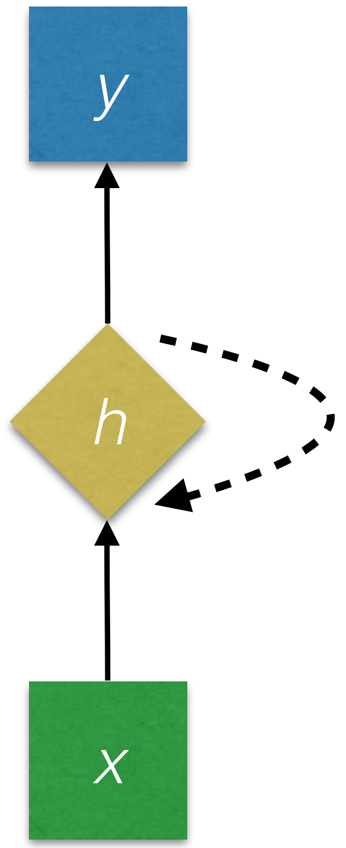
RNN pros



1. Visual and semantic features encoded into a common space.
2. In memory contains image content and predicted words.

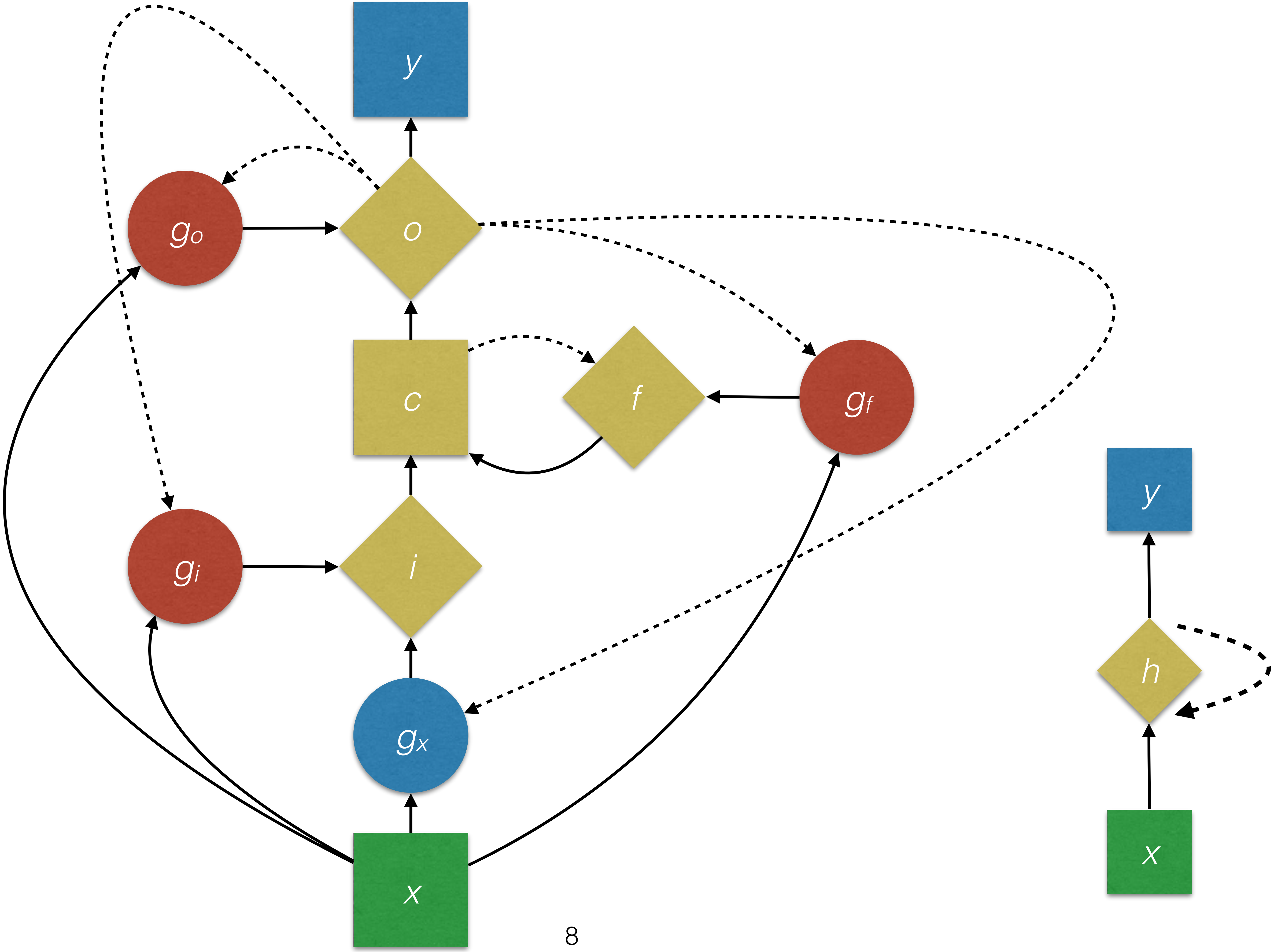
RNN-cons



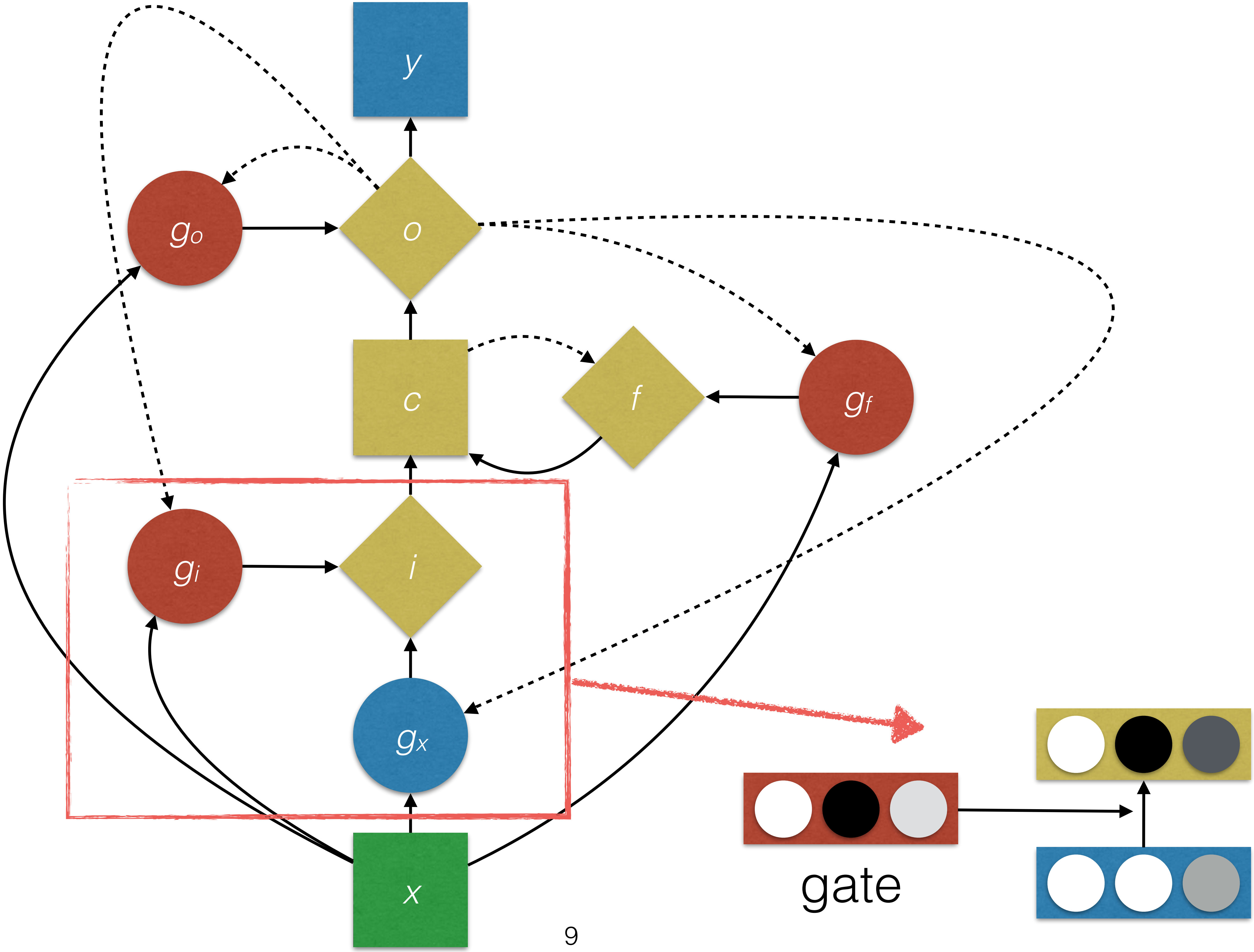


1. Numerical risks (GD) in training.
2. Predict a word using full memory?

LSTM-RNN

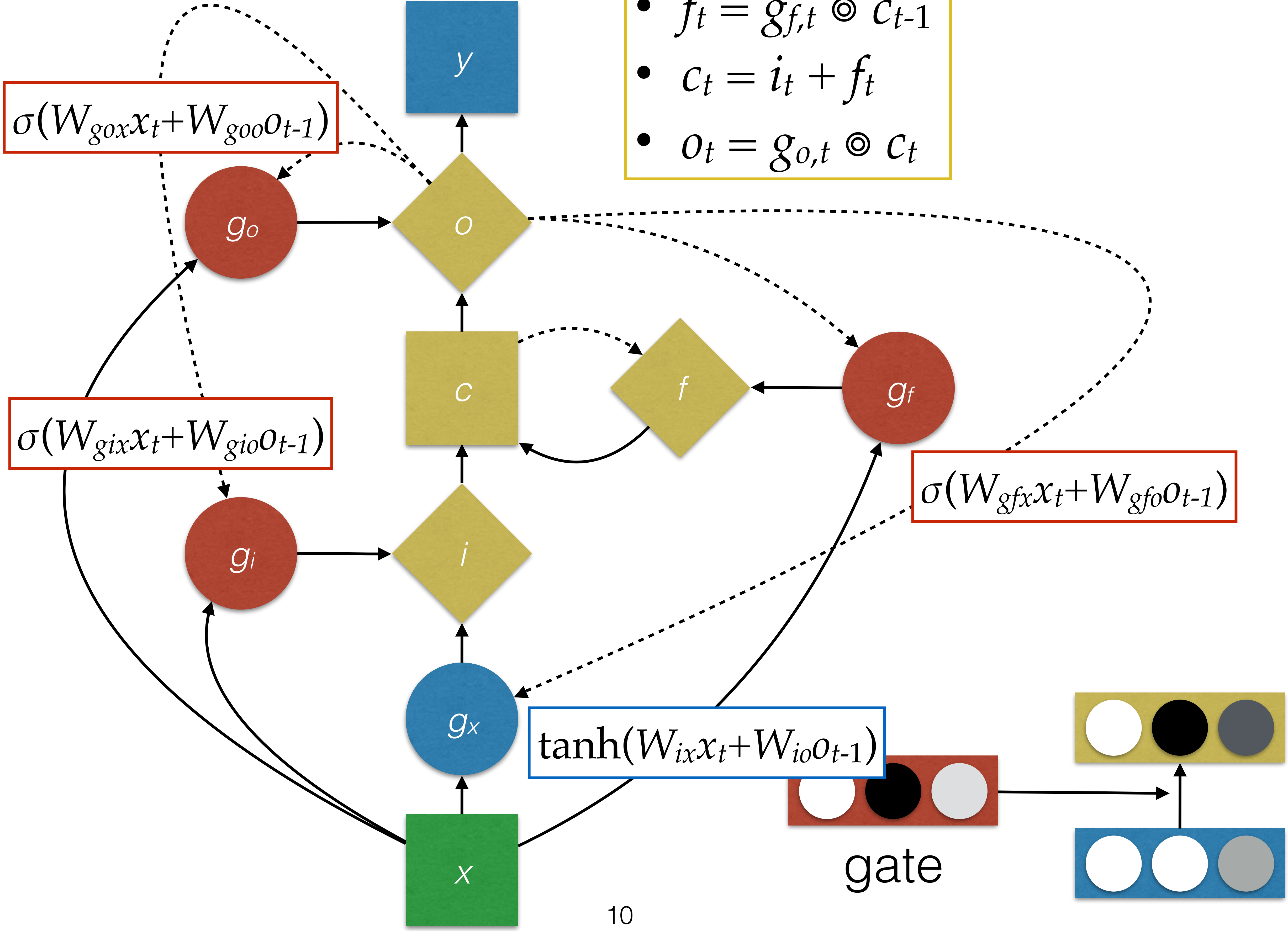


LSTM-RNN



LSTM-RNN

- $i_t = g_{i,t} \circledcirc g_{x,t}$
- $f_t = g_{f,t} \circledcirc c_{t\text{-}1}$
- $c_t = i_t + f_t$
- $o_t = g_{o,t} \circledcirc c_t$



$\sigma(W_{gox}x_t+W_{goo}o_{t\text{-}1})$

$\sigma(W_{gix}x_t+W_{gio}o_{t\text{-}1})$

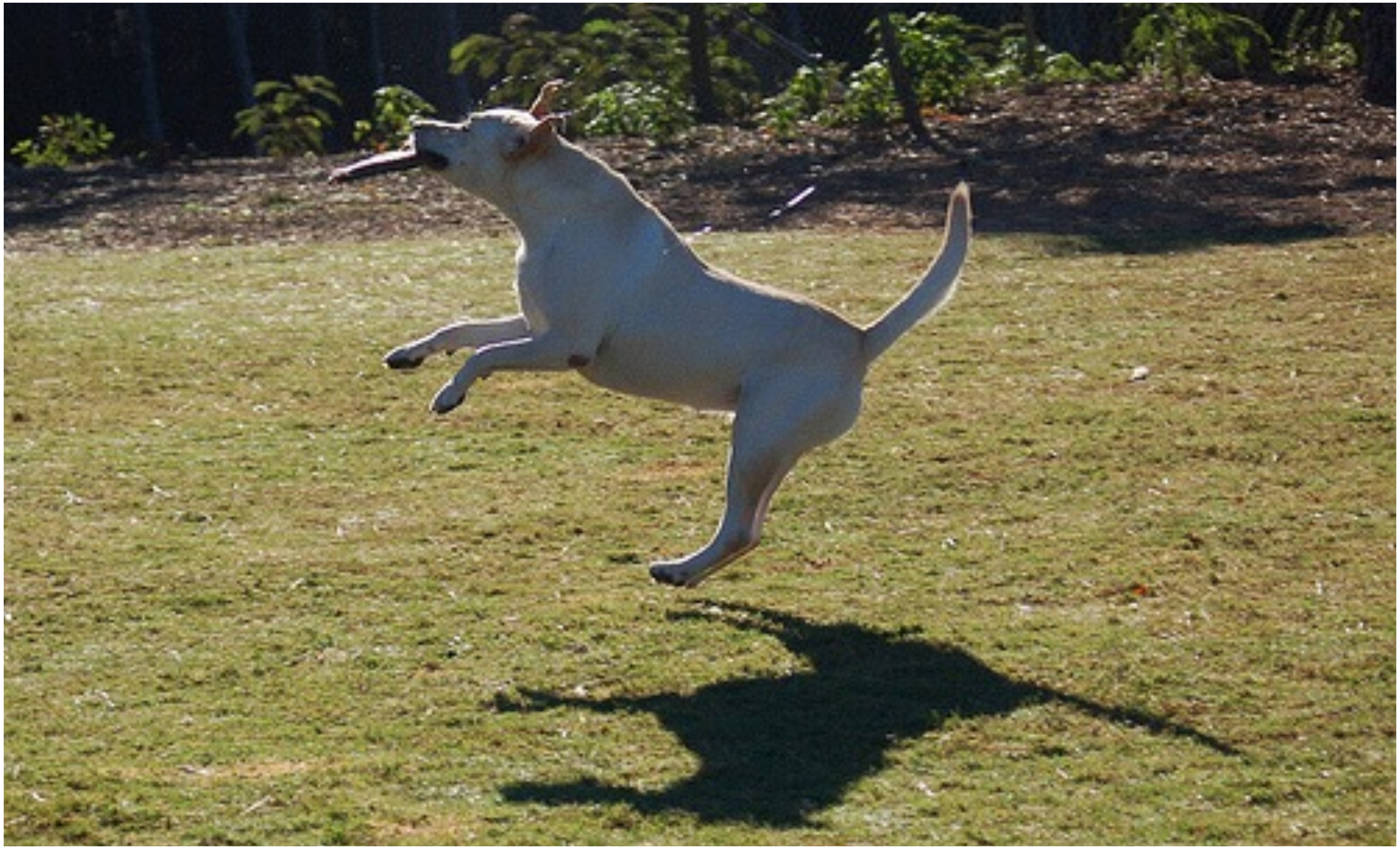
$\sigma(W_{gfx}x_t+W_{gfo}o_{t\text{-}1})$

$\tanh(W_{ix}x_t+W_{io}o_{t\text{-}1})$

gate

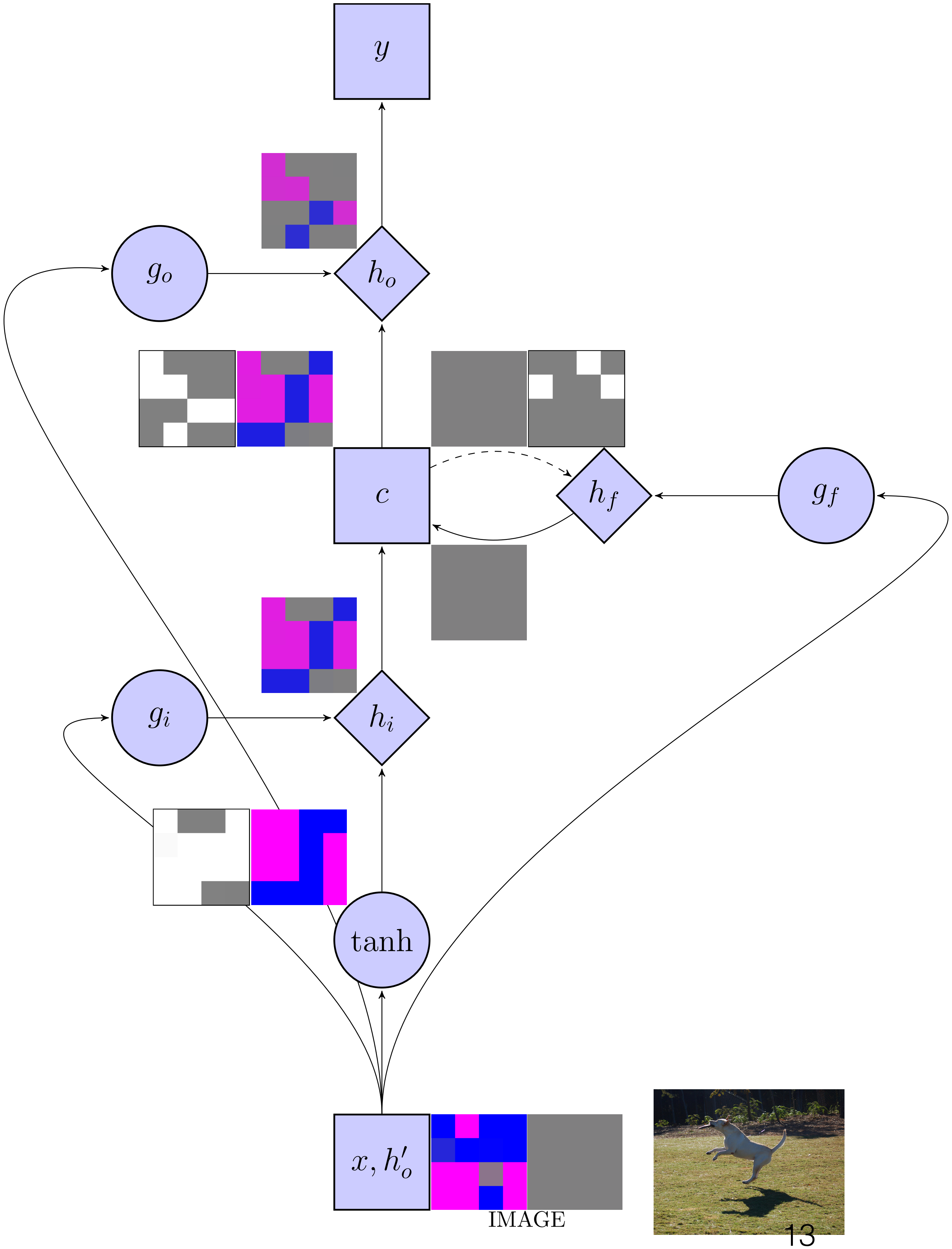
Toy Experiment

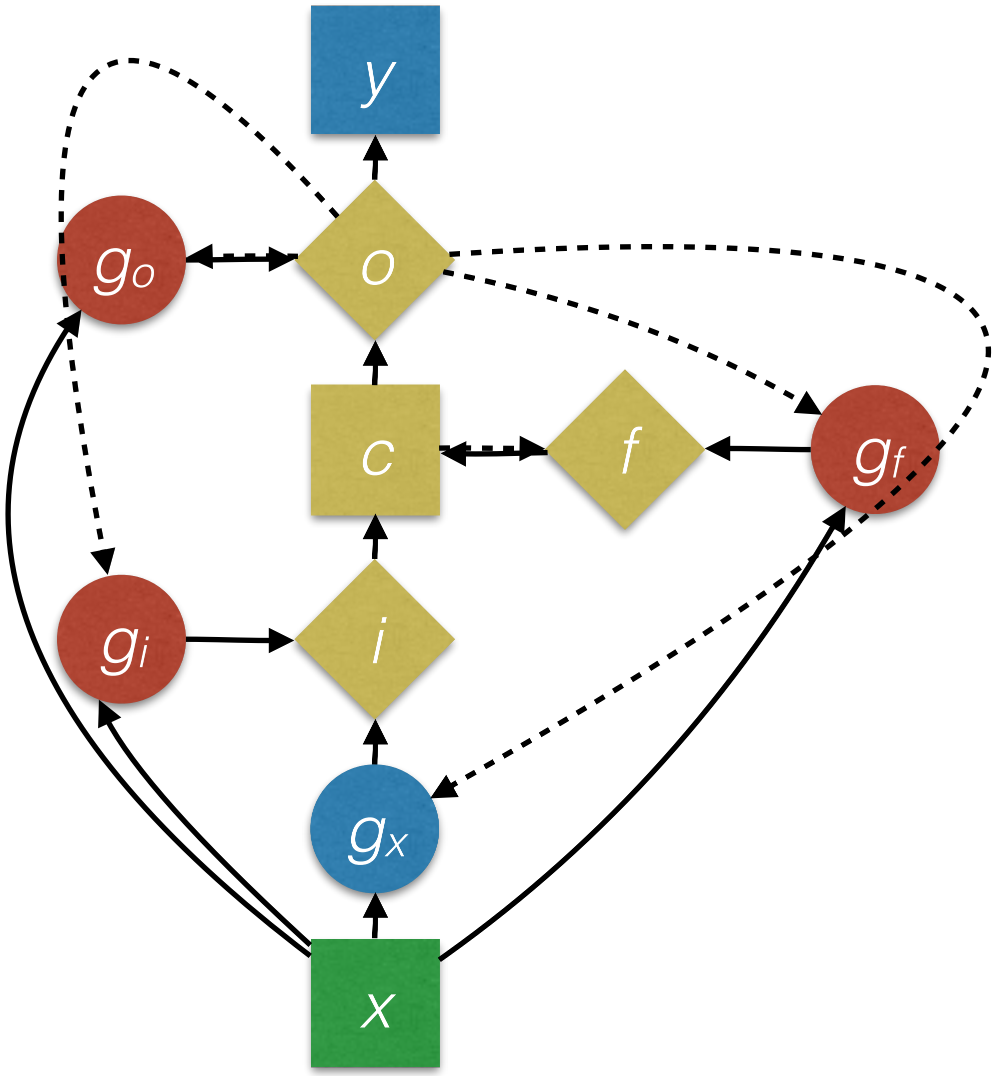
- Training set (407)
  - dog & frisbee: 59
  - man & ride: 324
  - kiss: 24



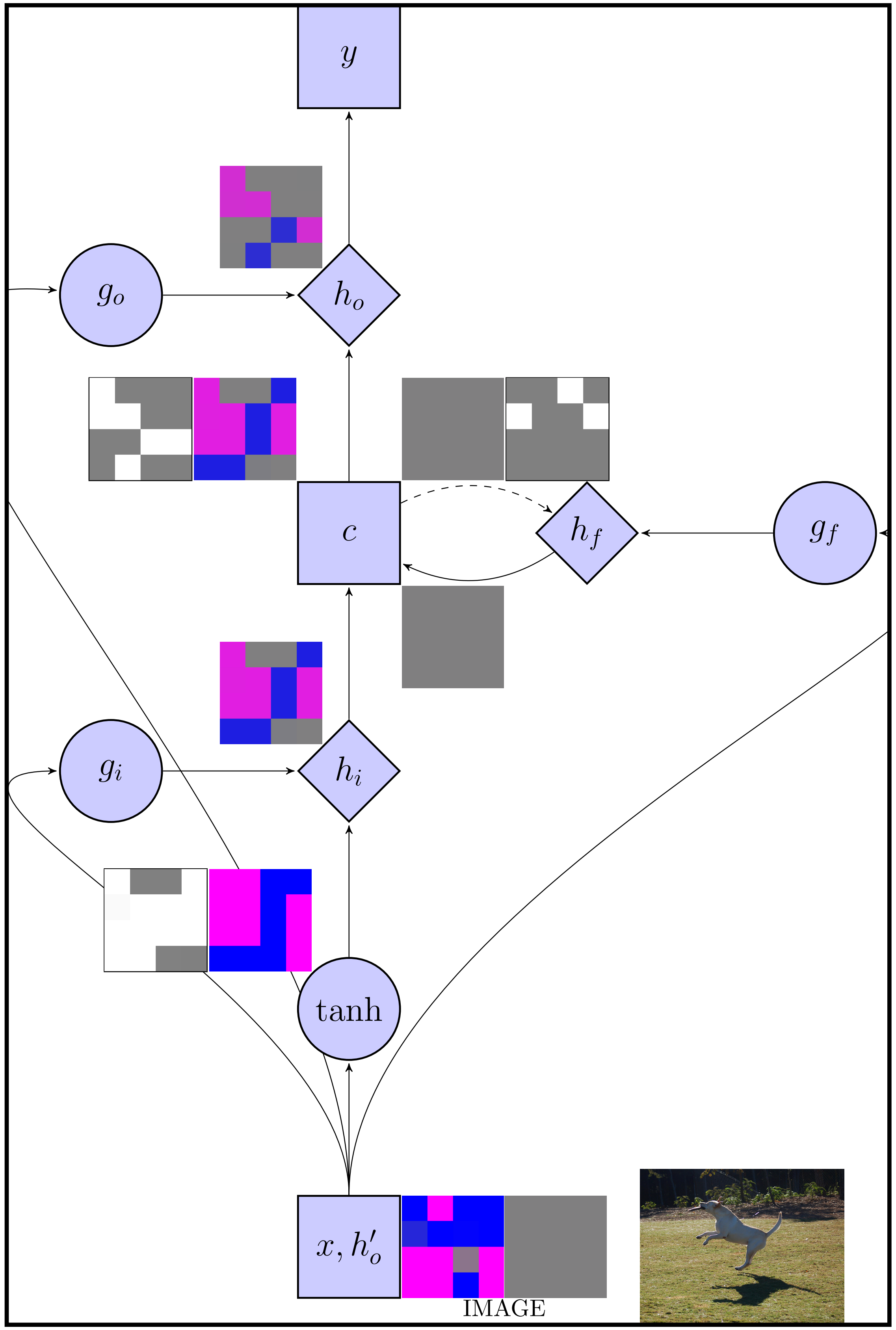
a dog jumps to catch a frisbee .

1674612291_7154c5ab61.jpg

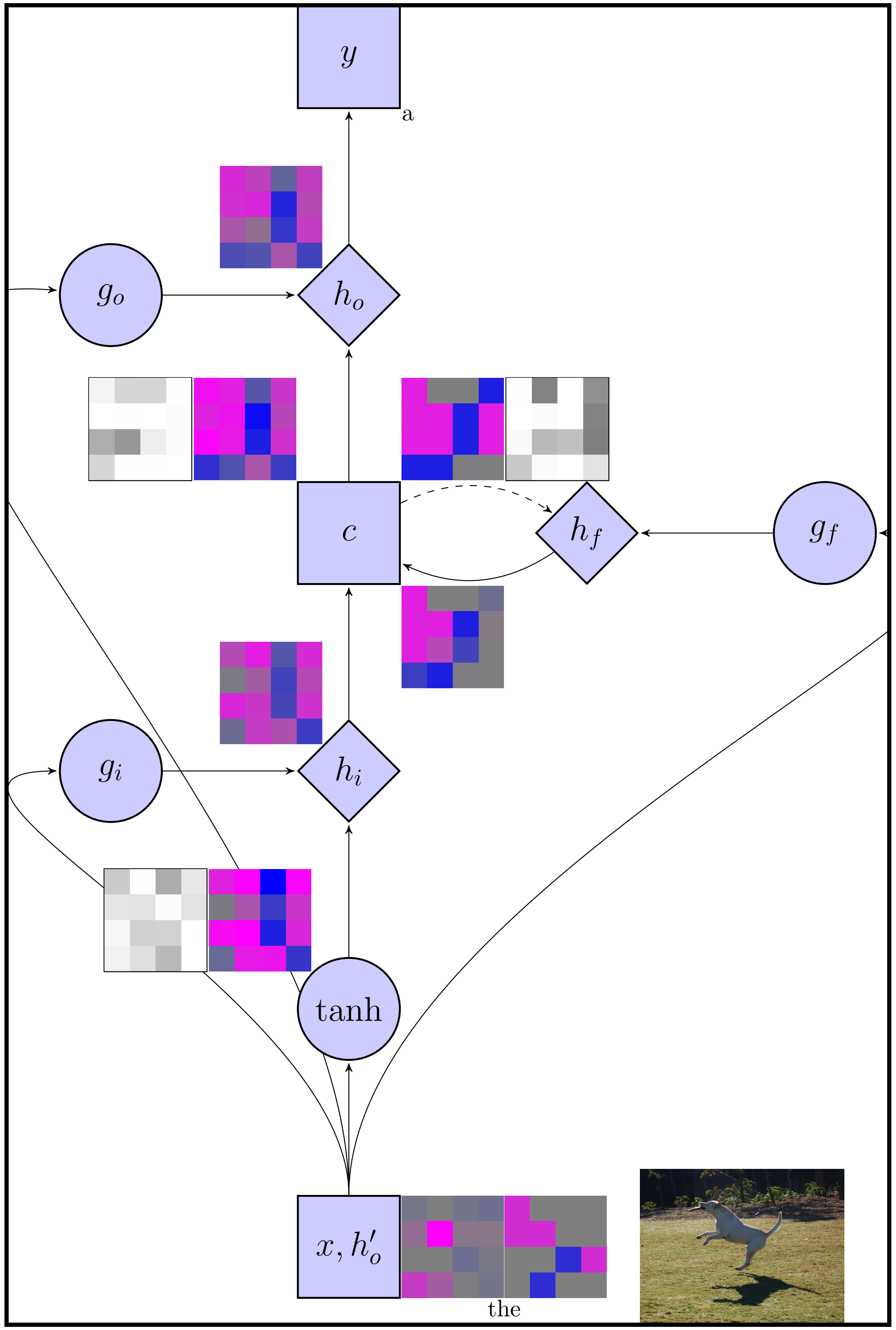

y
g_o
h_o
c
h_f
g_f
g_i
h_i
tanh
x, h'_o
IMAGE


y
go
o
c
f
gf
gi
i
gx
x

a dog jumps to catch a frisbee .

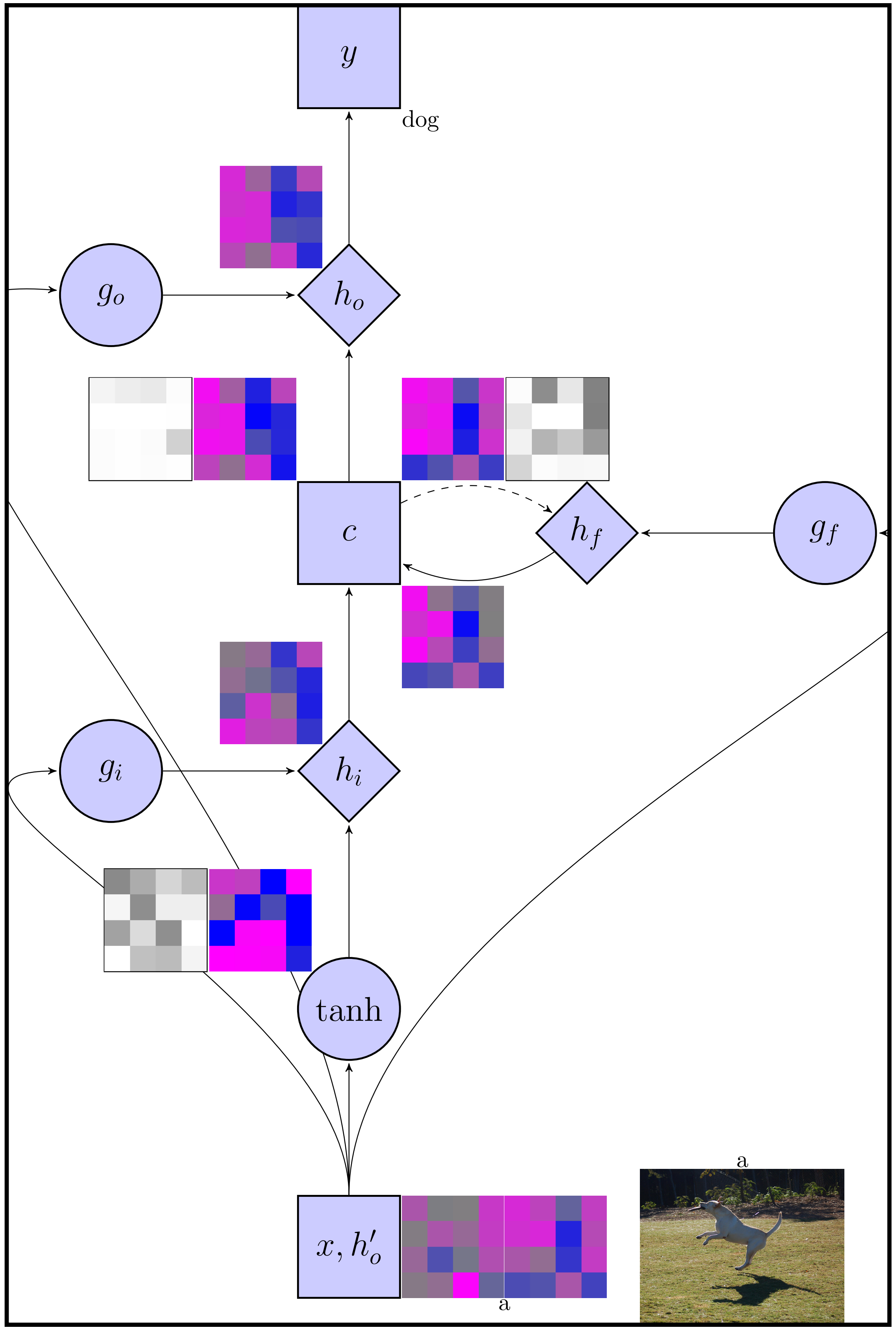


0

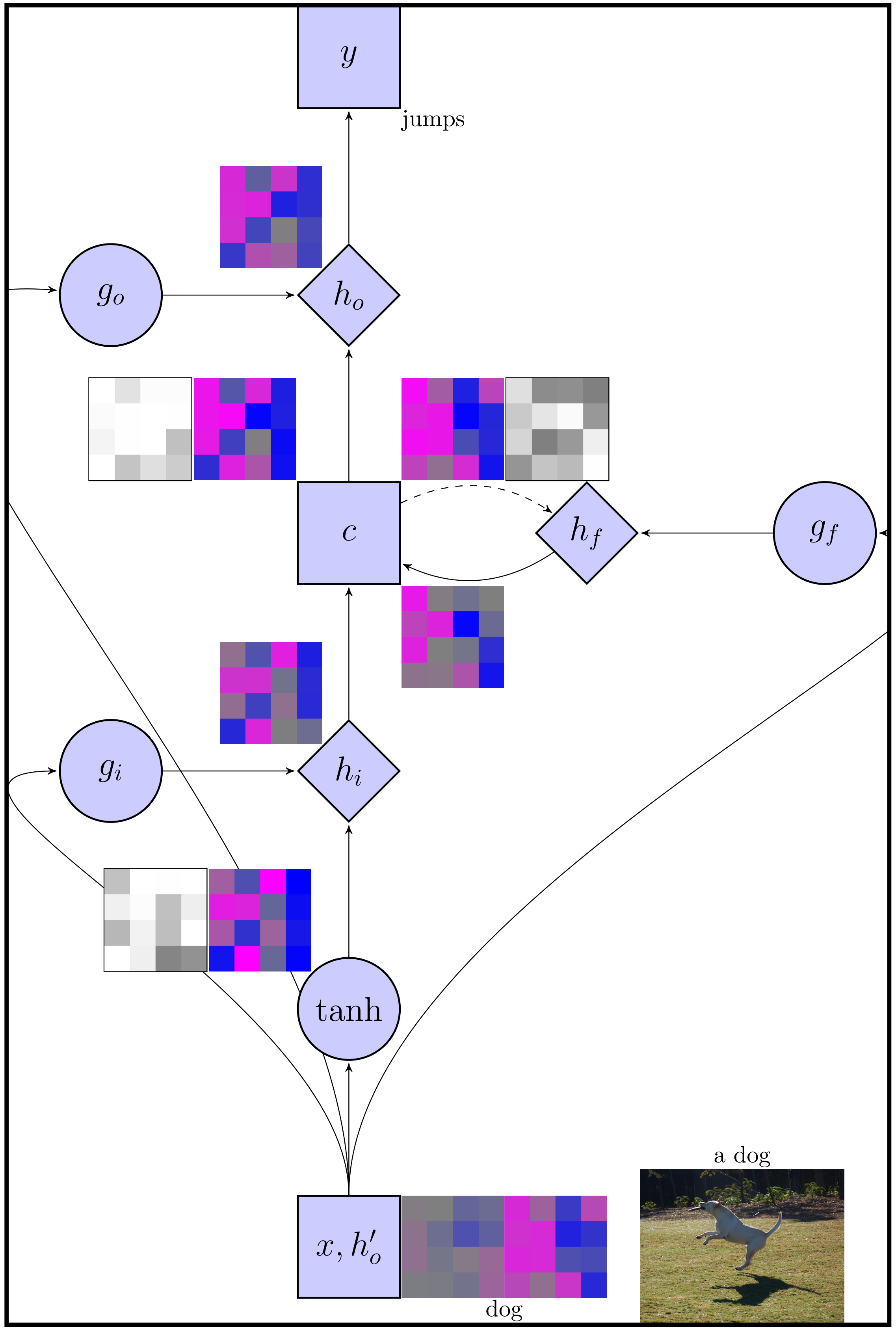


1

a dog jumps to catch a frisbee .

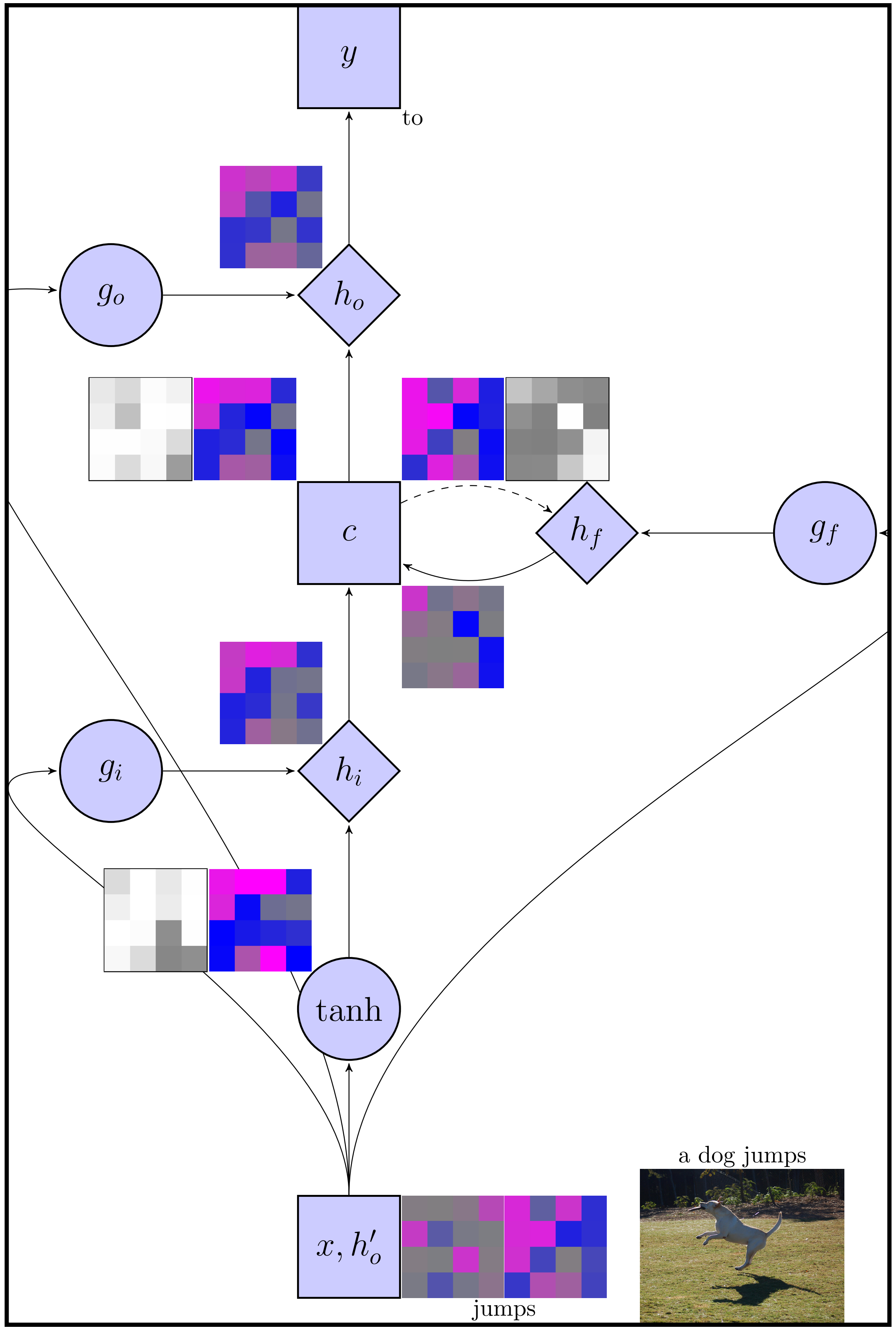


2

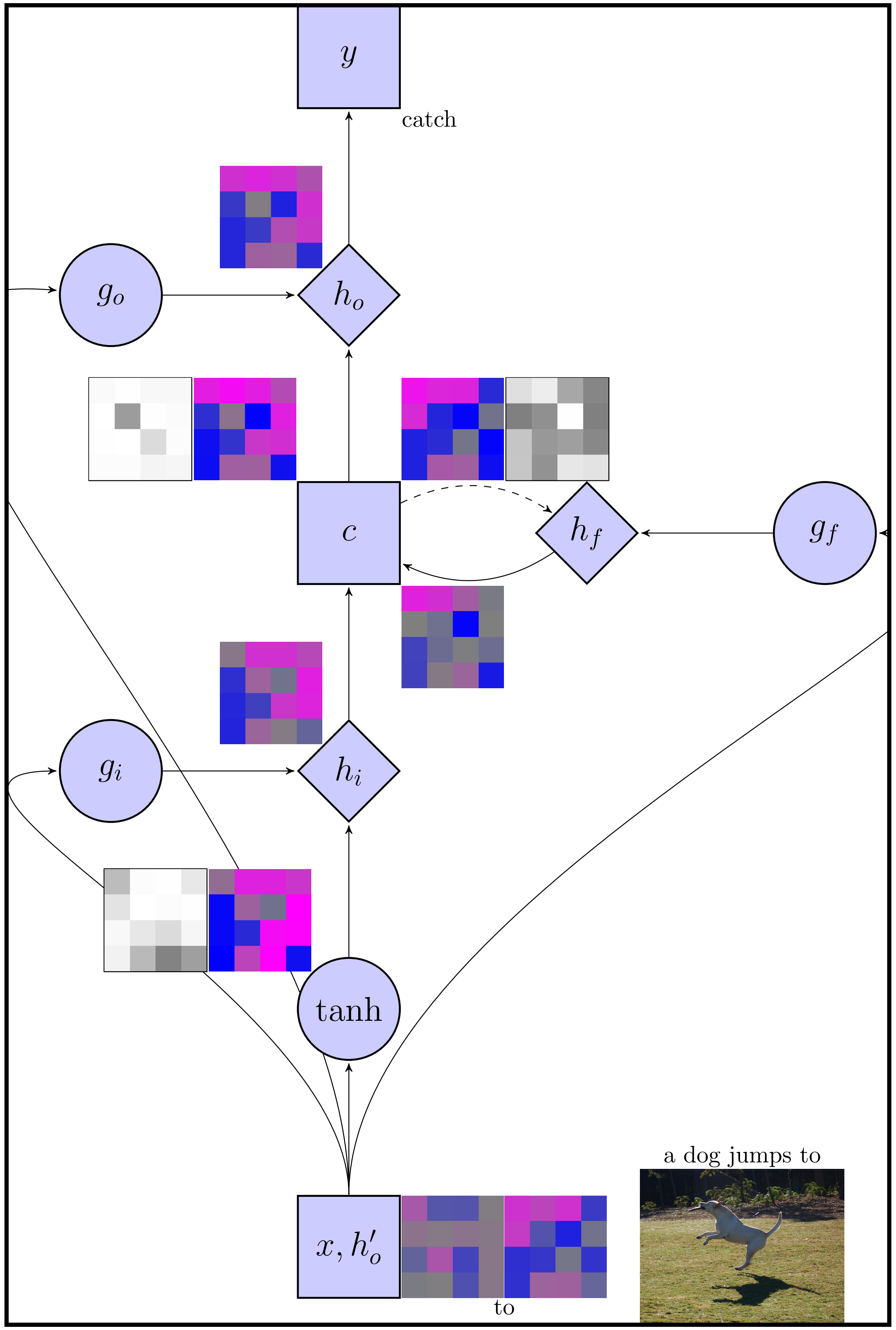


3

a dog jumps to catch a frisbee .

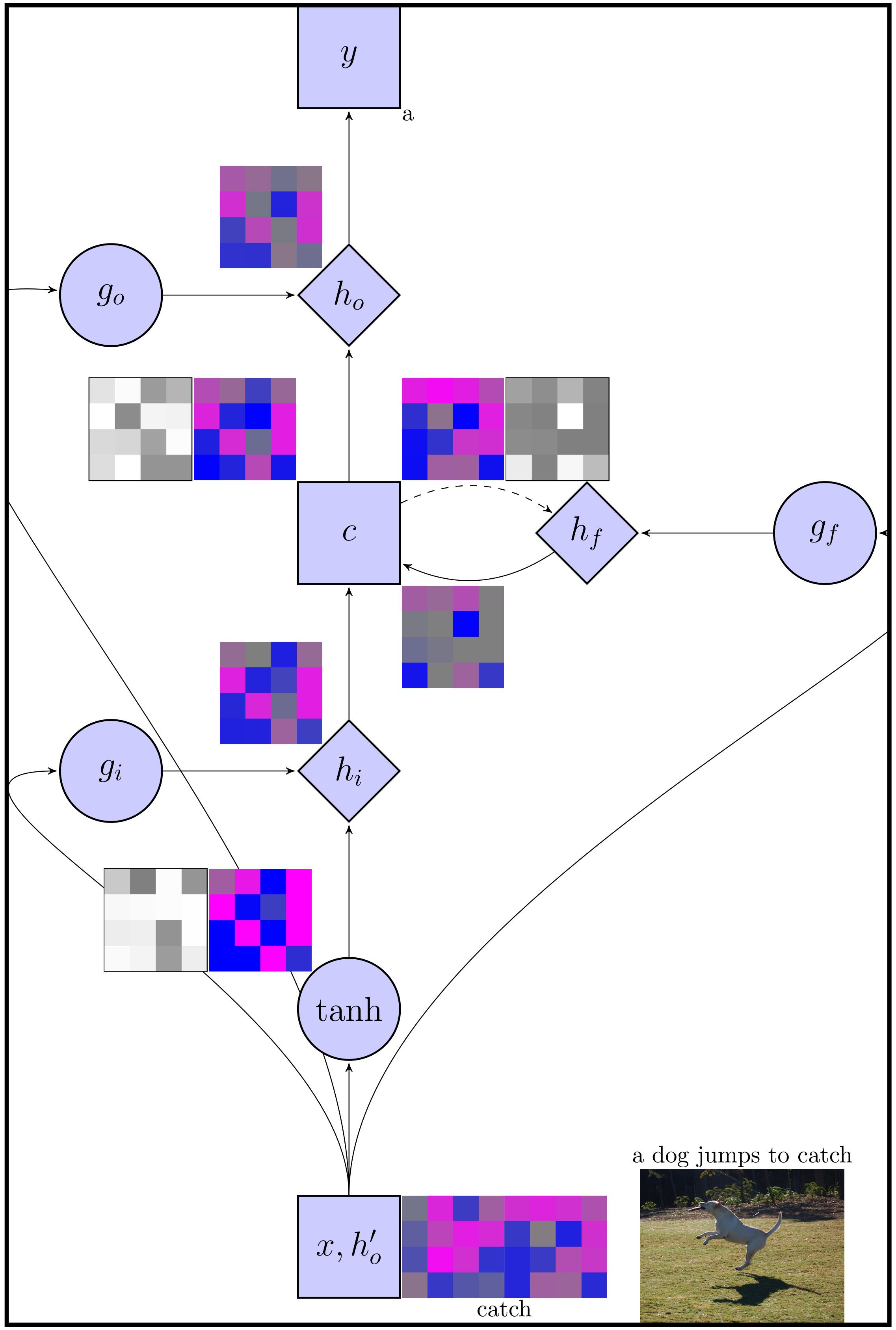


4

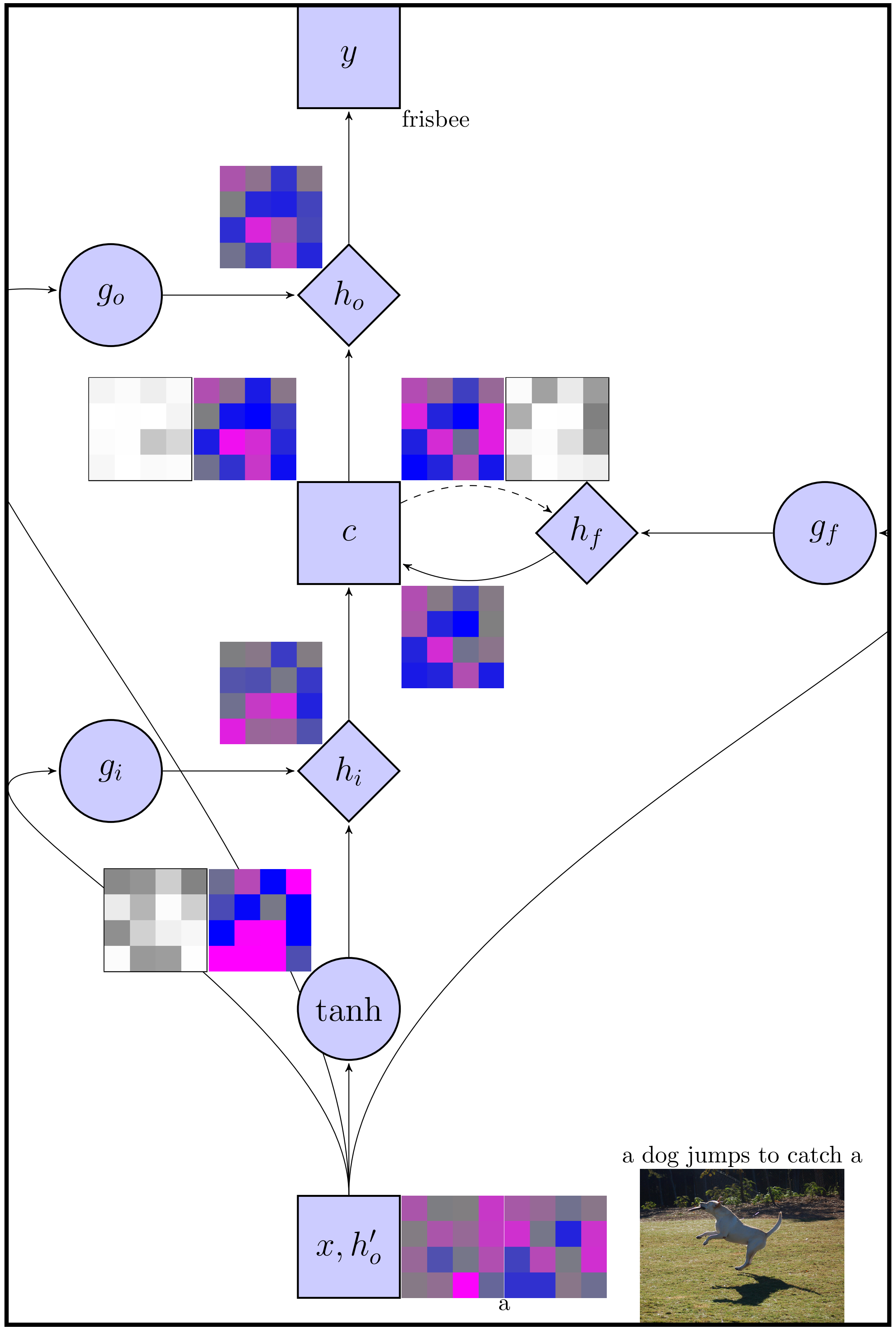


5

a dog jumps to catch a frisbee .

