Q?

A!

FLIPDIAL

A GENERATIVE MODEL FOR TWO-WAY VISUAL DIALOGUE

DANIELA MASSICETI, N. SIDDHARTH, PUNEET K. DOKANIA, PHILIP H.S. TORR

UNIVERSITY OF OXFORD

CVPR, SALT LAKE CITY 2018

TOWARD REAL-WORLD CONVERSATIONAL A.I.

TOWARD REAL-WORLD CONVERSATIONAL A.I.

"Where is the bus stop?

"On the road to your right"

TOWARD REAL-WORLD CONVERSATIONAL A.I.

TOWARD REAL-WORLD CONVERSATIONAL A.I.

"Where is the bus stop?

"On the road to your right"

Sensible response

Visually grounded

TOWARD REAL-WORLD CONVERSATIONAL A.I.

"Where is the bus stop?

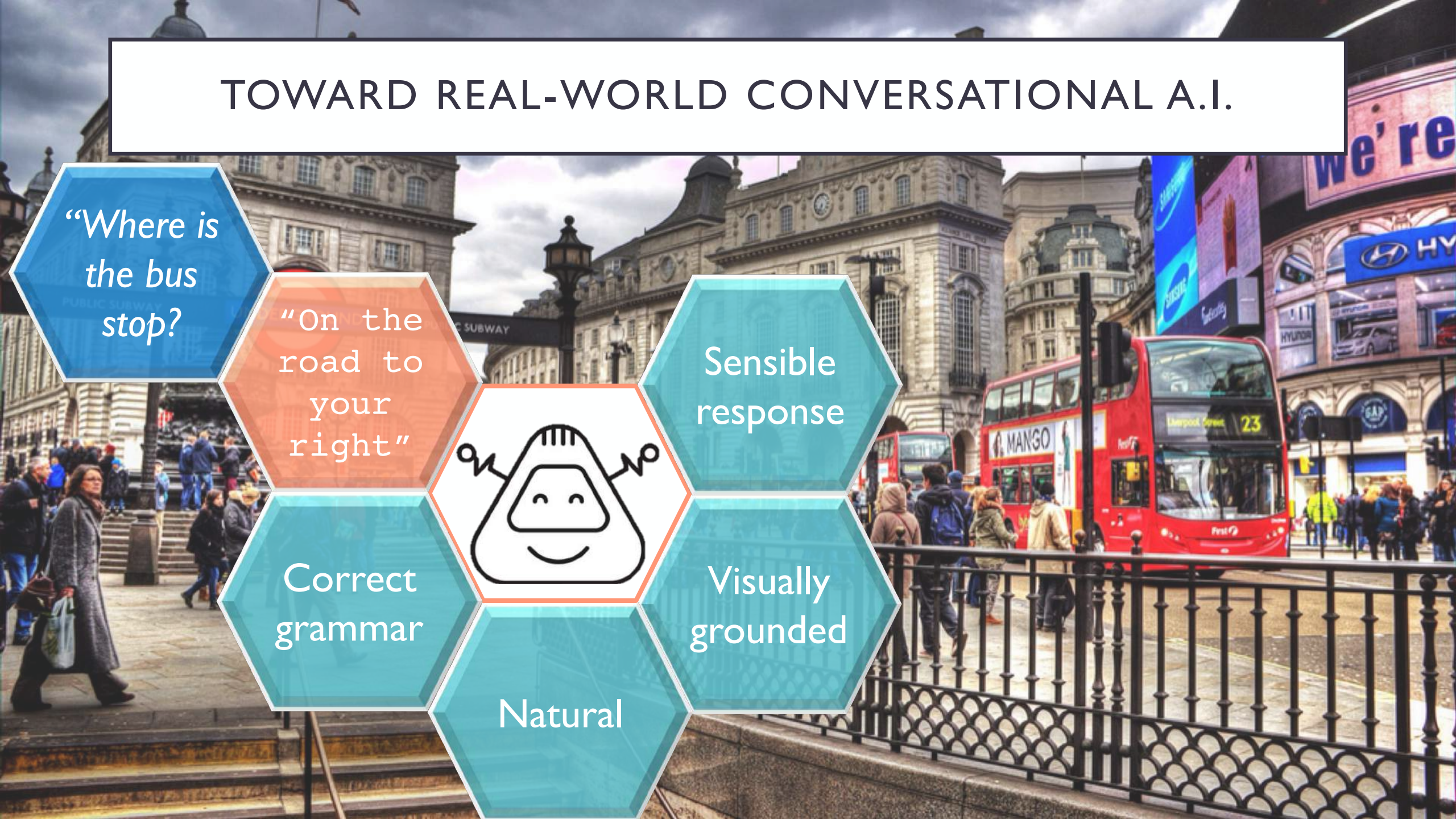
"On the road to your right"

Sensible response

Visually grounded

Natural

TOWARD REAL-WORLD CONVERSATIONAL A.I.