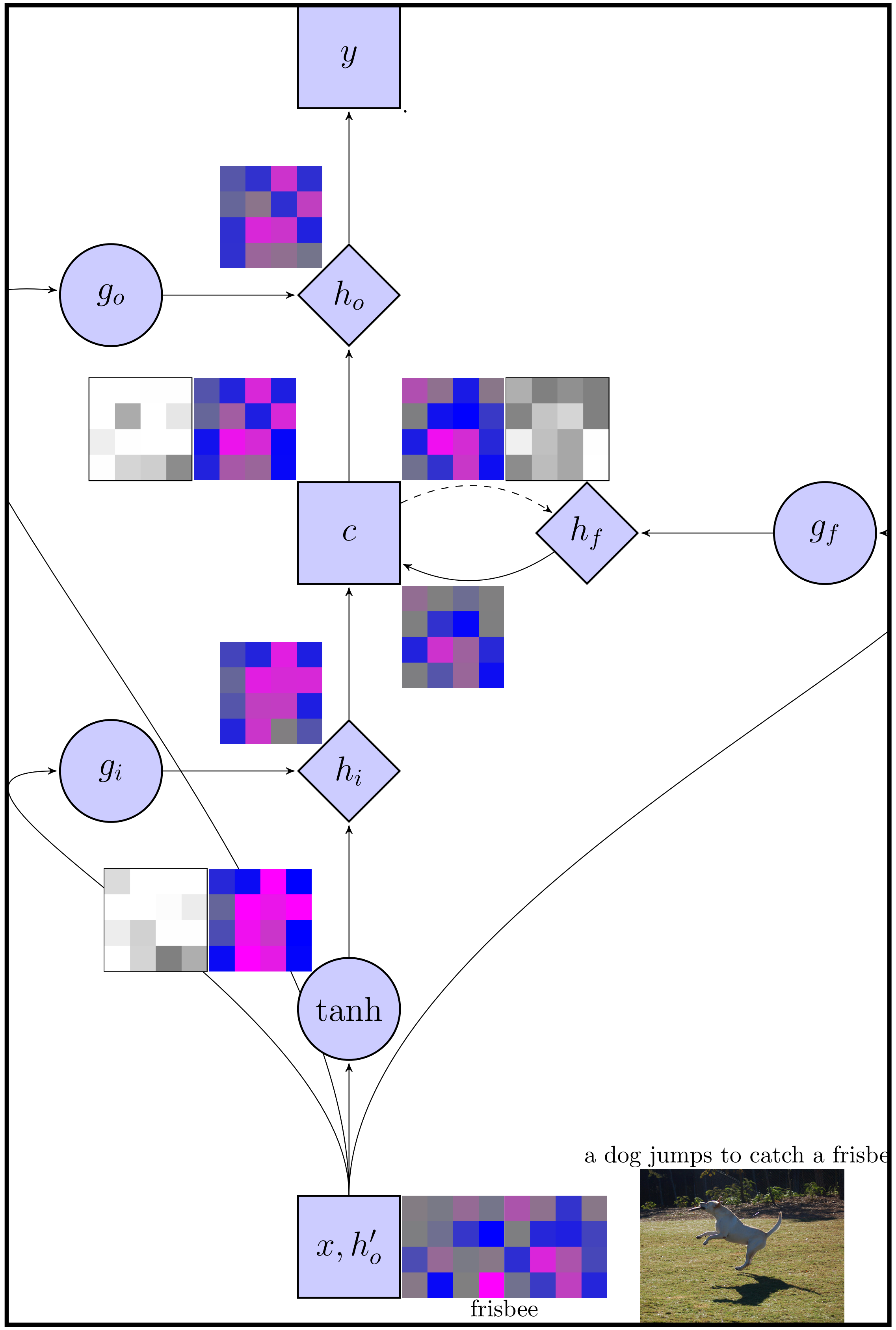


6



7

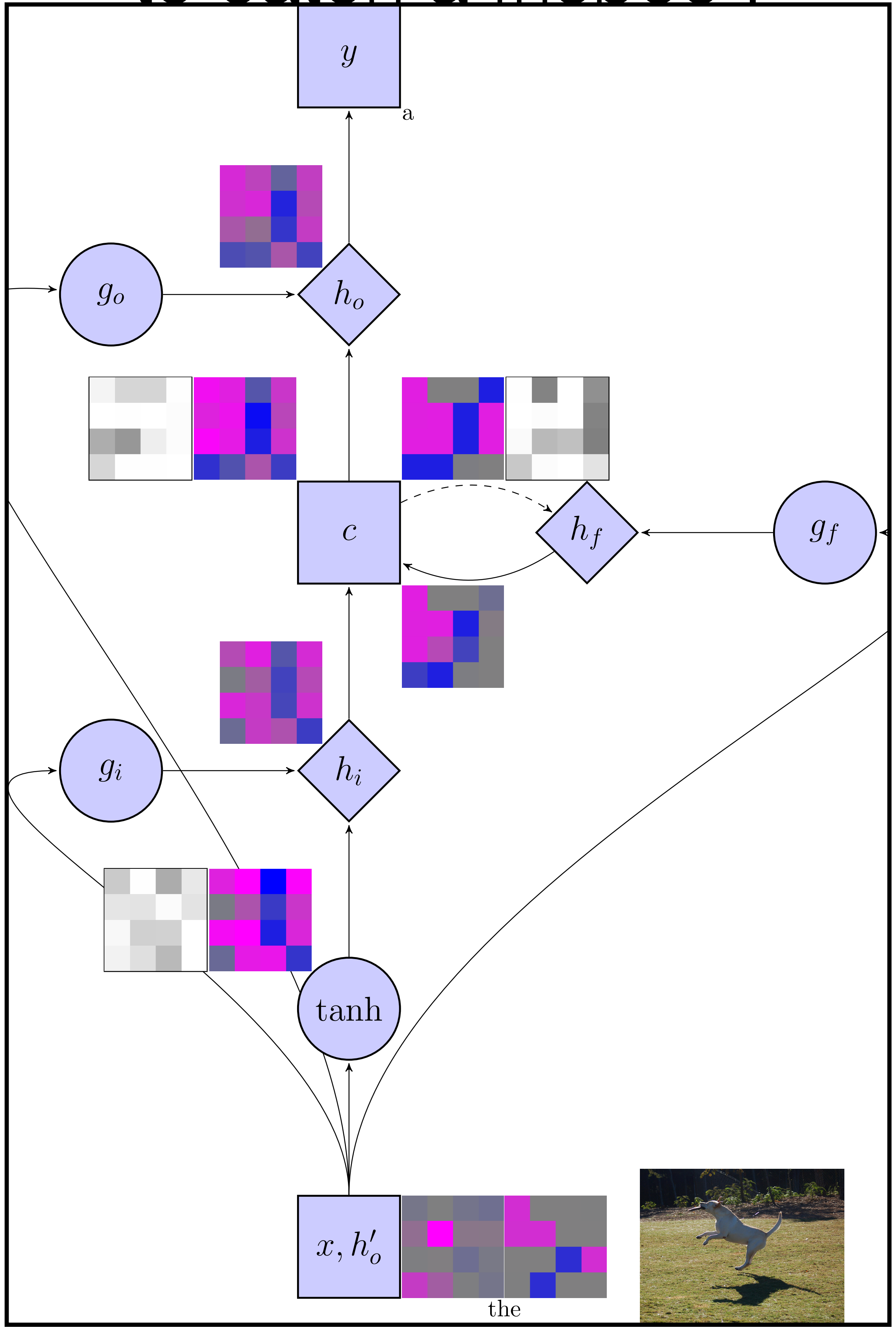
a dog jumps to catch a frisbee .



8

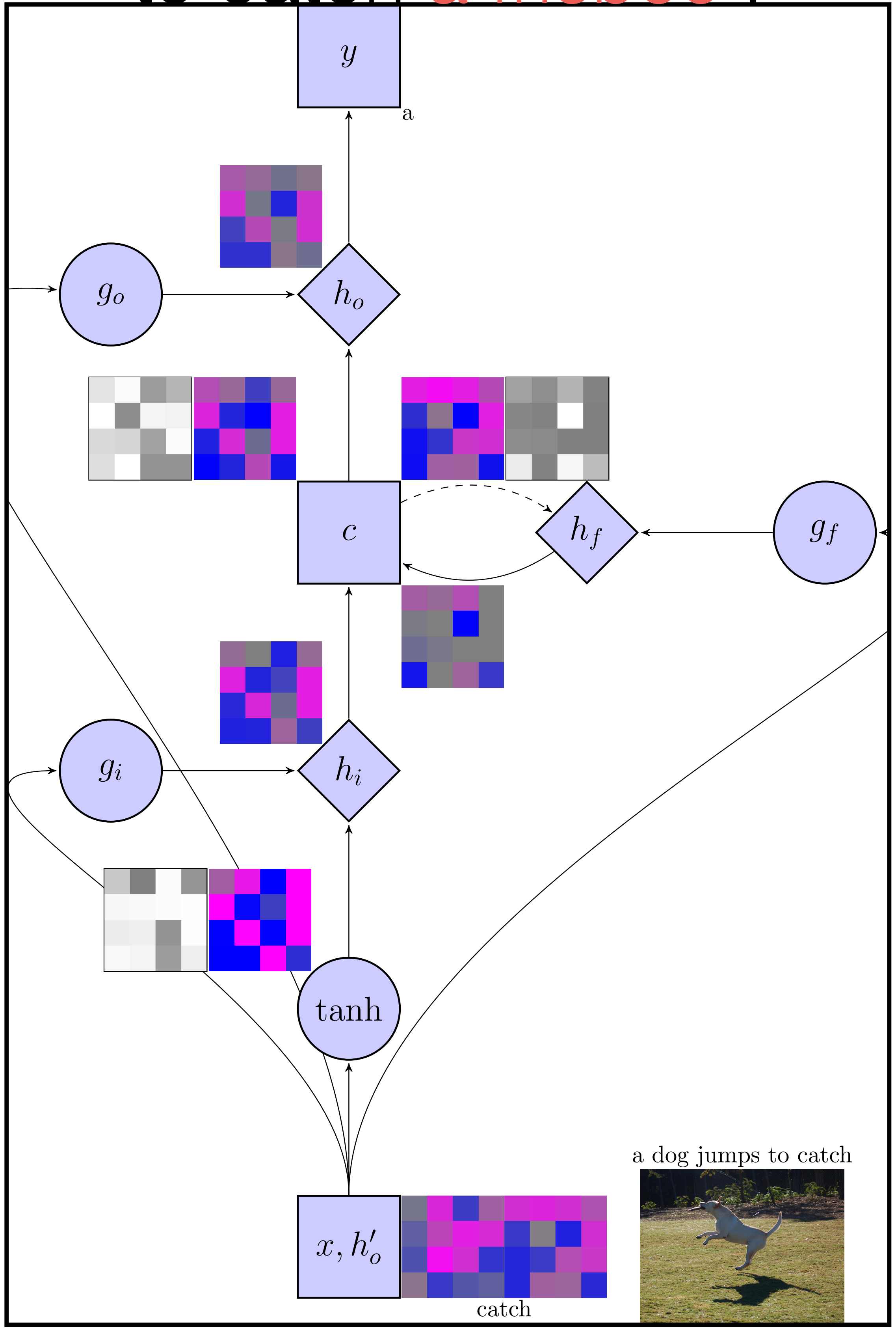
Inherit previous memory
Acknowledge previous word
Update current memory
Predict next word
Until all memory fades out

a dog jumps
to catch a frisbee .



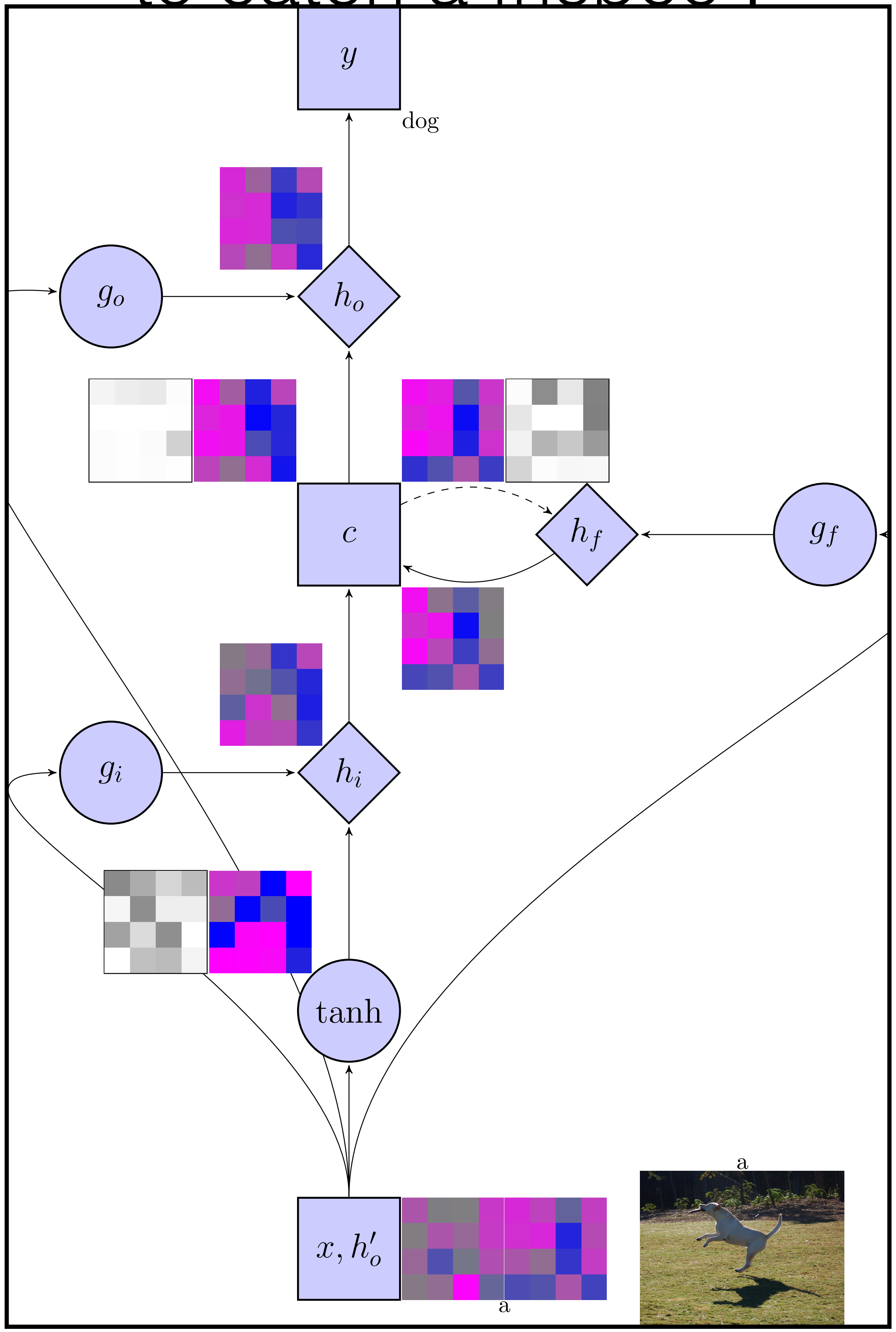
1

a dog jumps
to catch a frisbee .



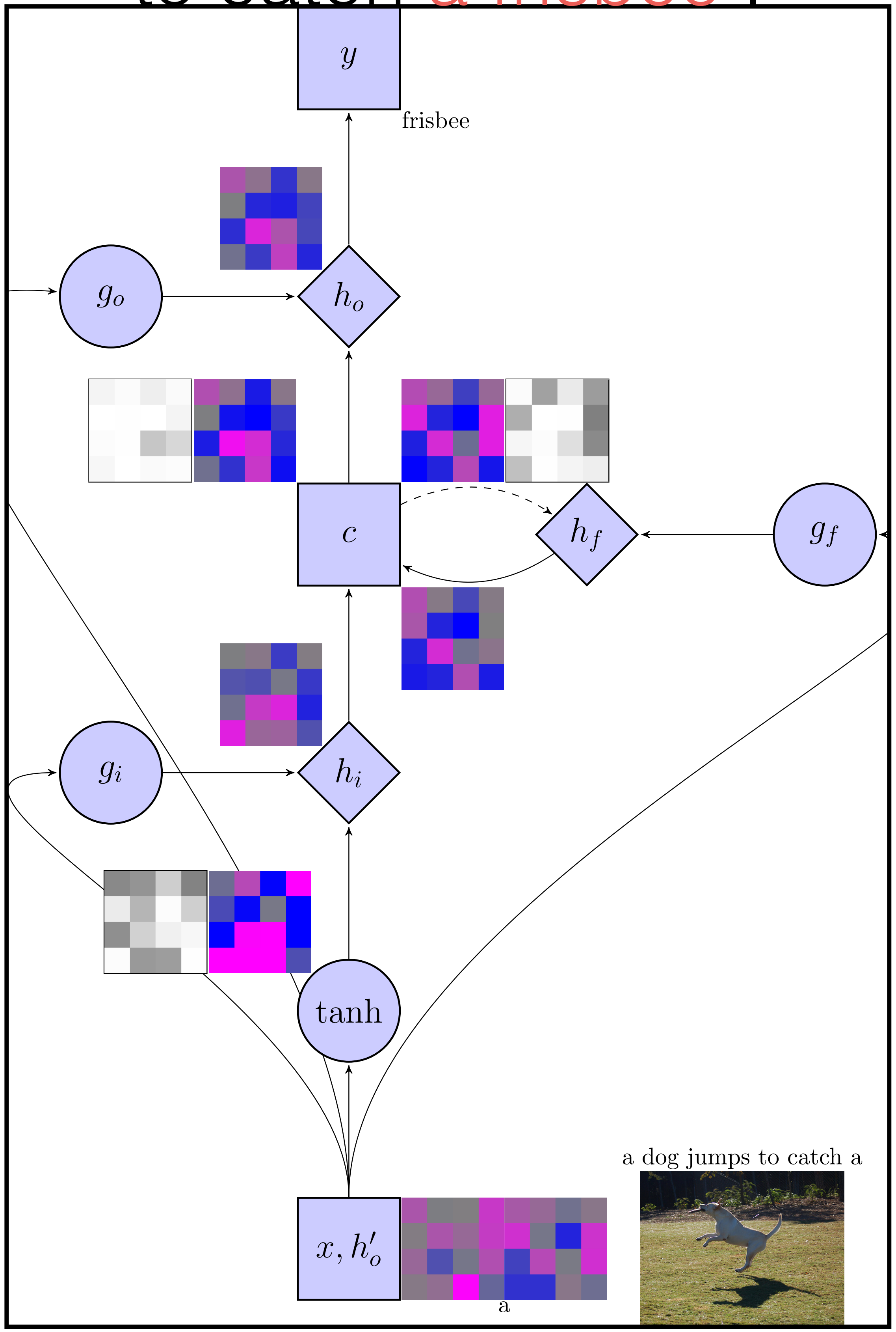
6

a dog jumps
to catch a frisbee .



2

a dog jumps
to catch a frisbee .



7

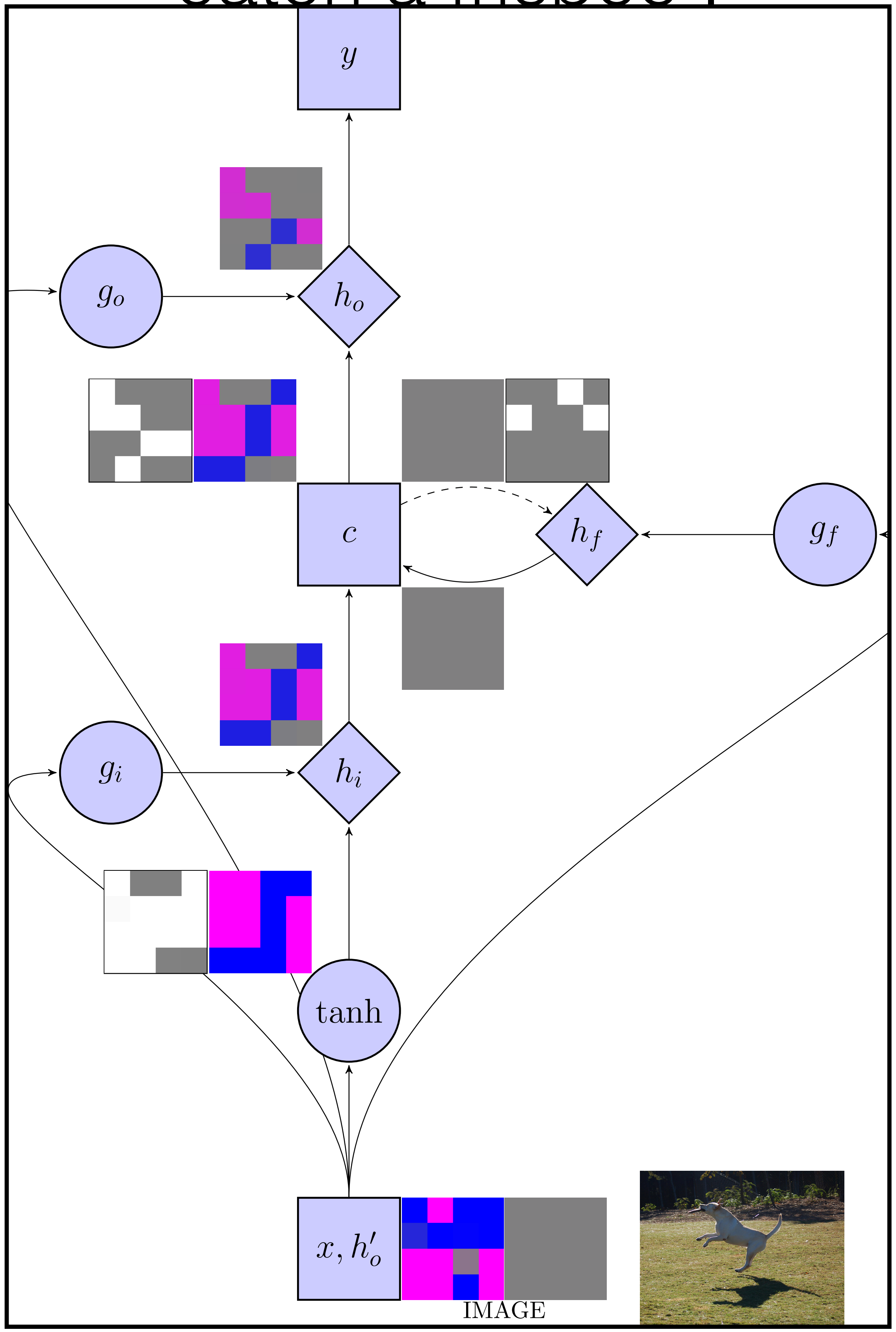
a dog jumps to catch a frisbee .

1674612291_7154c5ab61.jpg

a dog is jumping to catch a frisbee .

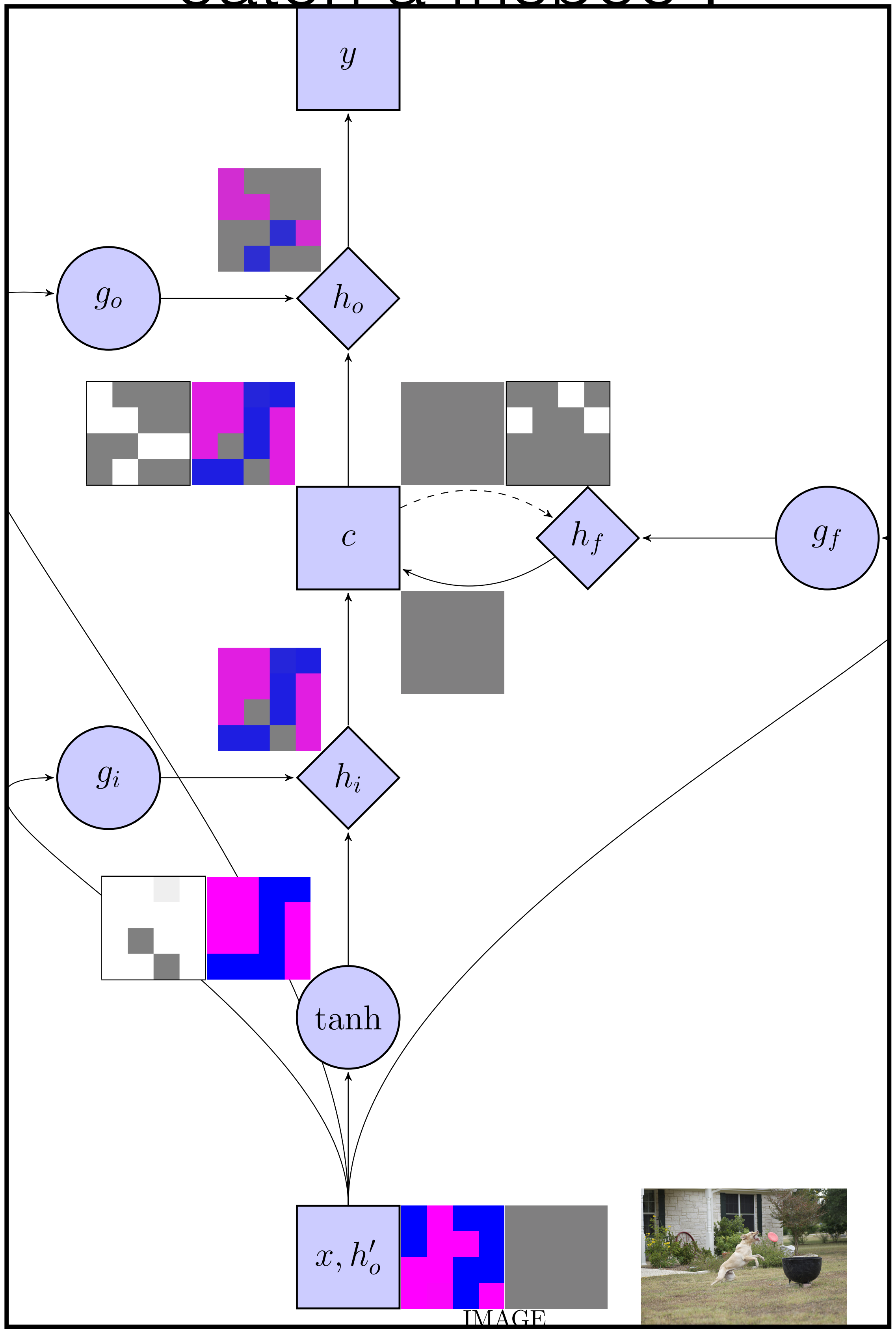
2945036454_280fa5b29f.jpg

a dog jumps to catch a frisbee .



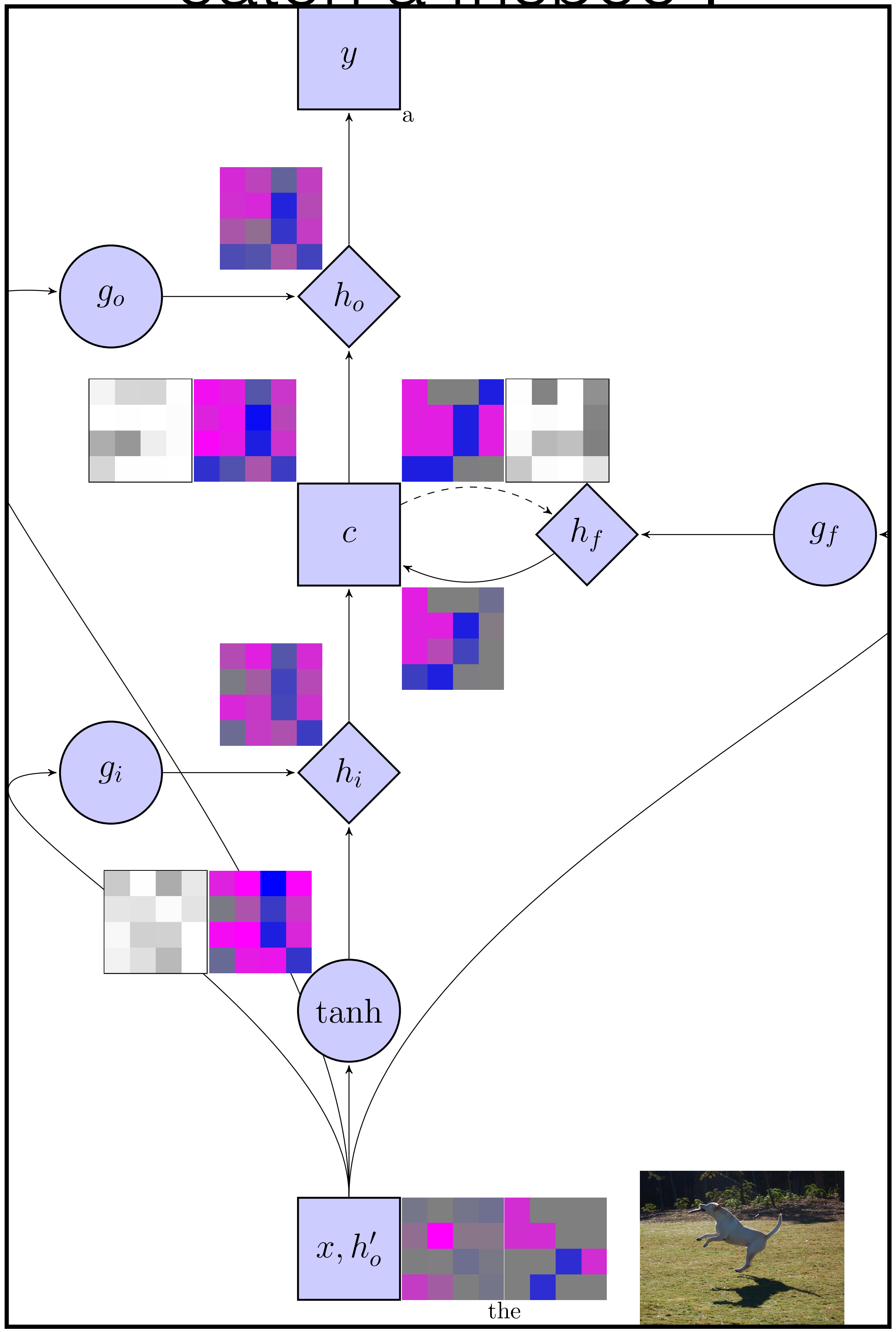
0

a dog is jumping to catch a frisbee .



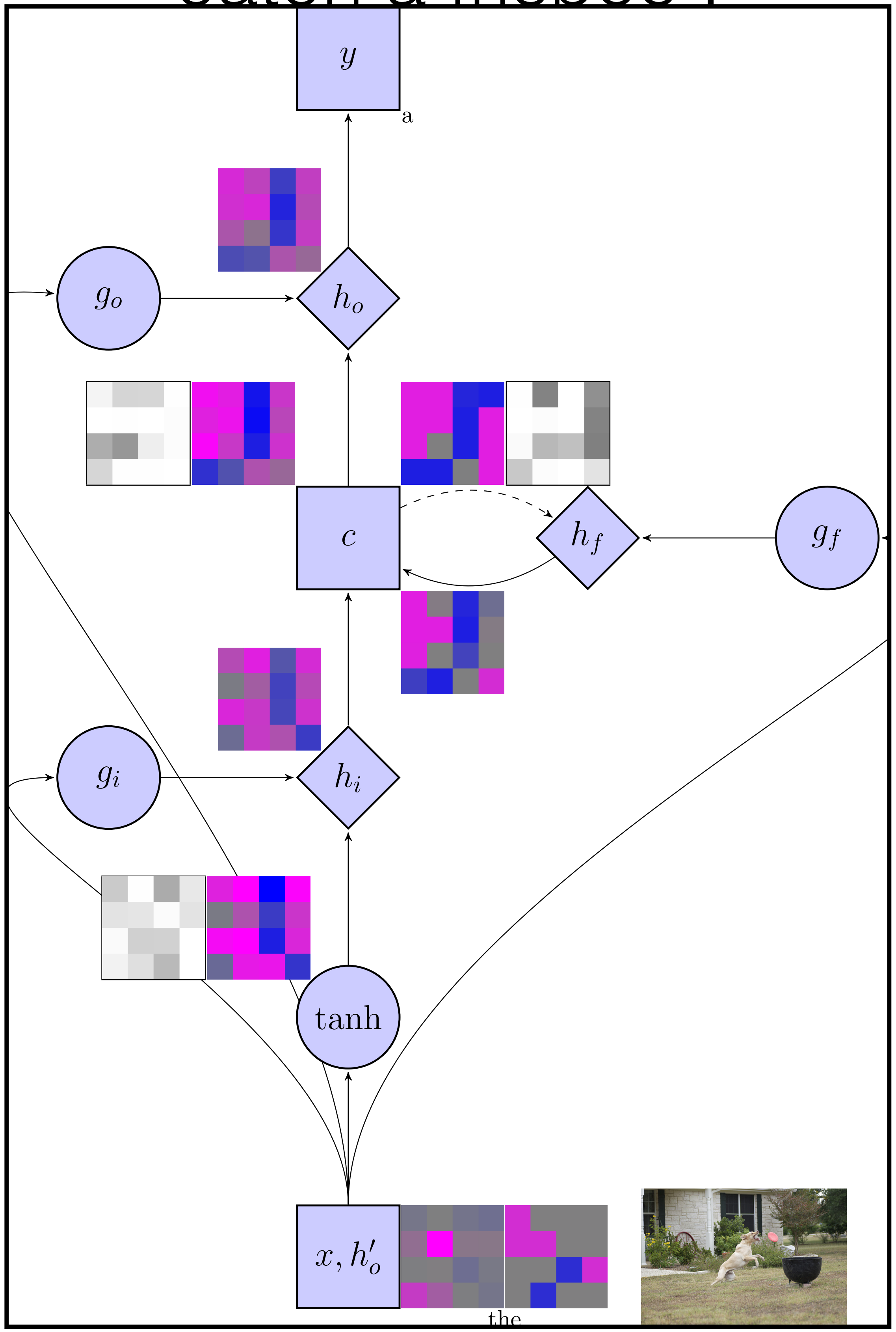
0

a dog jumps to catch a frisbee .



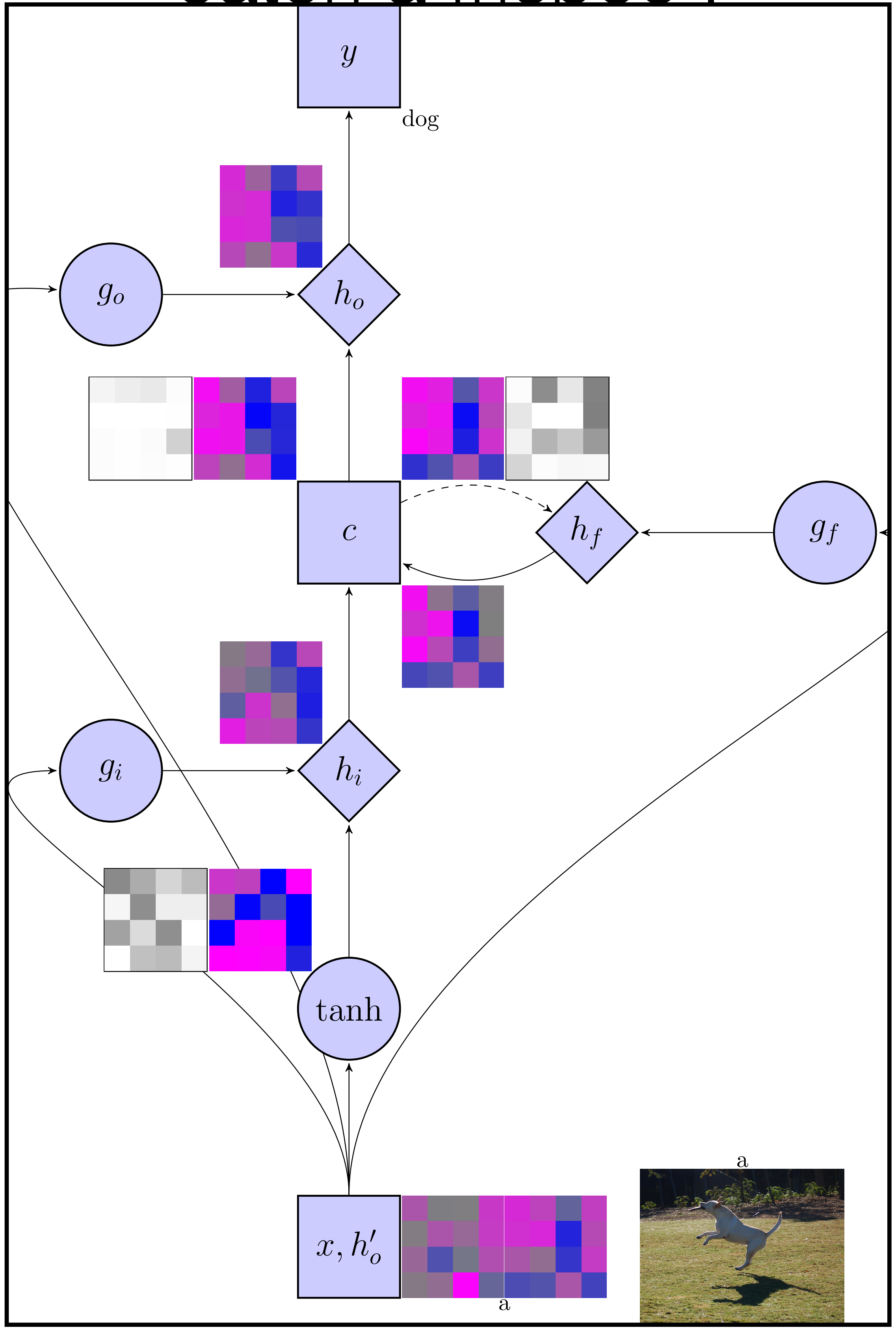
1

a dog is jumping to catch a frisbee .



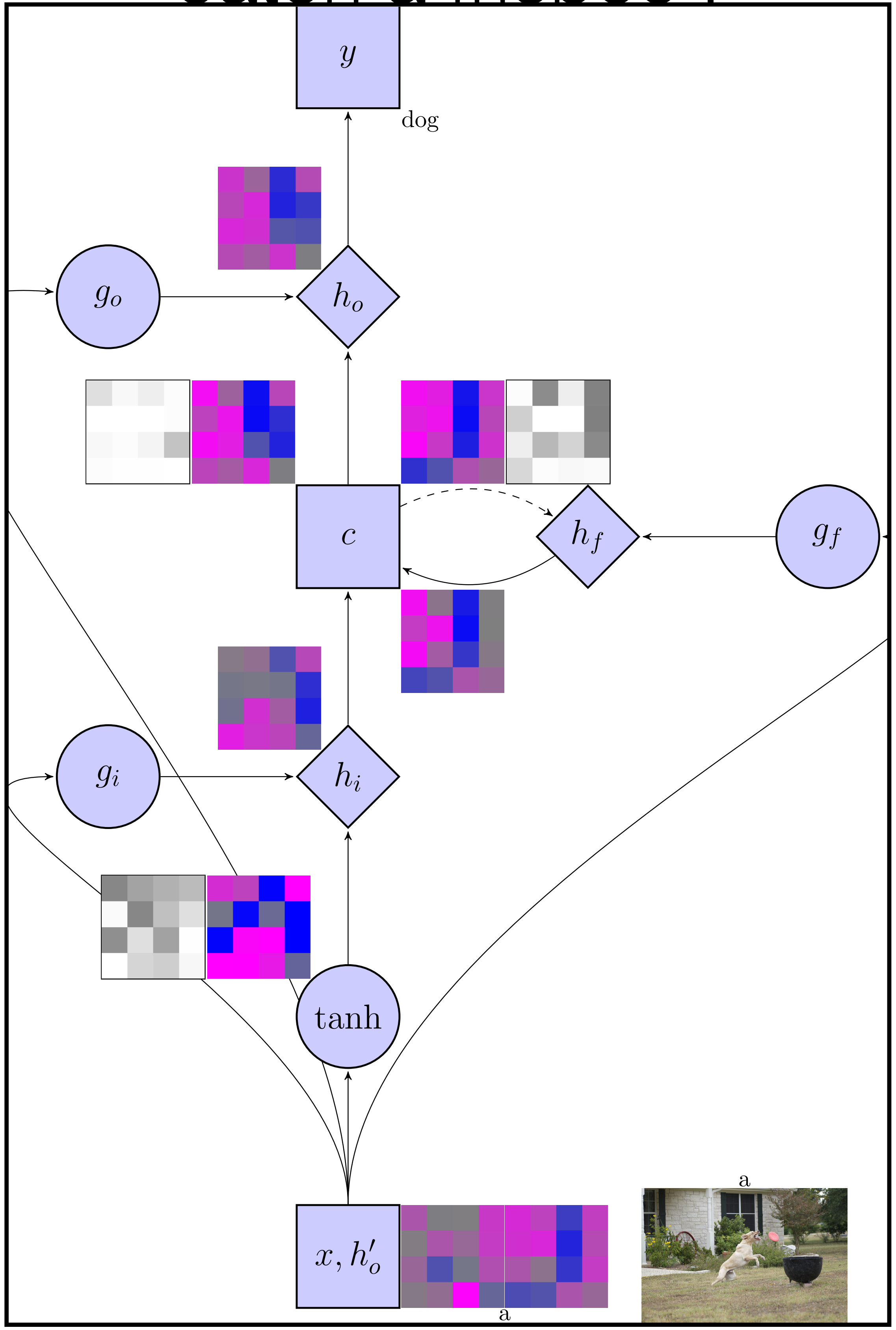
1

a dog jumps to catch a frisbee .



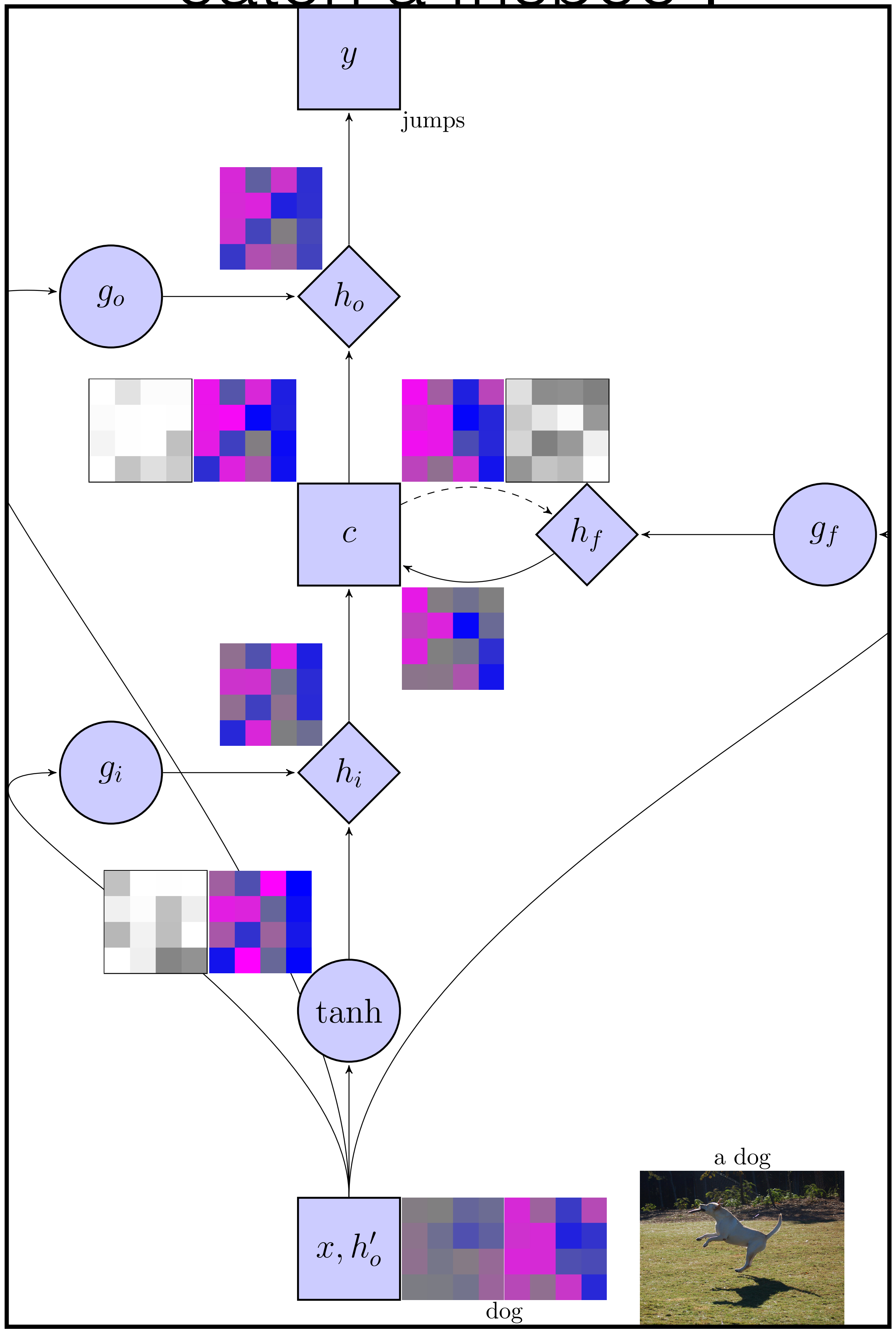
2

a dog is jumping to catch a frisbee .



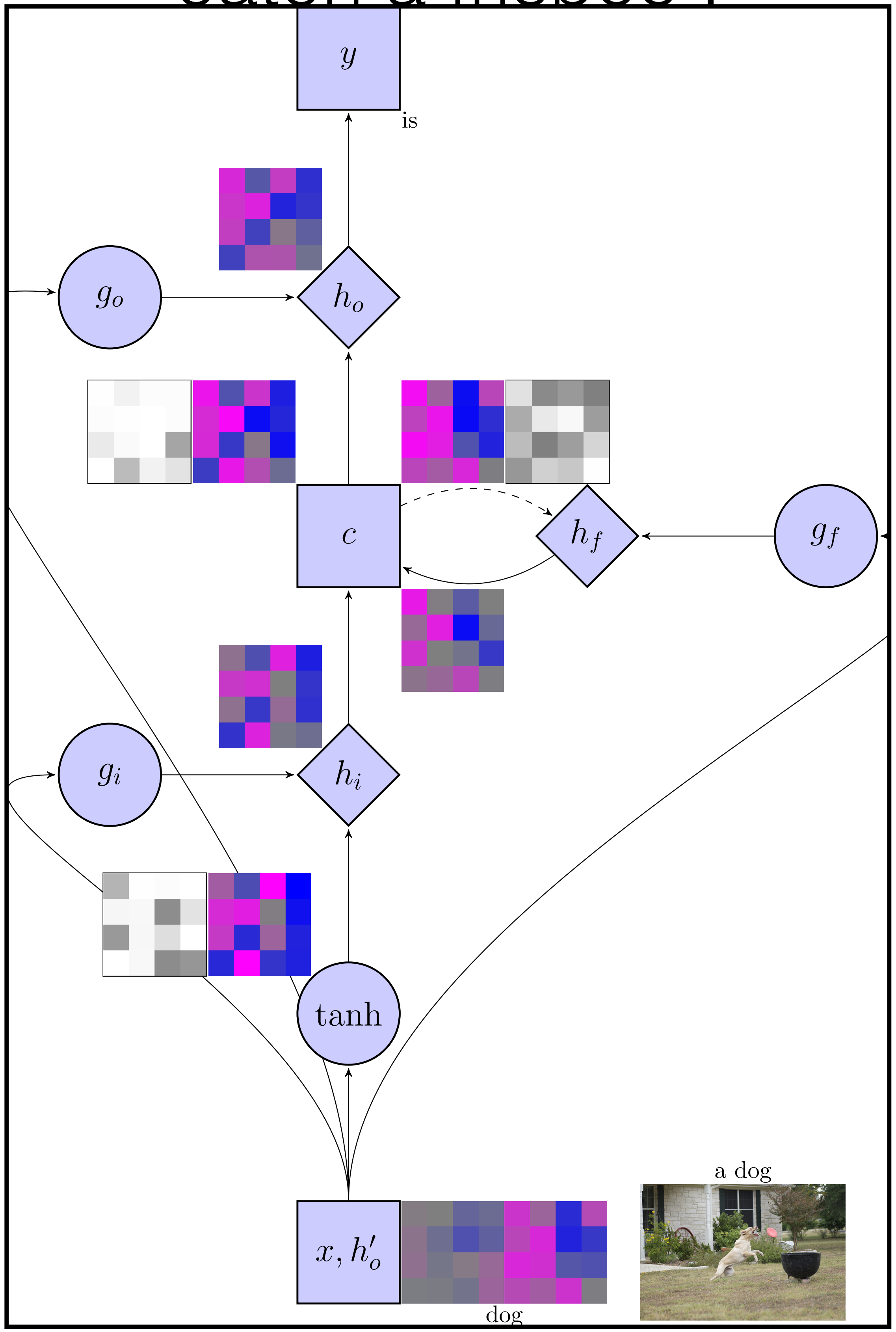
2

a dog jumps to catch a frisbee .



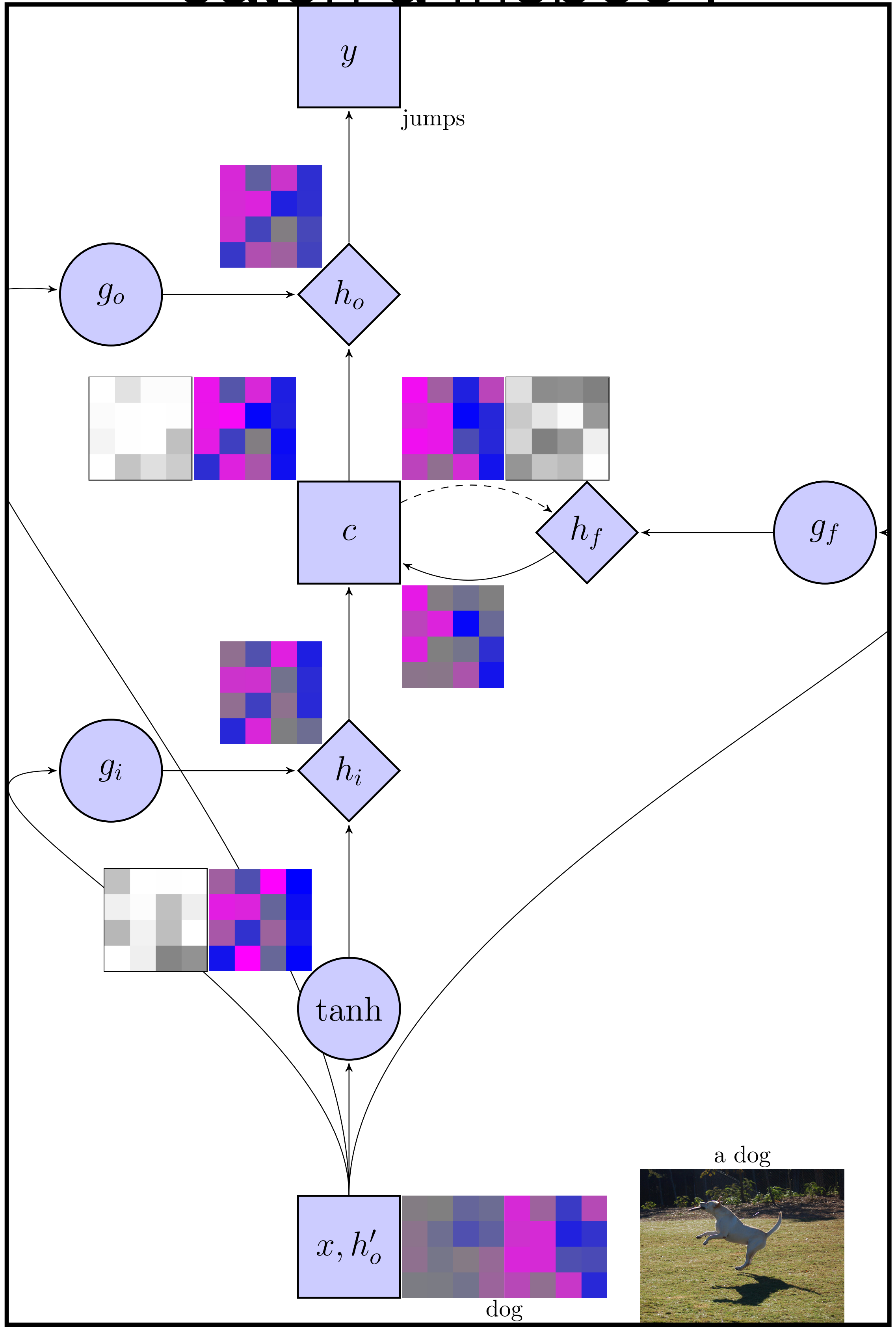
3

a dog is jumping to catch a frisbee .



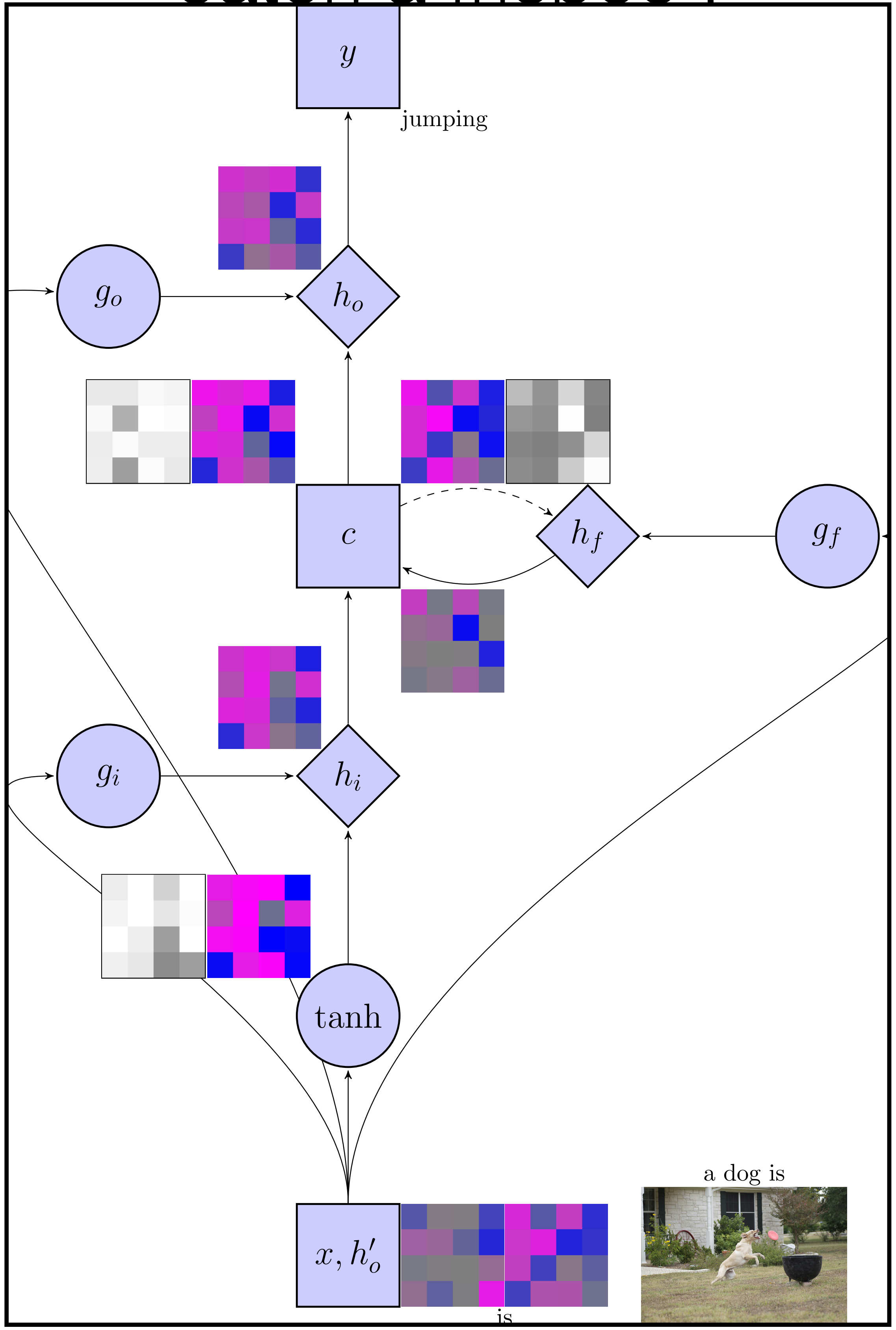
3

a dog jumps to catch a frisbee .



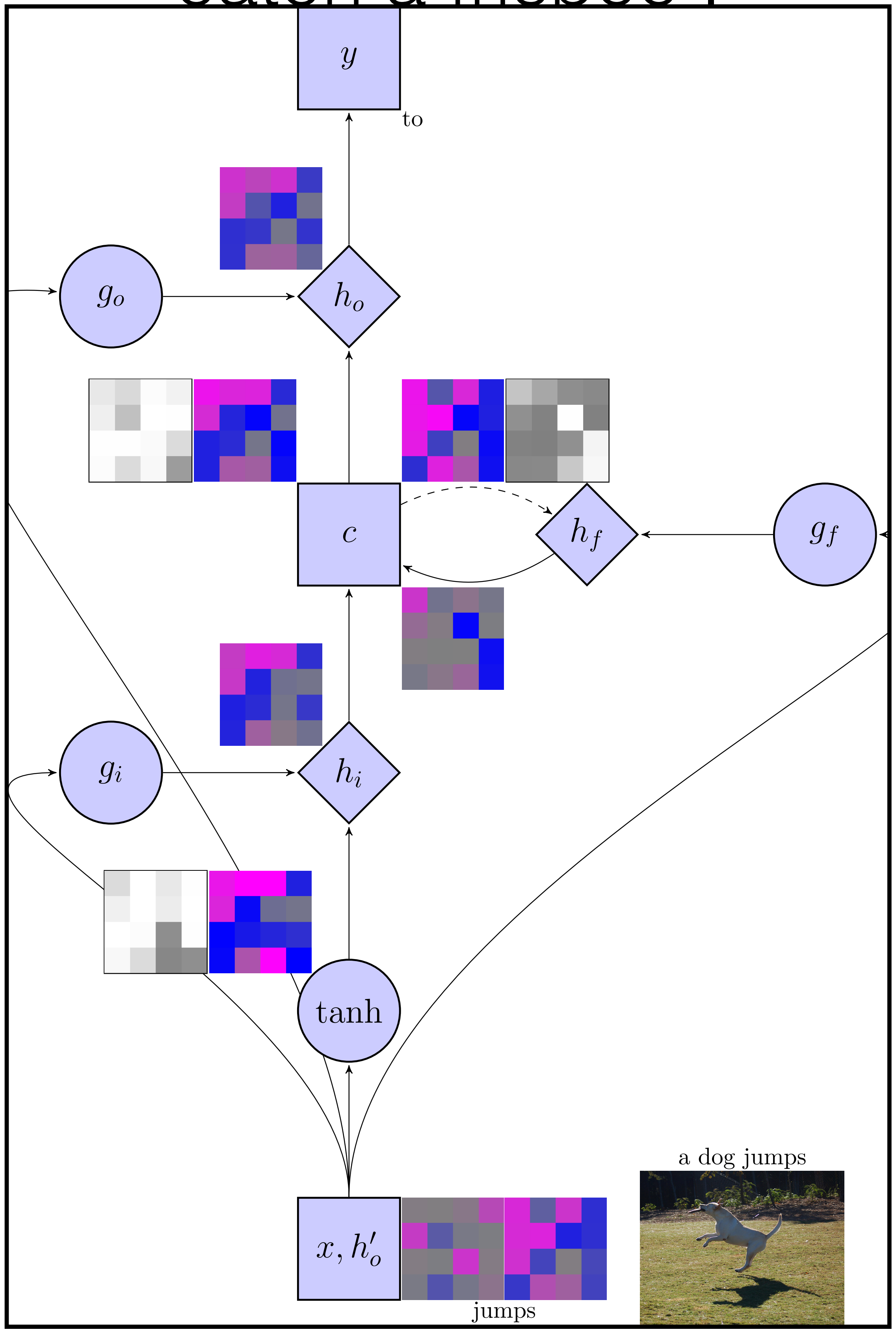
3

a dog is jumping to catch a frisbee .



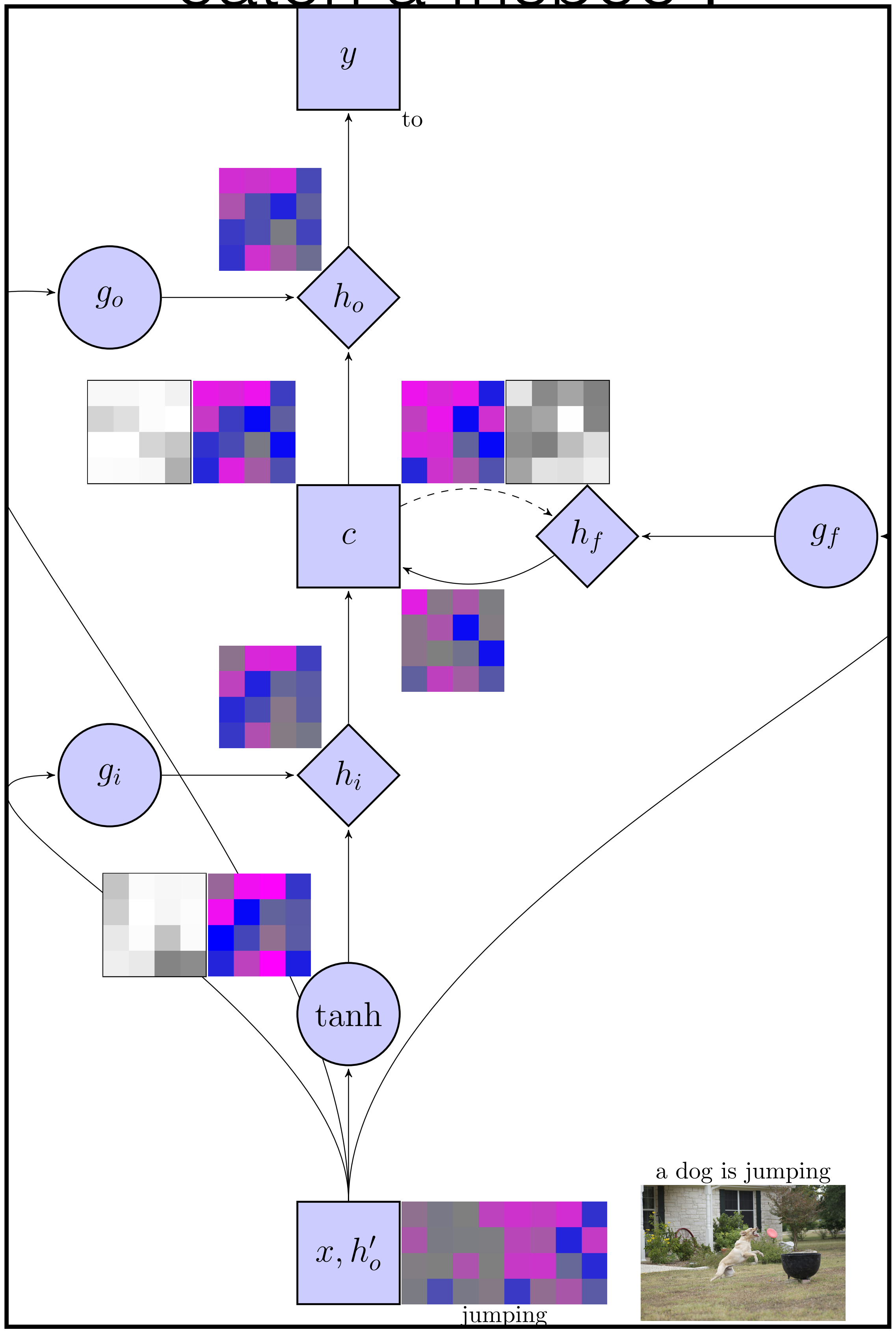
4

a dog jumps to catch a frisbee .



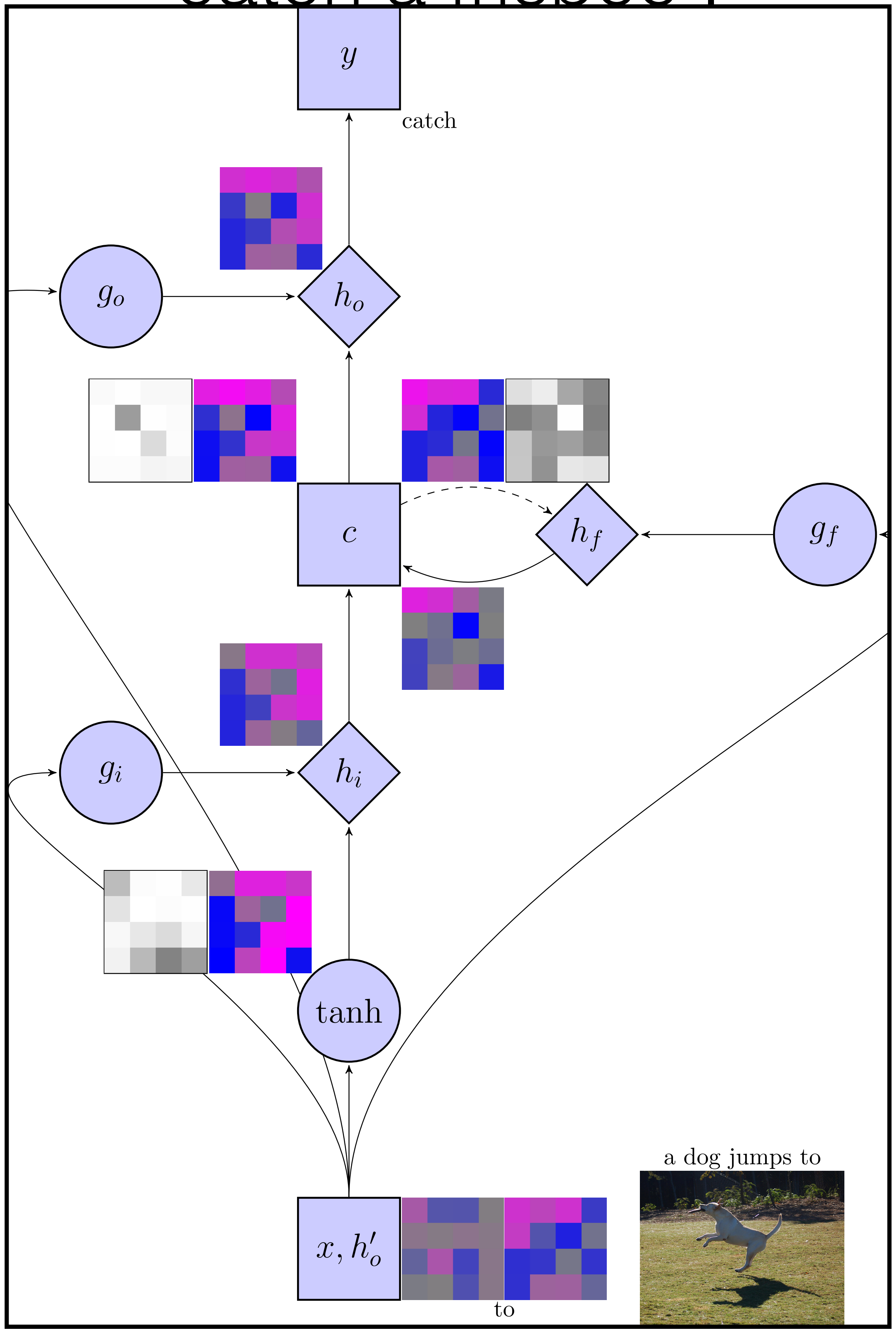
4

a dog is jumping to catch a frisbee .



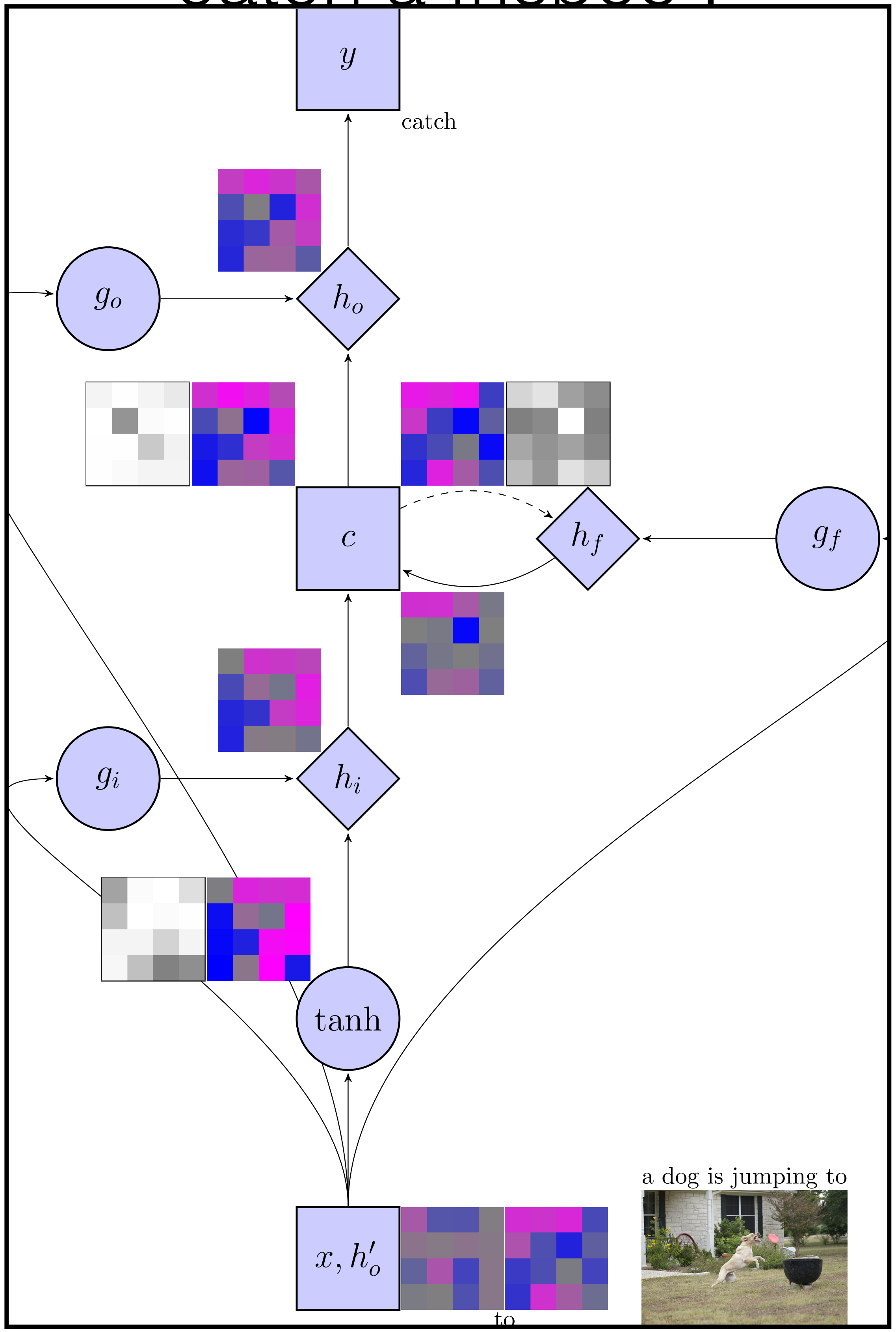
5

a dog jumps to catch a frisbee .



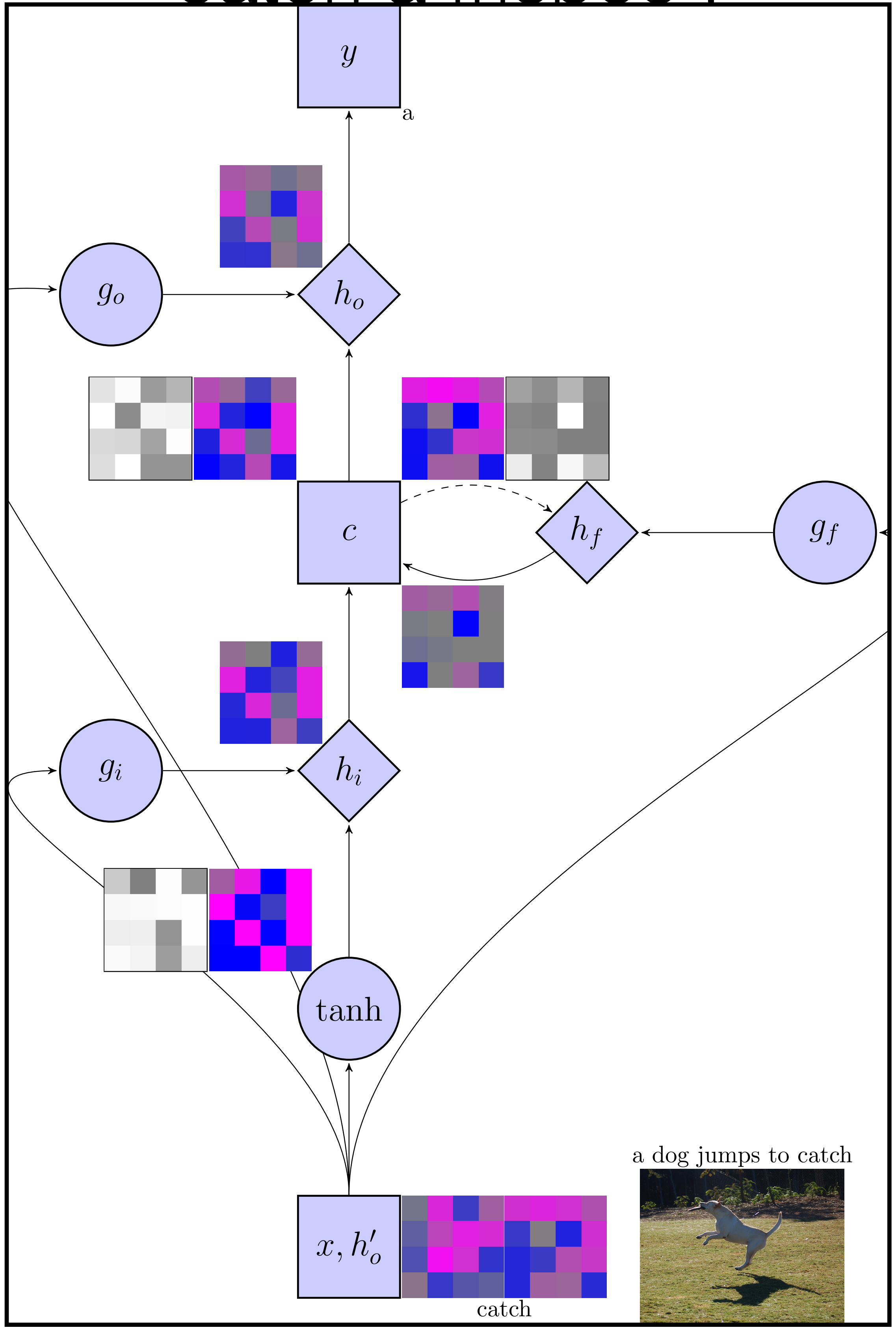
5

a dog is jumping to catch a frisbee .



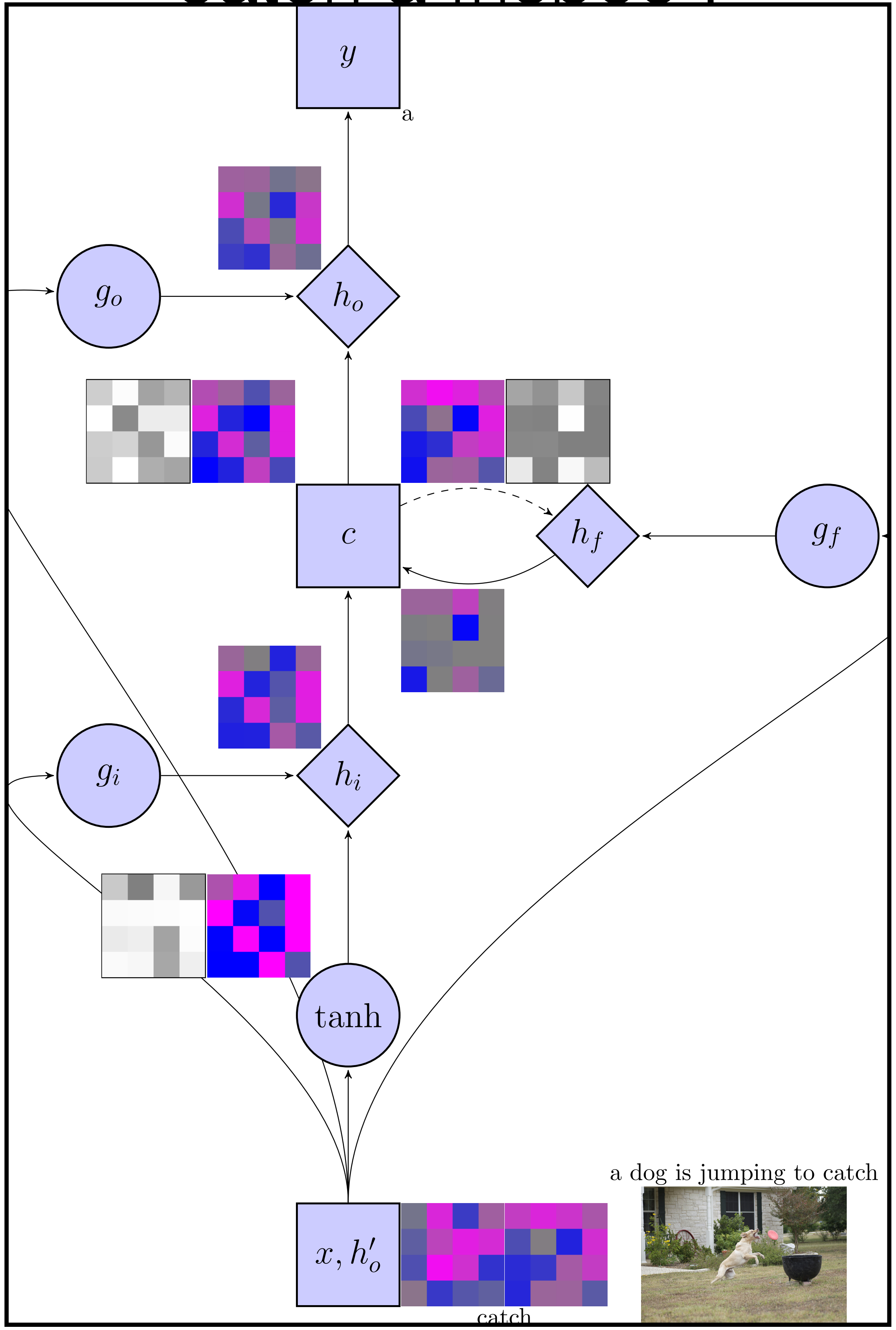
6

a dog jumps to catch a frisbee .



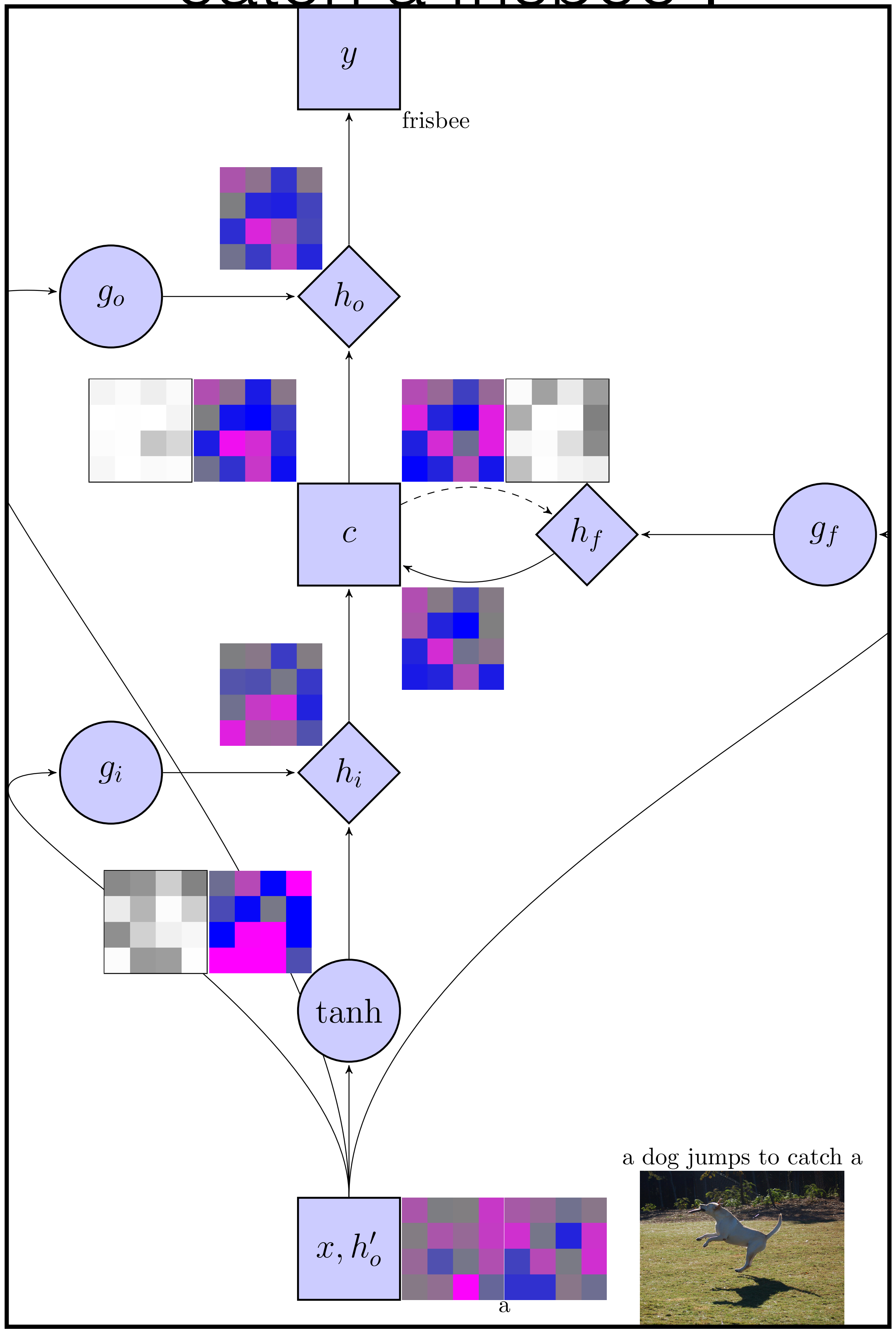
6

a dog is jumping to catch a frisbee .



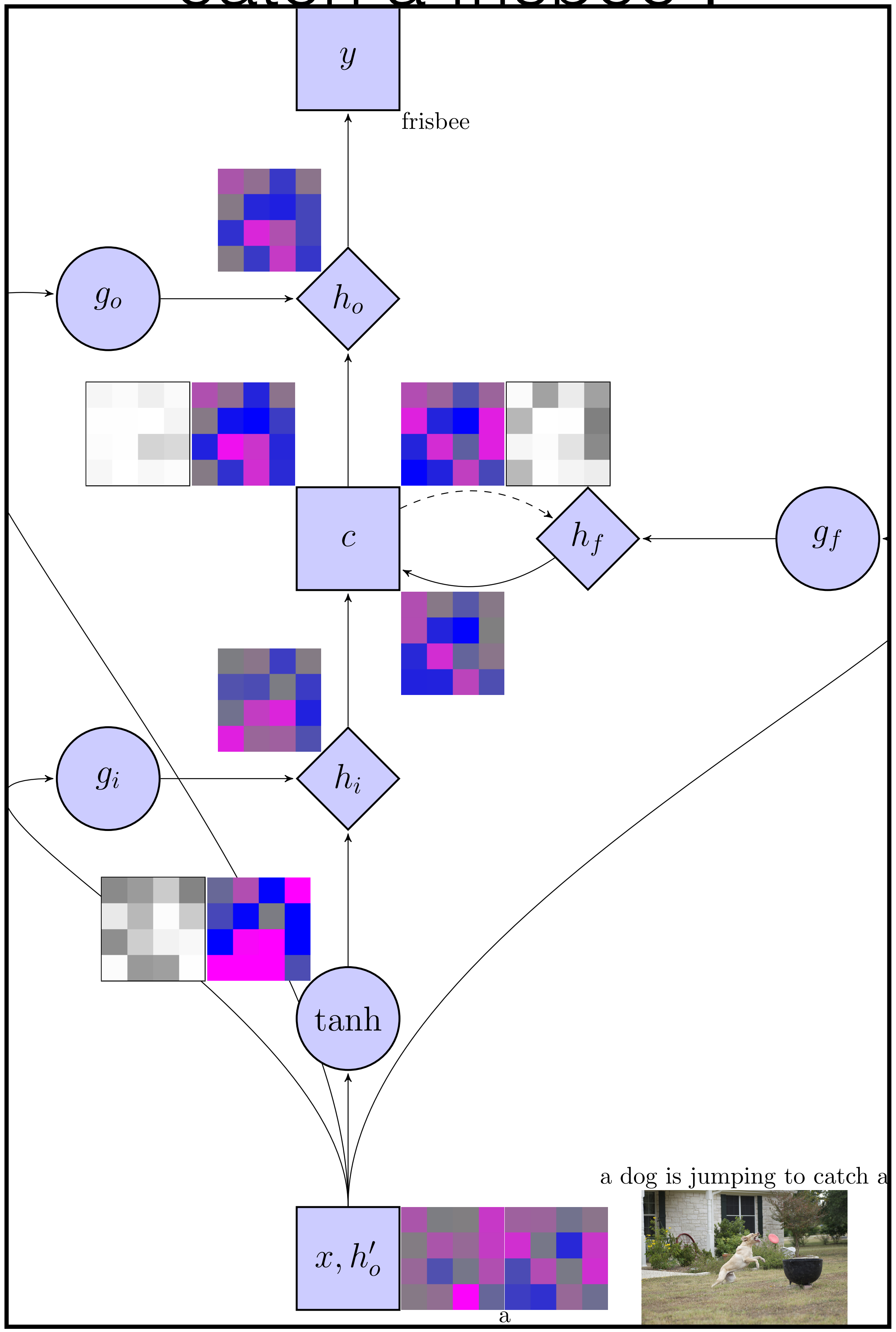
7

a dog jumps to catch a frisbee .



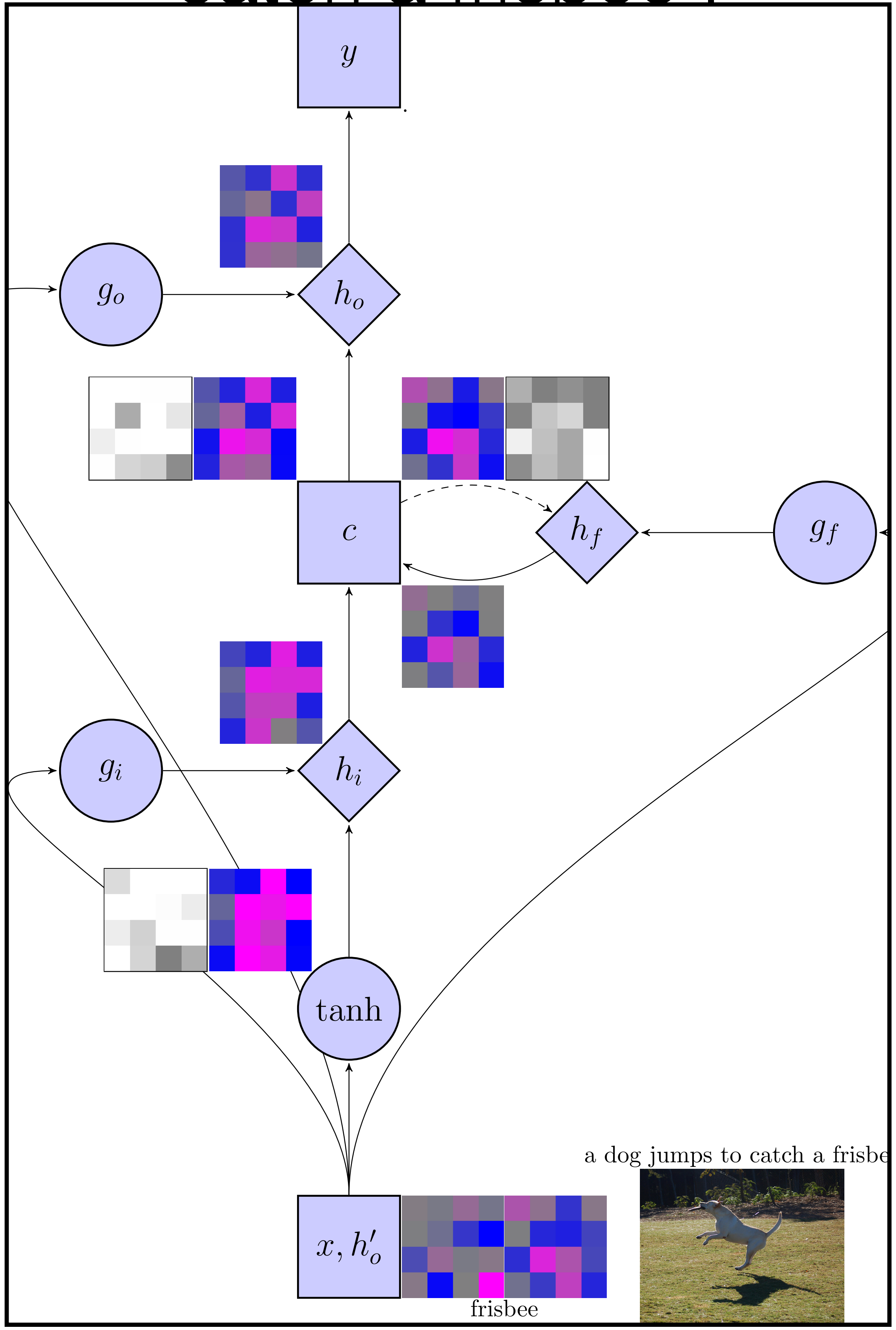
7

a dog is jumping to catch a frisbee .



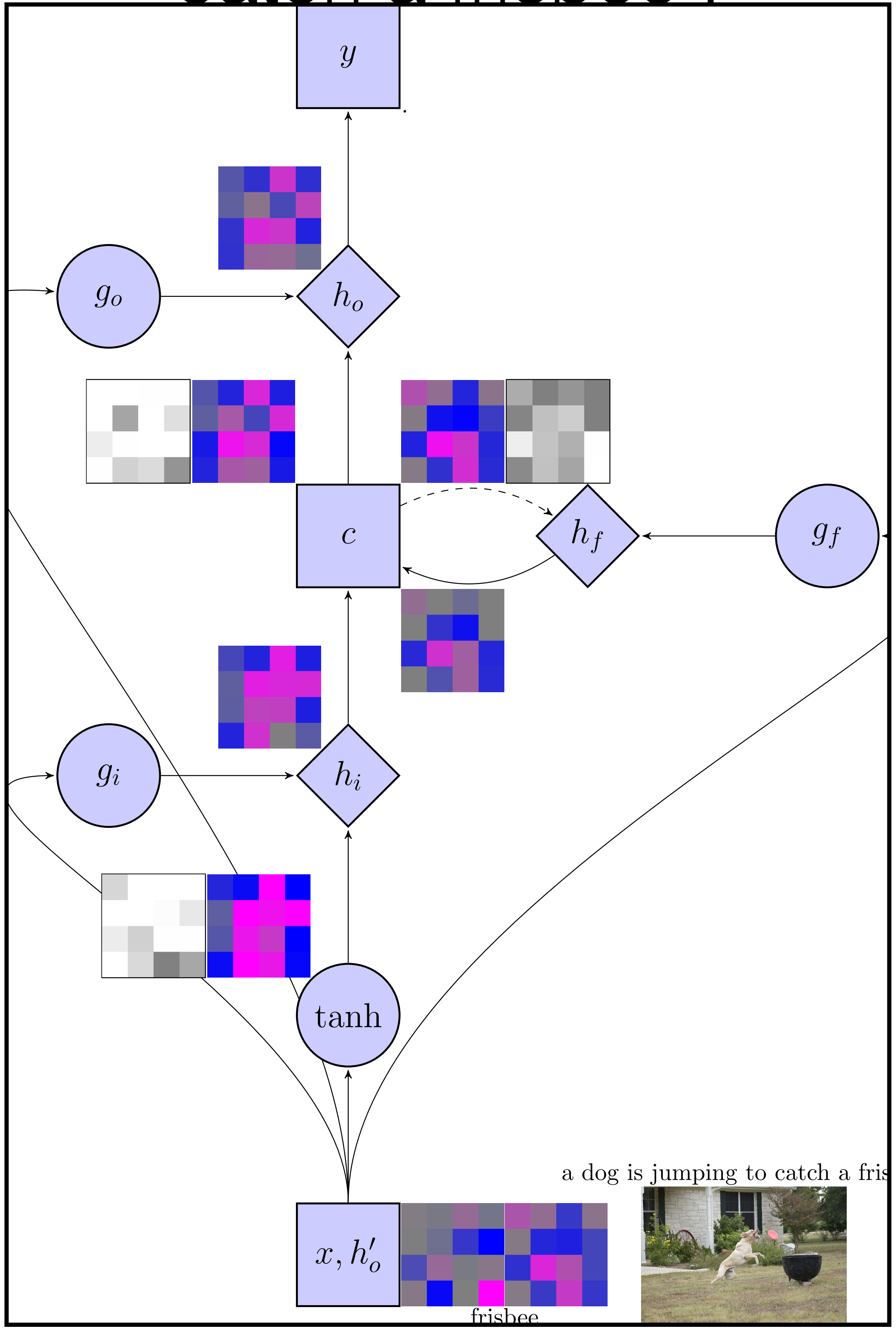
8

a dog jumps to catch a frisbee .



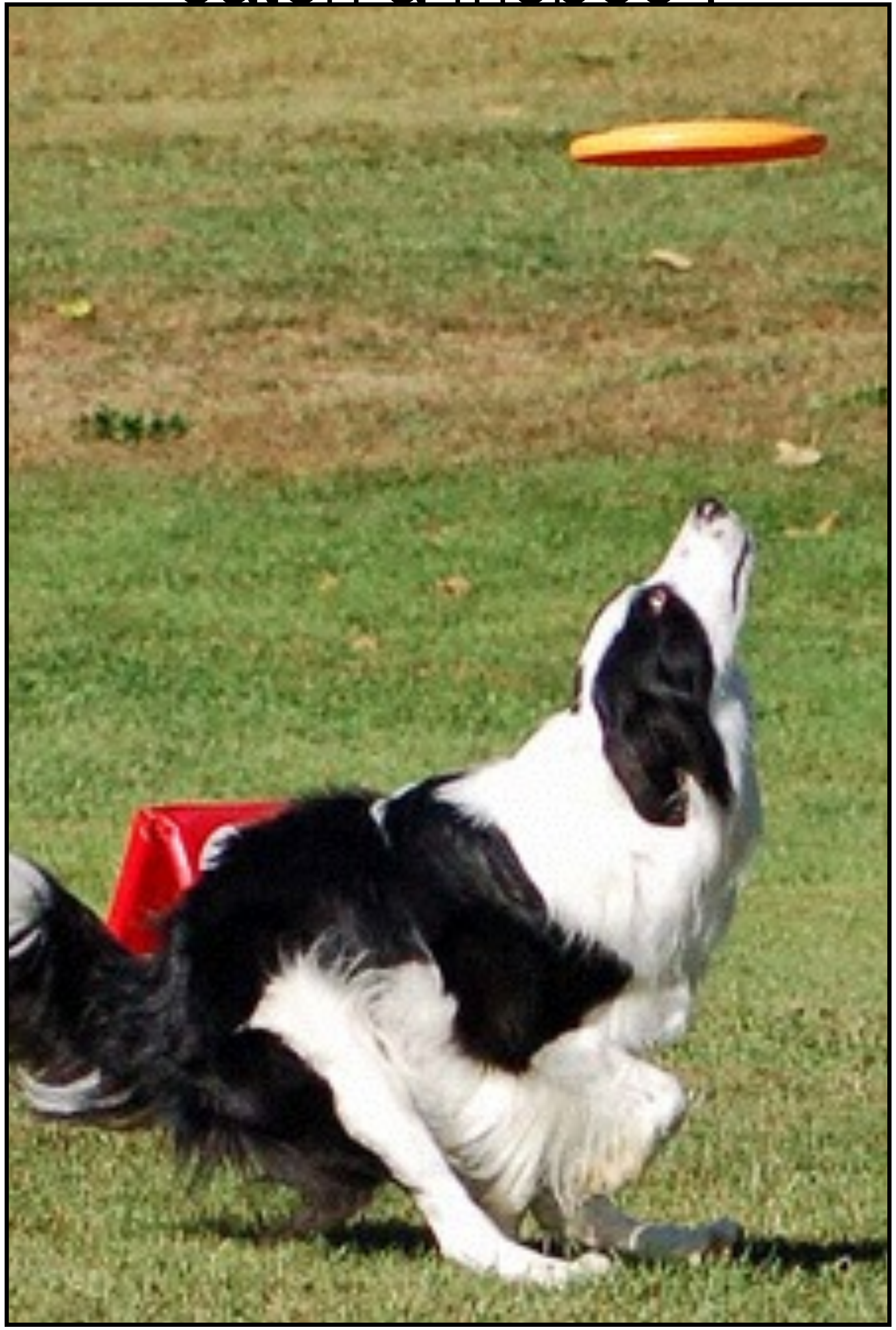
8

a dog is jumping to catch a frisbee .



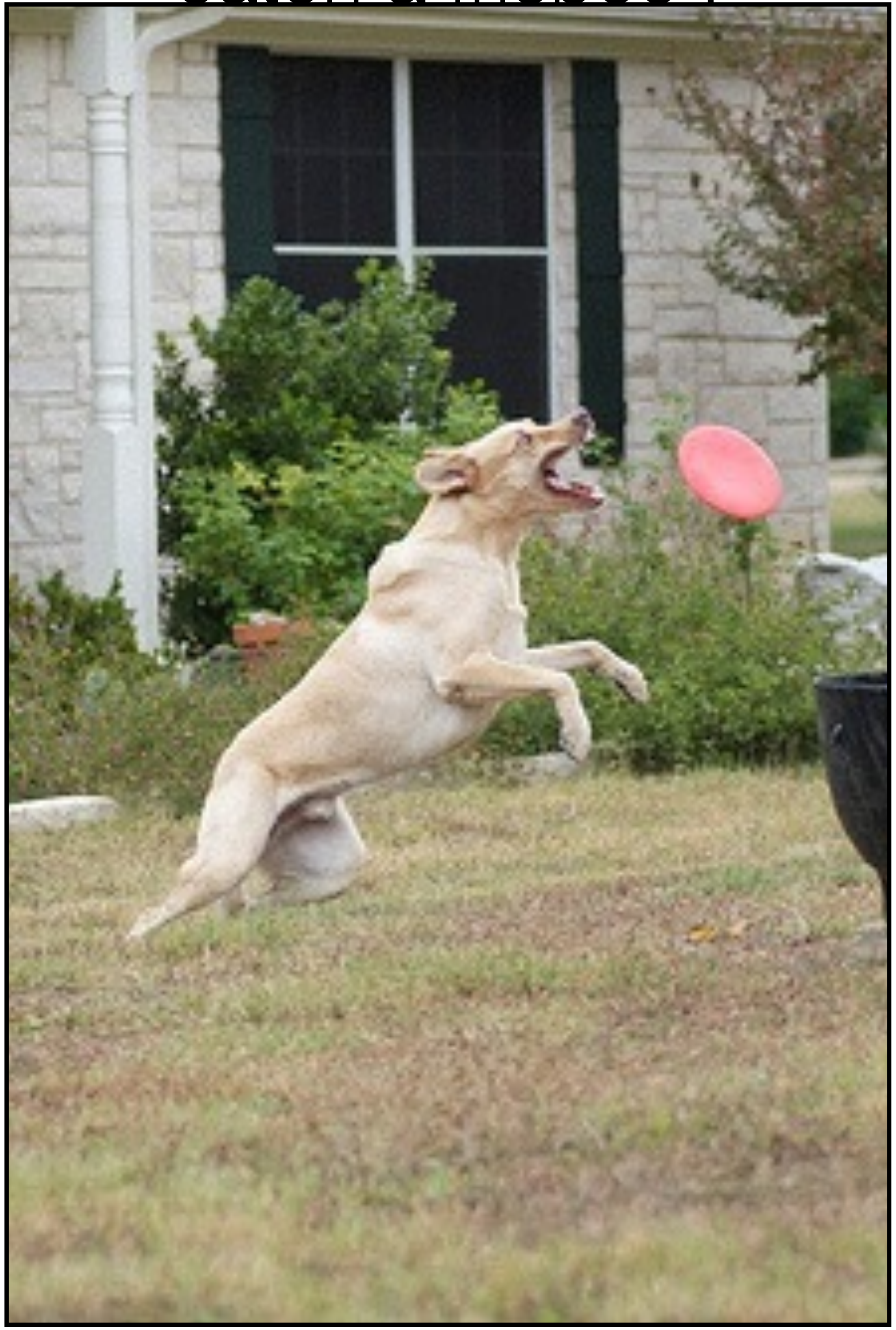
9

a black dog is jumping to catch a frisbee .