Where is the bus stop?
PUBLIC SUBWAY
UNDERGROUND
PUBLIC SUBWAY
MANGO
23
We're

Where is the bus stop?

On the road to your right

Where is the bus stop?

On the road to your right

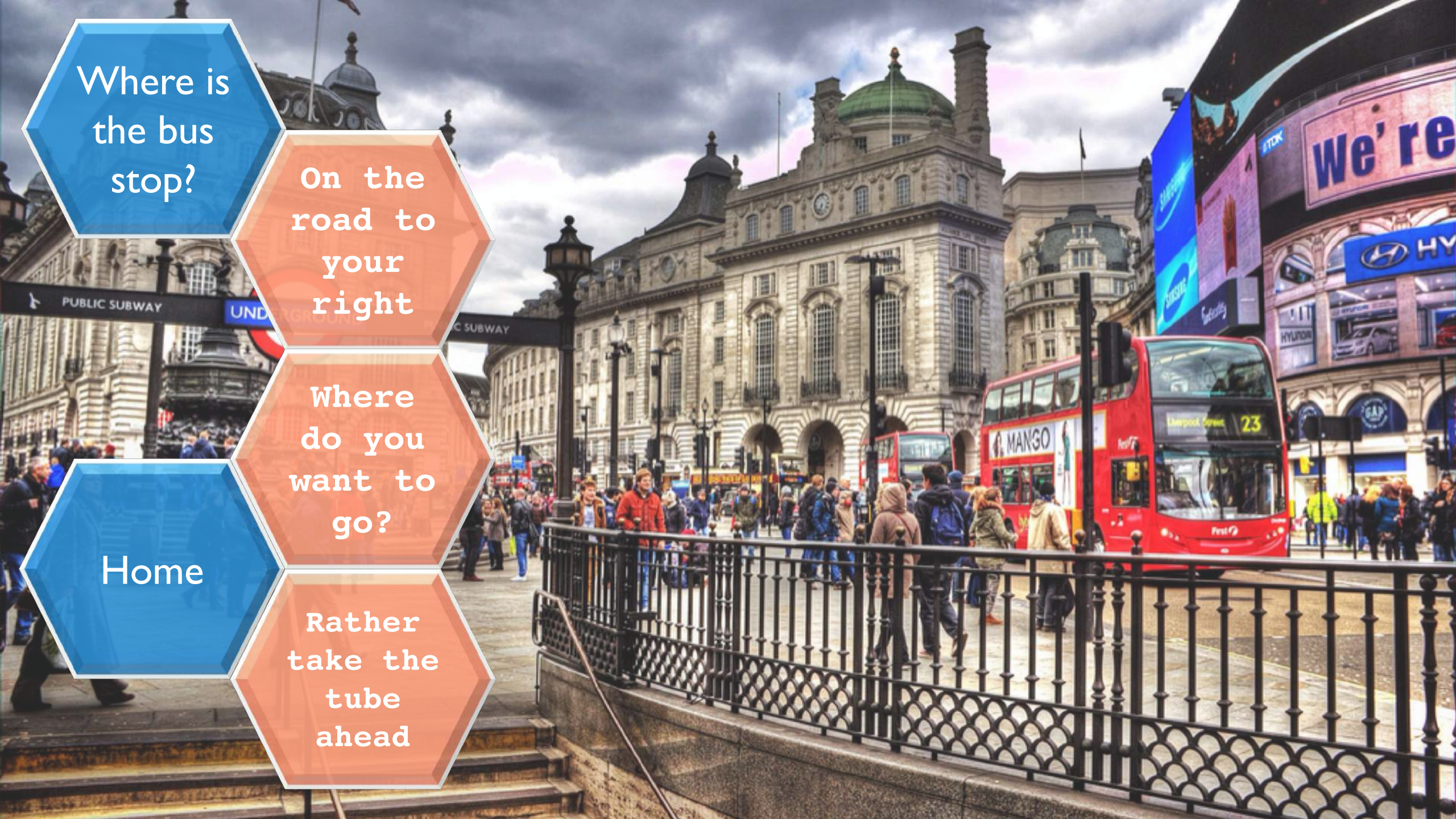
Where do you want to go?