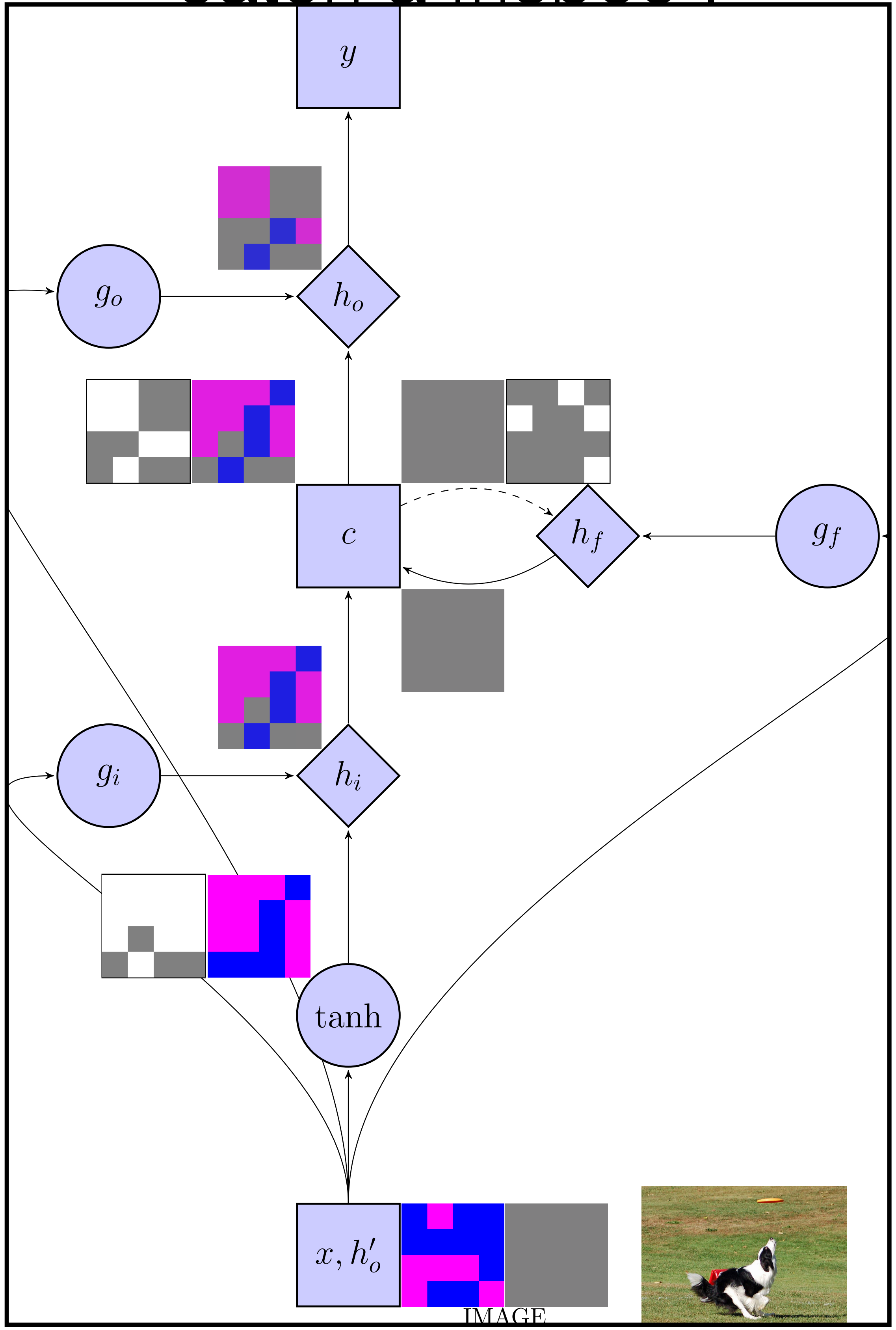
1626754053_81126b67b6.jpg

a dog is jumping to catch a frisbee .



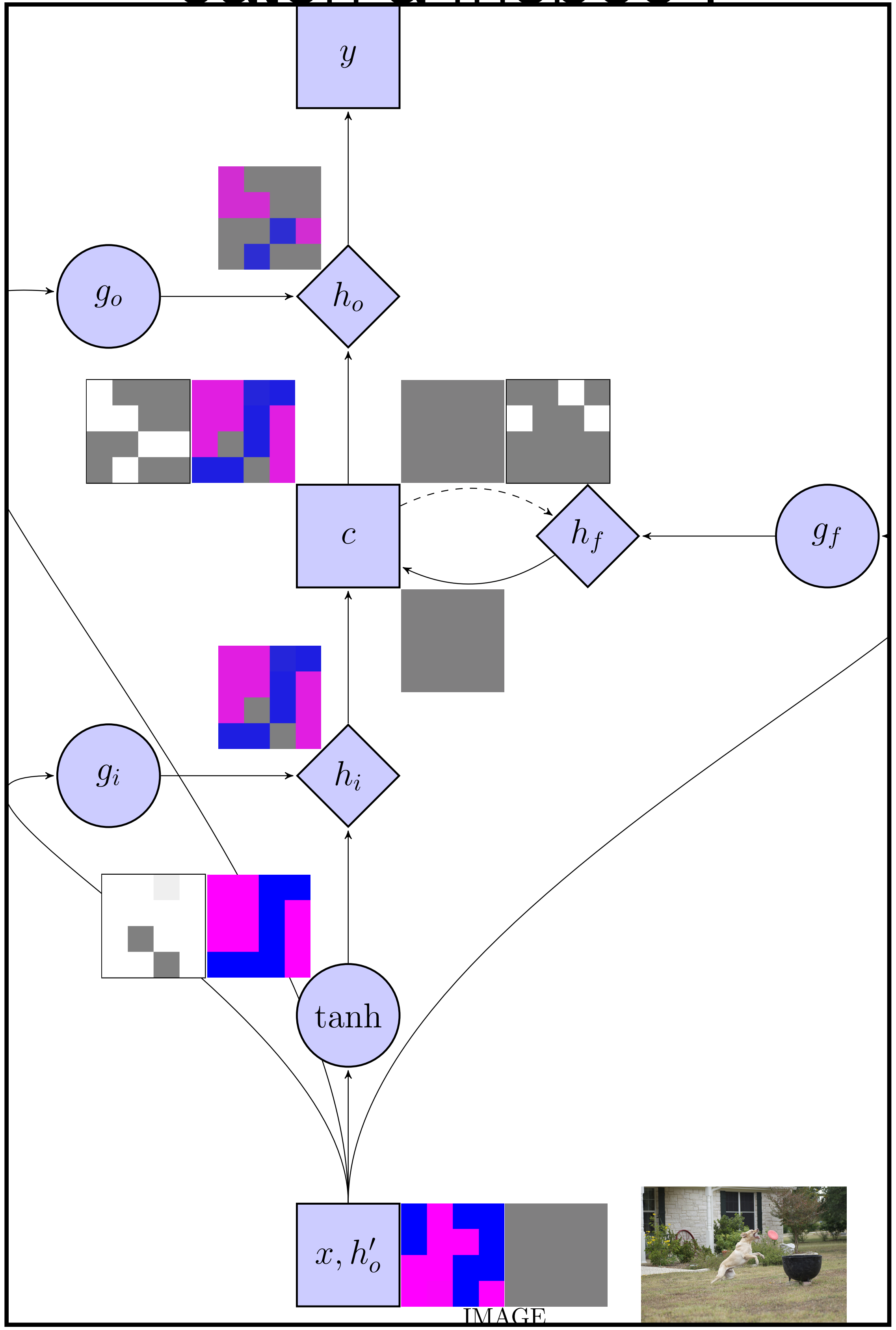
2945036454_280fa5b29f.jpg

a black dog is jumping to catch a frisbee .



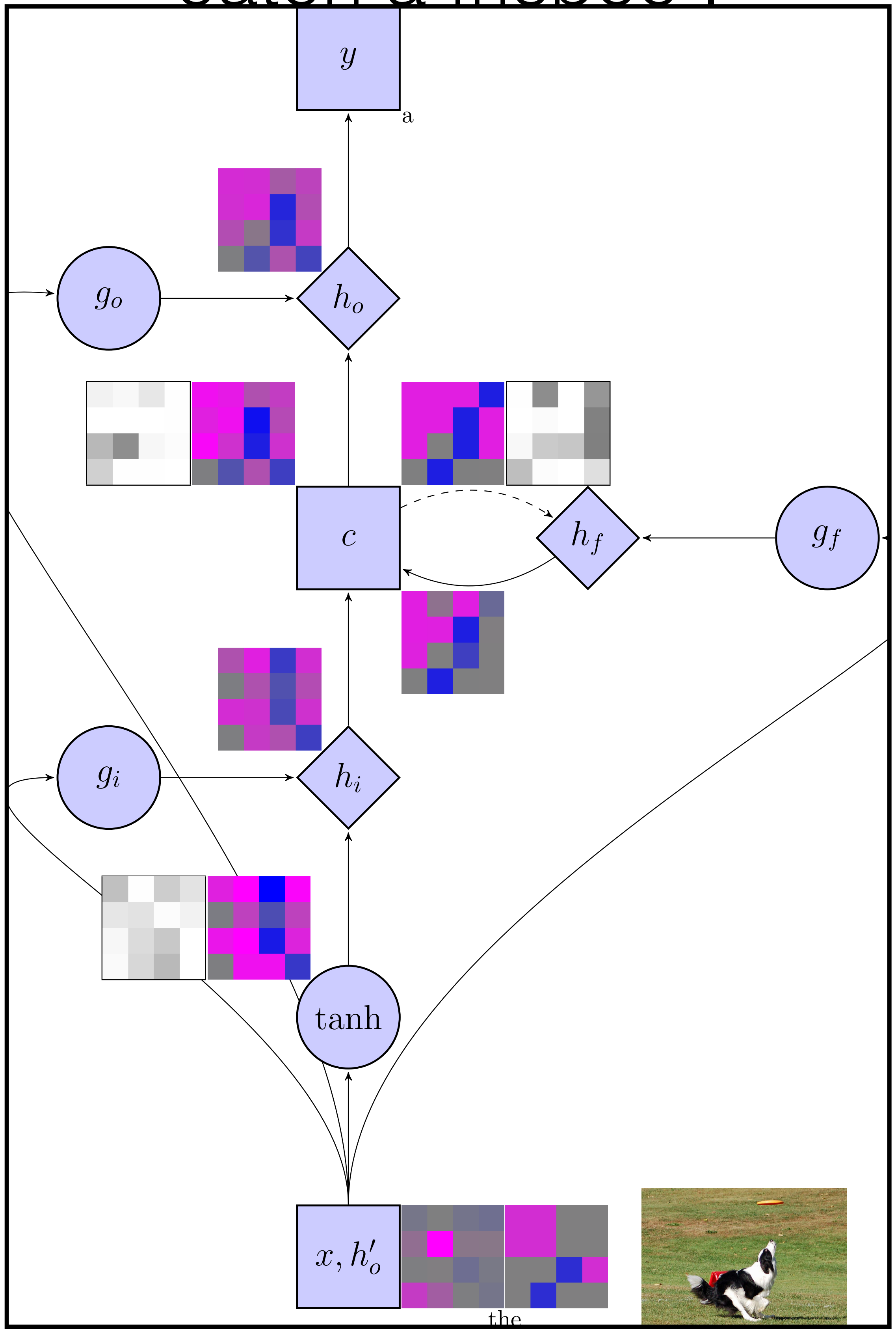
0

a dog is jumping to catch a frisbee .



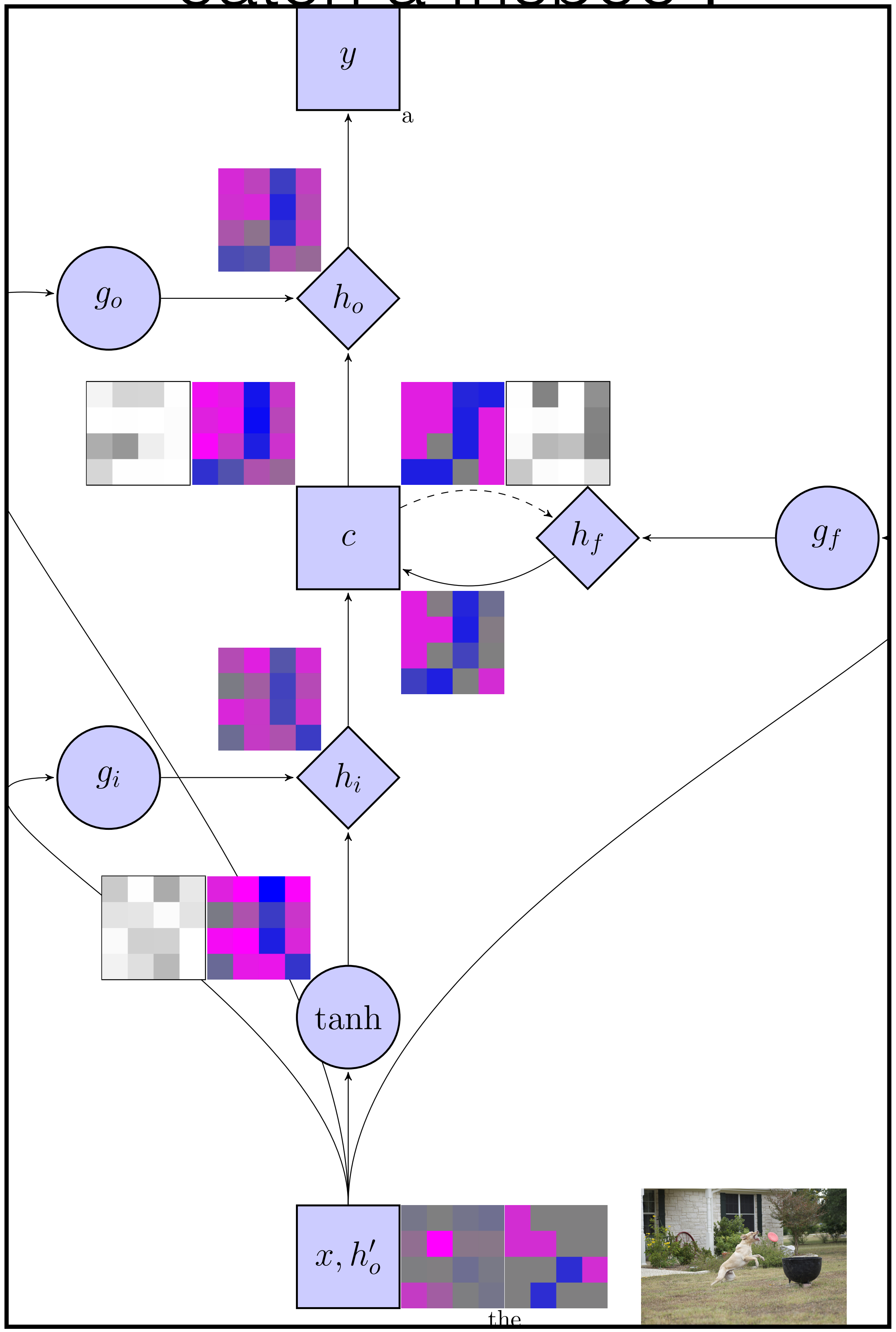
0

a black dog is jumping to catch a frisbee .



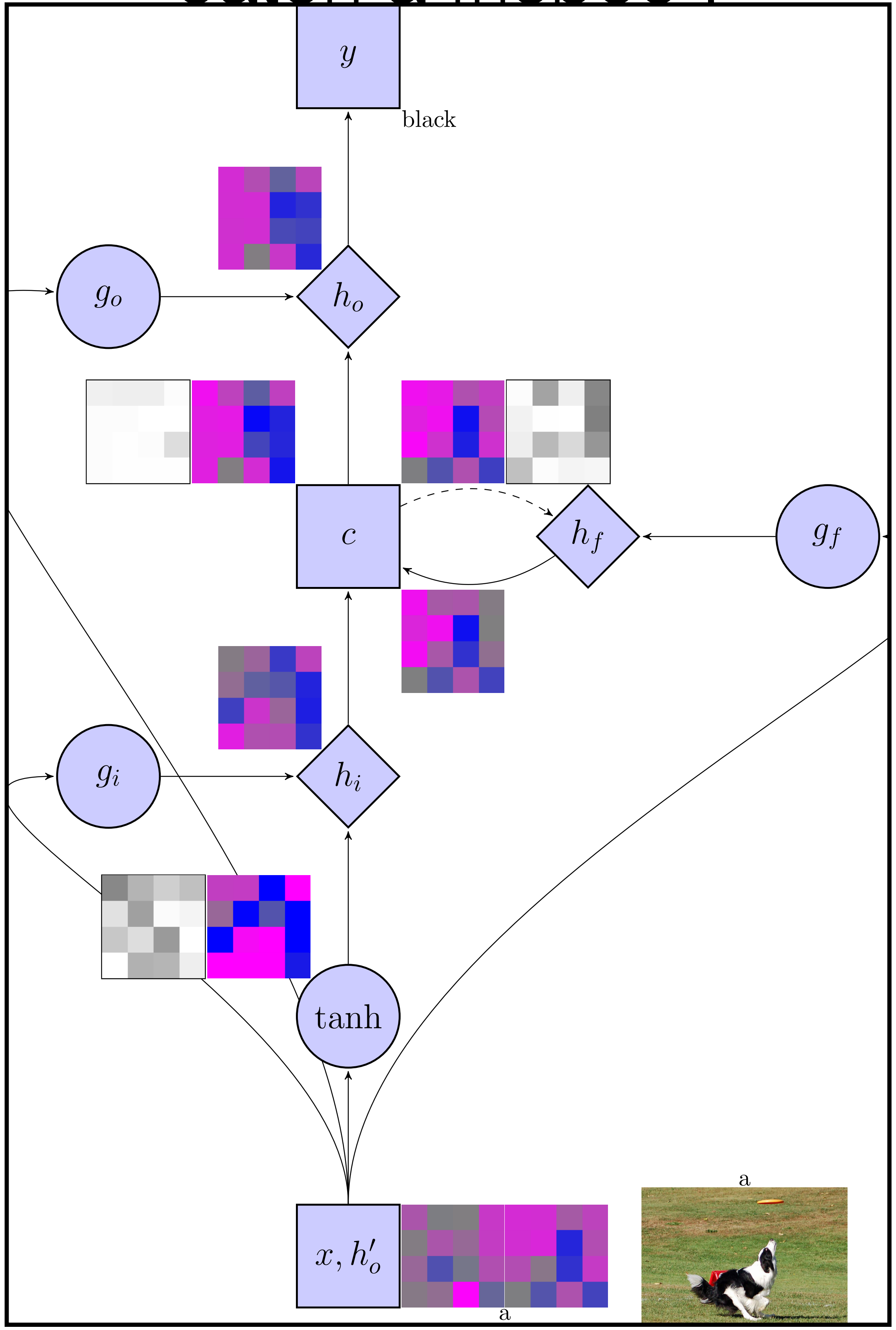
1

a dog is jumping to catch a frisbee .



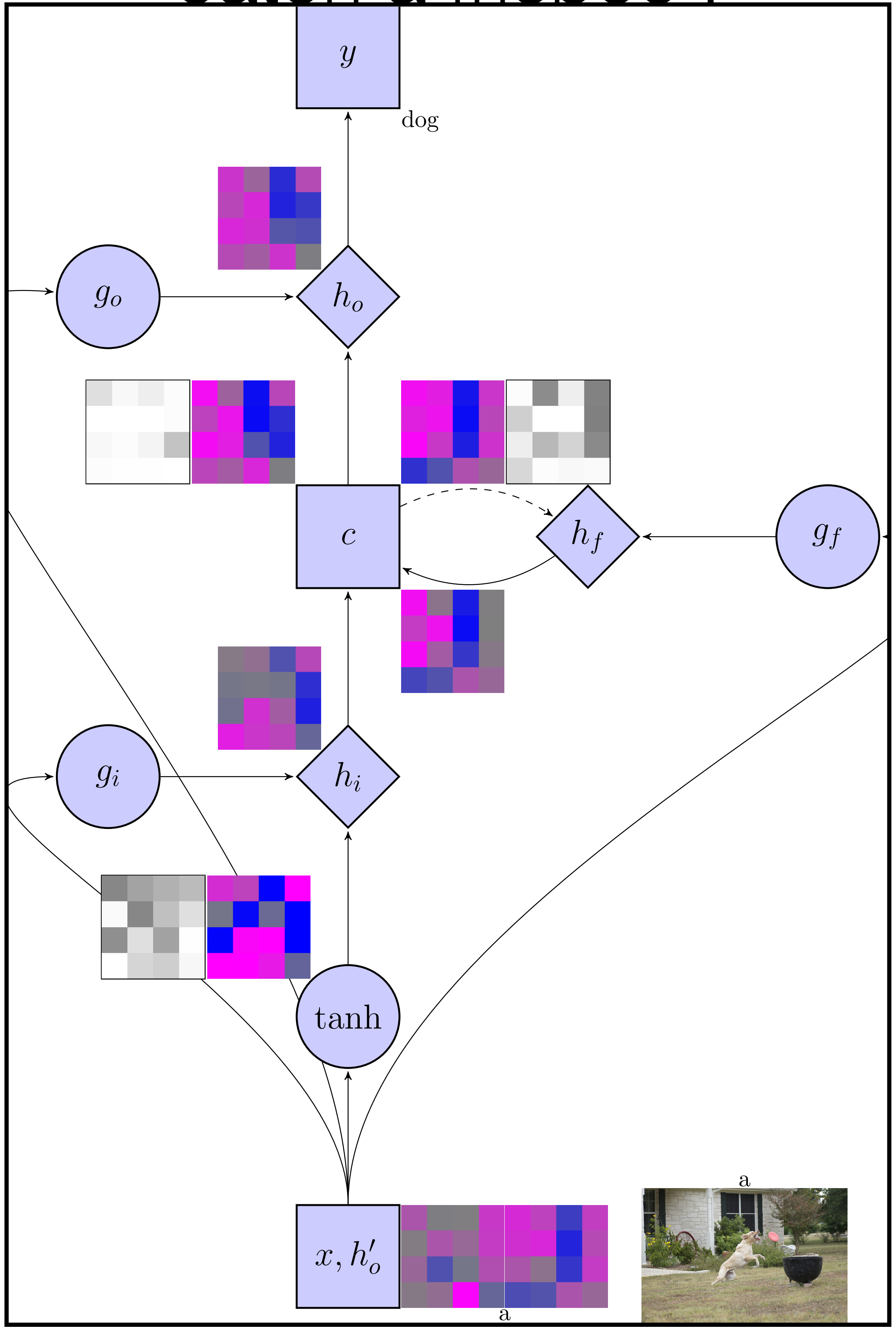
1

a black dog is jumping to catch a frisbee .



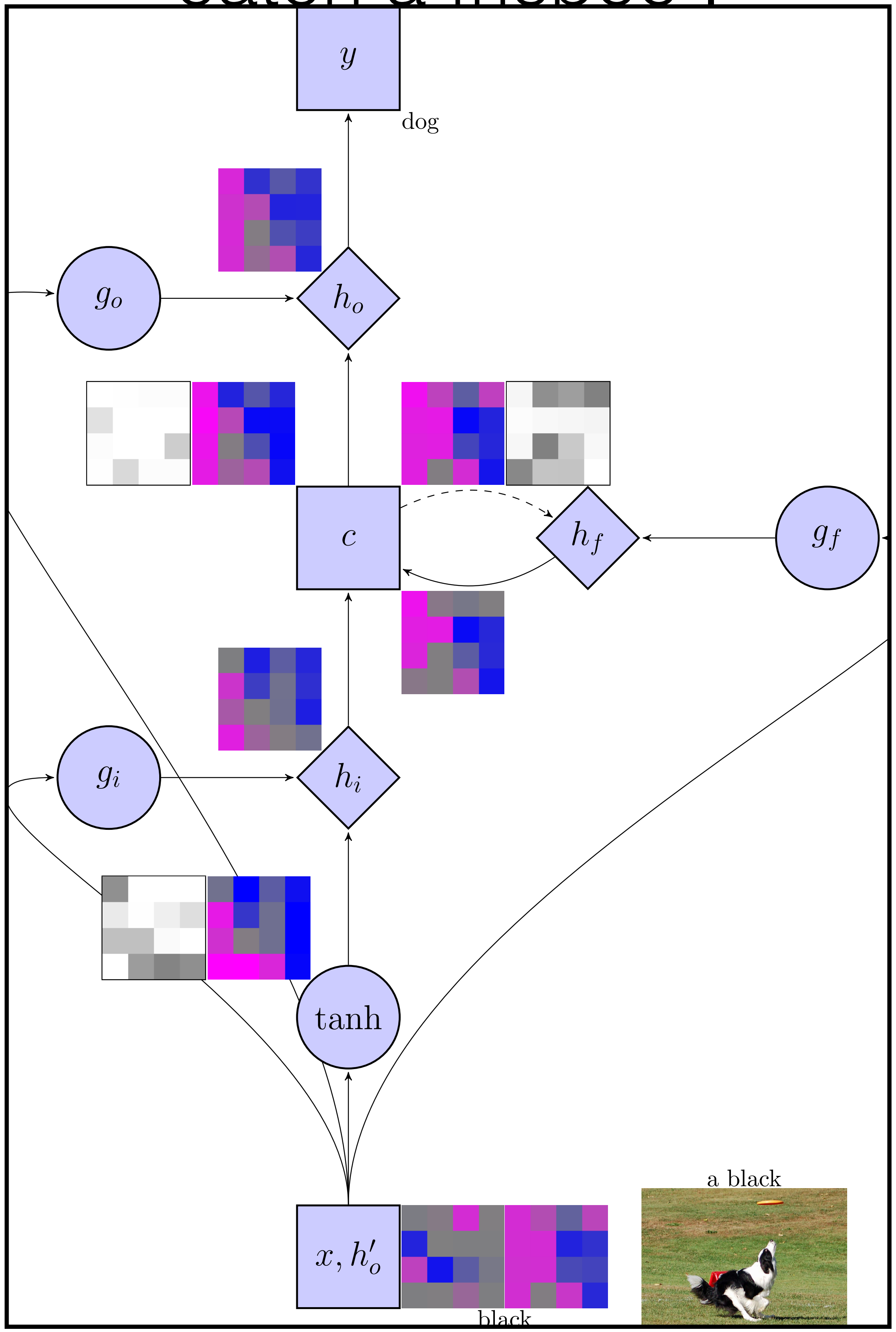
2

a dog is jumping to catch a frisbee .



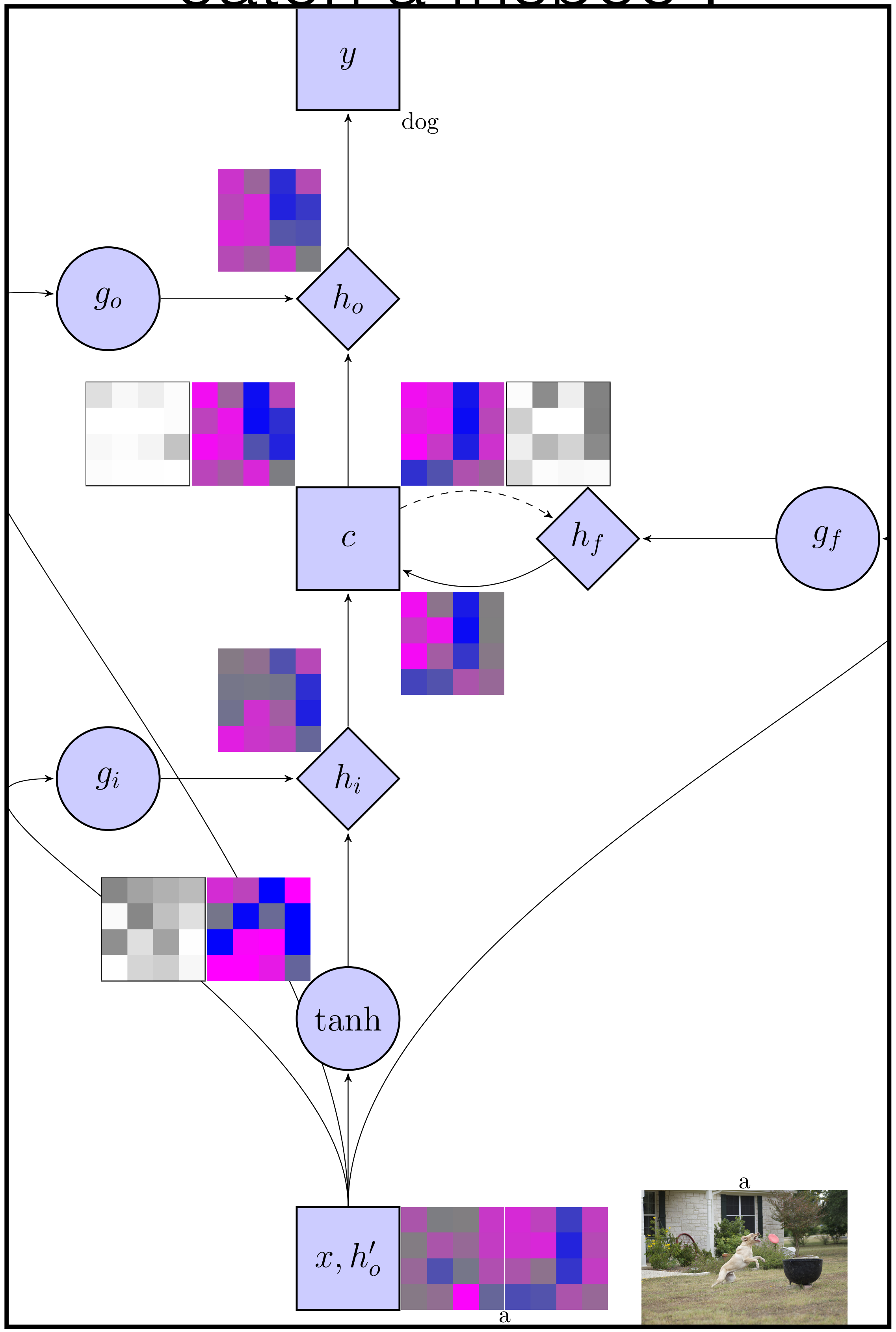
2

a black dog is jumping to catch a frisbee .



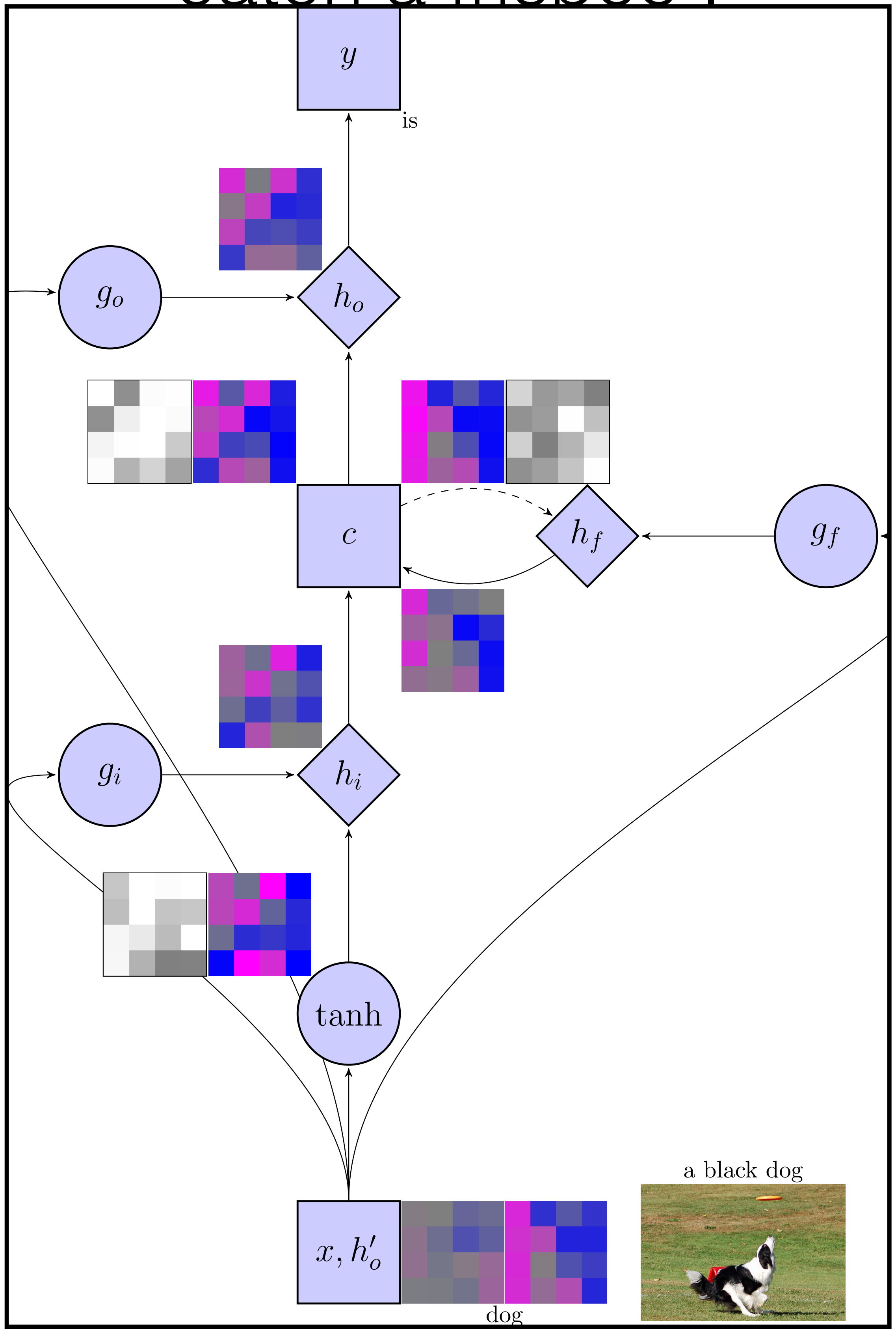
3

a dog is jumping to catch a frisbee .



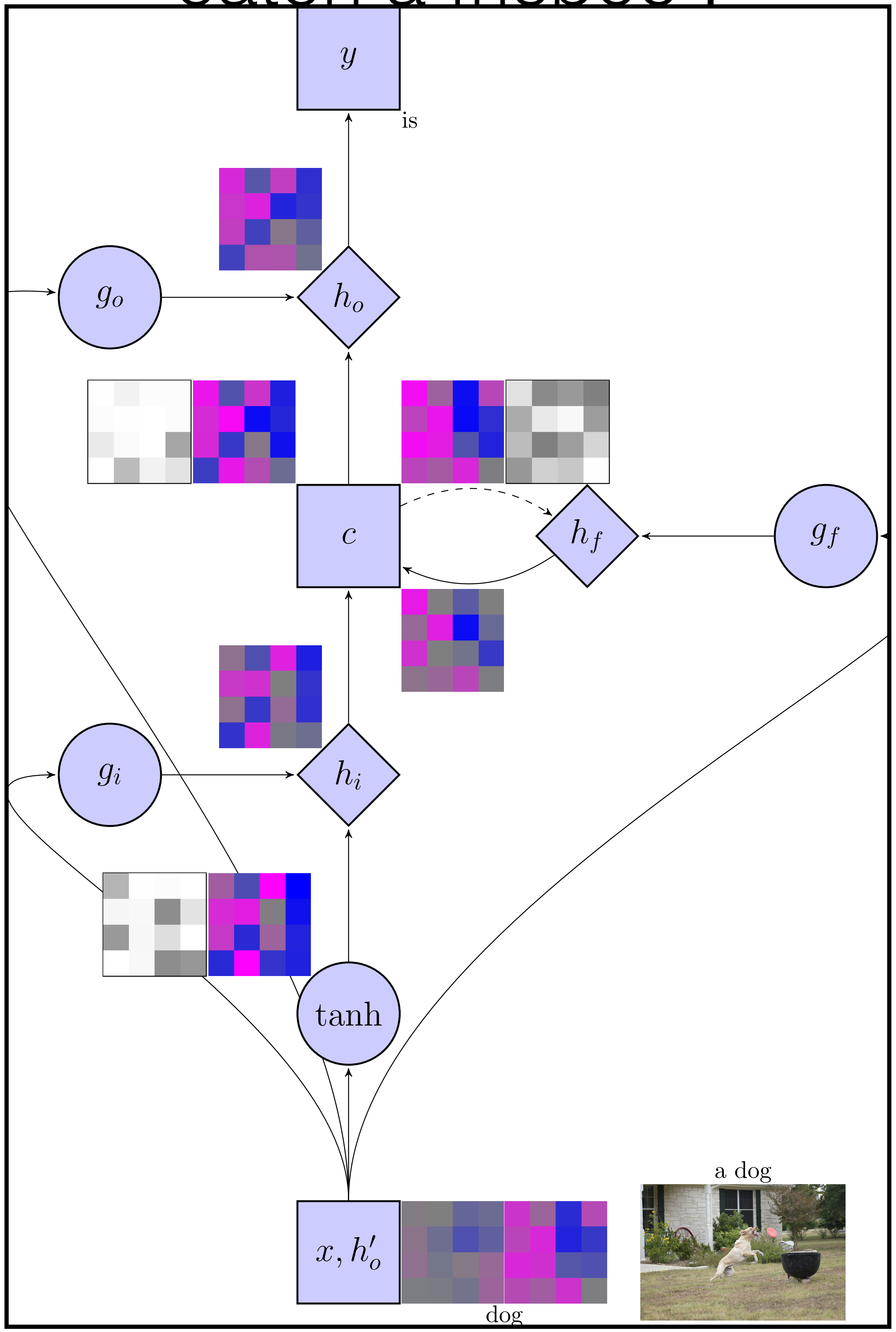
2

a black dog is jumping to catch a frisbee .



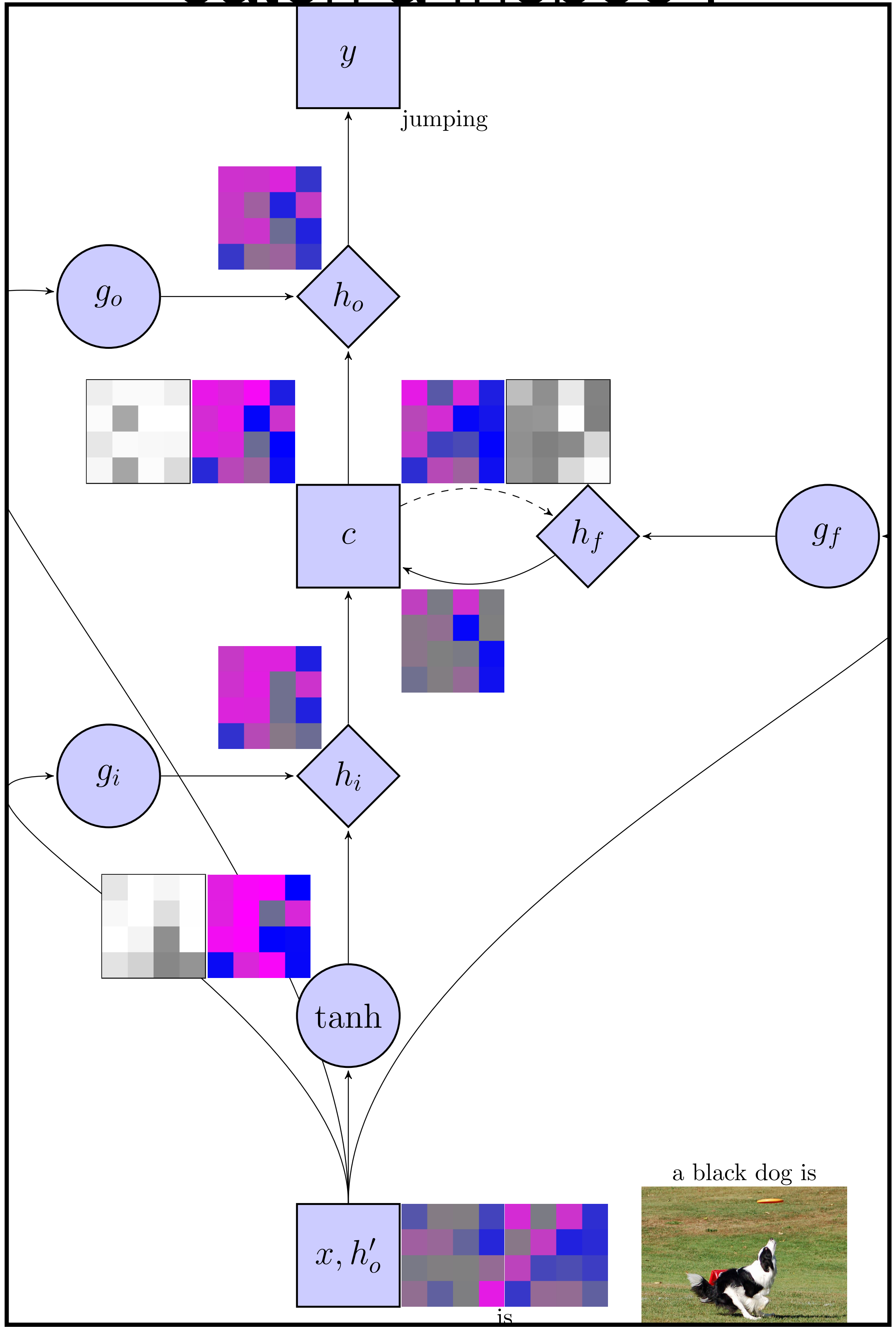
4

a dog is jumping to catch a frisbee .



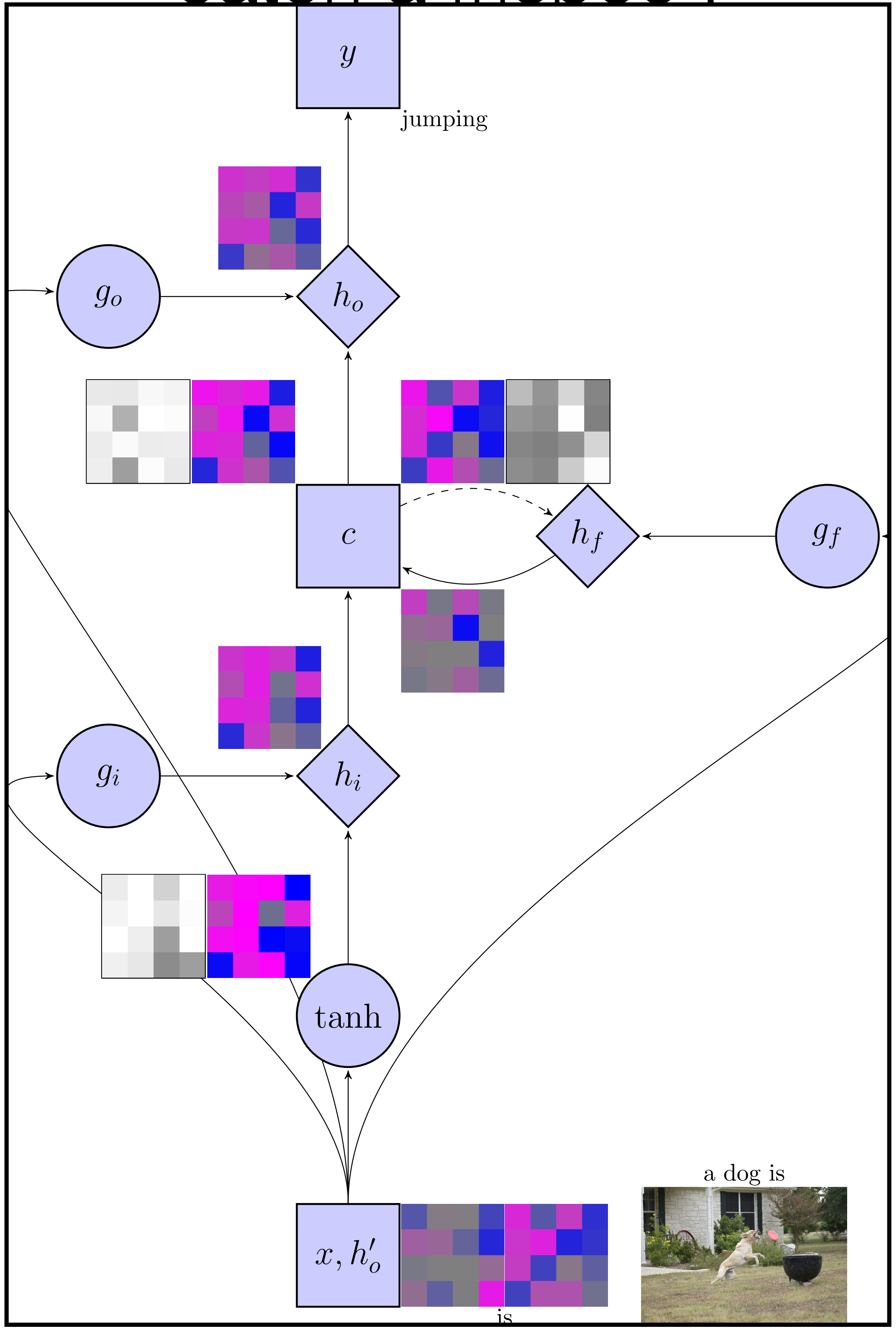
3

a black dog is jumping to catch a frisbee .



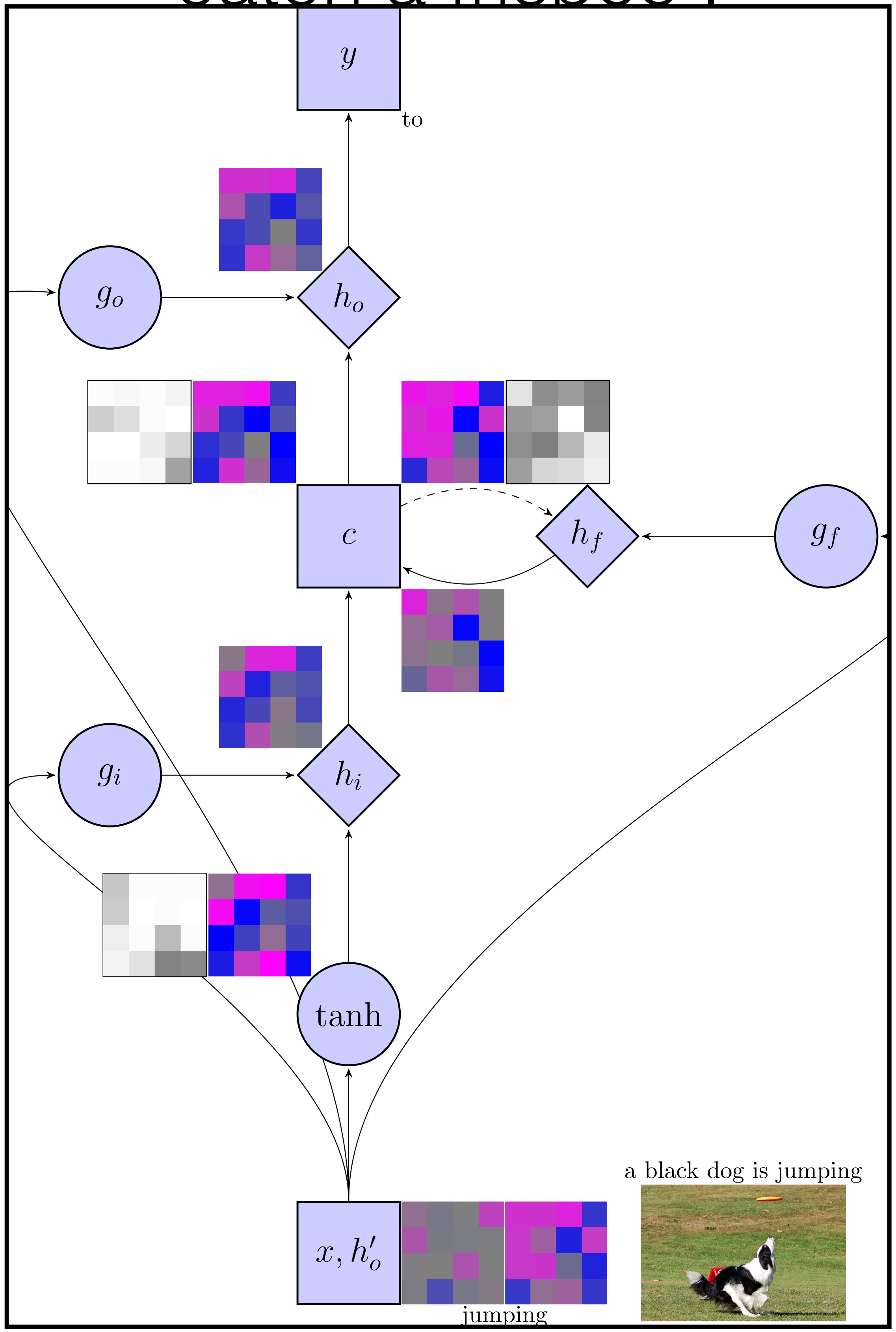
5

a dog is jumping to catch a frisbee .



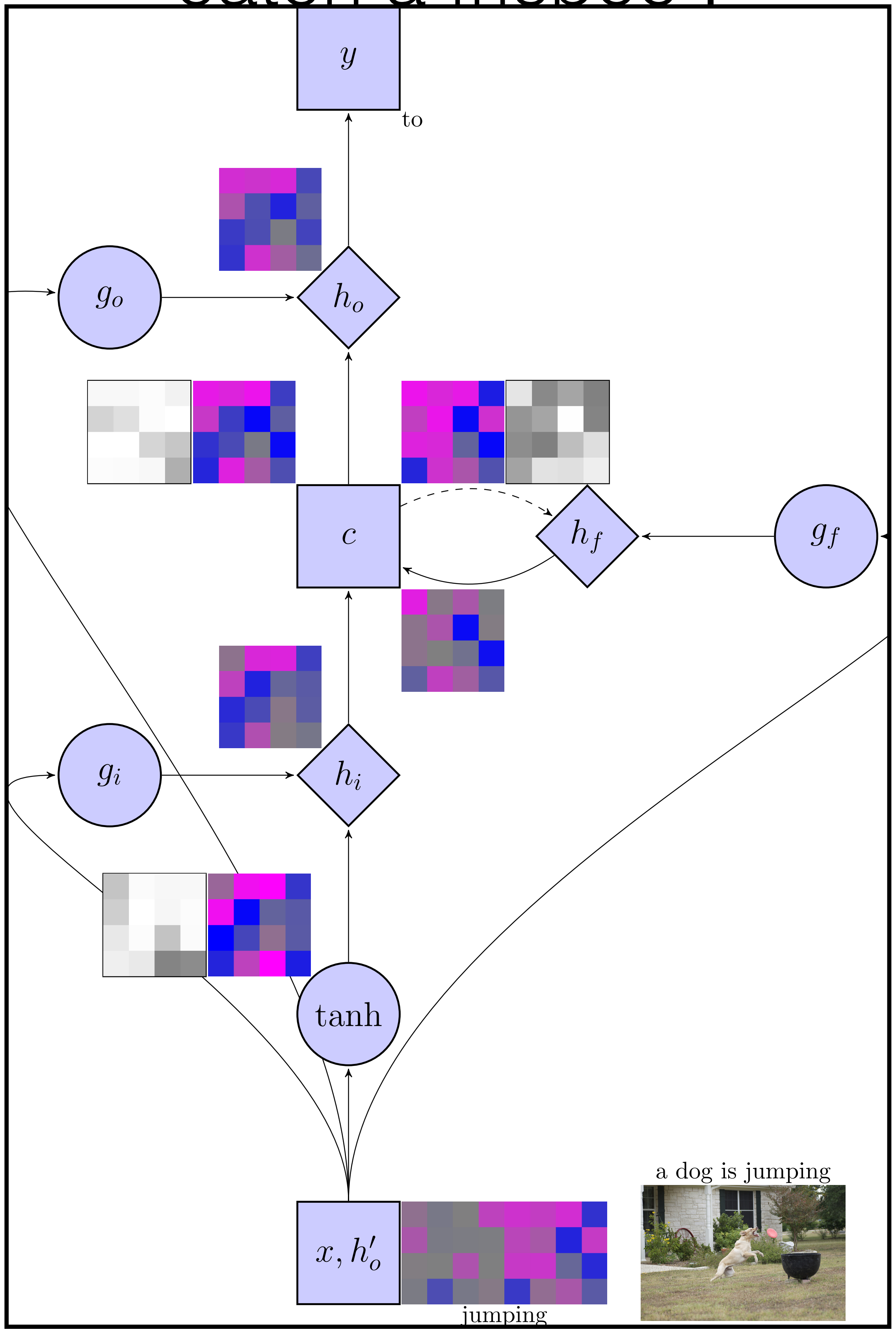
4

a black dog is jumping to catch a frisbee .