Where is the bus stop?

On the road to your right

Where do you want to go?

Home



Where is the bus stop?

On the road to your right

Where do you want to go?

Home

Rather take the tube ahead

Where is the bus stop?

On the road to your right

Where do you want to go?

Home

Rather take the tube ahead

The tube entrance is in front of you and has a direct line home

The tube is faster, so take that instead. The entrance is ahead

The tube is running. The entry is left around the barriers

Rather take the tube then. The entrance is ahead of you

ONE-WAY VISUAL DIALOGUE (1VD)

A street with people and balloons

TWO-WAY VISUAL DIALOGUE (2VD)

A street with people and balloons

ONE-WAY VISUAL DIALOGUE (1VD)

How many people are there?

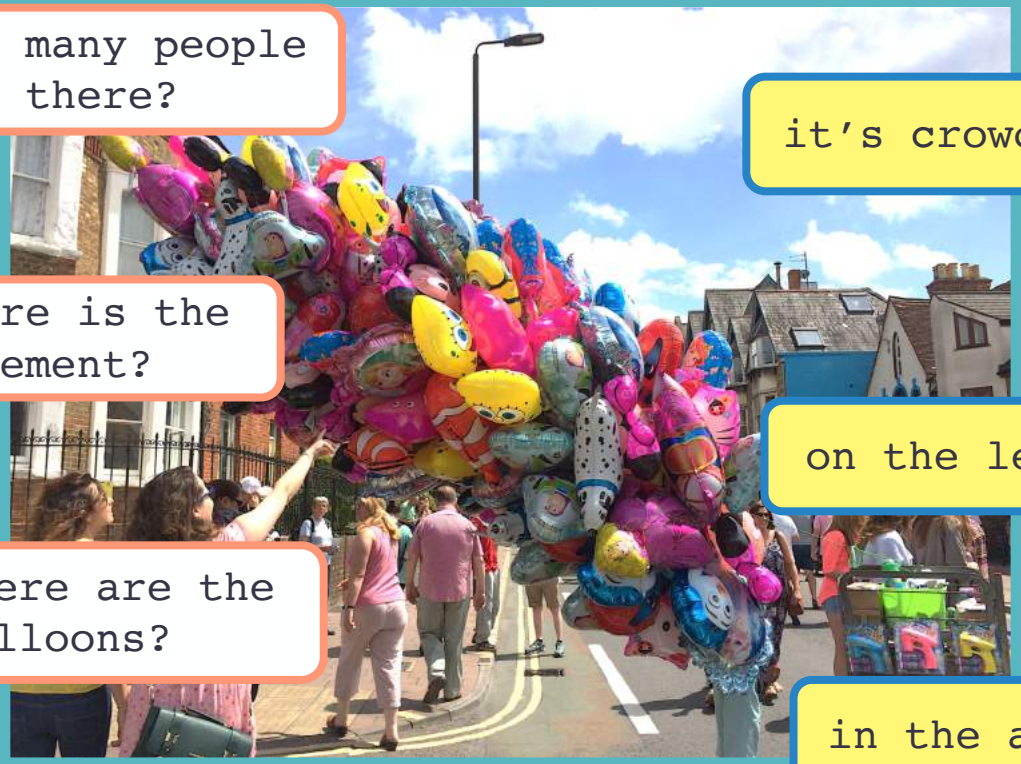
it's crowded

Where is the pavement?

on the left

Where are the balloons?