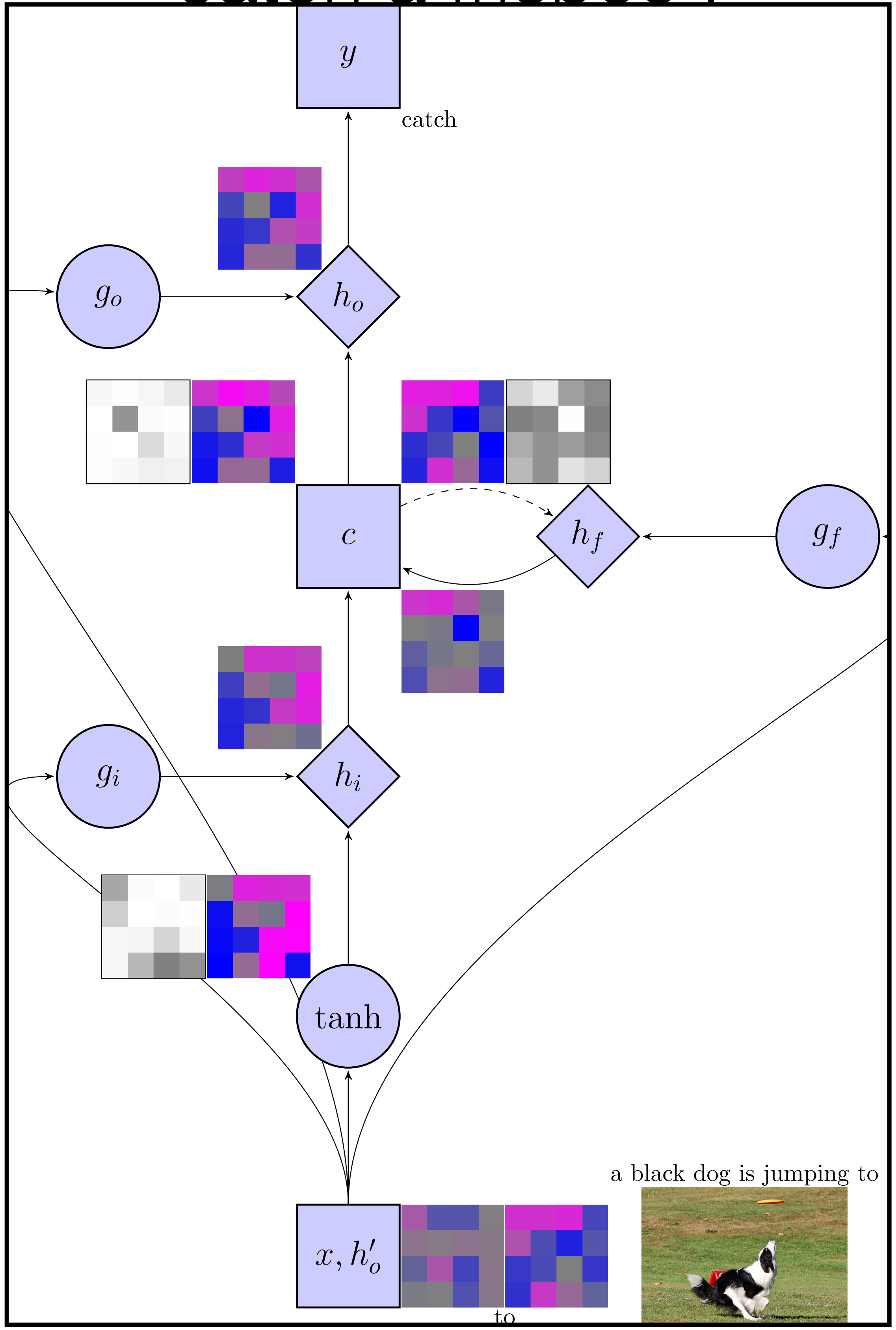
6

a dog is jumping to catch a frisbee .



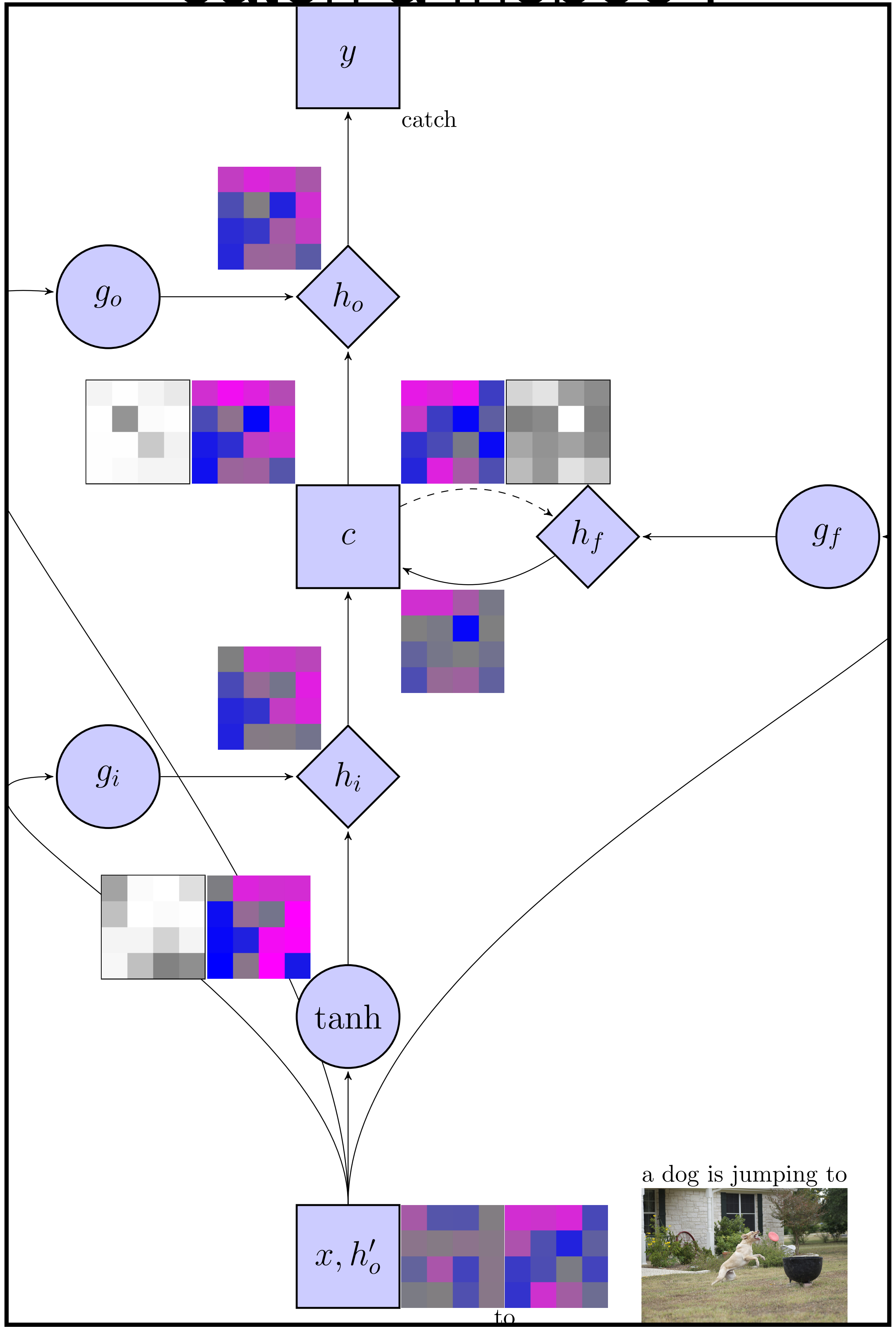
5

a black dog is jumping to catch a frisbee .



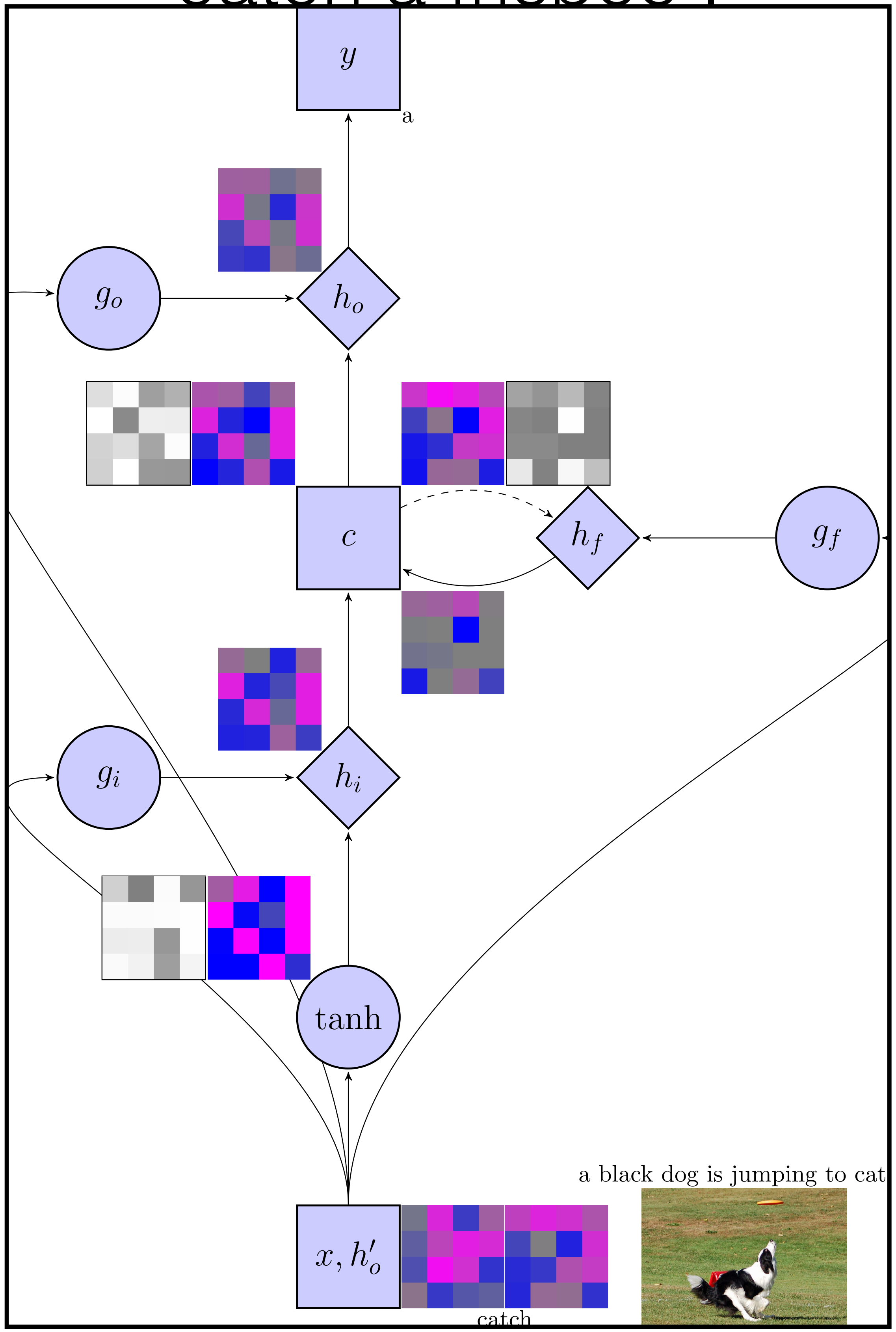
7

a dog is jumping to catch a frisbee .



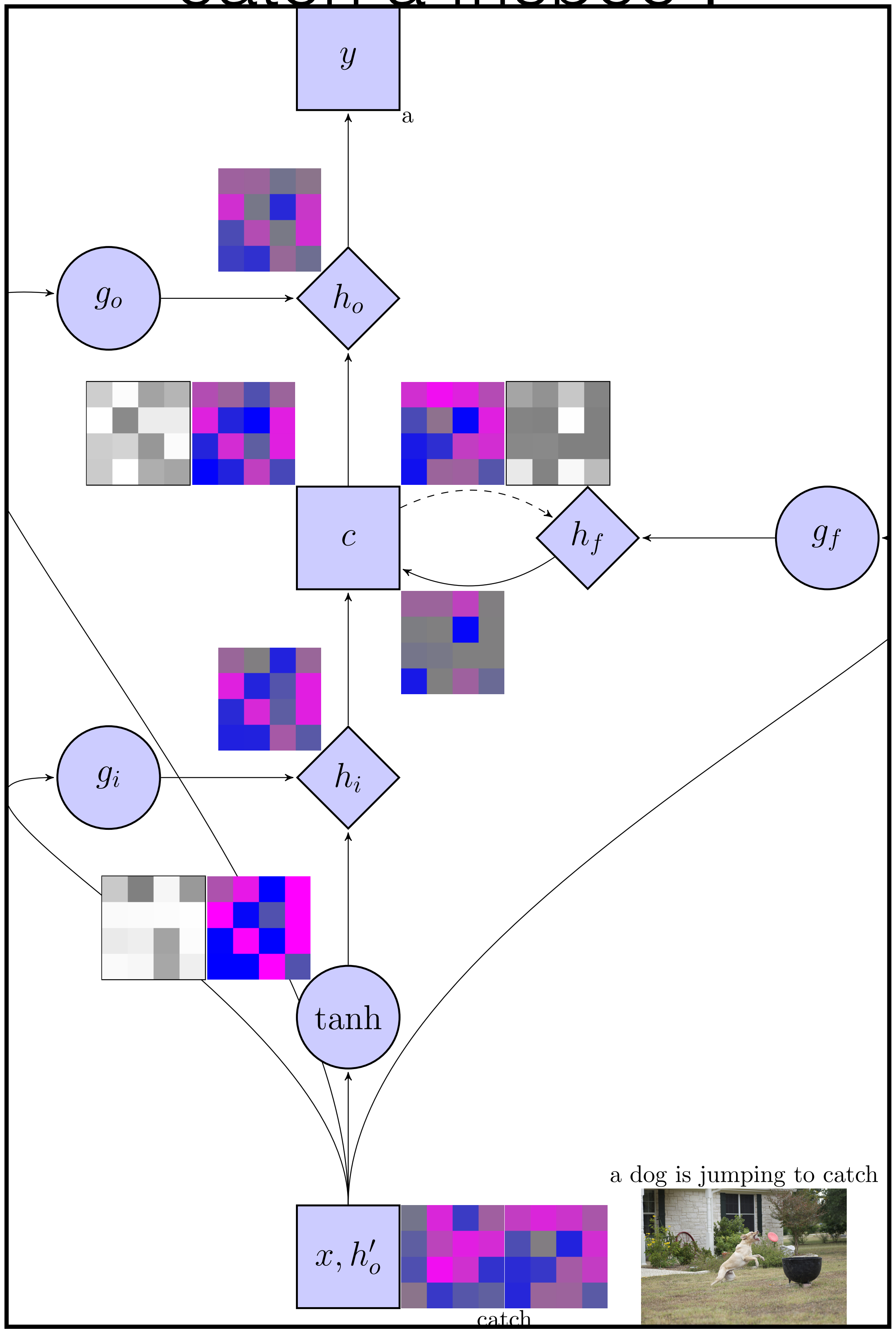
6

a black dog is jumping to catch a frisbee .



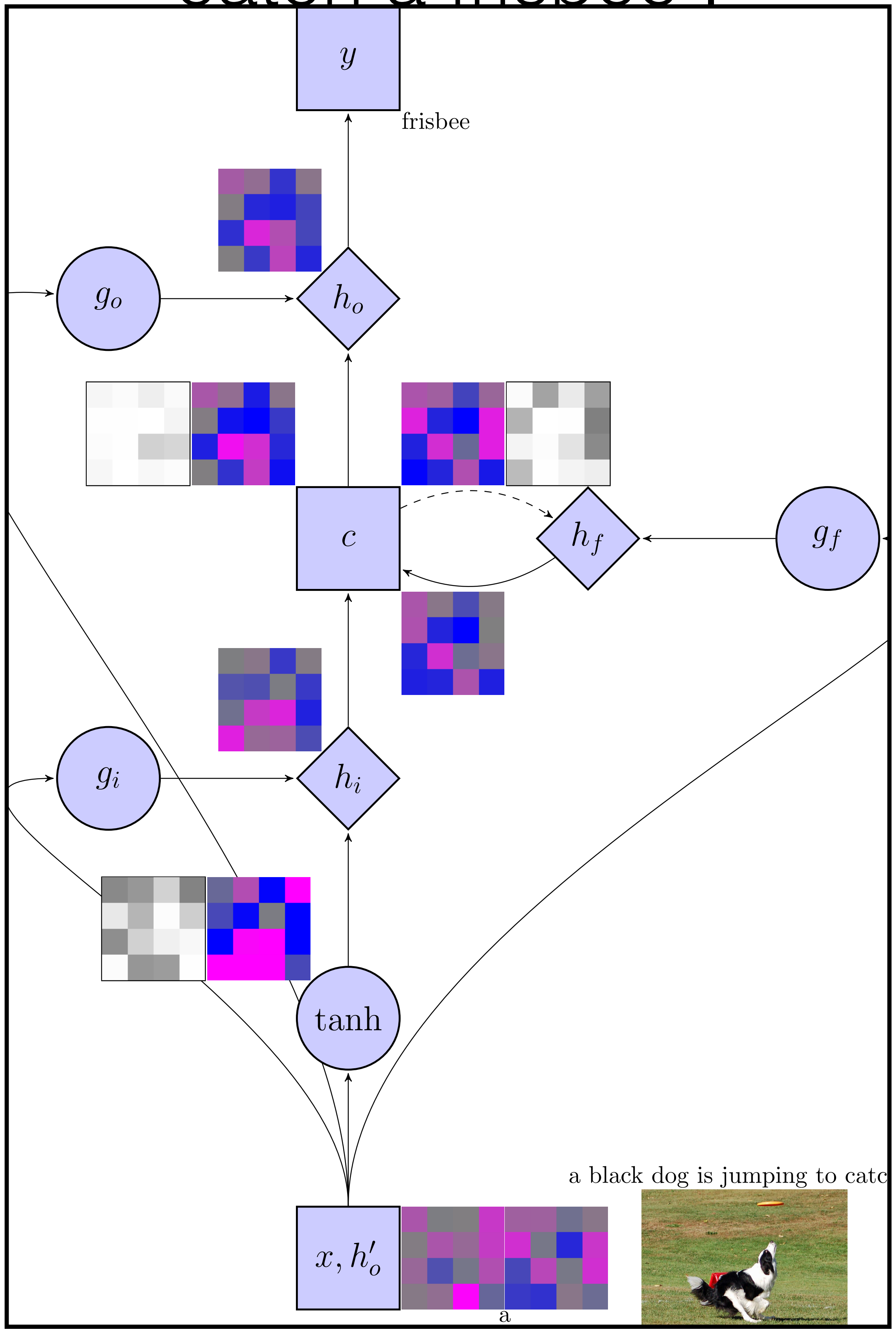
8

a dog is jumping to catch a frisbee .



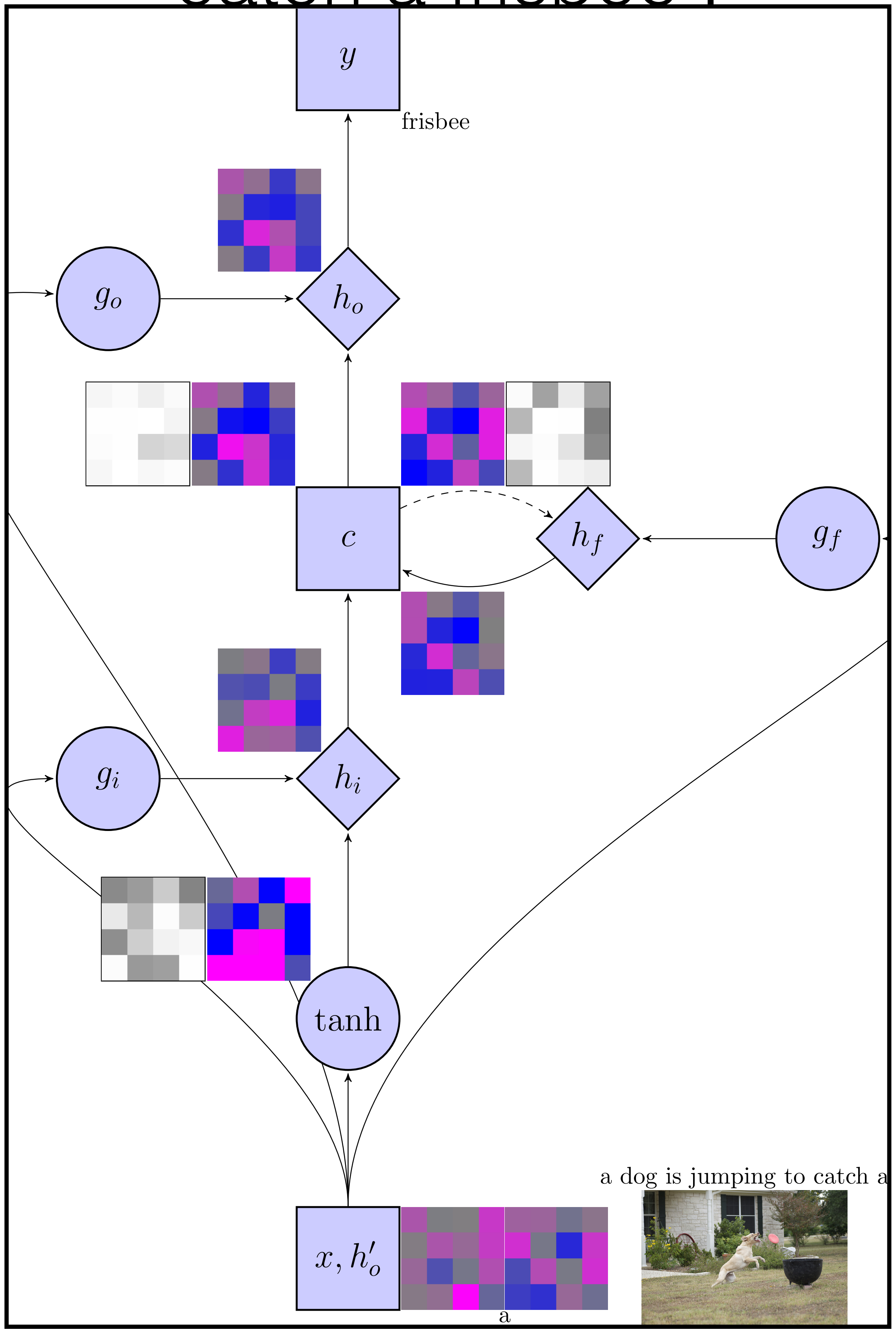
7

a black dog is jumping to catch a frisbee .



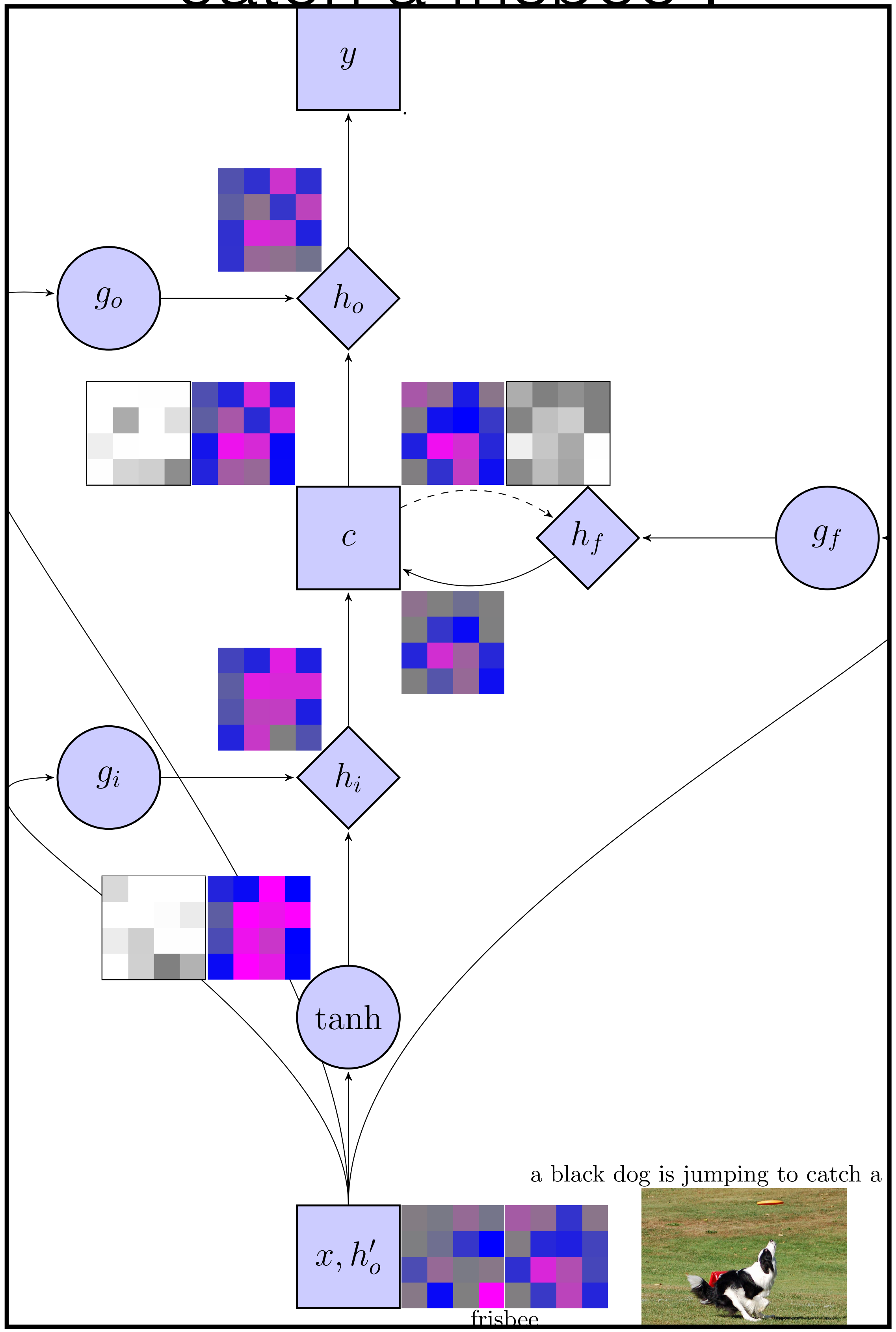
9

a dog is jumping to catch a frisbee .



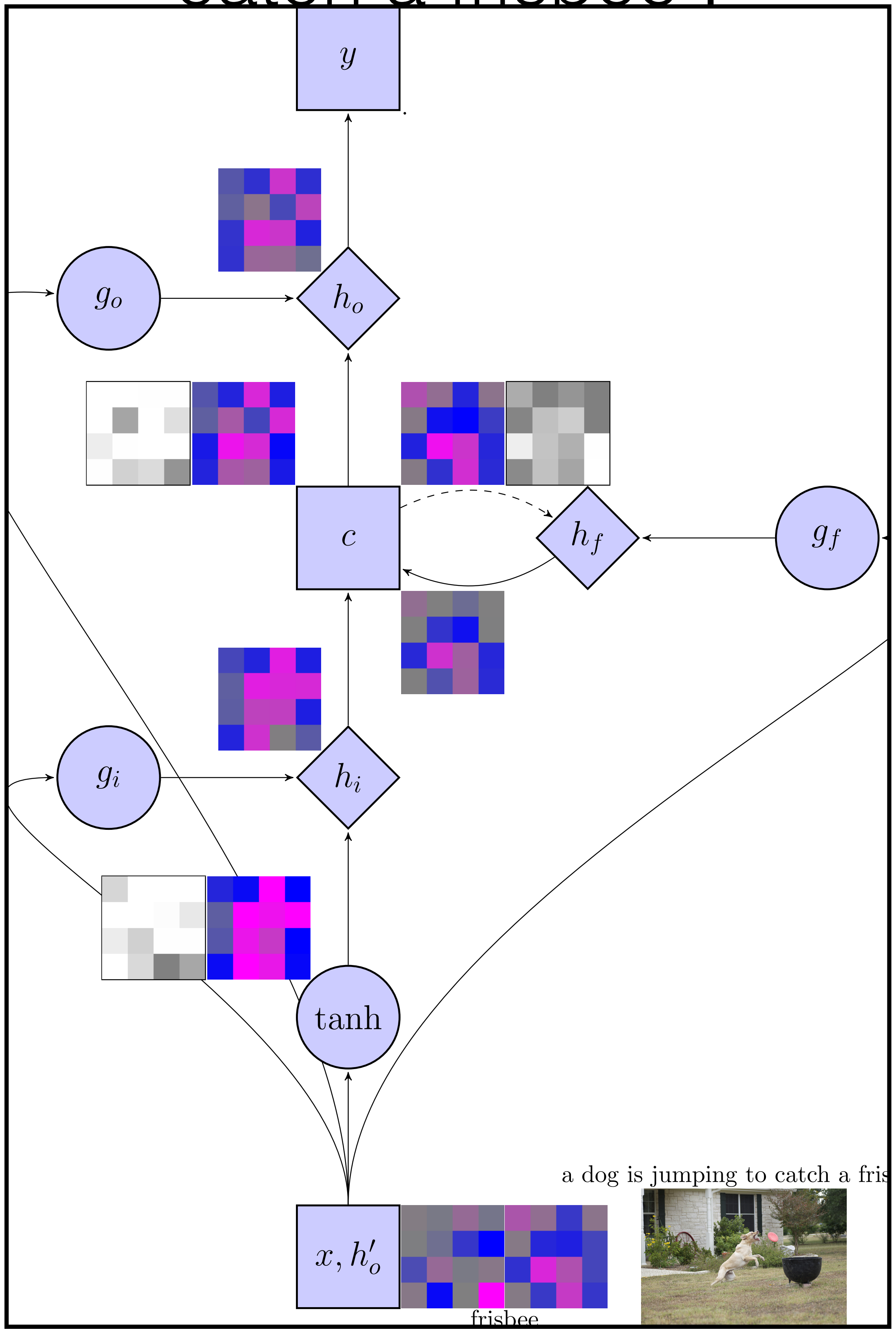
8

a black dog is jumping to catch a frisbee .



10

a dog is jumping to catch a frisbee .



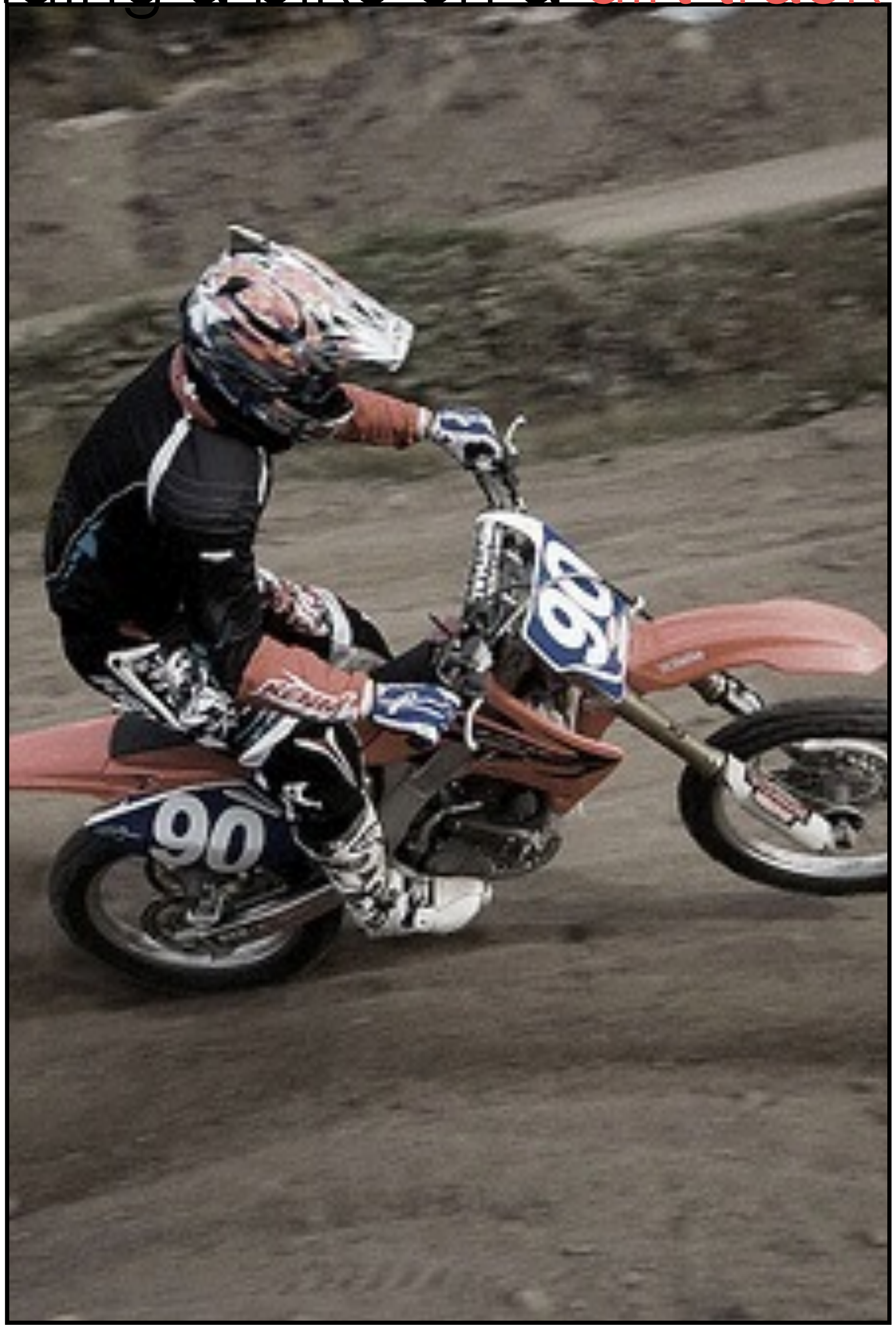
9

Why all dogs end with “frisbee”?

Count last word in training sentences with “dog” and “frisbee”:

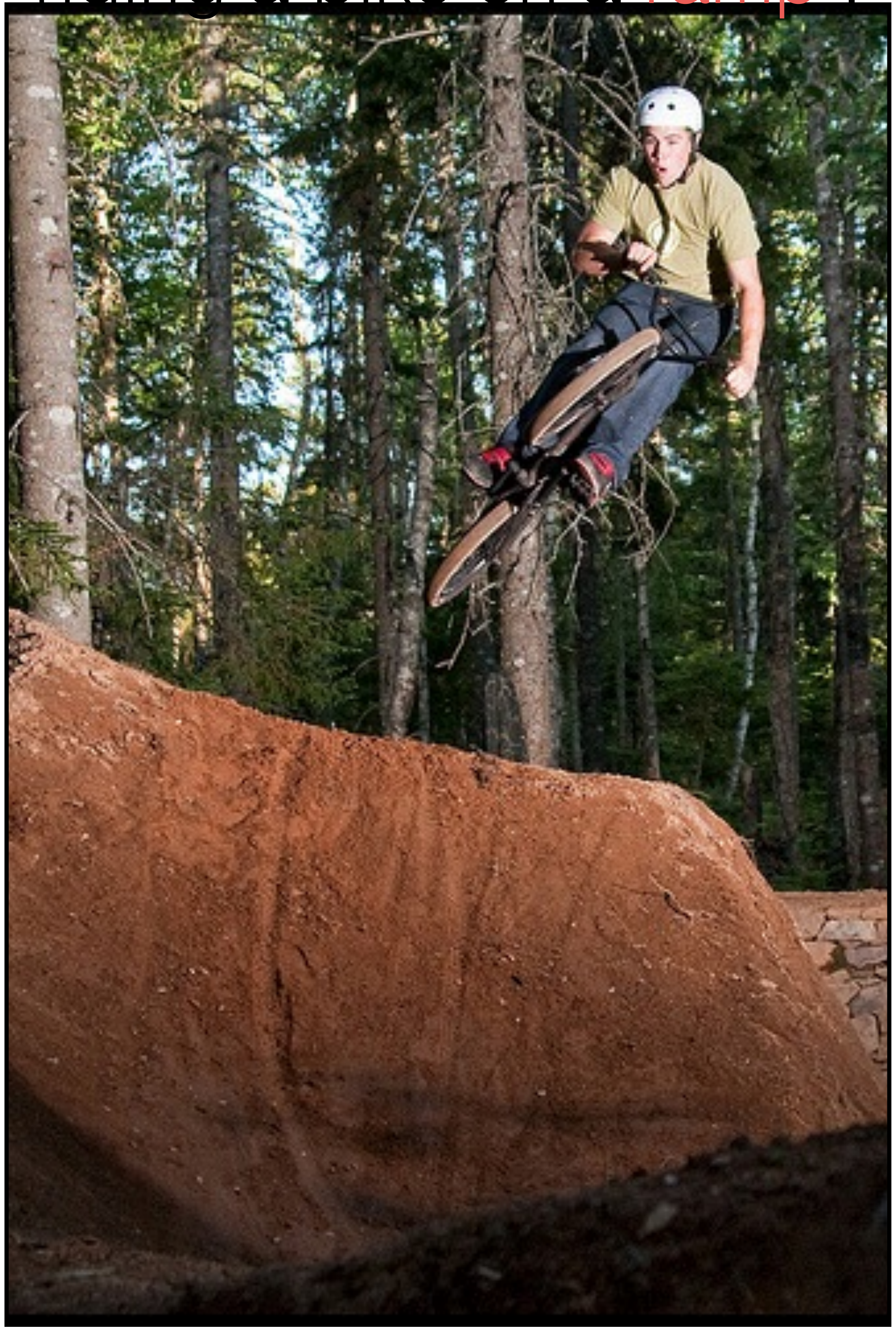
```
86 frisbee
30 mouth
15 snow
15 grass
11 field
11 dog
 8 toy
 7 water
6 yard
6 disc
6 air
5 watches
5 midair
5 background
4 watch
4 park
4 it
4 ground
4 fence
4 beach
3 road
3 object
3 boat
3 ball
2 other
2 mouths
2 man
2 legs
2 hand
2 dogs
1 underfoot
1 …
```

a man in a blue shirt is riding a bike on a dirt track .



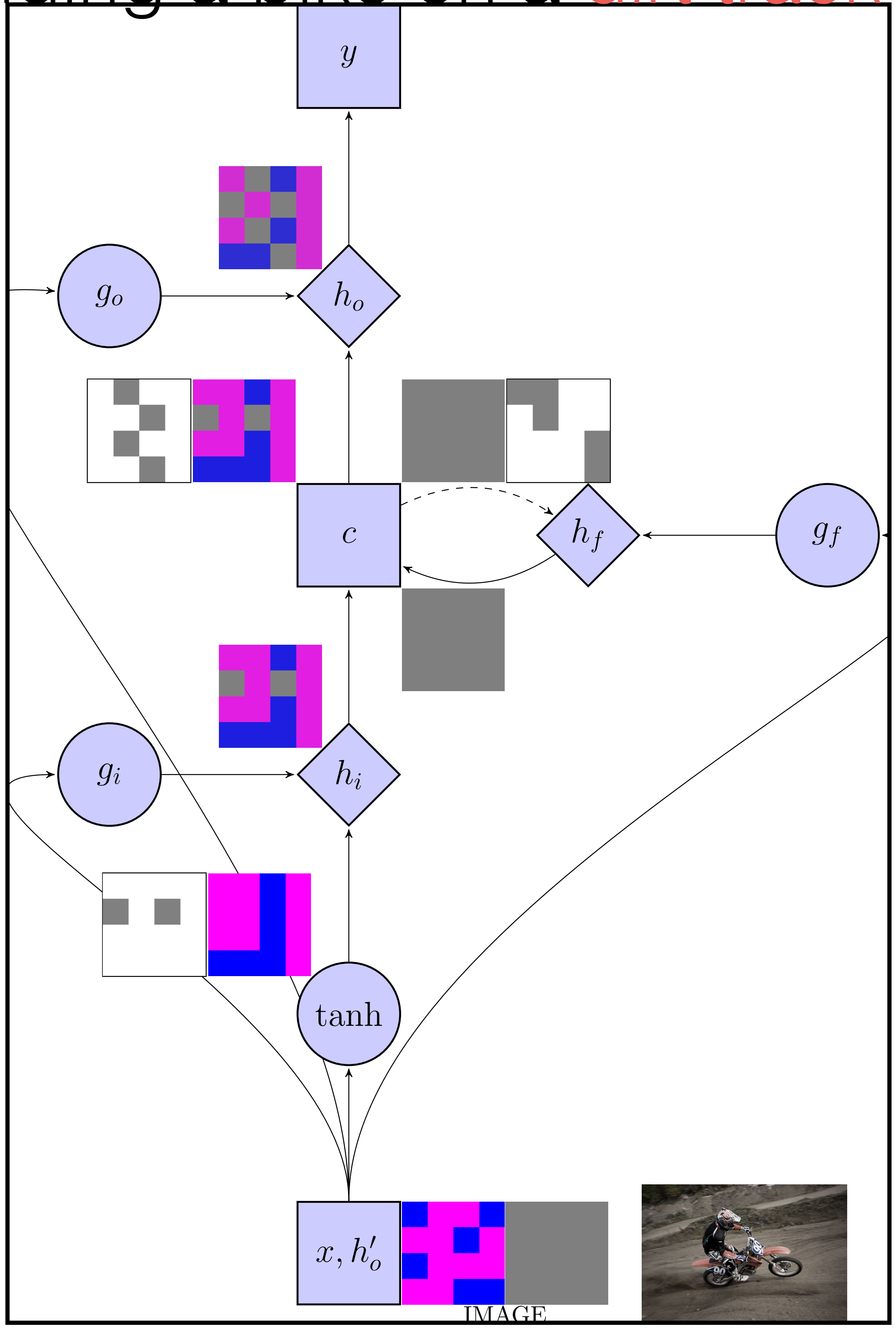
2891617125_f939f604c7.jpg

a man in a blue shirt is riding a bike on a ramp .



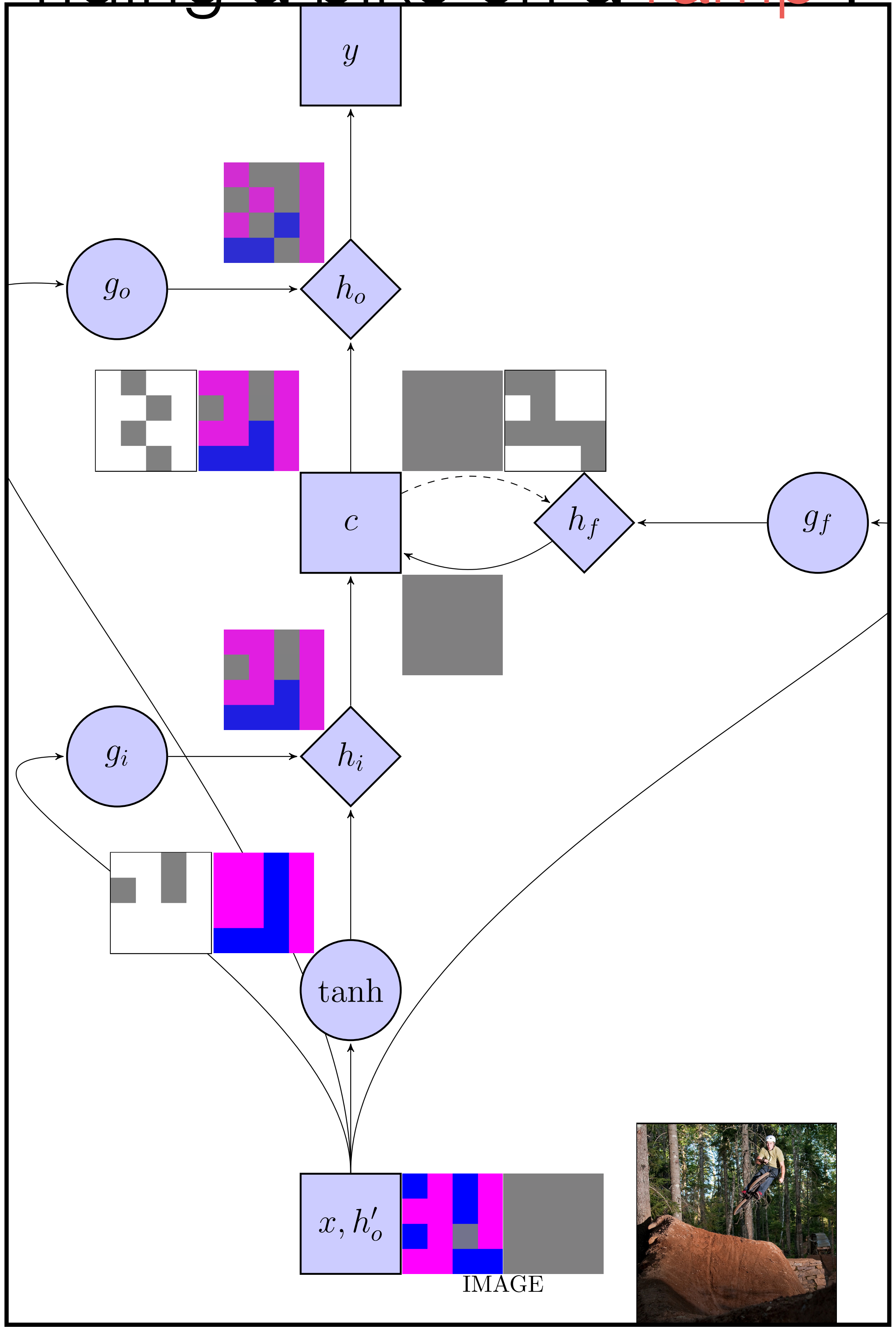
3640422448_a0f42e4559.jpg

a man in a blue shirt is riding a bike on a dirt track .



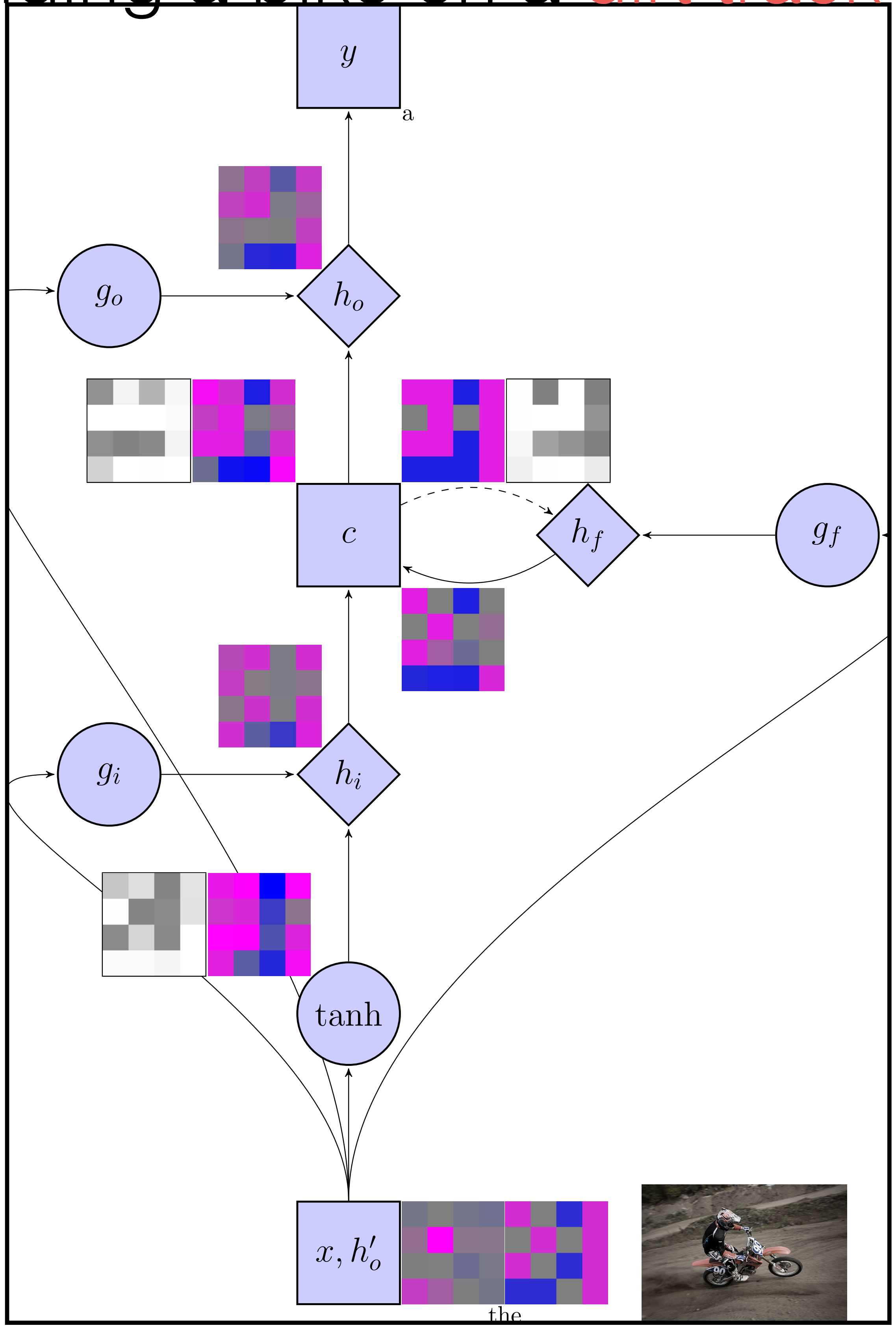
0

a man in a blue shirt is riding a bike on a ramp .



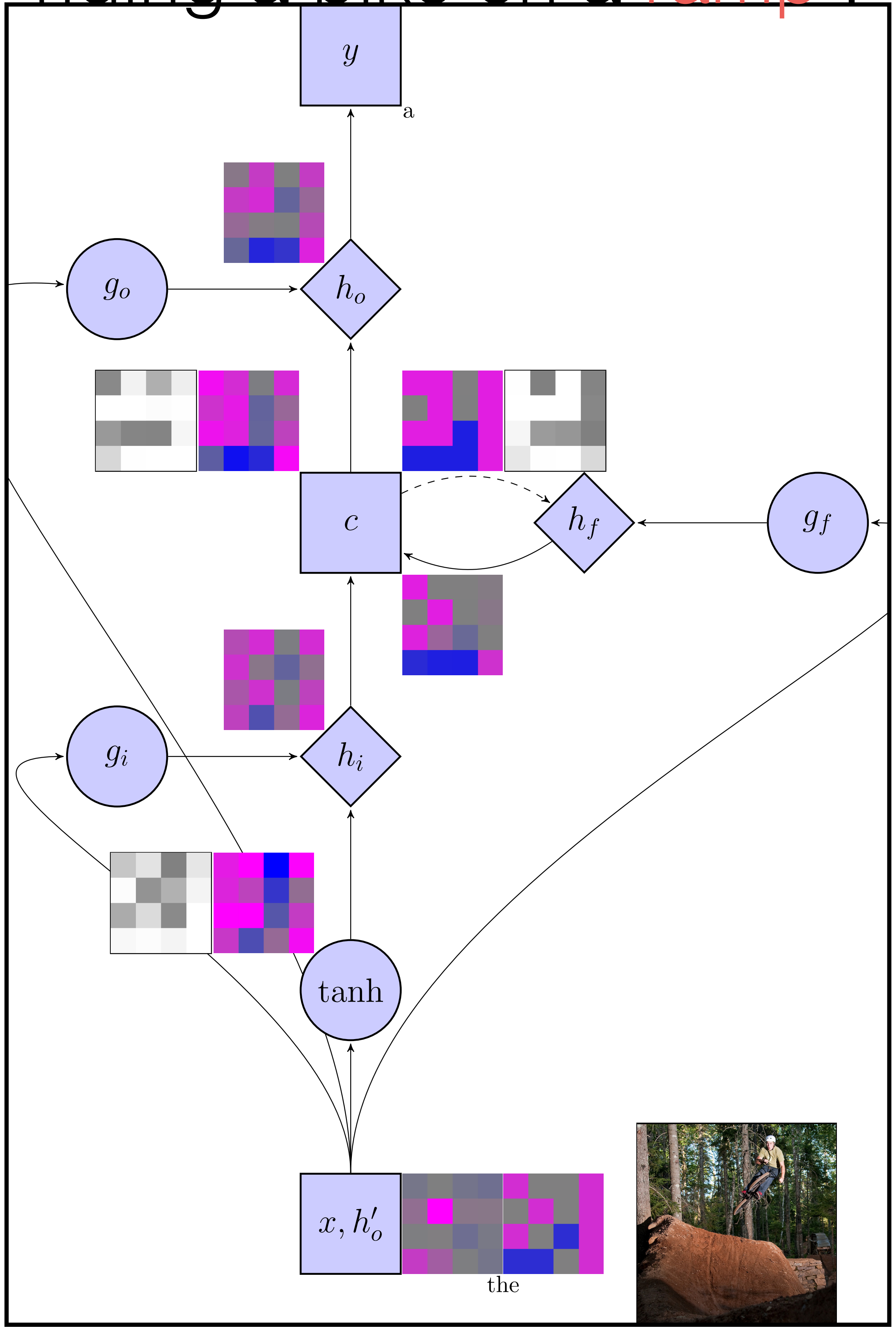
0

a man in a blue shirt is riding a bike on a dirt track .