in the air

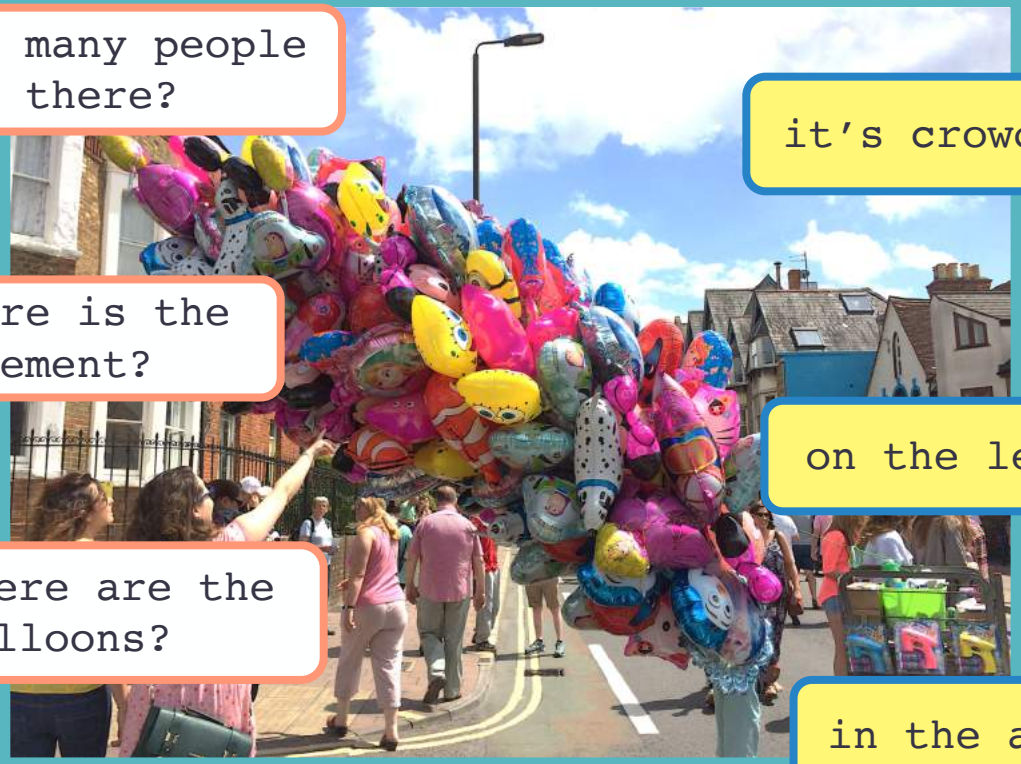


A street with people and balloons

TWO-WAY VISUAL DIALOGUE (2VD)

A street with people and balloons

ONE-WAY VISUAL DIALOGUE (1VD)



How many people are there?

it's crowded

Where is the pavement?

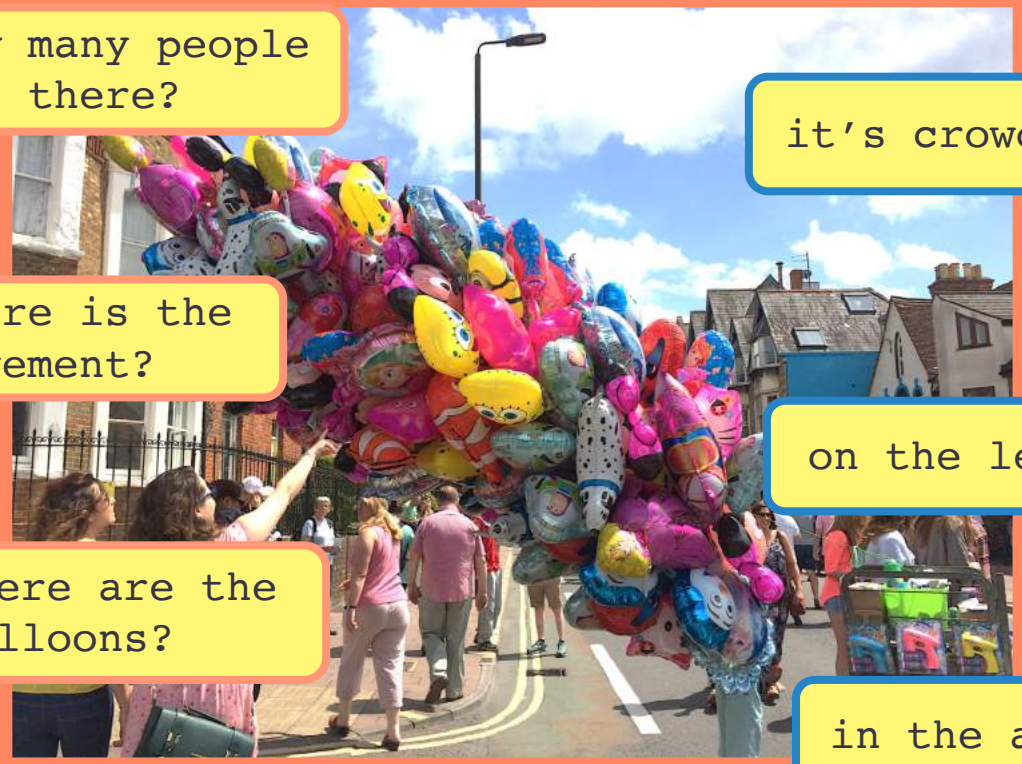
on the left

Where are the balloons?