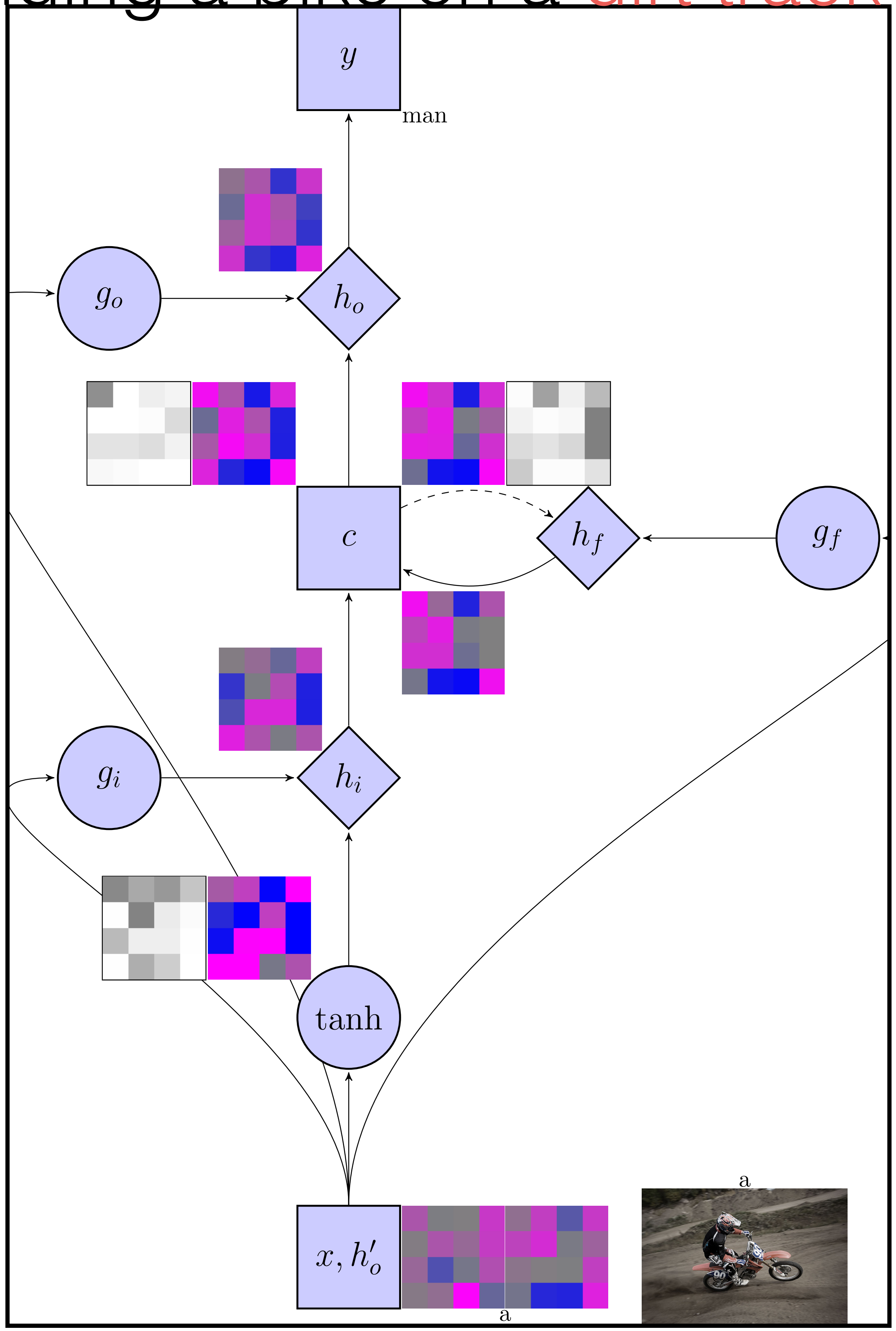
0

a man in a blue shirt is riding a bike on a ramp .



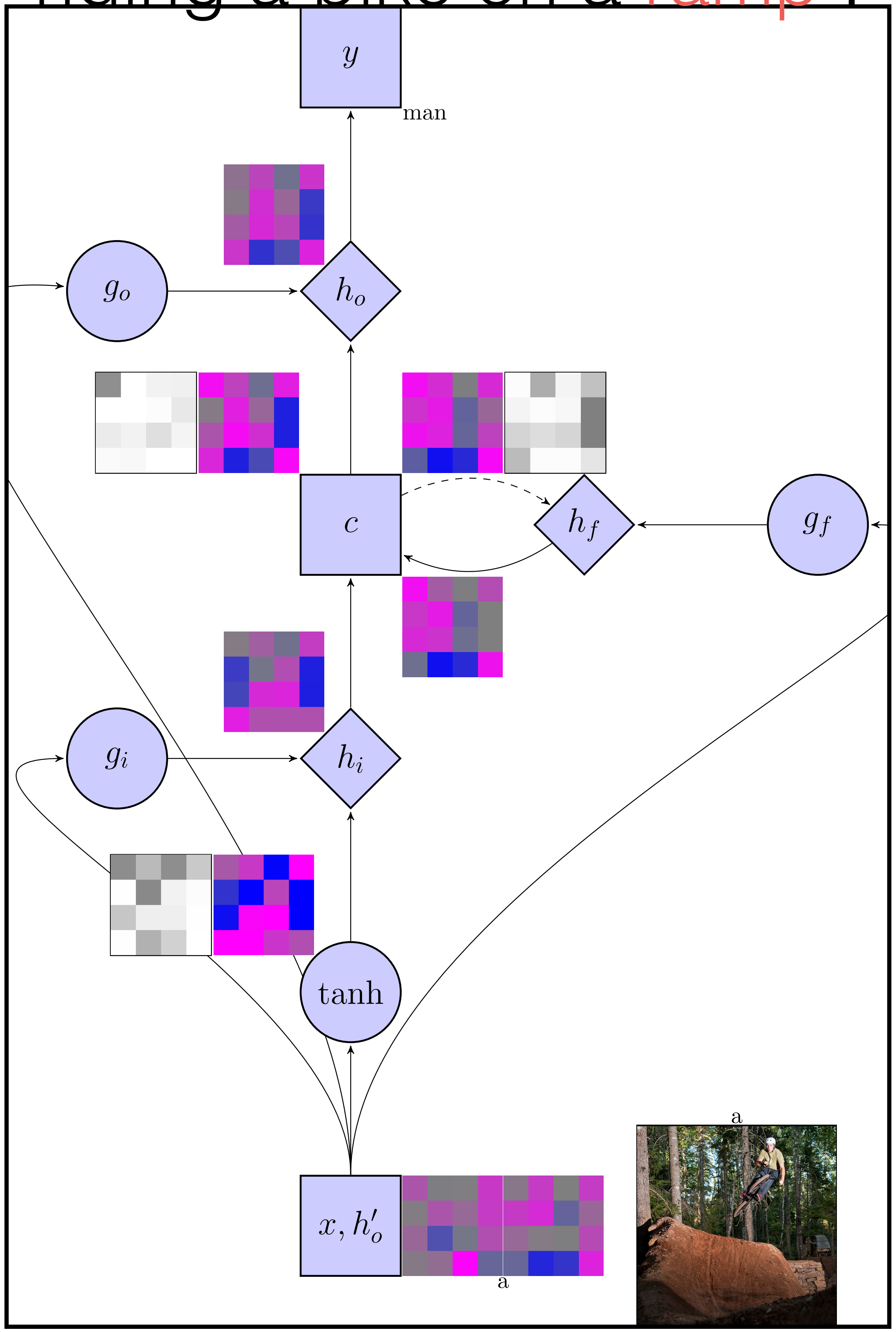
0

a man in a blue shirt is riding a bike on a dirt track .



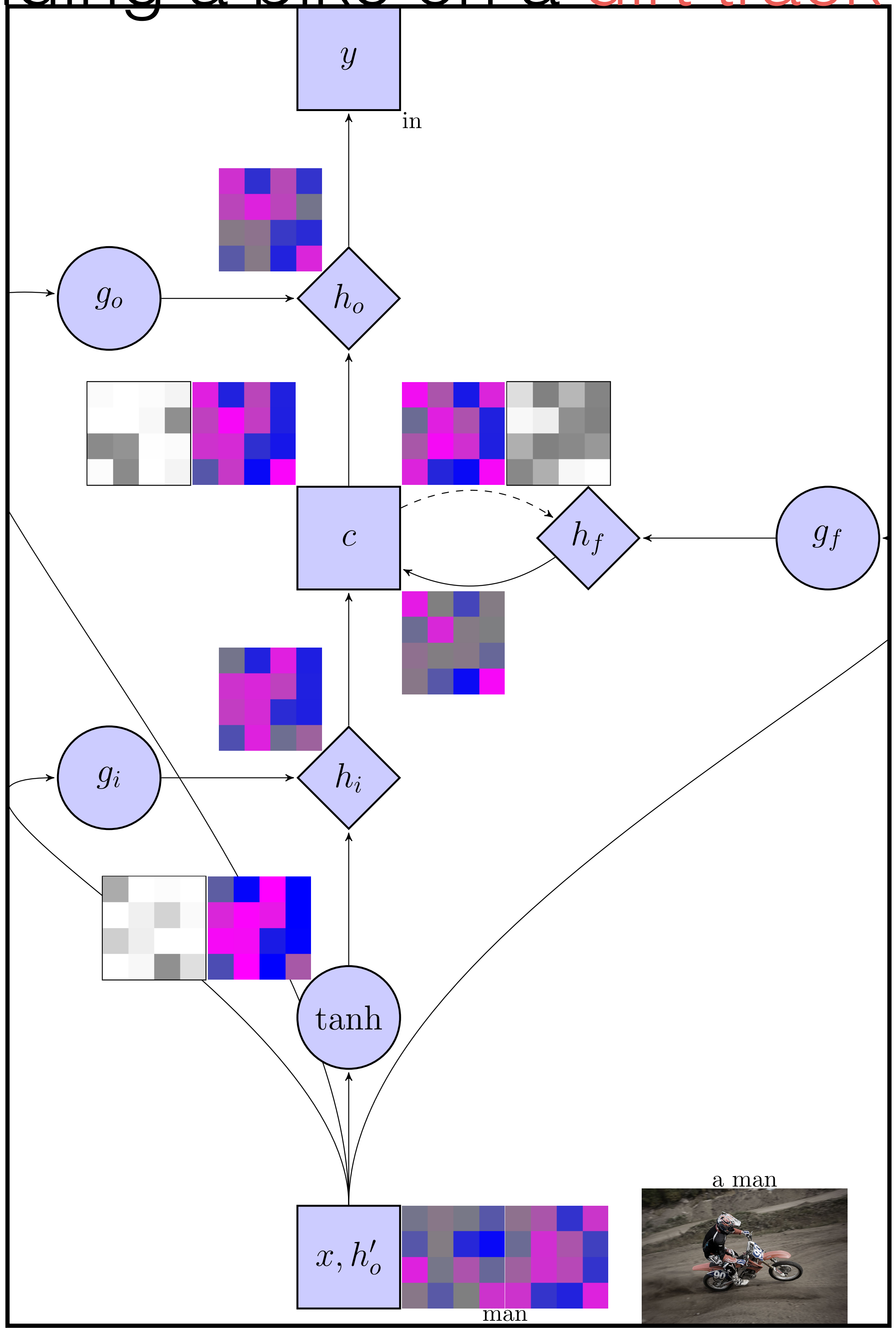
0

a man in a blue shirt is riding a bike on a ramp .



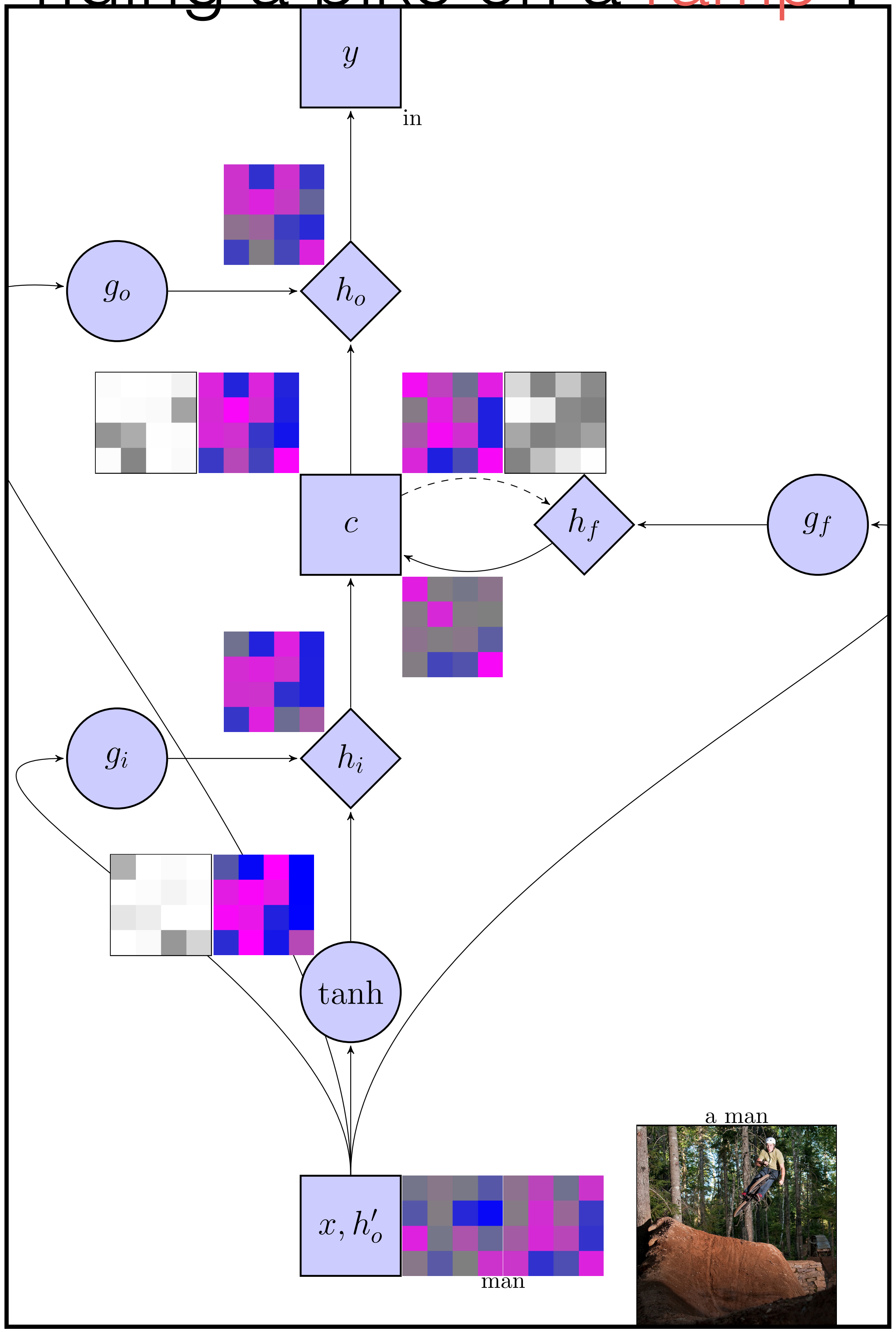
0

a man in a blue shirt is riding a bike on a dirt track .



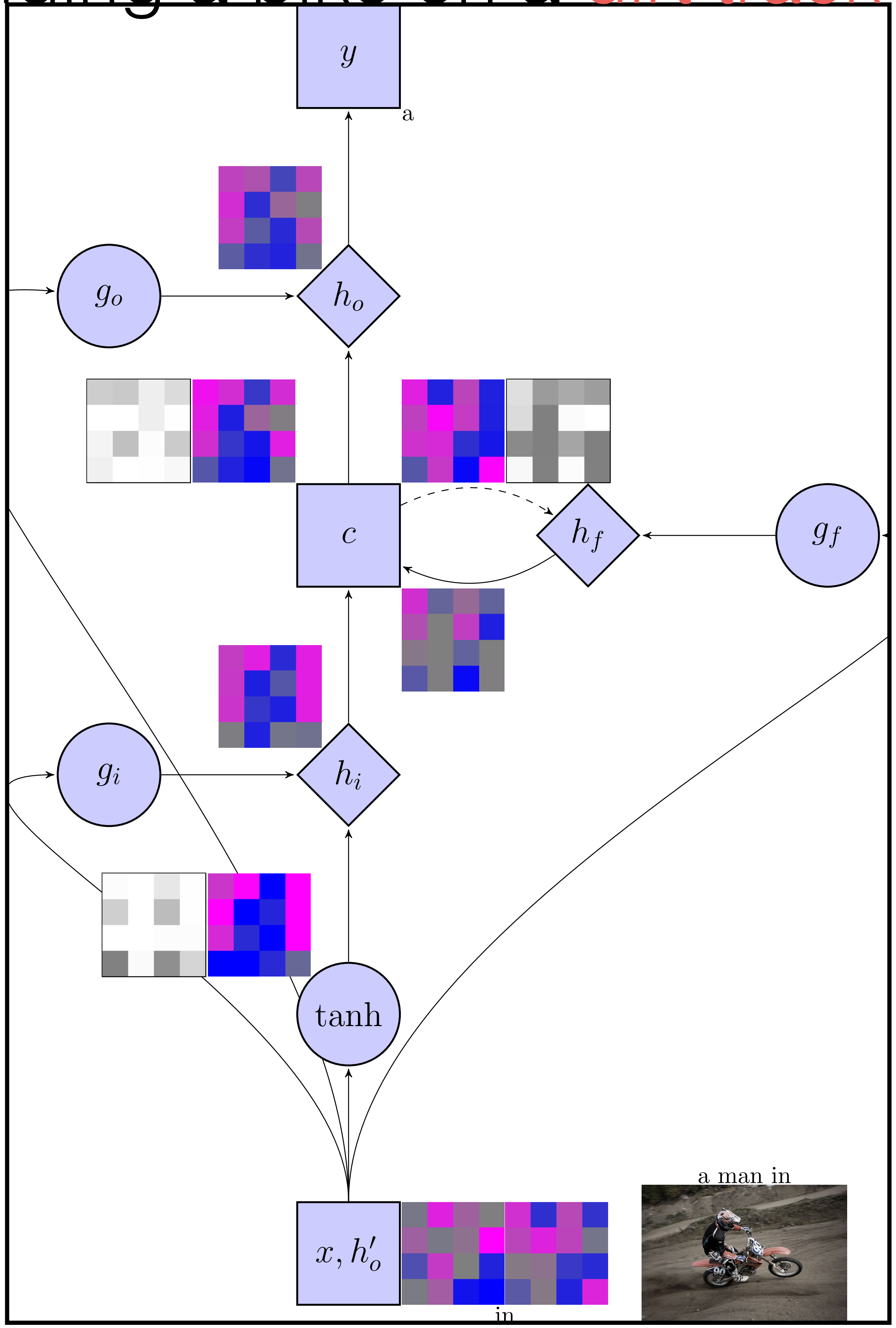
0

a man in a blue shirt is riding a bike on a ramp .



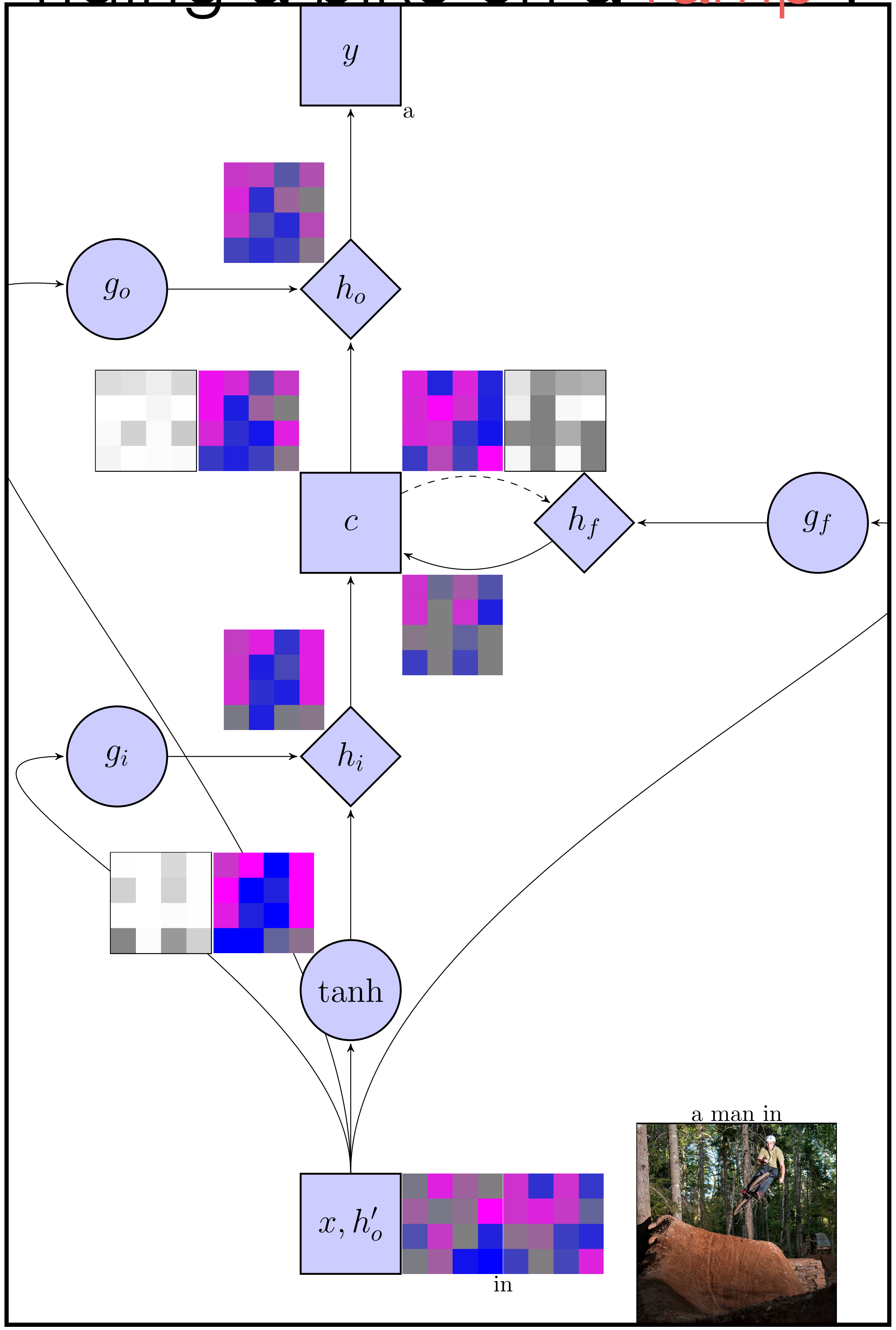
0

a man in a blue shirt is riding a bike on a dirt track .



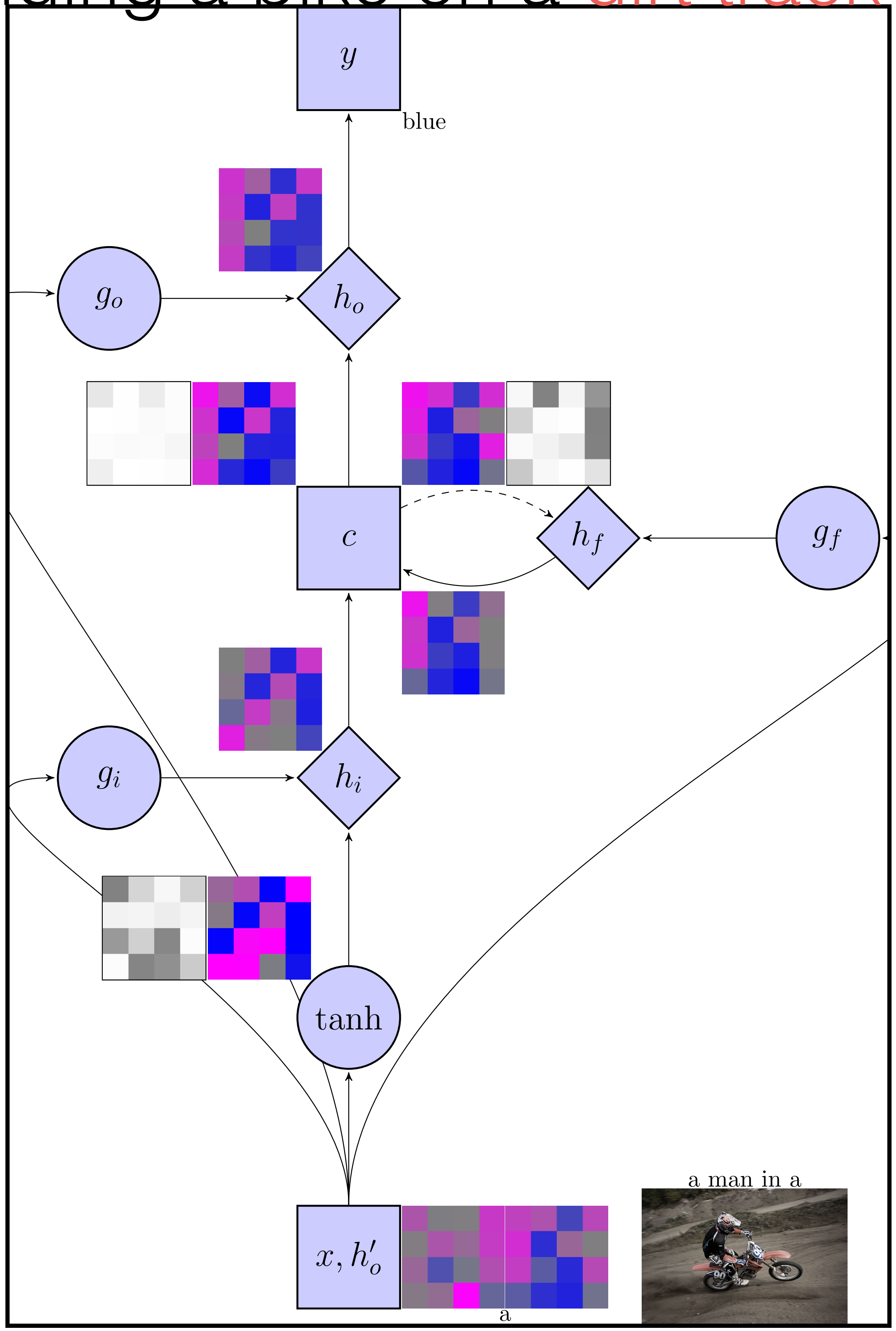
0

a man in a blue shirt is riding a bike on a ramp .



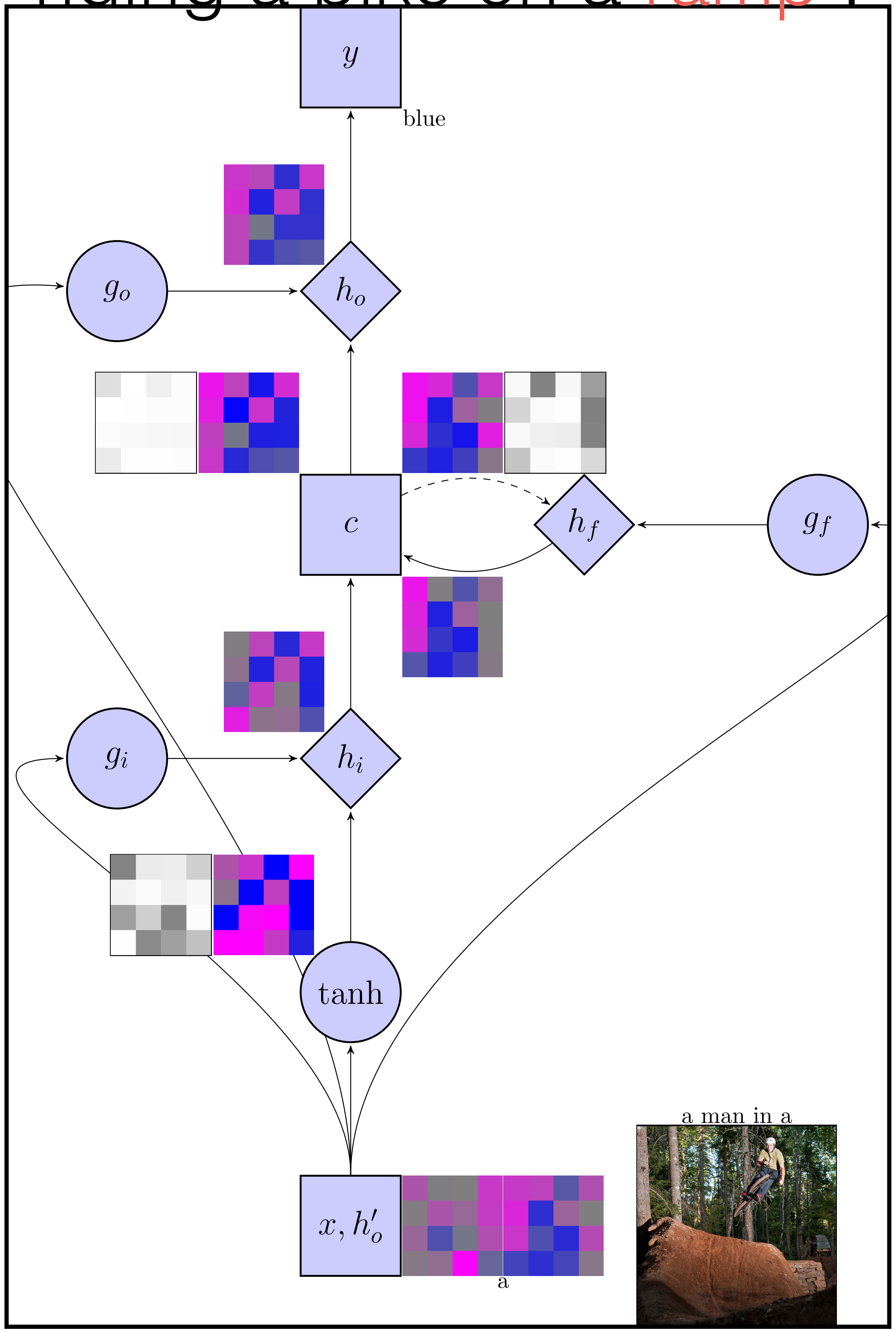
0

a man in a blue shirt is riding a bike on a dirt track .



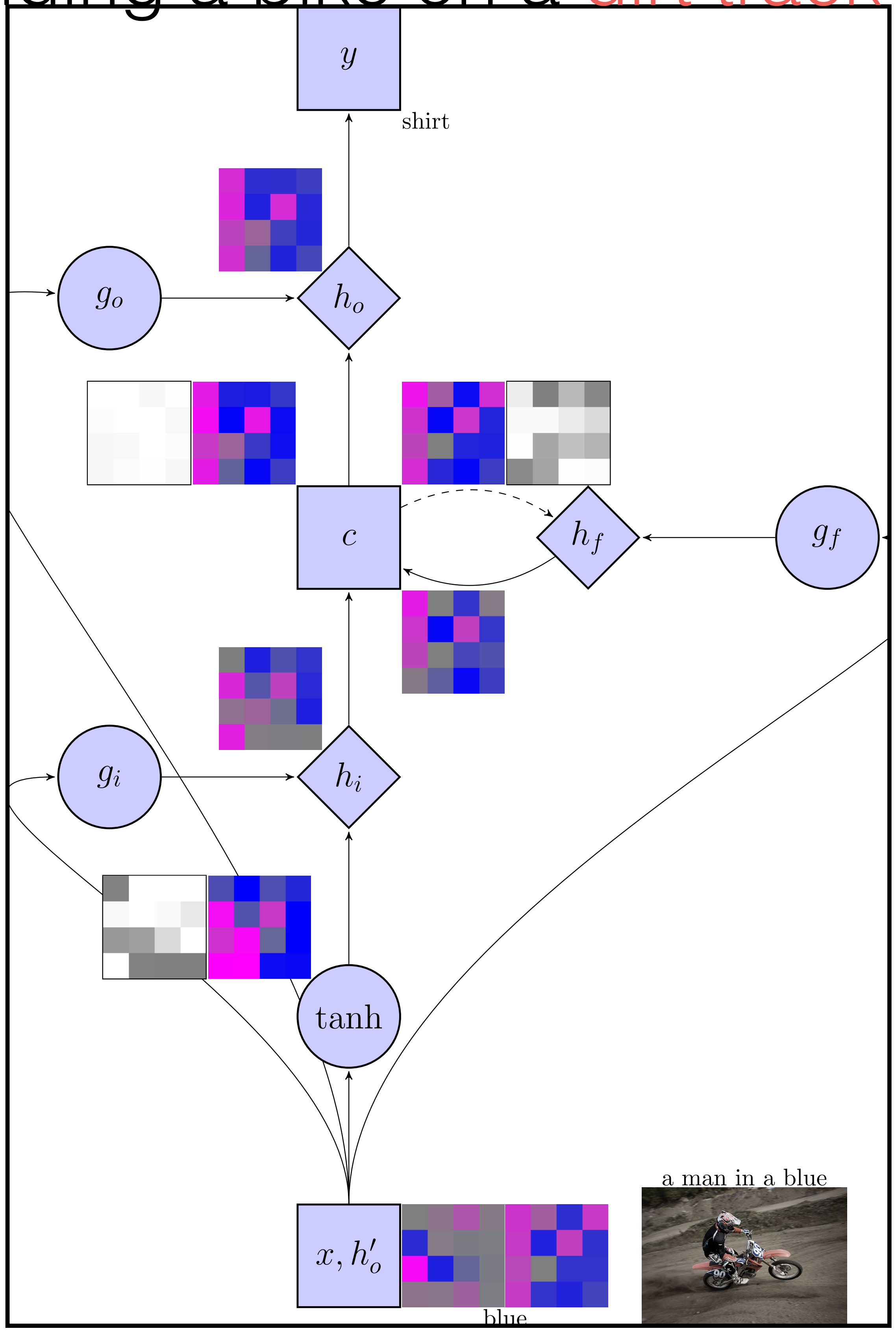
0

a man in a blue shirt is riding a bike on a ramp .



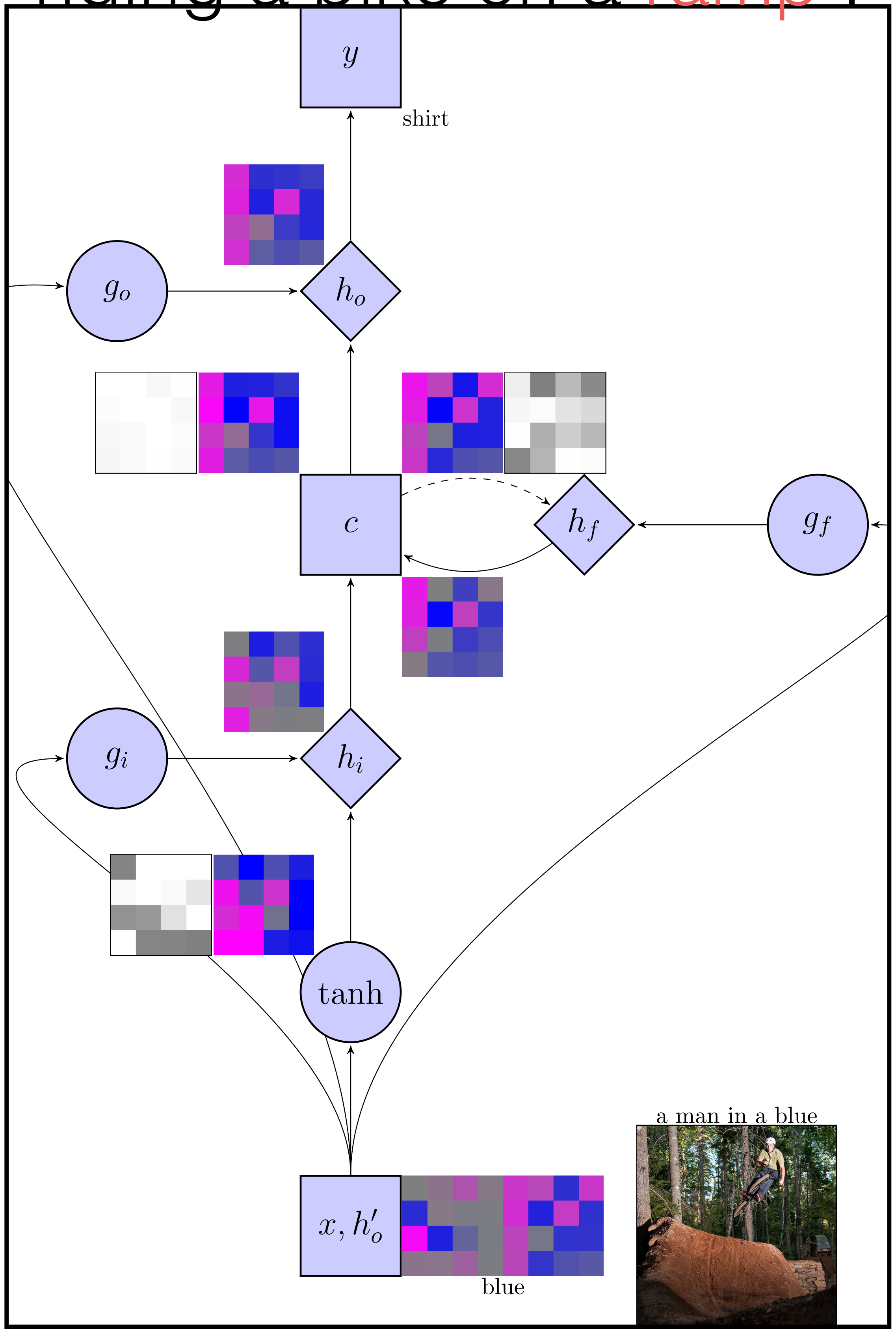
0

a man in a blue shirt is riding a bike on a dirt track .



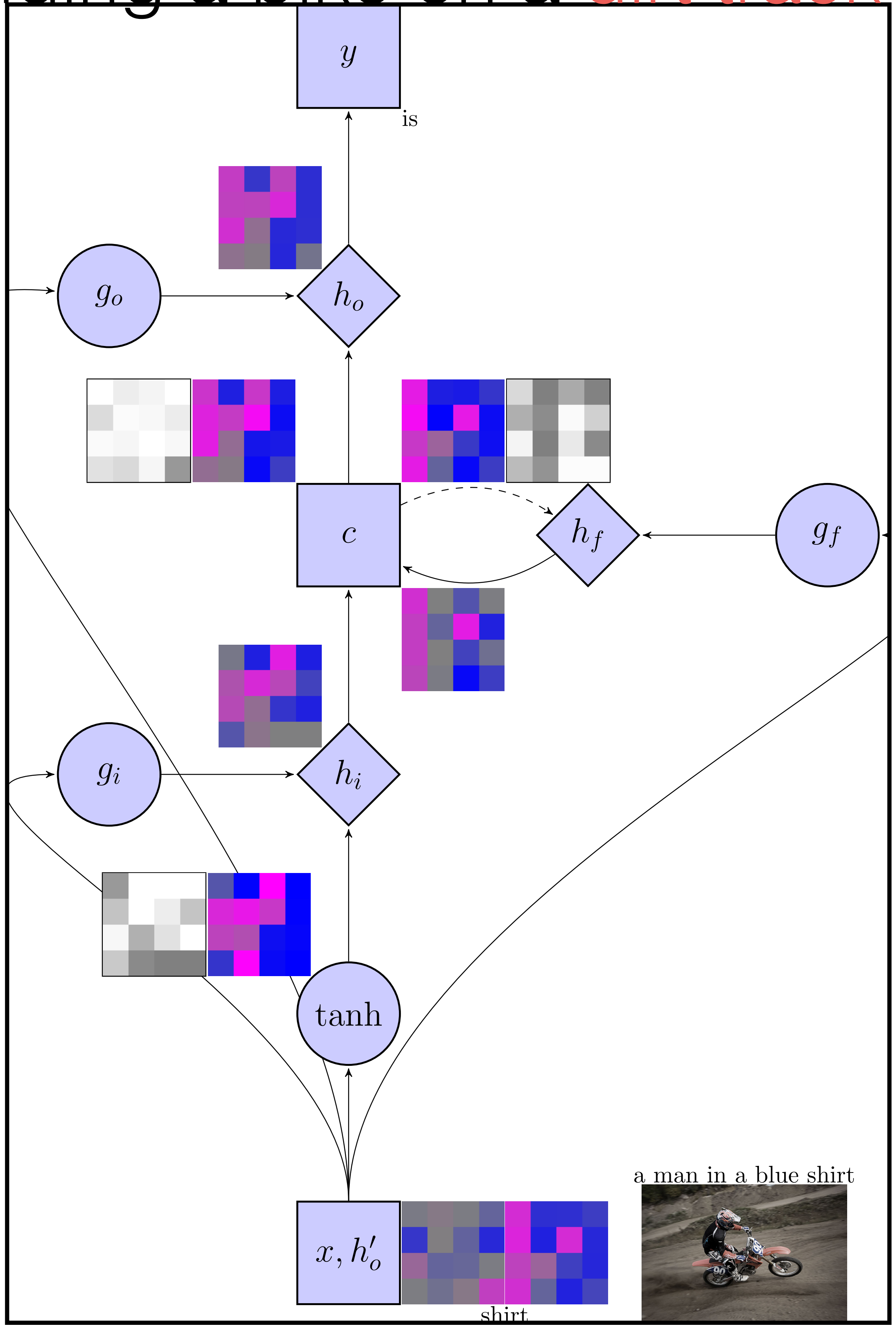
0

a man in a blue shirt is riding a bike on a ramp .



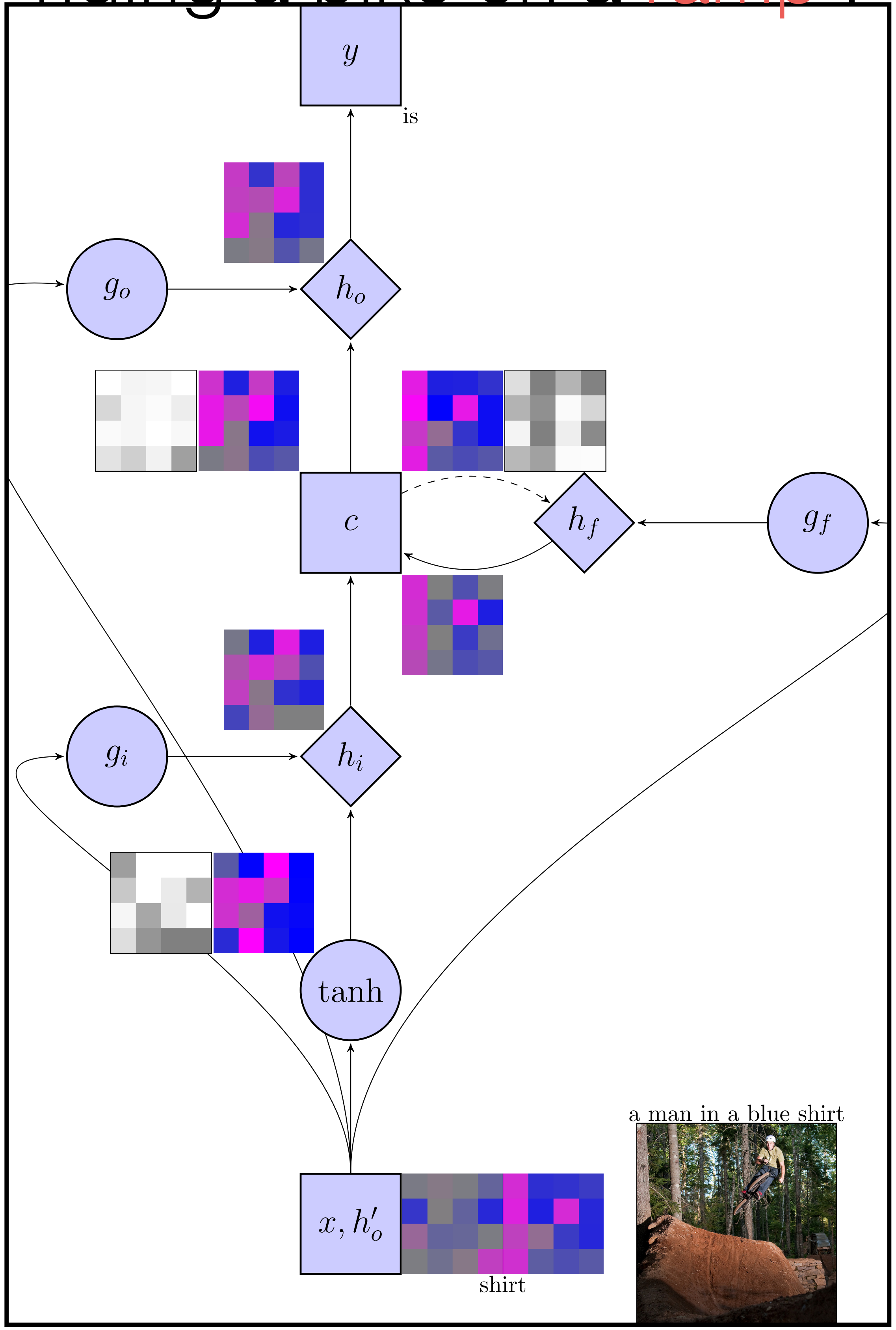
0

a man in a blue shirt is riding a bike on a dirt track .



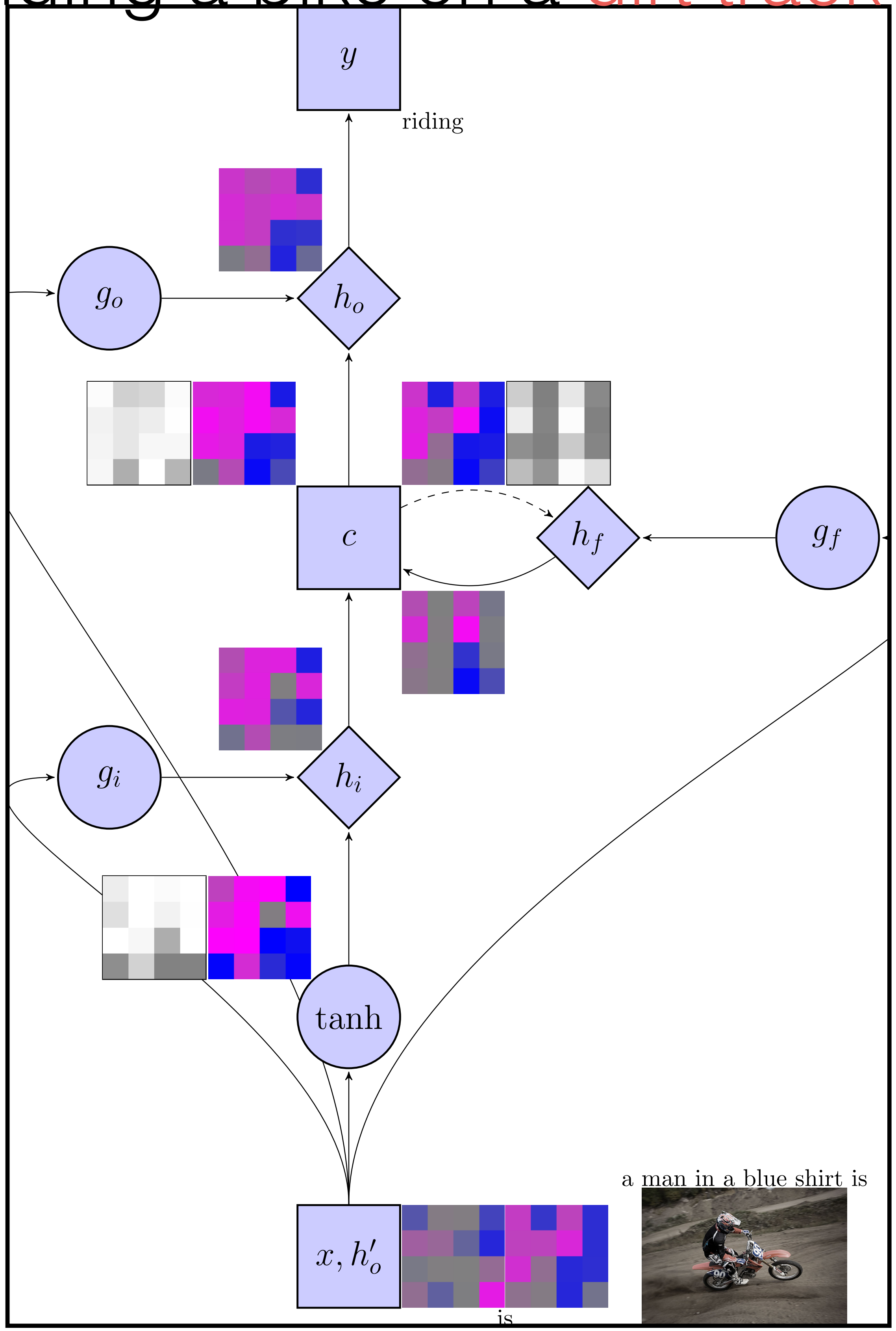
0

a man in a blue shirt is riding a bike on a ramp .



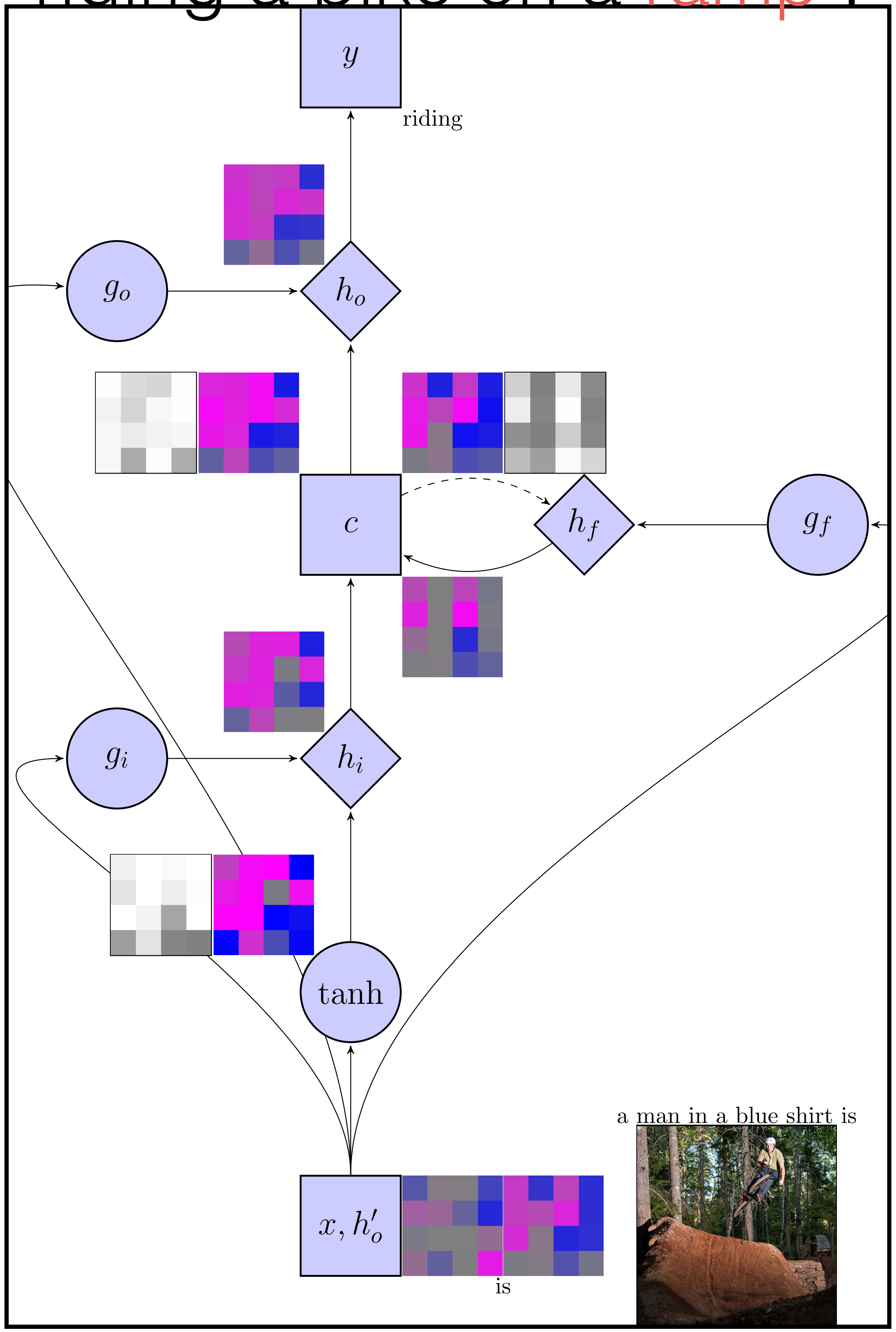
0

a man in a blue shirt is riding a bike on a dirt track .