in the air

A street with people and balloons

TWO-WAY VISUAL DIALOGUE (2VD)



How many people are there?

it's crowded

Where is the pavement?

on the left

Where are the balloons?

in the air

A street with people and balloons

CONDITIONAL VAE FOR IVD

Sohn et al. (2015), Kingma & Welling (2014)

image $\boldsymbol{i}$

A girl in a pink shirt eating a frosted donut

caption $\boldsymbol{c}$

How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	

dialogue history $\boldsymbol{h}_t^+ = \{\boldsymbol{h}_{t-1}, \boldsymbol{q}_t\}$

$$\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)$$

Blonde

CONDITIONAL VAE FOR IVD

Sohn et al. (2015), Kingma & Welling (2014)

image $\boldsymbol{i}$

A girl in a pink shirt eating a frosted donut

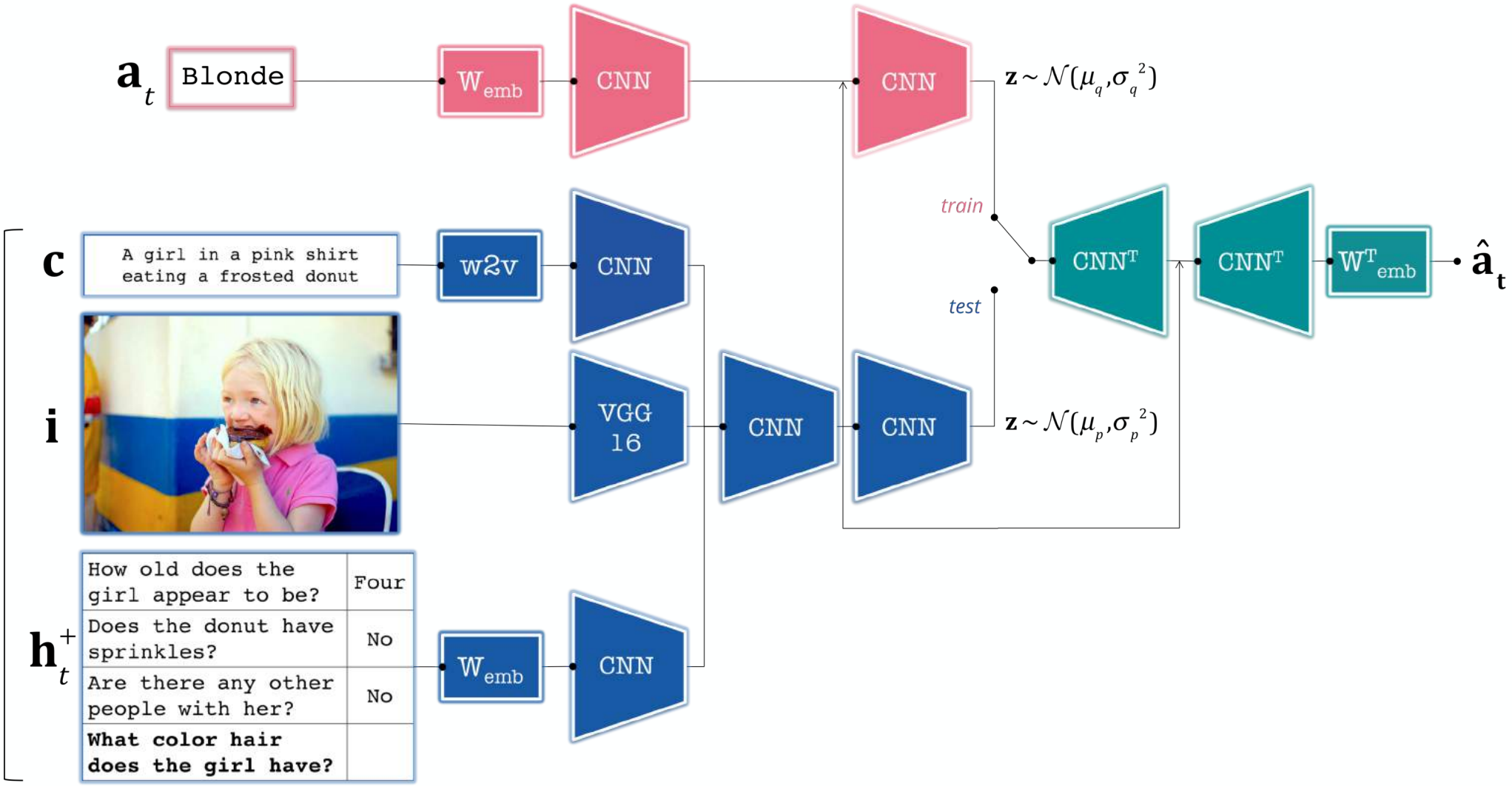
caption $\boldsymbol{c}$

How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	

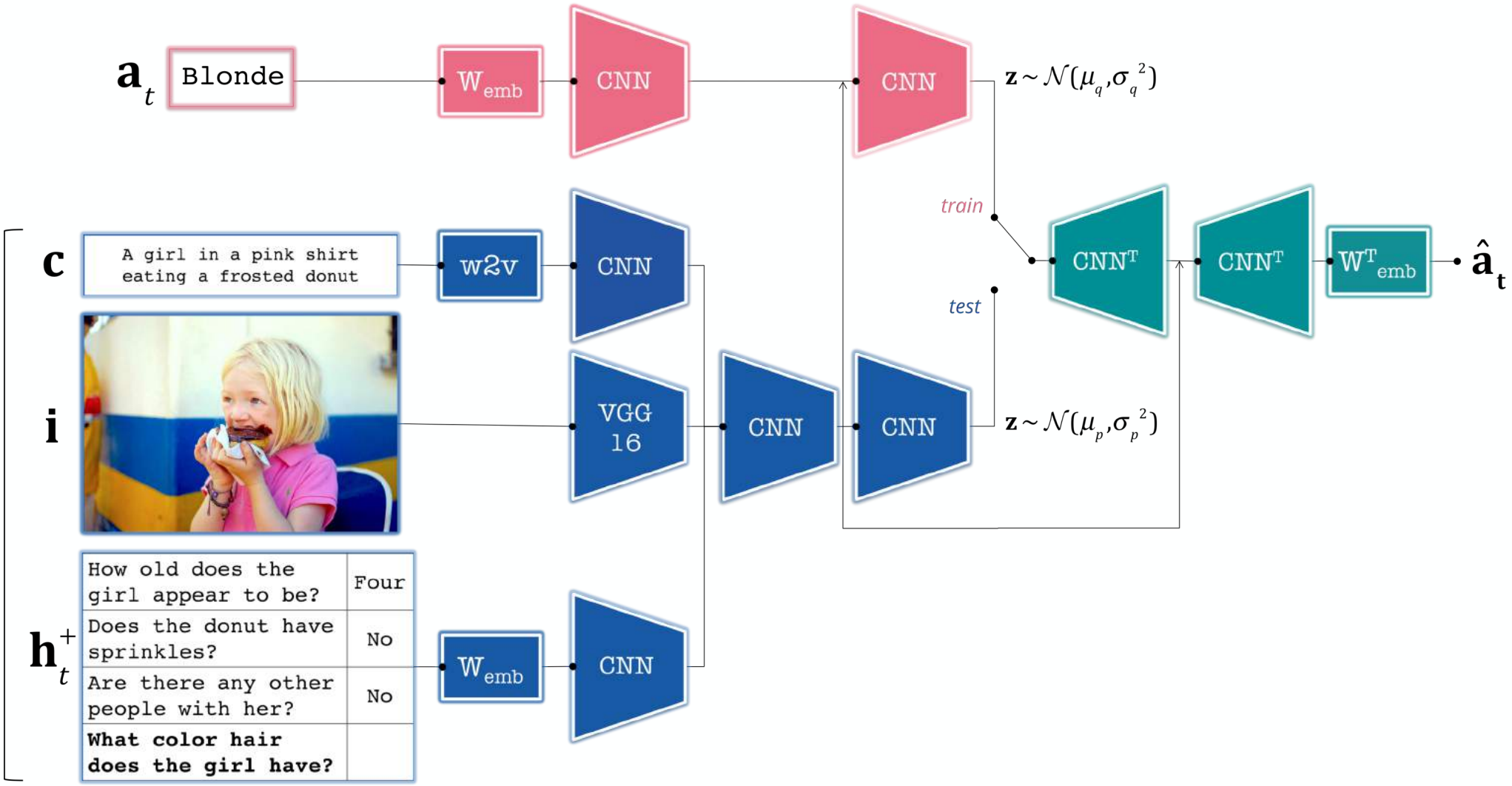
dialogue history $\boldsymbol{h}_t^+ = \{\boldsymbol{h}_{t-1}, \boldsymbol{q}_t\}$