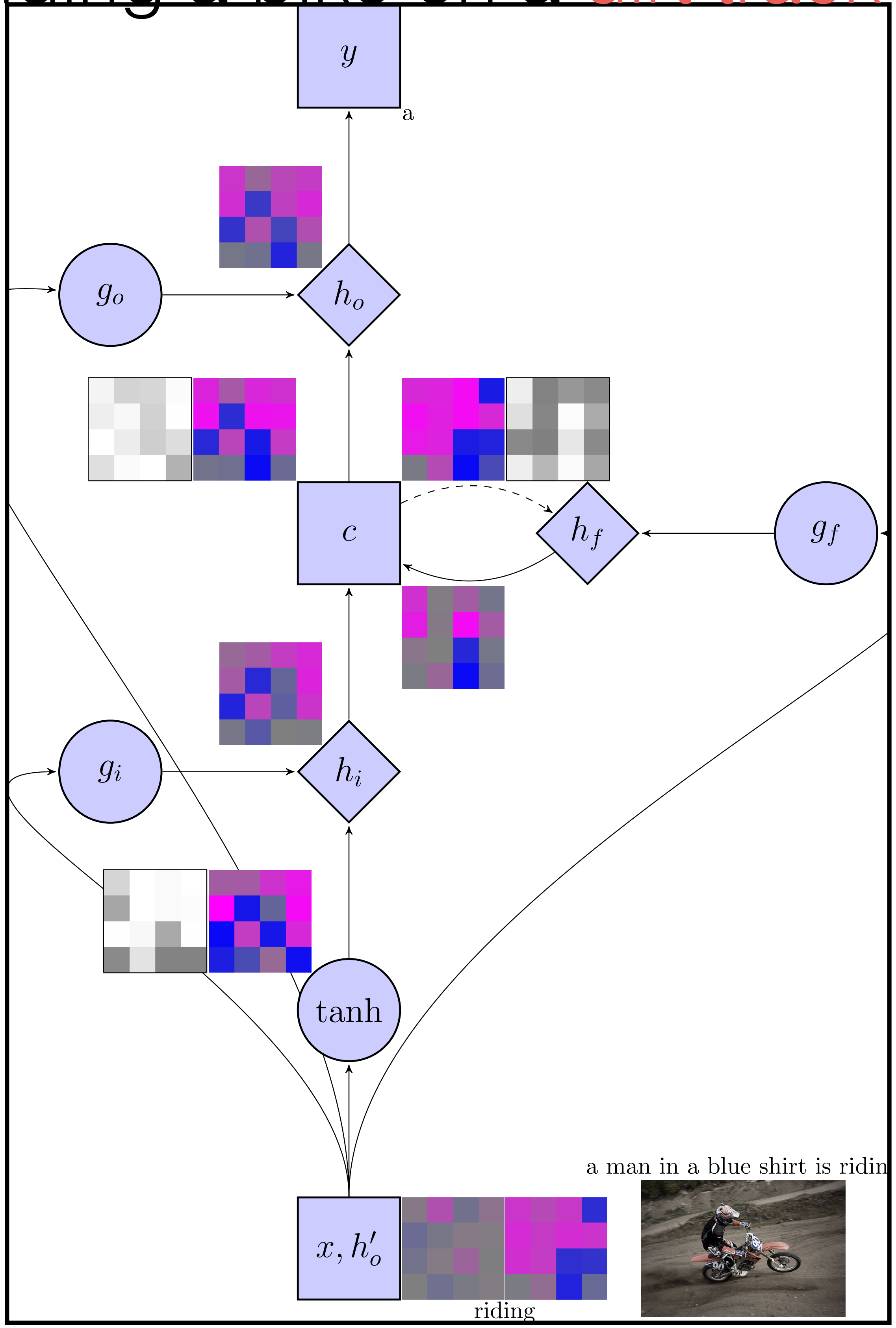
0

a man in a blue shirt is riding a bike on a ramp .



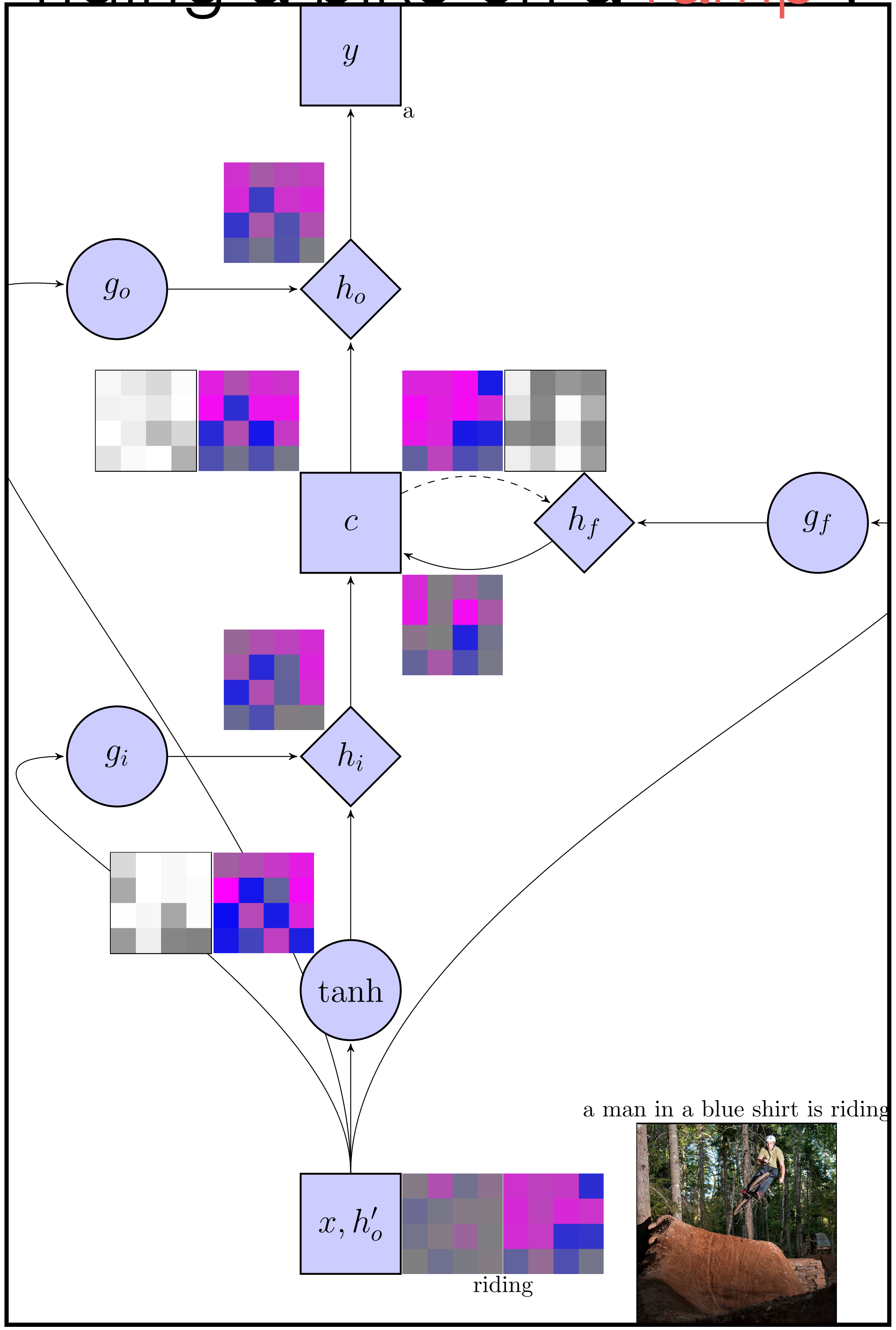
0

a man in a blue shirt is riding a bike on a dirt track .



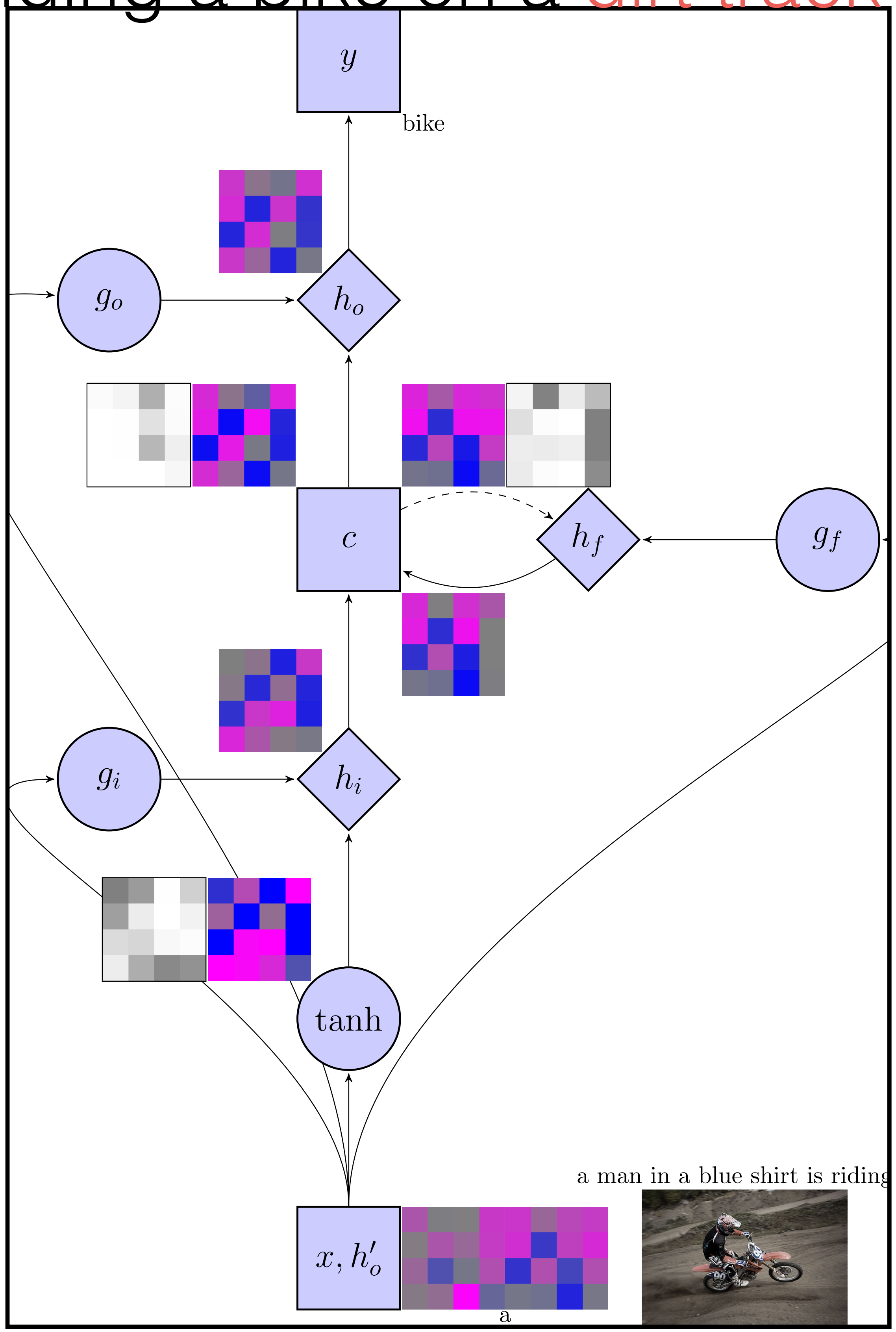
0

a man in a blue shirt is riding a bike on a ramp .



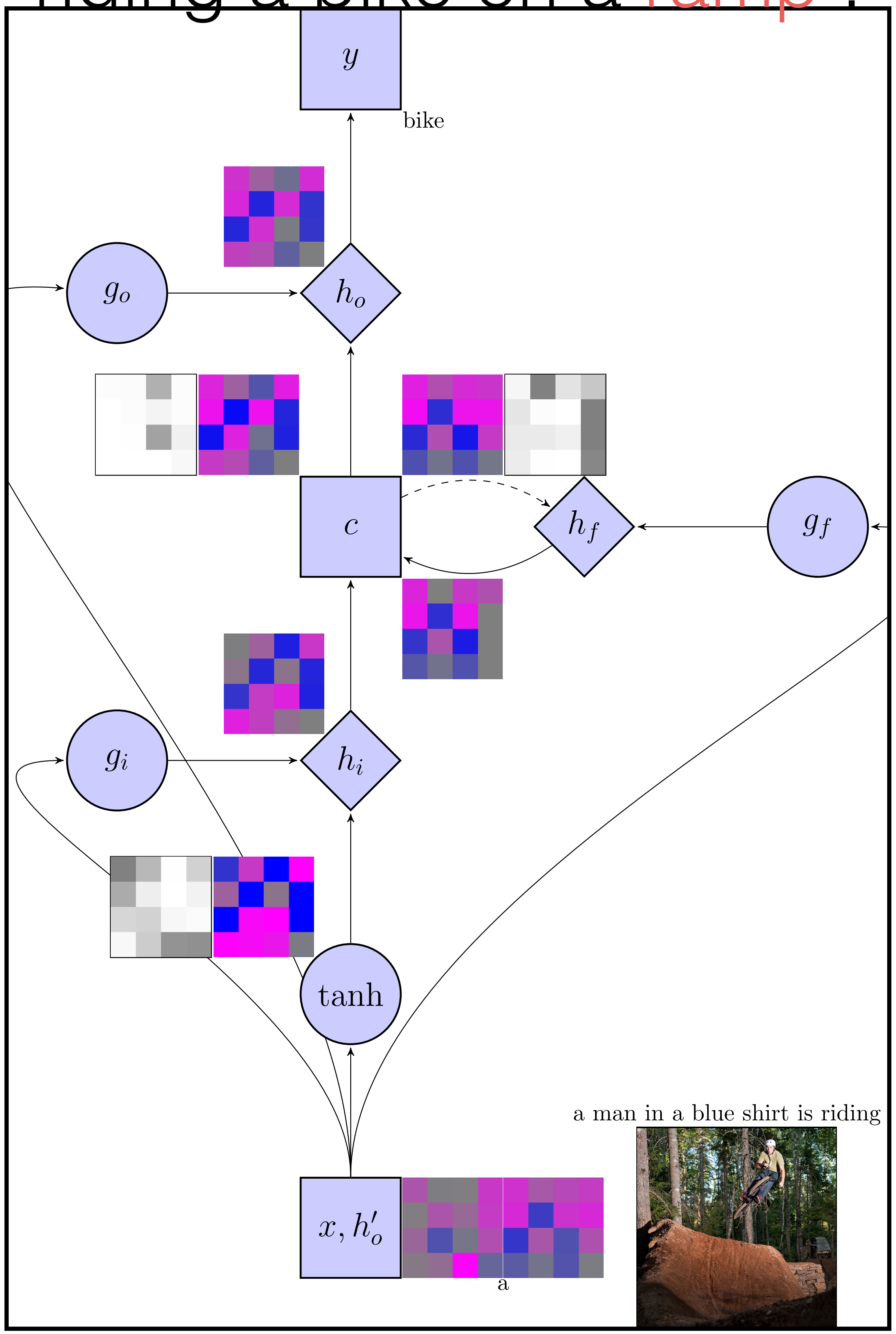
0

a man in a blue shirt is riding a bike on a dirt track .



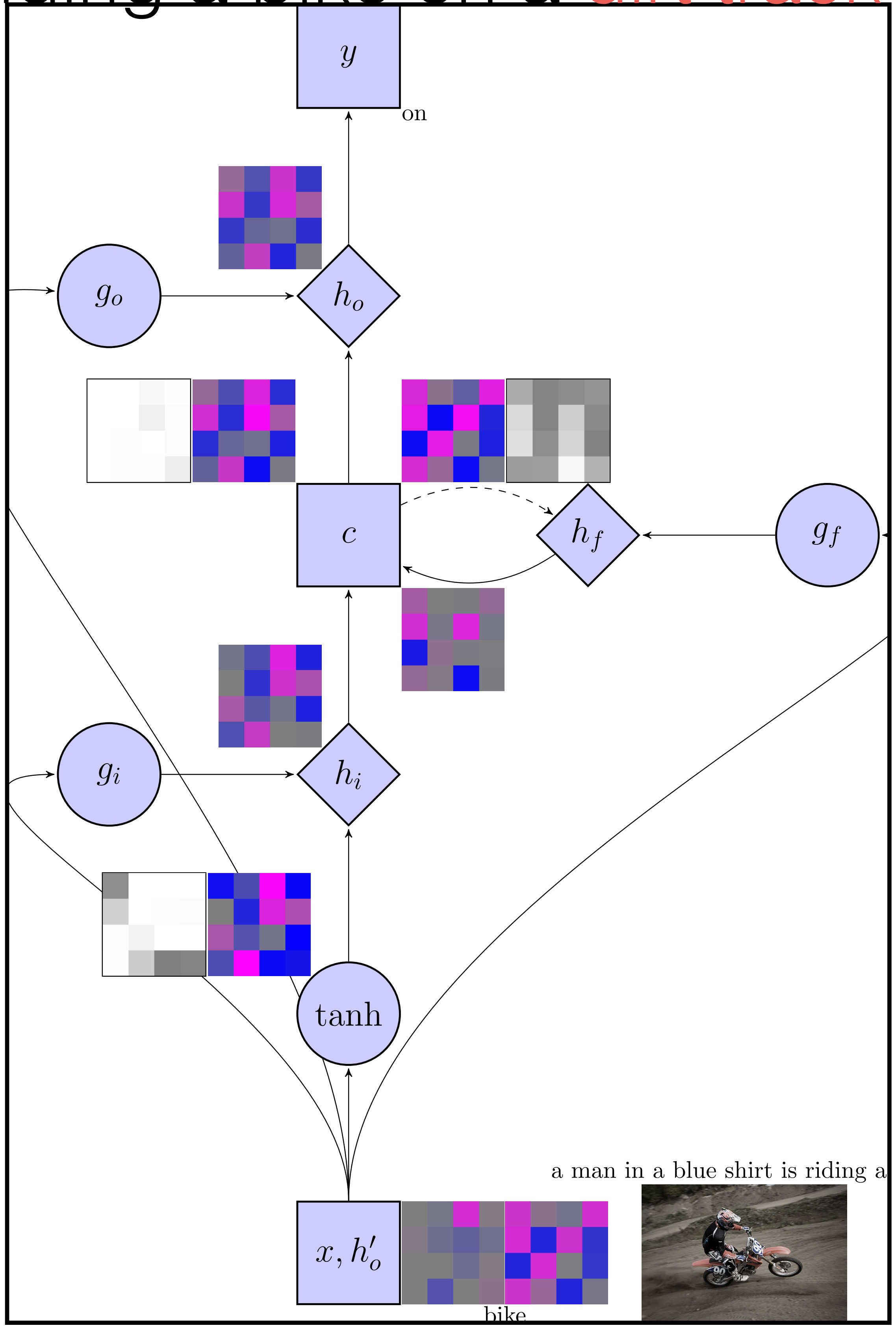
0

a man in a blue shirt is riding a bike on a ramp .



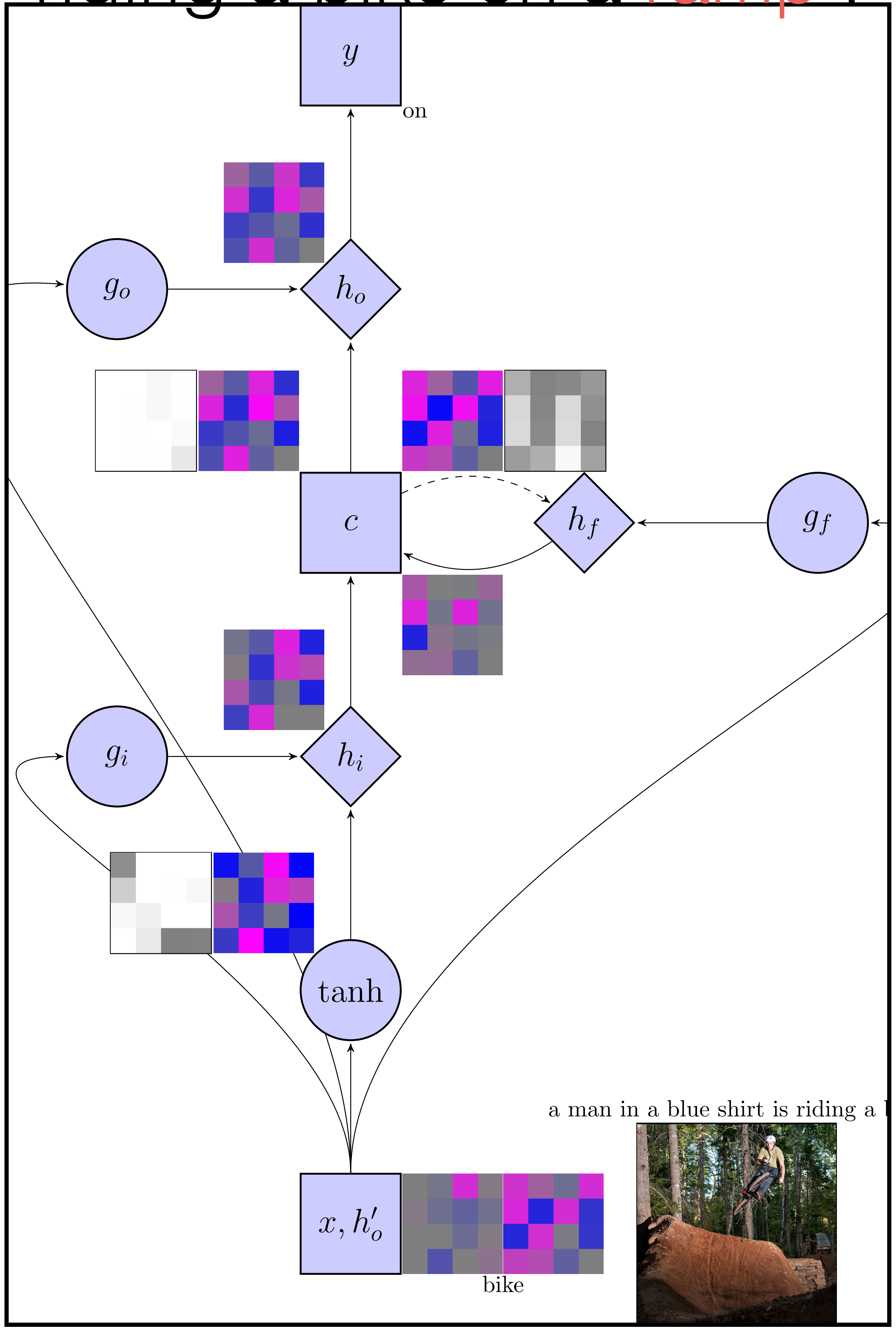
0

a man in a blue shirt is riding a bike on a dirt track .



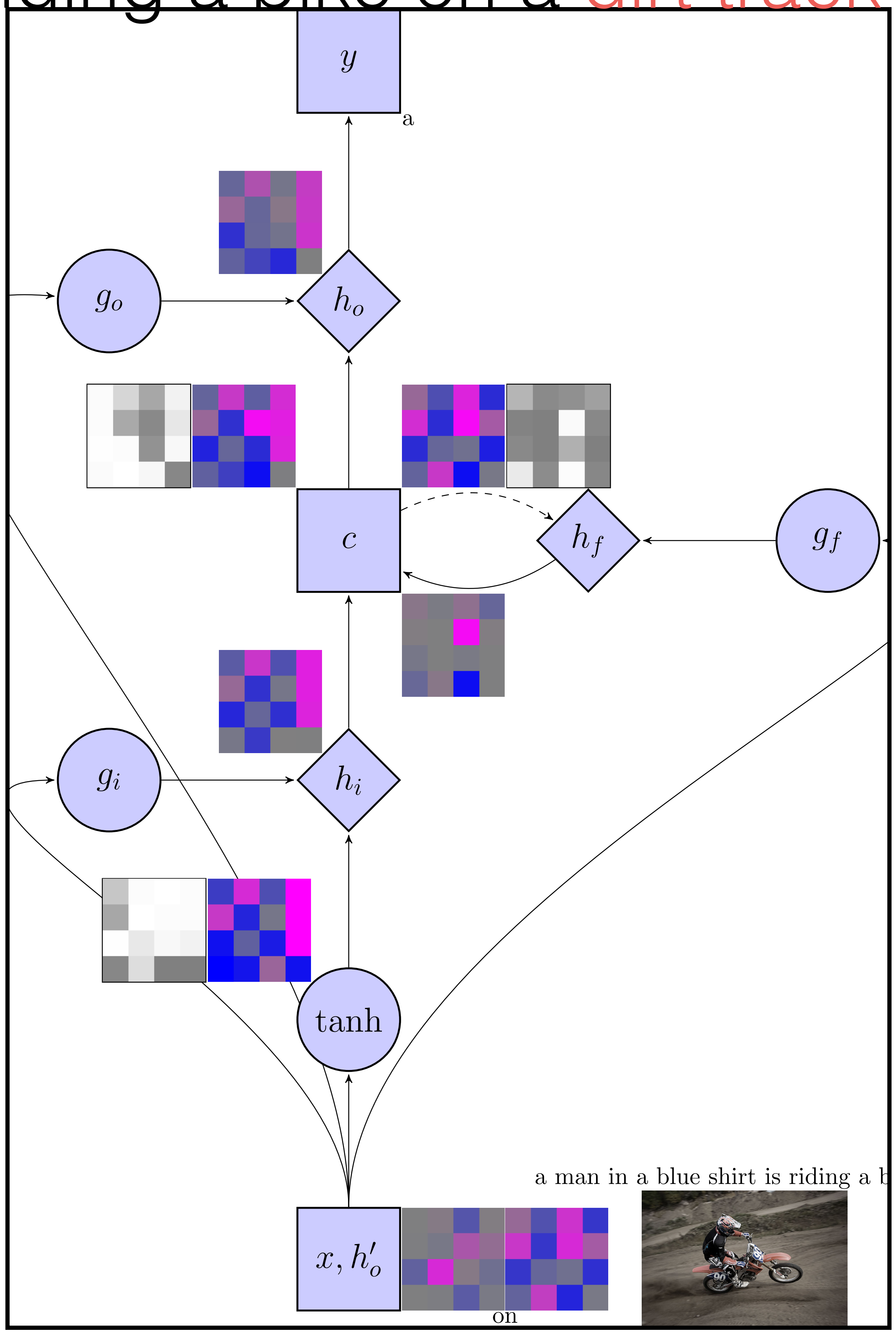
0

a man in a blue shirt is riding a bike on a ramp .



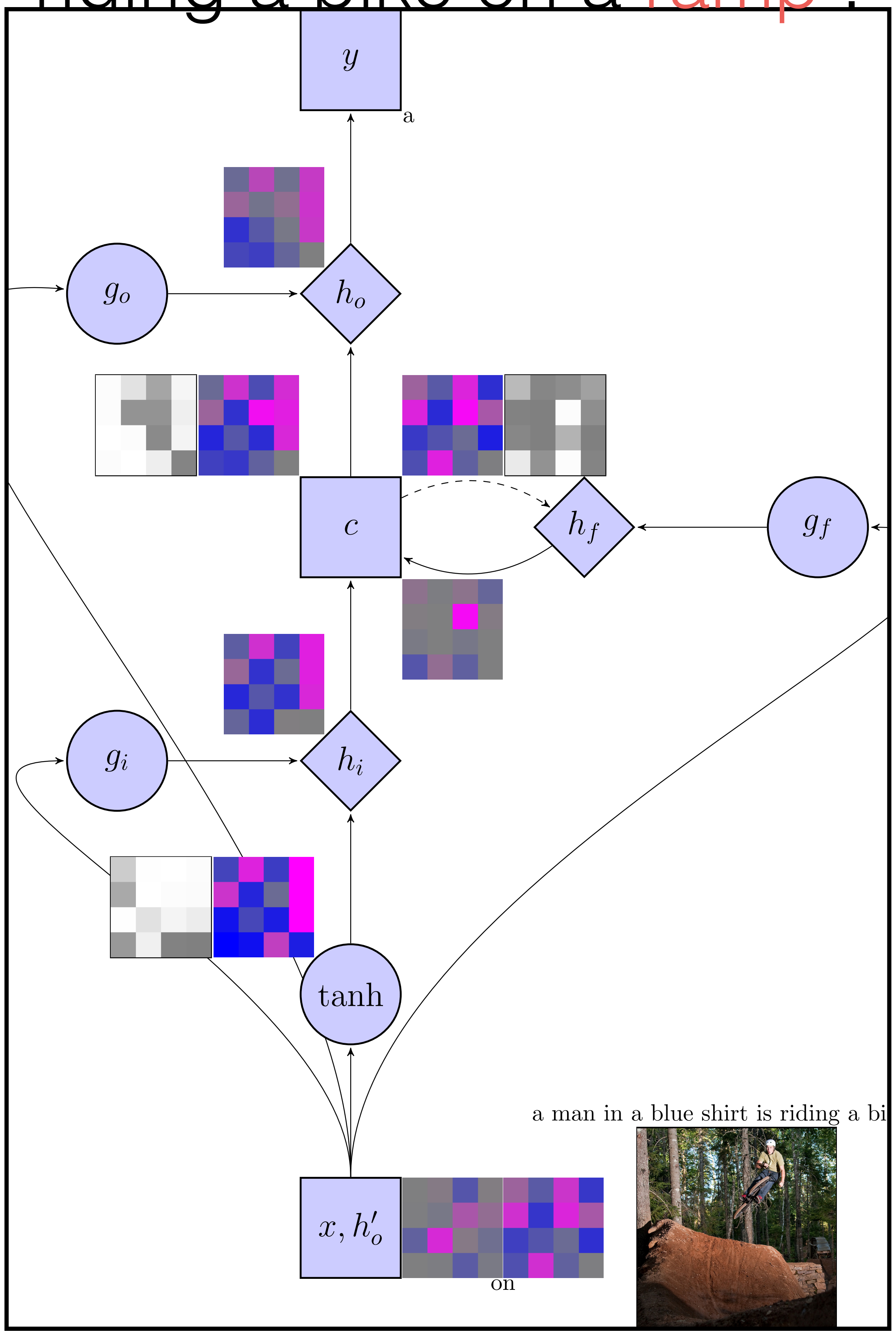
0

a man in a blue shirt is riding a bike on a dirt track .



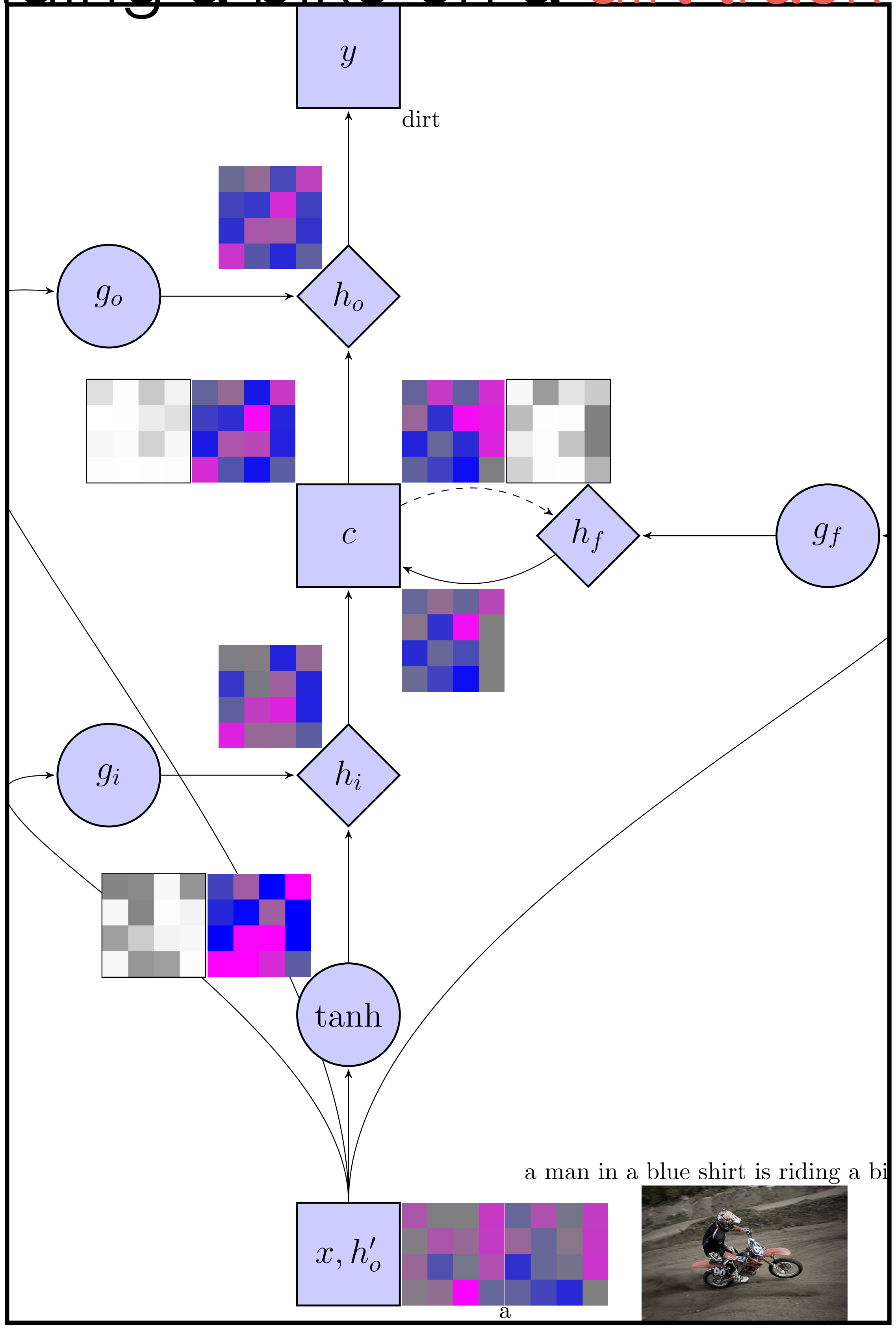
0

a man in a blue shirt is riding a bike on a ramp .



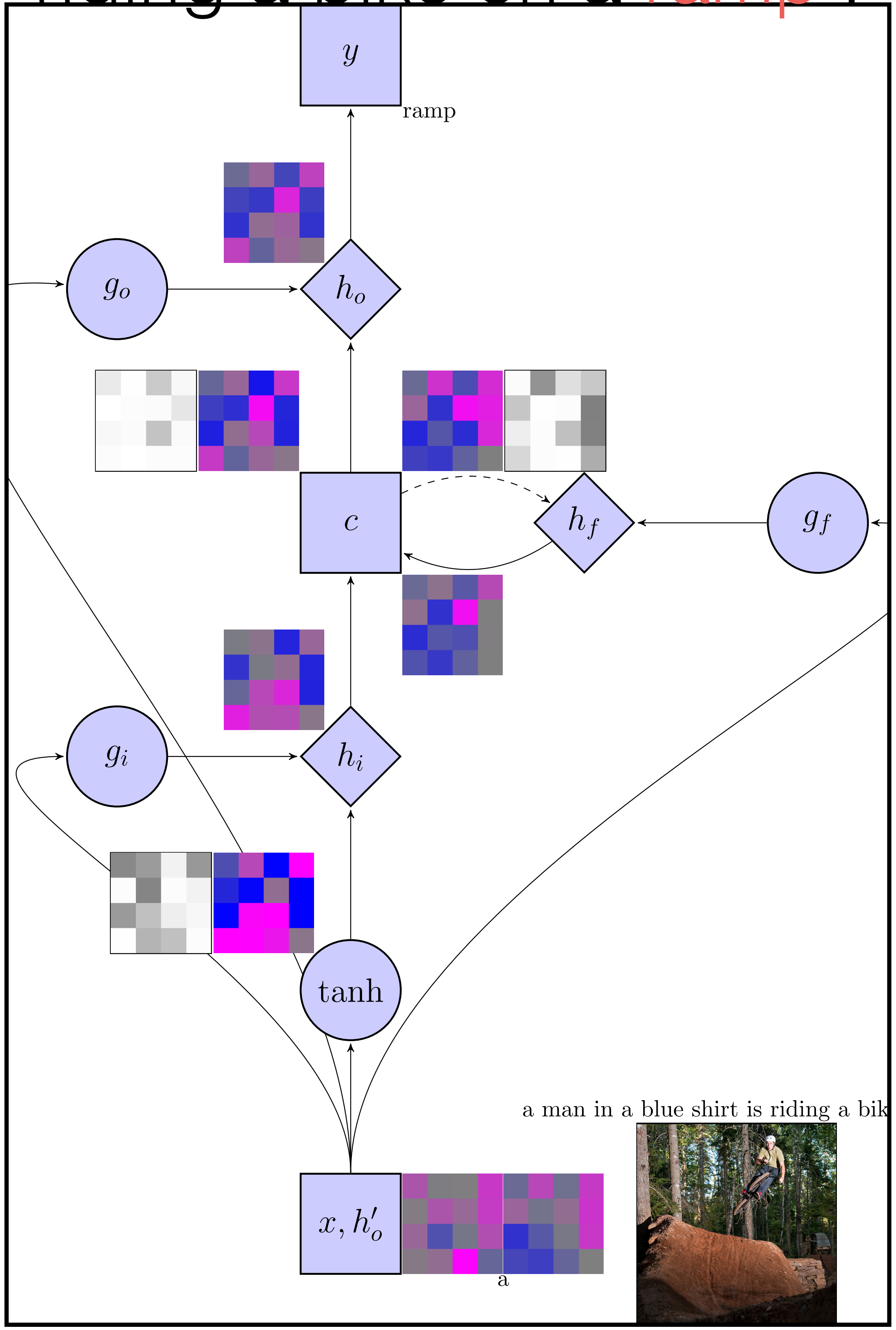
0

a man in a blue shirt is riding a bike on a dirt track .



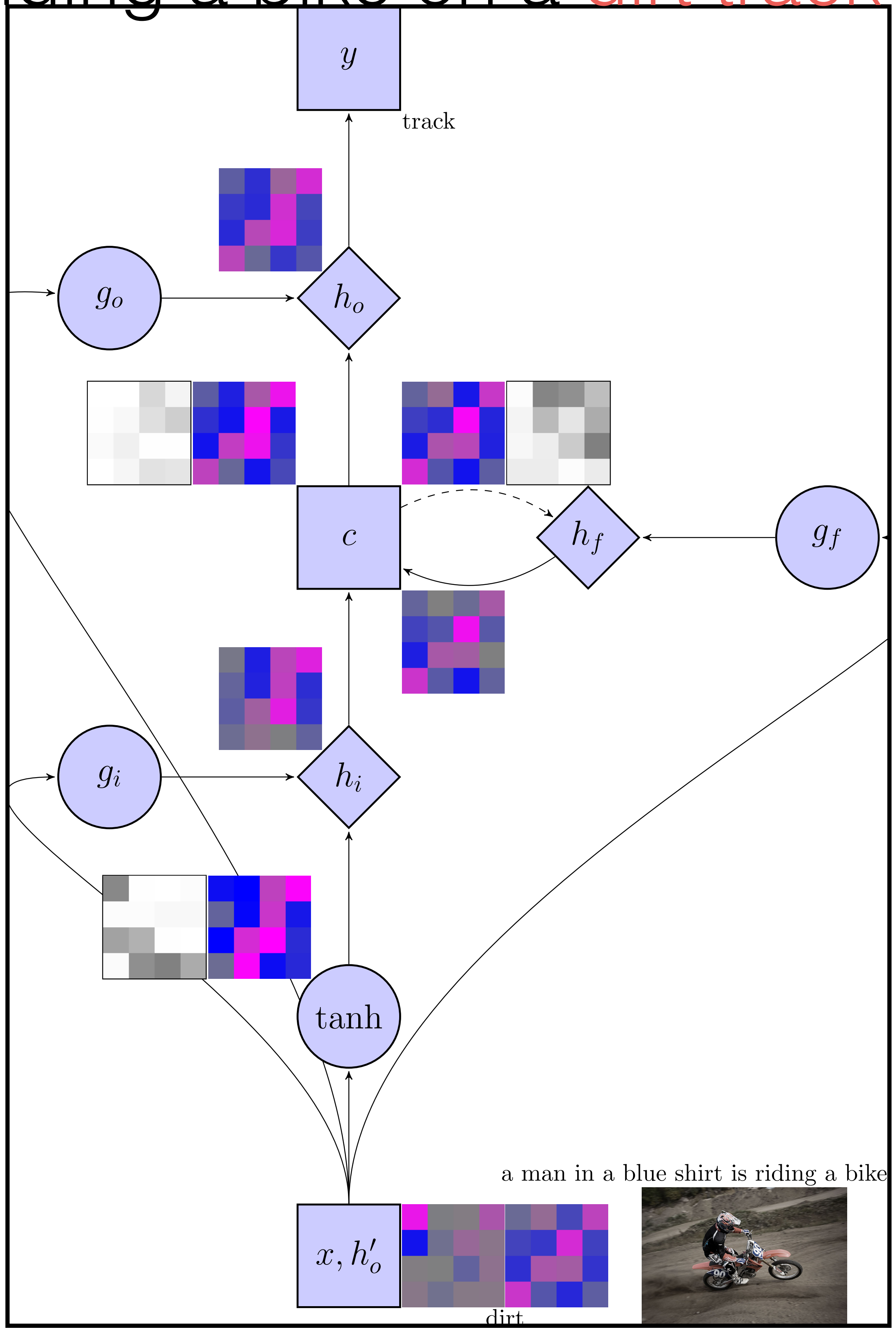
0

a man in a blue shirt is riding a bike on a ramp .



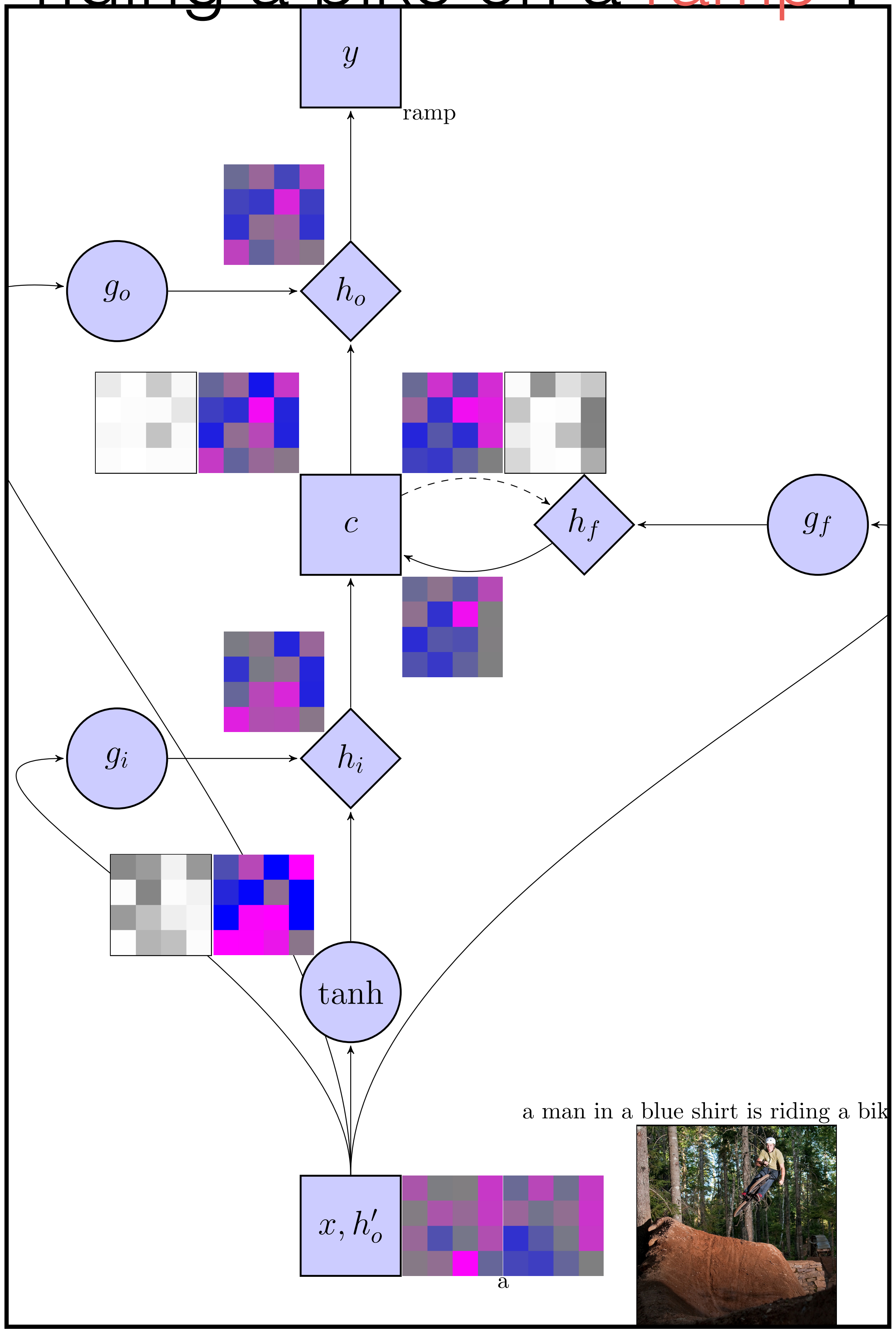
0

a man in a blue shirt is riding a bike on a dirt track .



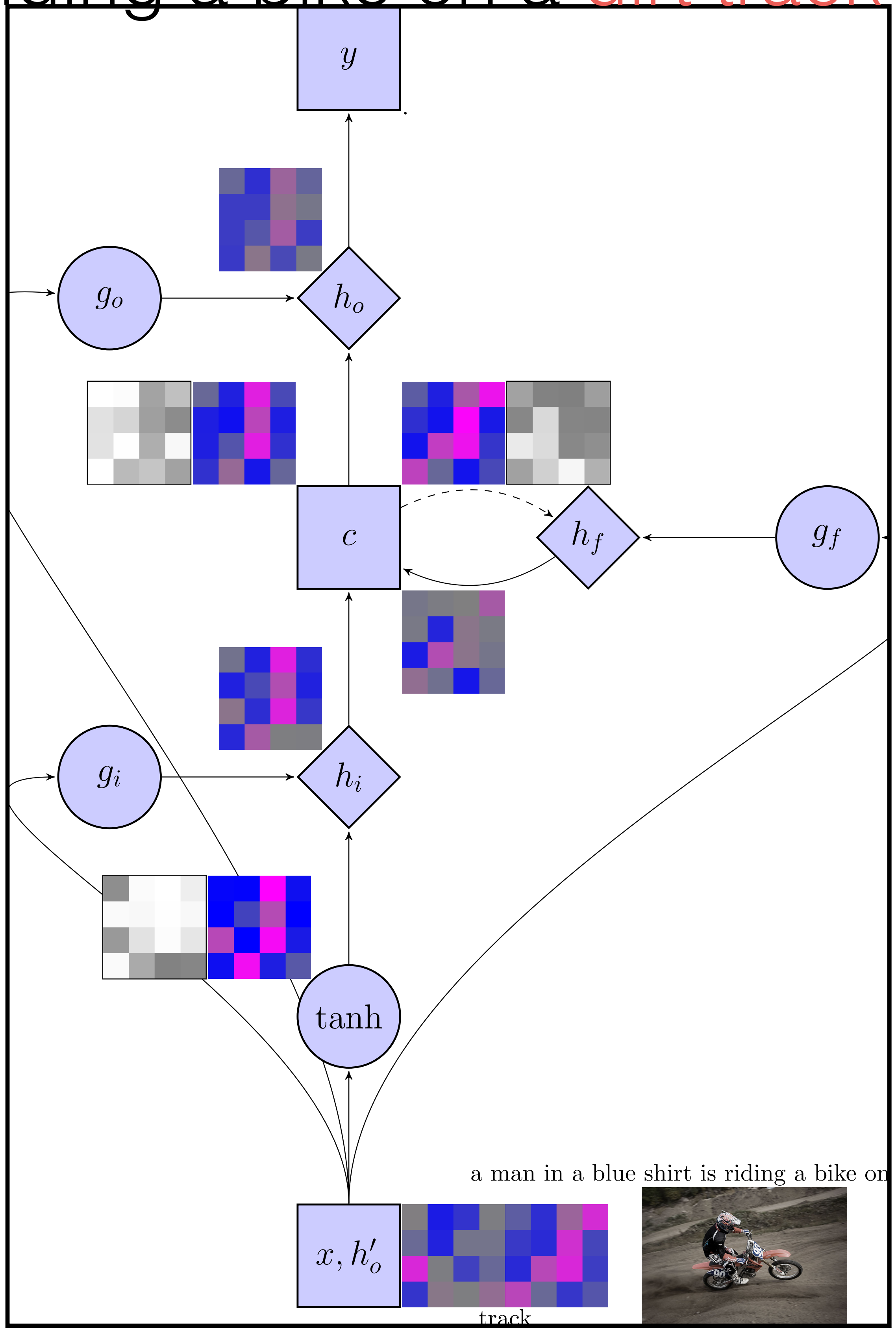
0

a man in a blue shirt is riding a bike on a ramp .



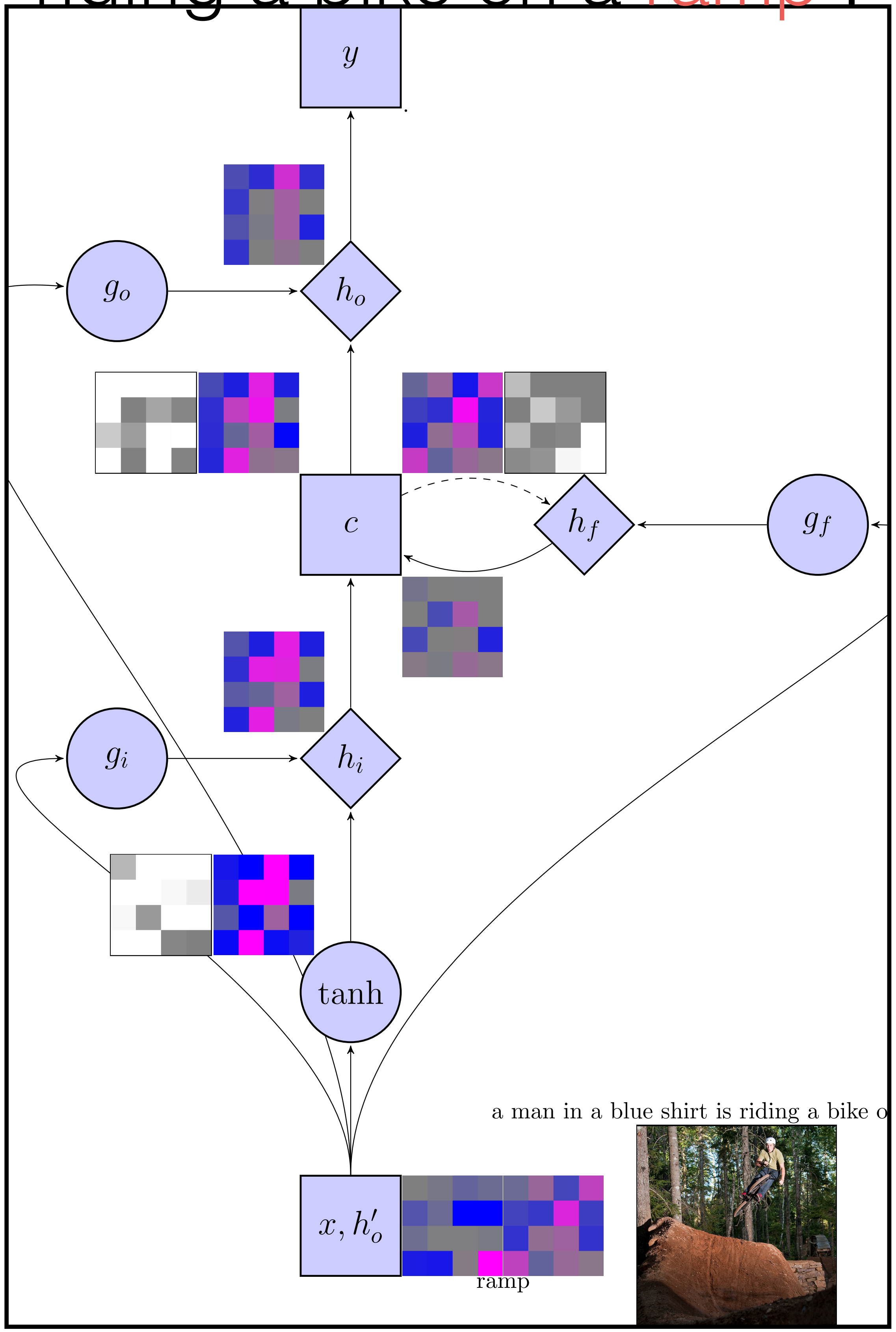
0

a man in a blue shirt is riding a bike on a dirt track .



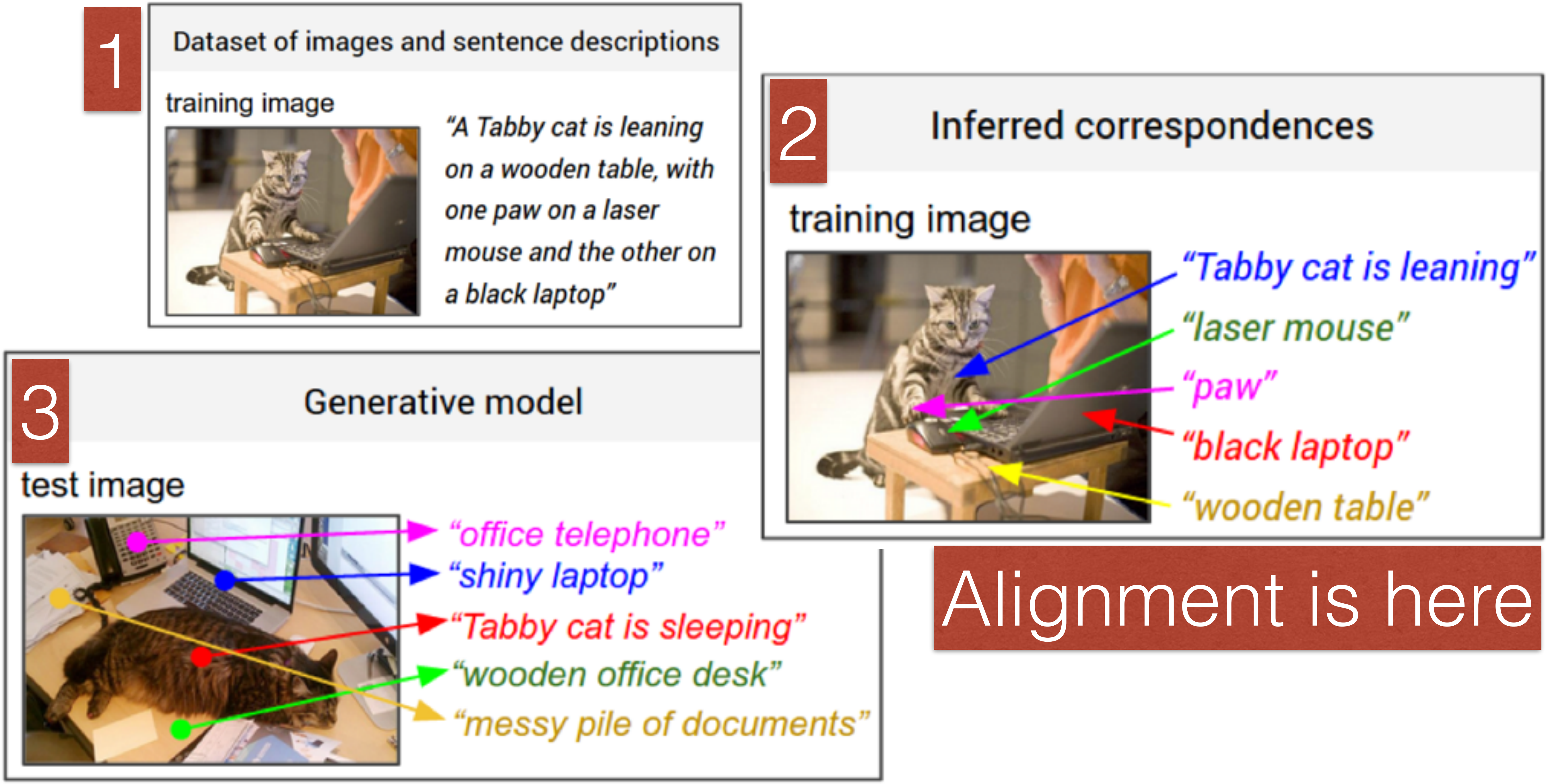
0

a man in a blue shirt is riding a bike on a ramp .



0

Generating descriptions for the regional images



1

Dataset of images and sentence descriptions

training image

"A Tabby cat is leaning on a wooden table, with one paw on a laser mouse and the other on a black laptop"

2

Inferred correspondences

training image

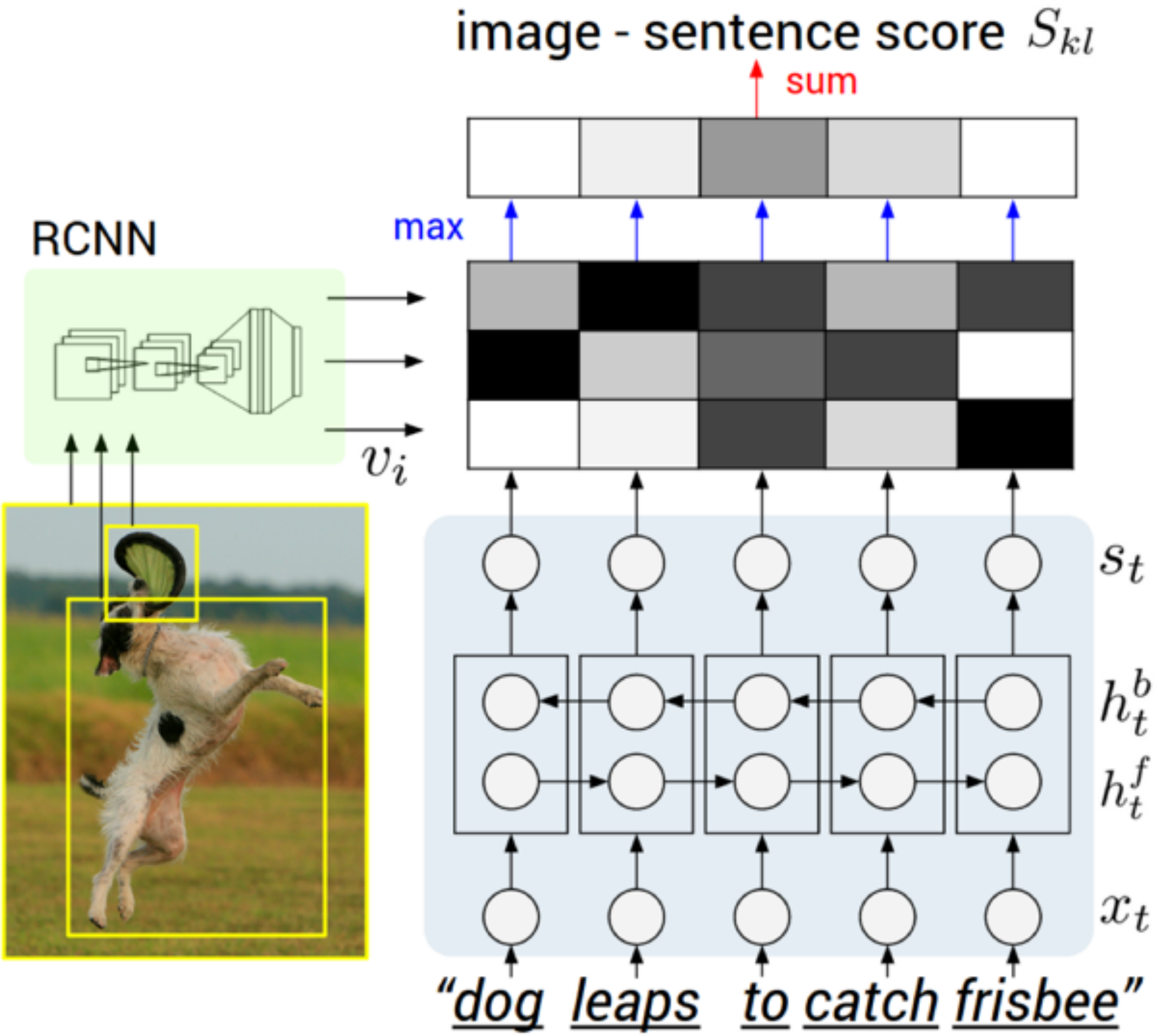
3

Generative model

test image

Alignment is here

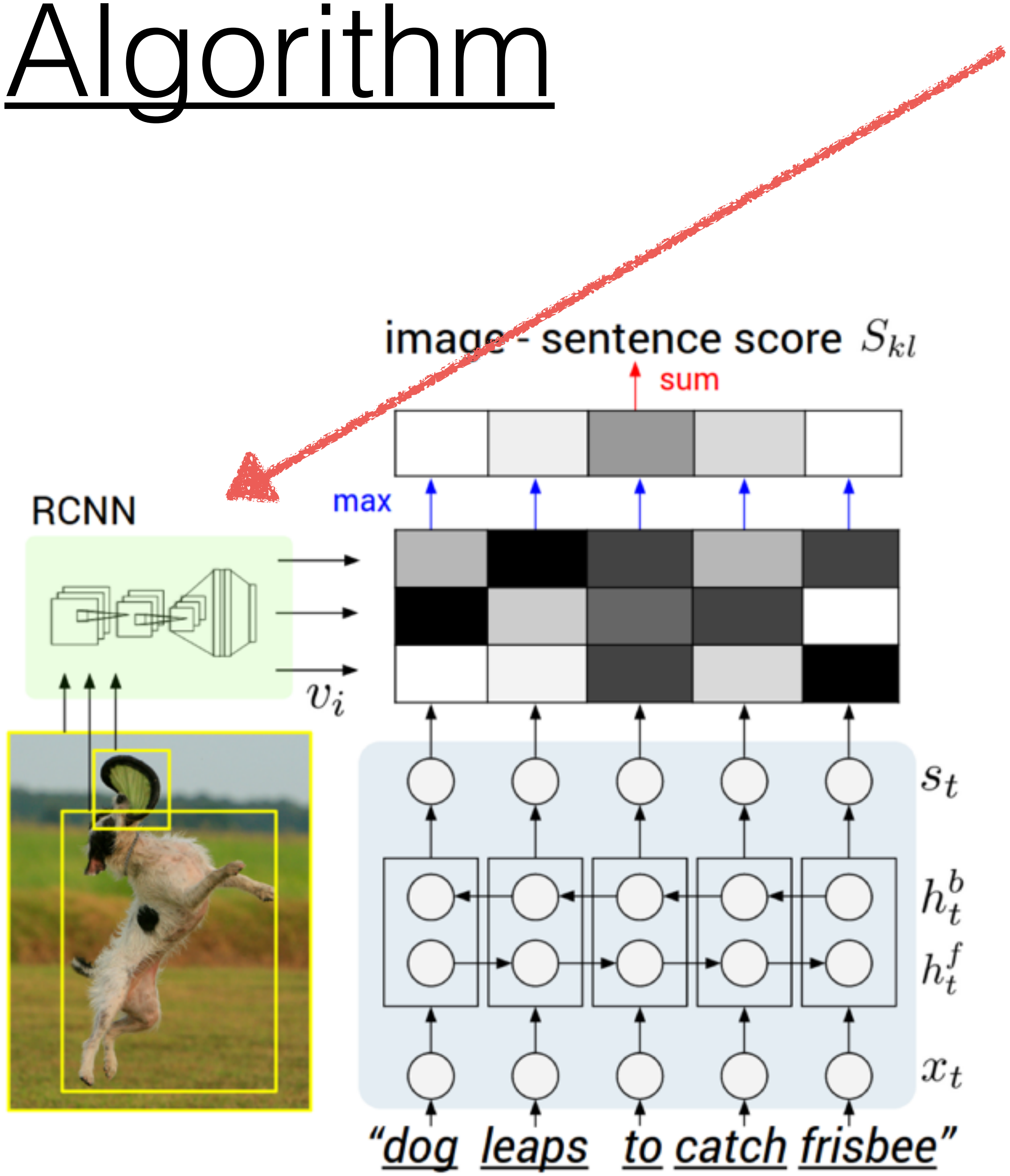
Alignment model



Algorithm

image embedding

$$v = \boxed{W_m}[CNN_{\theta_c}(I_b)] + b_m$$



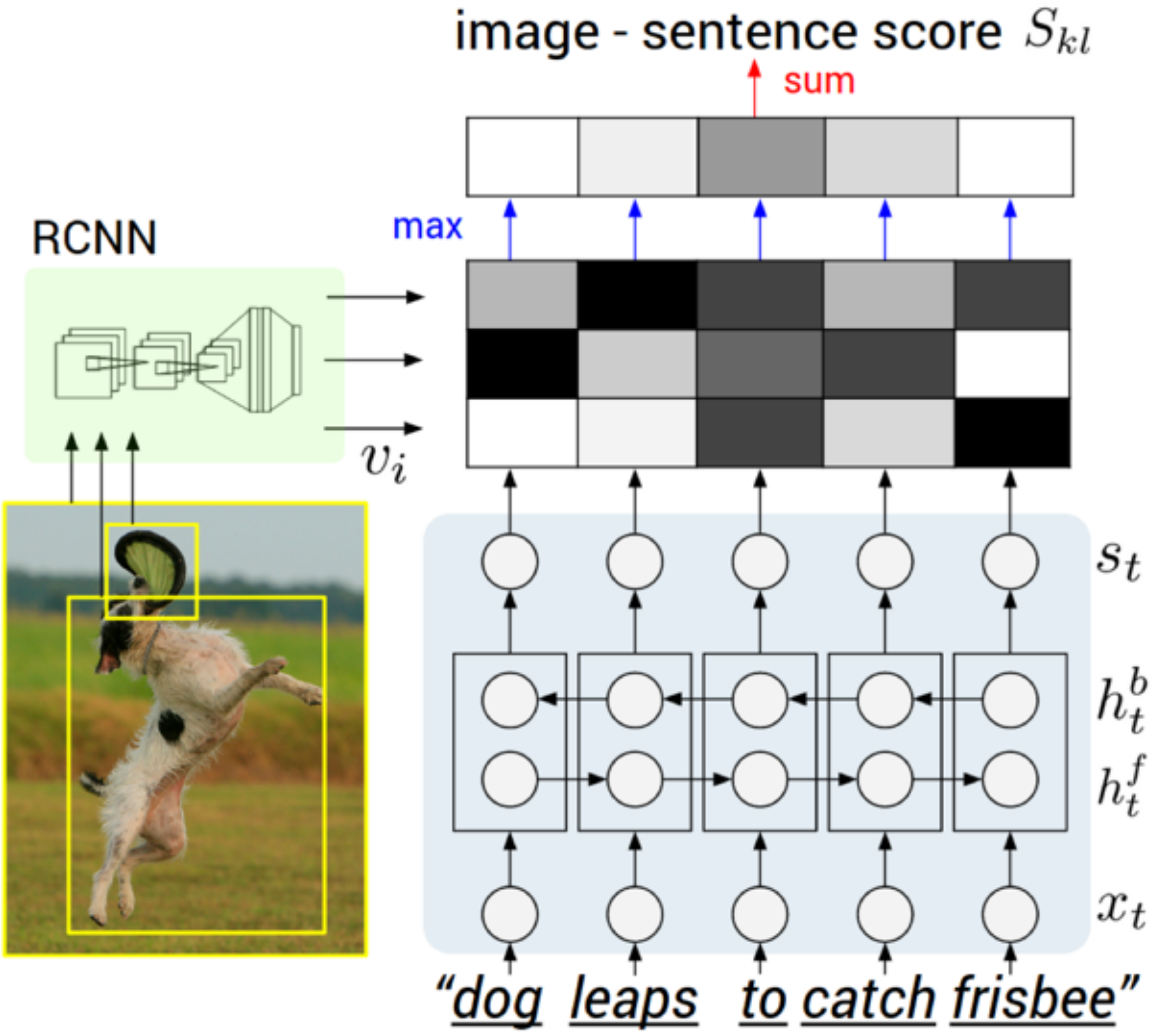
Algorithm

image embedding

$$v = \boxed{W_m}[CNN_{\theta_c}(I_b)] + b_m$$

word embedding

$$\begin{aligned}
x_t &= W_w \mathbb{I}_t \\
e_t &= f(\boxed{W_e} x_t + b_e) \\
h_t^f &= f(e_t + \boxed{W_f} h_{t-1}^f + b_f) \\
h_t^b &= f(e_t + \boxed{W_b} h_{t+1}^b + b_b) \\
s_t &= f(\boxed{W_d}(h_t^f + h_t^b) + b_d).
\end{aligned}$$

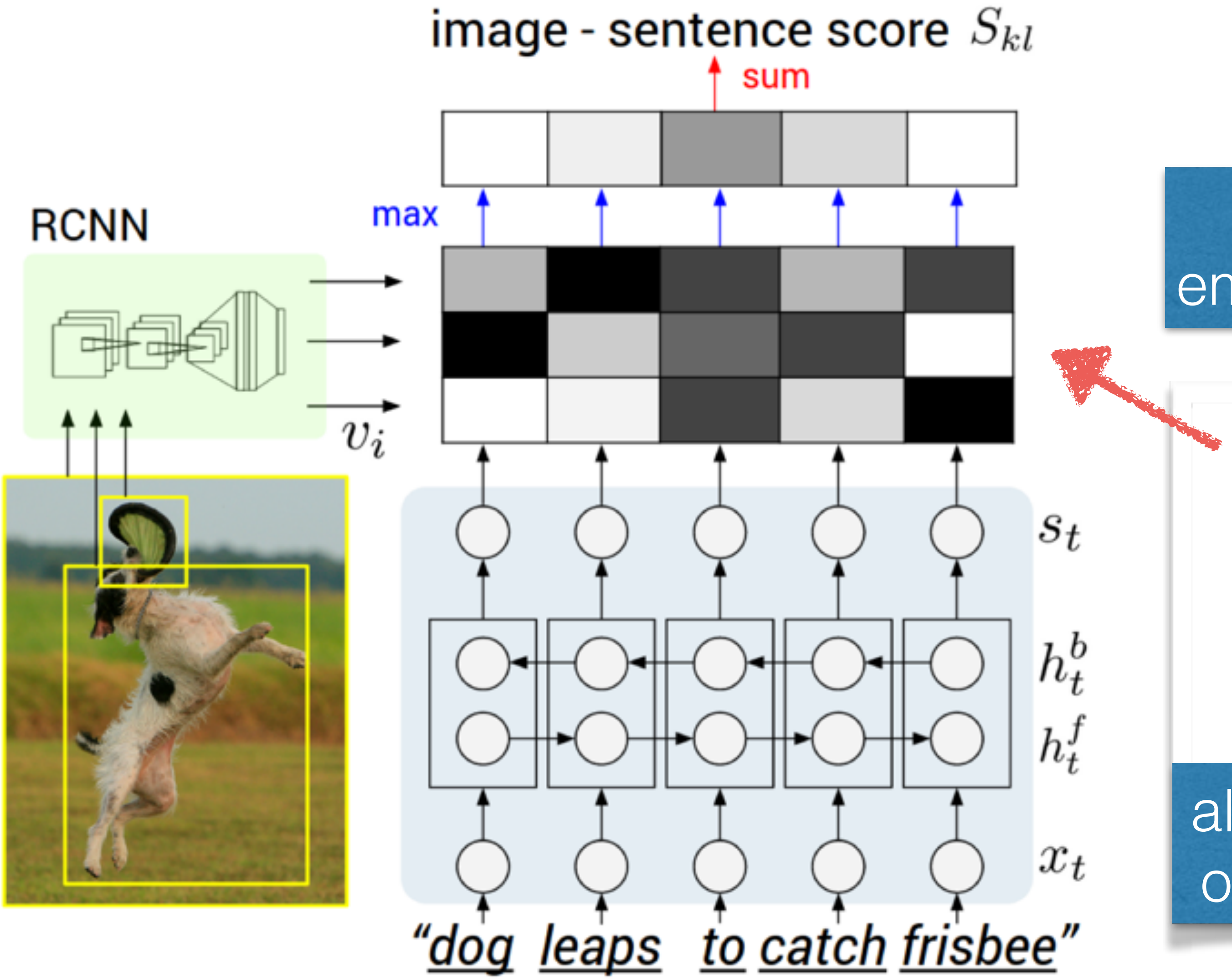


Algorithm

image embedding

$$v = \boxed{W_m}[CNN_{\theta_c}(I_b)] + b_m$$

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t.$$



word embedding

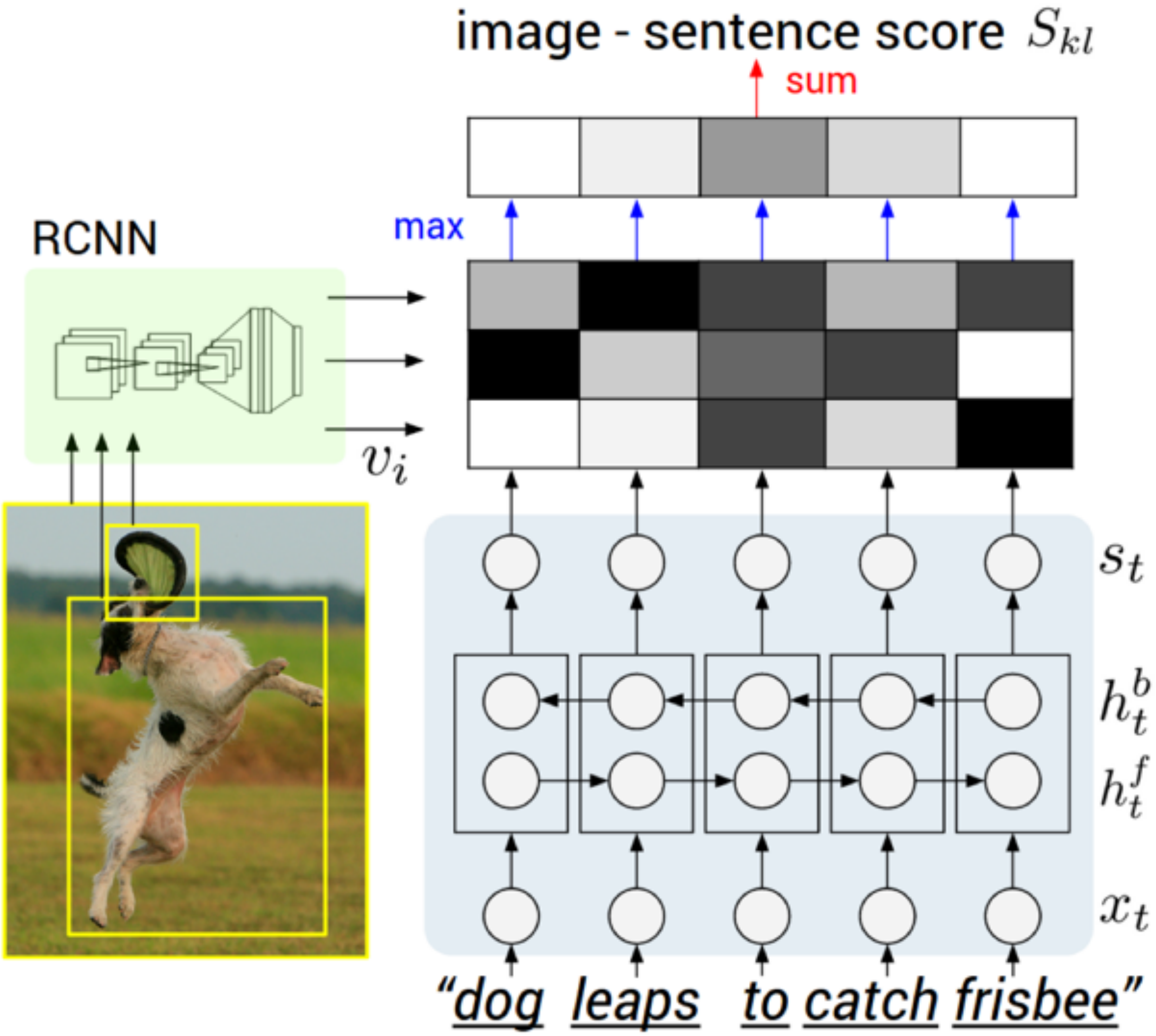
$$\begin{aligned}
x_t &= W_w \mathbb{I}_t \\
e_t &= f(\boxed{W_e} x_t + b_e) \\
h_t^f &= f(e_t + \boxed{W_f} h_{t-1}^f + b_f) \\
h_t^b &= f(e_t + \boxed{W_b} h_{t+1}^b + b_b) \\
s_t &= f(\boxed{W_d}(h_t^f + h_t^b) + b_d).
\end{aligned}$$

alignment objective

$$\mathcal{C}(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big].$$

A ranking model that makes similarity scores of matching pairs higher than those of mis-matches.

Algorithm



Encourage neighbour words to align to the same region.

image embedding

$$v = \boxed{W_m}[CNN_{\theta_c}(I_b)] + b_m$$

word embedding

$$\begin{aligned} x_t &= W_w \mathbb{I}_t \\ e_t &= f(\boxed{W_e} x_t + b_e) \\ h_t^f &= f(e_t + \boxed{W_f} h_{t-1}^f + b_f) \\ h_t^b &= f(e_t + \boxed{W_b} h_{t+1}^b + b_b) \\ s_t &= f(\boxed{W_d}(h_t^f + h_t^b) + b_d). \end{aligned}$$

alignment objective

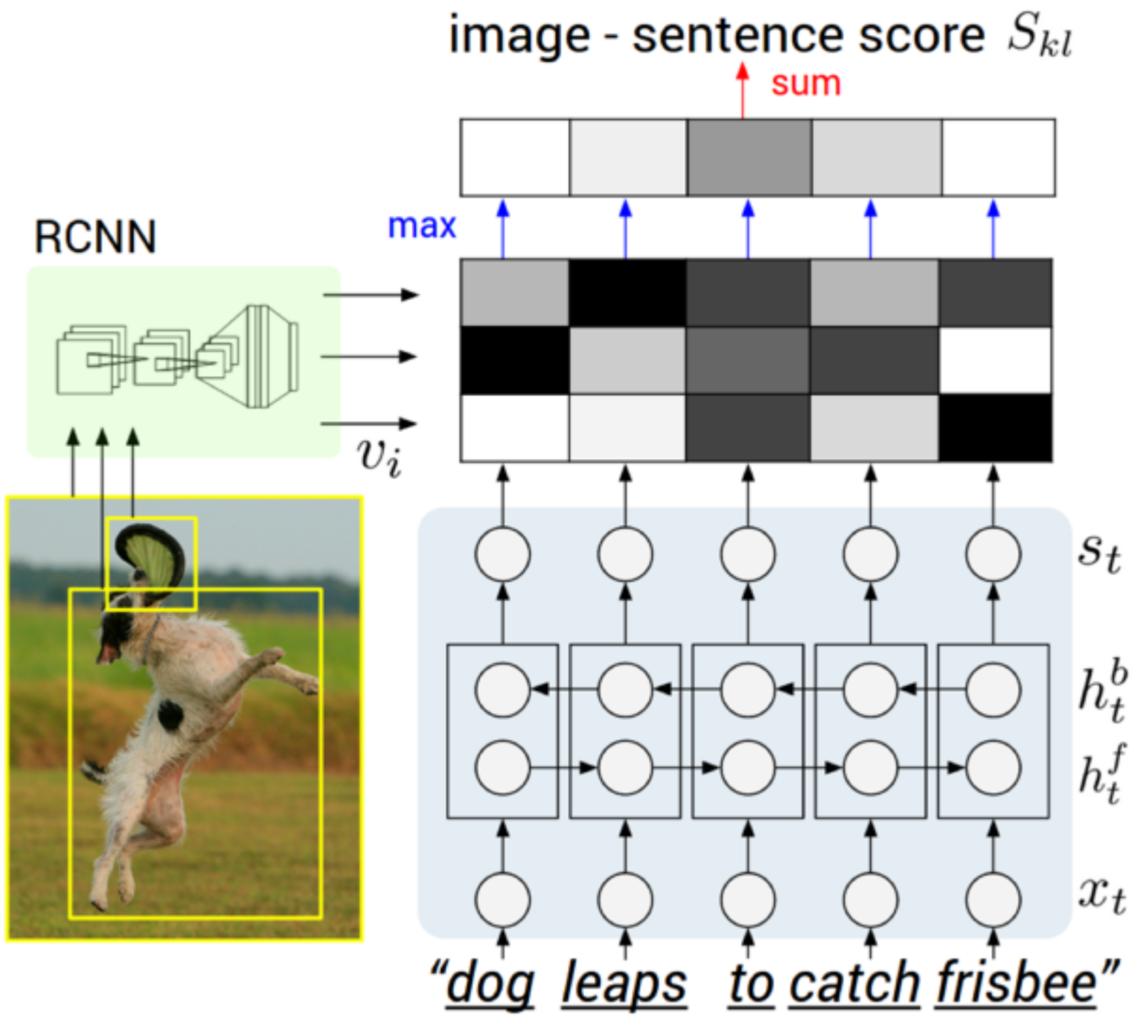
$$C(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big].$$

MRF in decoding

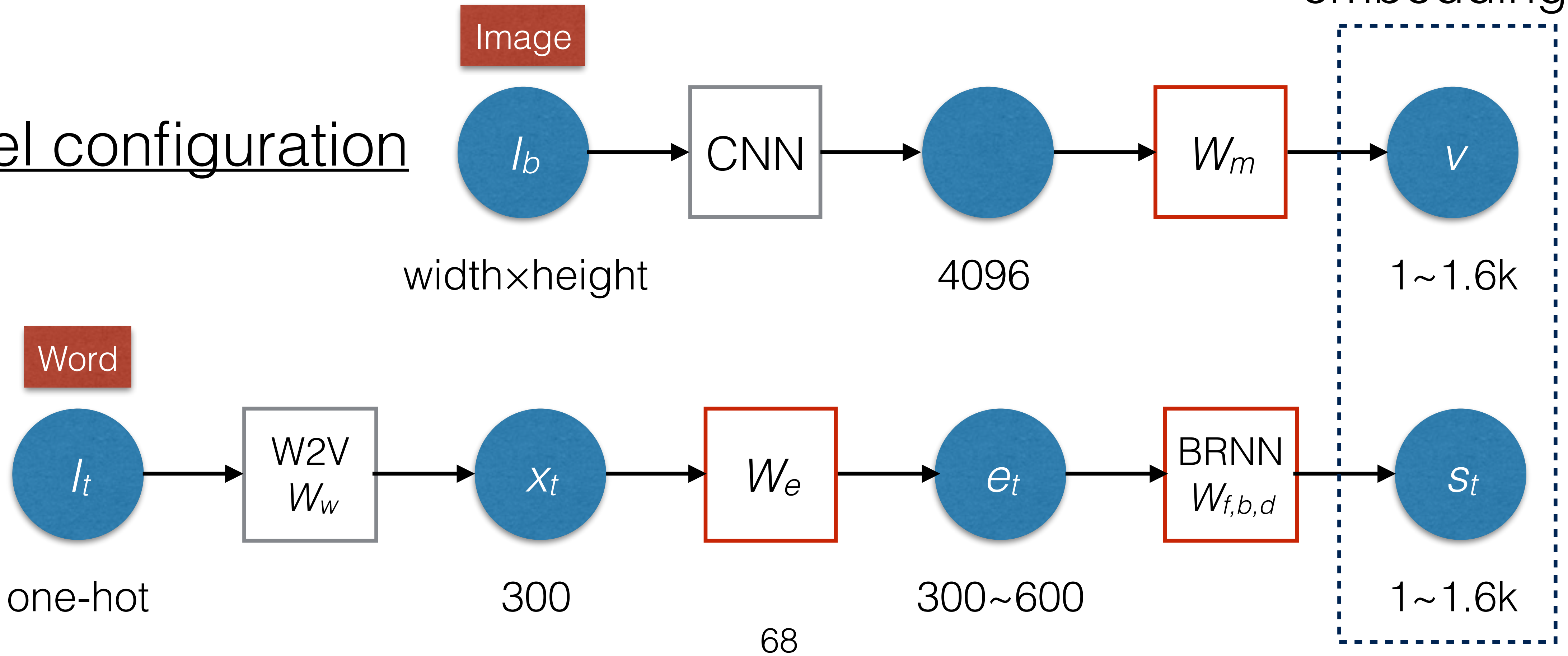
$$E(\mathbf{a}) = \sum_{j=1\ldots N} \psi_j^U(a_j) + \sum_{j=1\ldots N-1} \psi_j^B(a_j, a_{j+1})$$

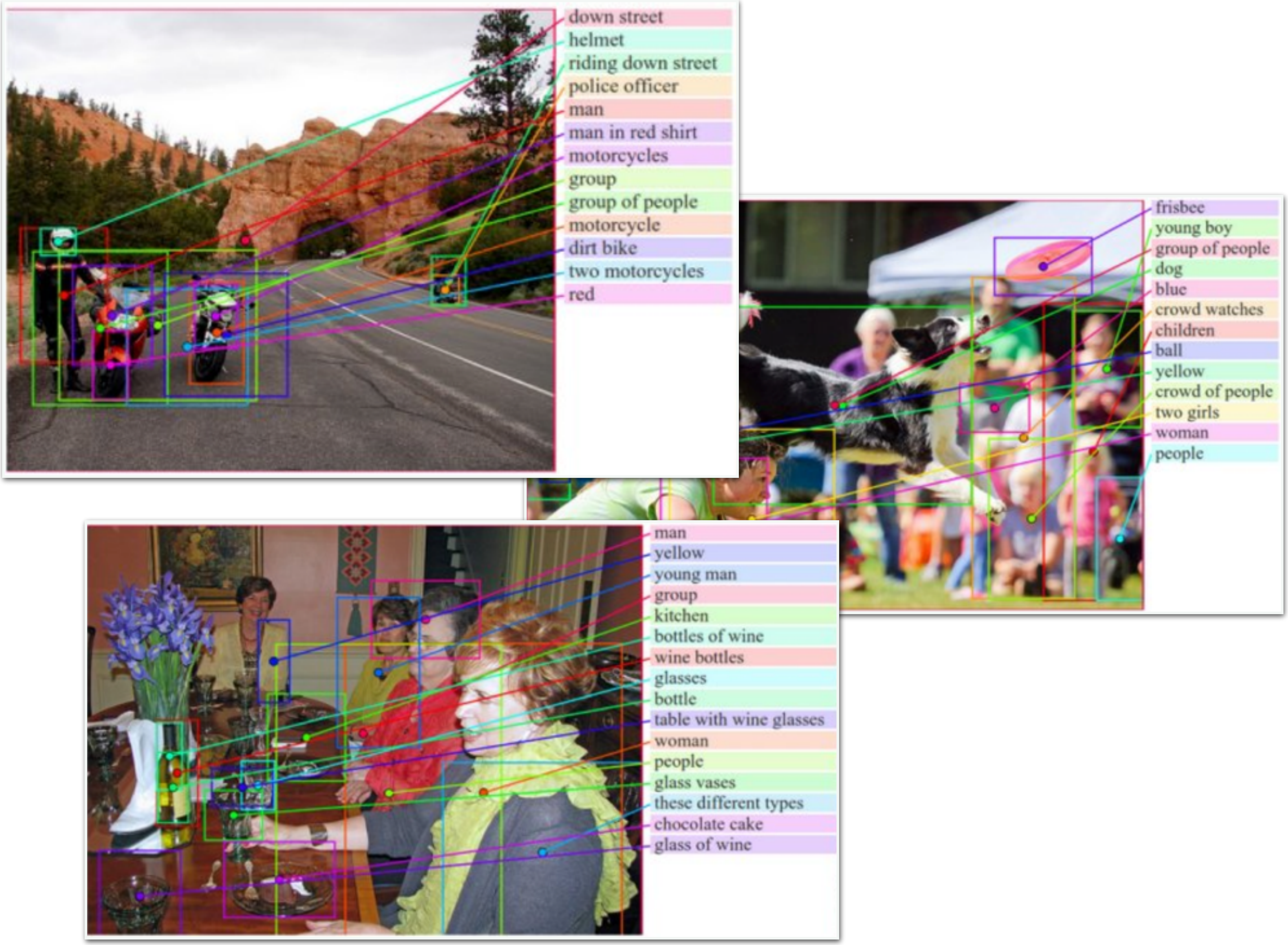
$$\psi_j^U(a_j = t) = v_i^T s_t$$

$$\psi_j^B(a_j, a_{j+1}) = \beta \mathbb{1}[a_j = a_{j+1}].$$



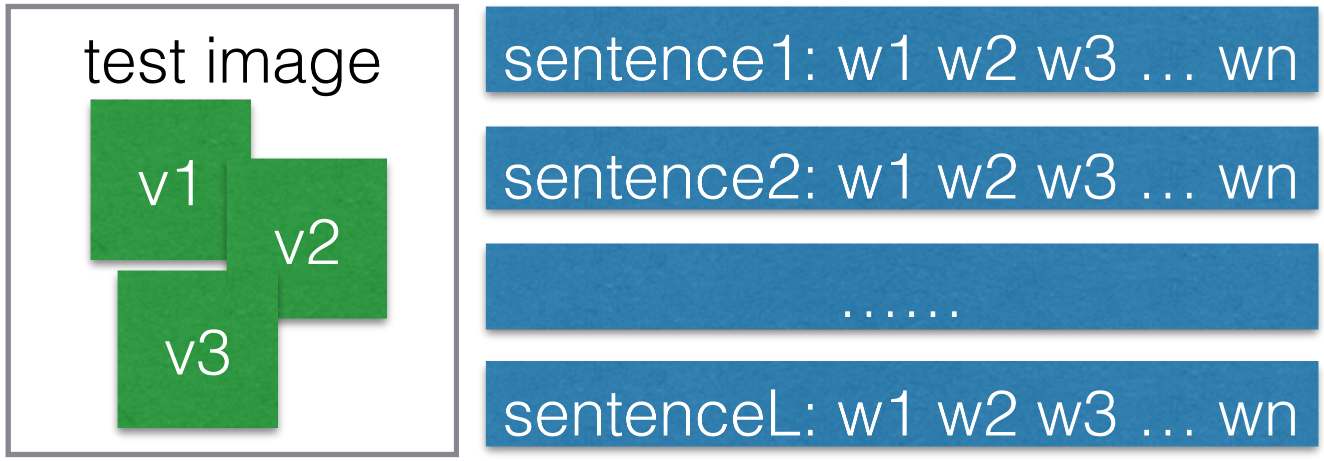
Model configuration




down street
helmet
riding down street
police officer
man
man in red shirt
motorcycles
group
group of people
motorcycle
dirt bike
two motorcycles
red
frisbee
young boy
group of people
dog
blue
crowd watches
children
ball
yellow
crowd of people
two girls
woman
people
man
yellow
young man
group
kitchen
bottles of wine
wine bottles
glasses
bottle
table with wine glasses
woman
people
glass vases
these different types
chocolate cake
glass of wine

Evaluation - Alignment

- Image annotation



test image

v1

v2

v3

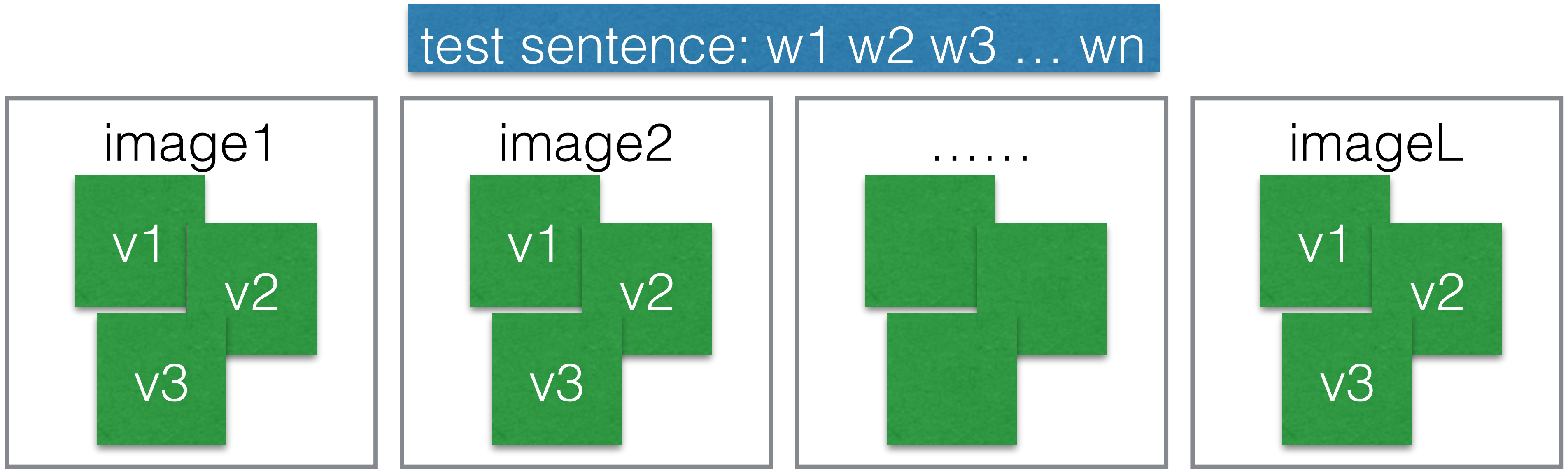
sentence1: w1 w2 w3 … wn

sentence2: w1 w2 w3 … wn

……

sentenceL: w1 w2 w3 … wn

- Image search



test sentence: w1 w2 w3 … wn

image1

v1

v2

v3

image2

v1

v2

v3

……

imageL

v1

v2

v3

Evaluation - Alignment

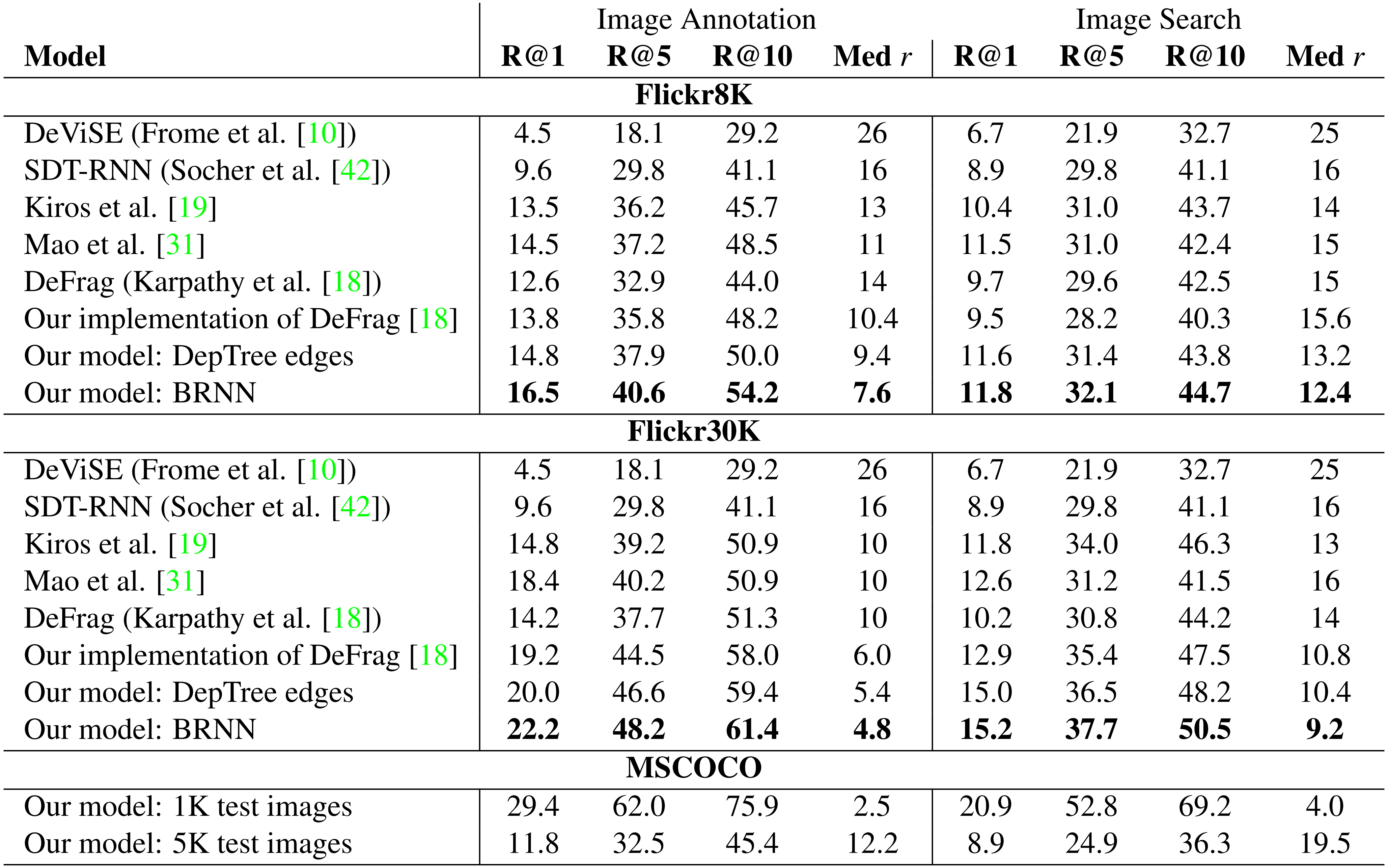
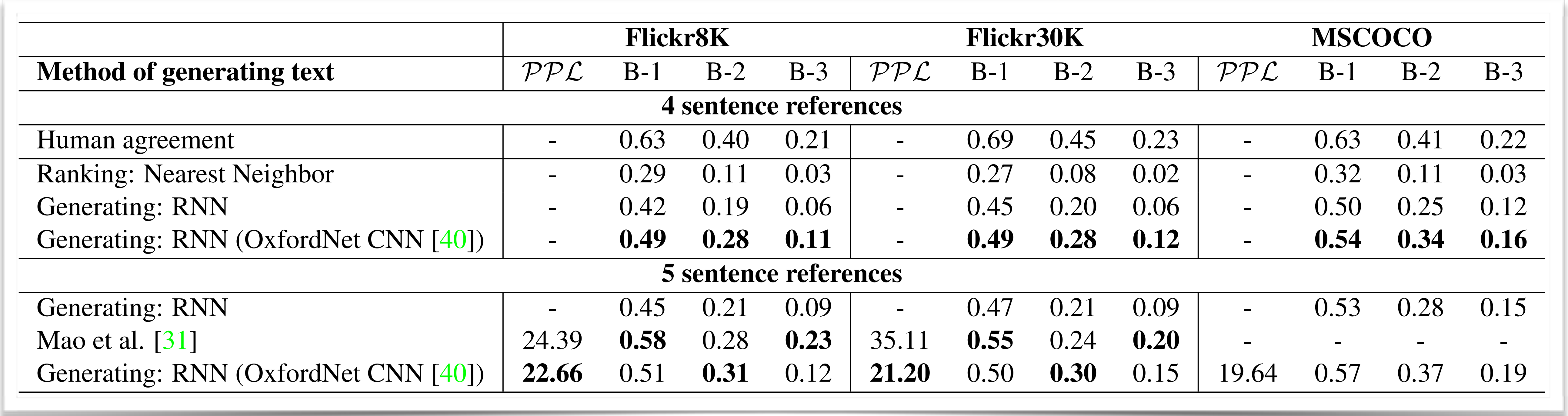
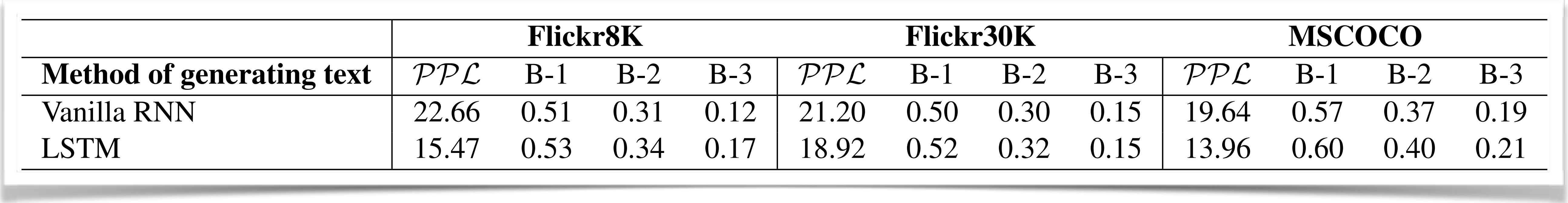


	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med** *r*	**R@1**	**R@5**	**R@10**	**Med** *r*
				Flickr8K				
DeViSE (Frome et al. [10])	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN (Socher et al. [42])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [19]	13.5	36.2	45.7	13	10.4	31.0	43.7	14
Mao et al. [31]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
DeFrag (Karpathy et al. [18])	12.6	32.9	44.0	14	9.7	29.6	42.5	15
Our implementation of DeFrag [18]	13.8	35.8	48.2	10.4	9.5	28.2	40.3	15.6
Our model: DepTree edges	14.8	37.9	50.0	9.4	11.6	31.4	43.8	13.2
Our model: BRNN	**16.5**	**40.6**	**54.2**	**7.6**	**11.8**	**32.1**	**44.7**	**12.4**
				Flickr30K				
DeViSE (Frome et al. [10])	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN (Socher et al. [42])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [19]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
Mao et al. [31]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
DeFrag (Karpathy et al. [18])	14.2	37.7	51.3	10	10.2	30.8	44.2	14
Our implementation of DeFrag [18]	19.2	44.5	58.0	6.0	12.9	35.4	47.5	10.8
Our model: DepTree edges	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
Our model: BRNN	**22.2**	**48.2**	**61.4**	**4.8**	**15.2**	**37.7**	**50.5**	**9.2**
				MSCOCO				
Our model: 1K test images	29.4	62.0	75.9	2.5	20.9	52.8	69.2	4.0
Our model: 5K test images	11.8	32.5	45.4	12.2	8.9	24.9	36.3	19.5

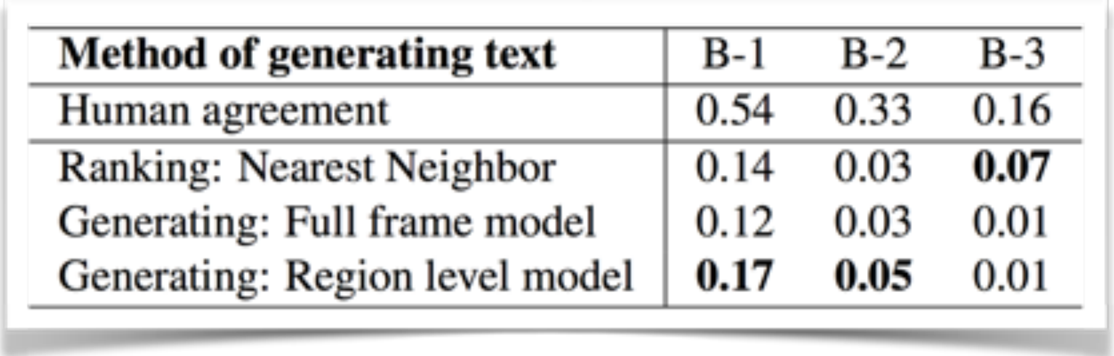
Evaluation - Translation



	Flickr8K				Flickr30K				MSCOCO			
Method of generating text	$\mathcal{PPL}$	B-1	B-2	B-3	$\mathcal{PPL}$	B-1	B-2	B-3	$\mathcal{PPL}$	B-1	B-2	B-3
4 sentence references												
Human agreement	-	0.63	0.40	0.21	-	0.69	0.45	0.23	-	0.63	0.41	0.22
Ranking: Nearest Neighbor	-	0.29	0.11	0.03	-	0.27	0.08	0.02	-	0.32	0.11	0.03
Generating: RNN	-	0.42	0.19	0.06	-	0.45	0.20	0.06	-	0.50	0.25	0.12
Generating: RNN (OxfordNet CNN [40])	-	**0.49**	**0.28**	**0.11**	-	**0.49**	**0.28**	**0.12**	-	**0.54**	**0.34**	**0.16**
5 sentence references												
Generating: RNN	-	0.45	0.21	0.09	-	0.47	0.21	0.09	-	0.53	0.28	0.15
Mao et al. [31]	24.39	**0.58**	0.28	**0.23**	35.11	**0.55**	0.24	**0.20**	-	-	-	-
Generating: RNN (OxfordNet CNN [40])	**22.66**	0.51	**0.31**	0.12	**21.20**	0.50	**0.30**	0.15	19.64	0.57	0.37	0.19



	Flickr8K				Flickr30K				MSCOCO			
Method of generating text	$\mathcal{PPL}$	B-1	B-2	B-3	$\mathcal{PPL}$	B-1	B-2	B-3	$\mathcal{PPL}$	B-1	B-2	B-3
Vanilla RNN	22.66	0.51	0.31	0.12	21.20	0.50	0.30	0.15	19.64	0.57	0.37	0.19
LSTM	15.47	0.53	0.34	0.17	18.92	0.52	0.32	0.15	13.96	0.60	0.40	0.21



Method of generating text	B-1	B-2	B-3
Human agreement	0.54	0.33	0.16
Ranking: Nearest Neighbor	0.14	0.03	**0.07**
Generating: Full frame model	0.12	0.03	0.01
Generating: Region level model	**0.17**	**0.05**	0.01

Reference

- Karpathy, A. & Fei-Fei, L., 2014. Deep Visual-Semantic Alignments for Generating Image Descriptions. arXiv.org, cs.CV.

- Vinyals, O. et al., 2014. Show and Tell: A Neural Image Caption Generator. arXiv.org, cs.CV.