$$\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \geq \mathbb{E}_{q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)}[\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)]$$
$$- \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+))$$

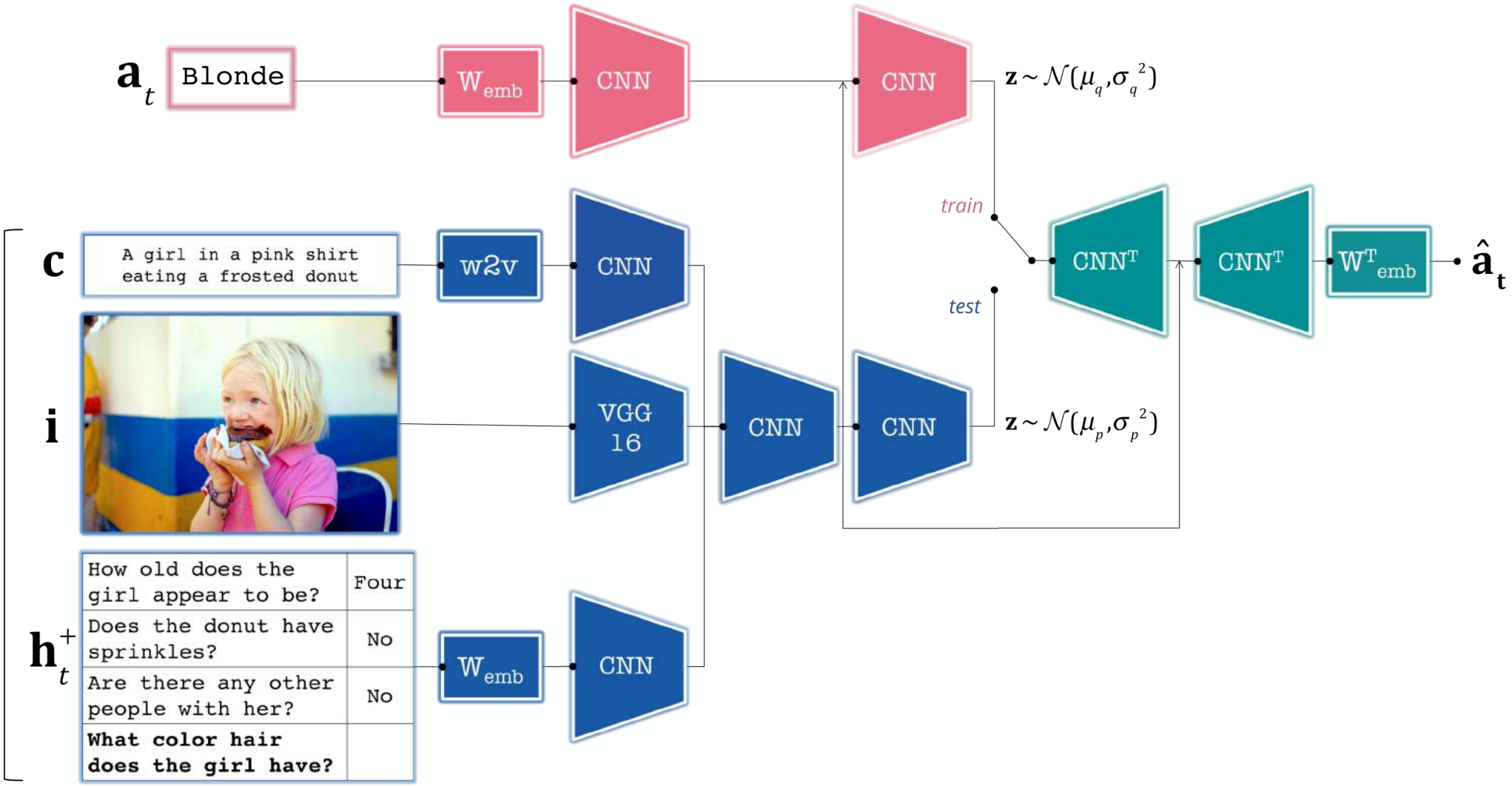
Blonde



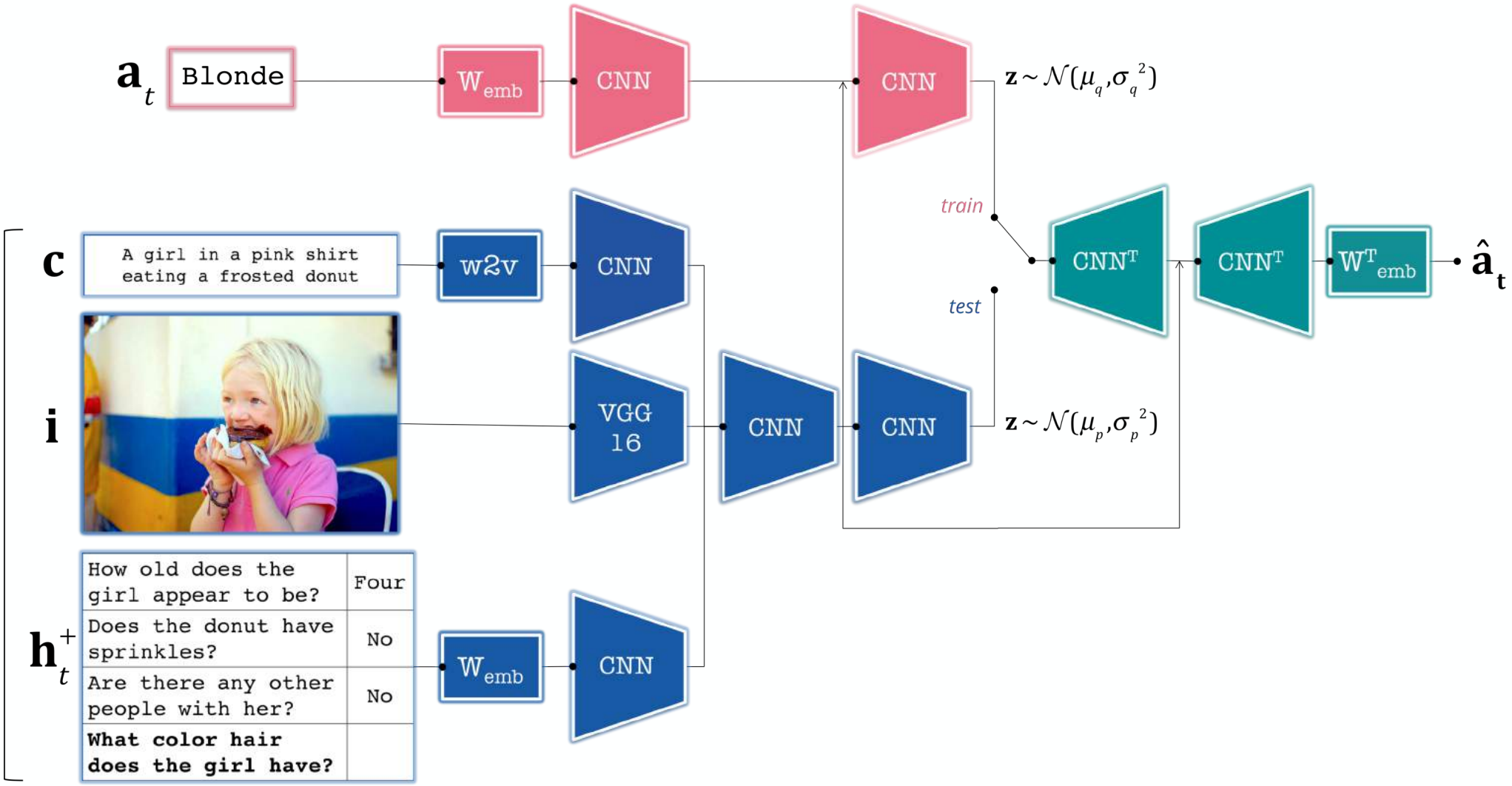
$$\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)}[\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)]$$
$$- \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+))$$



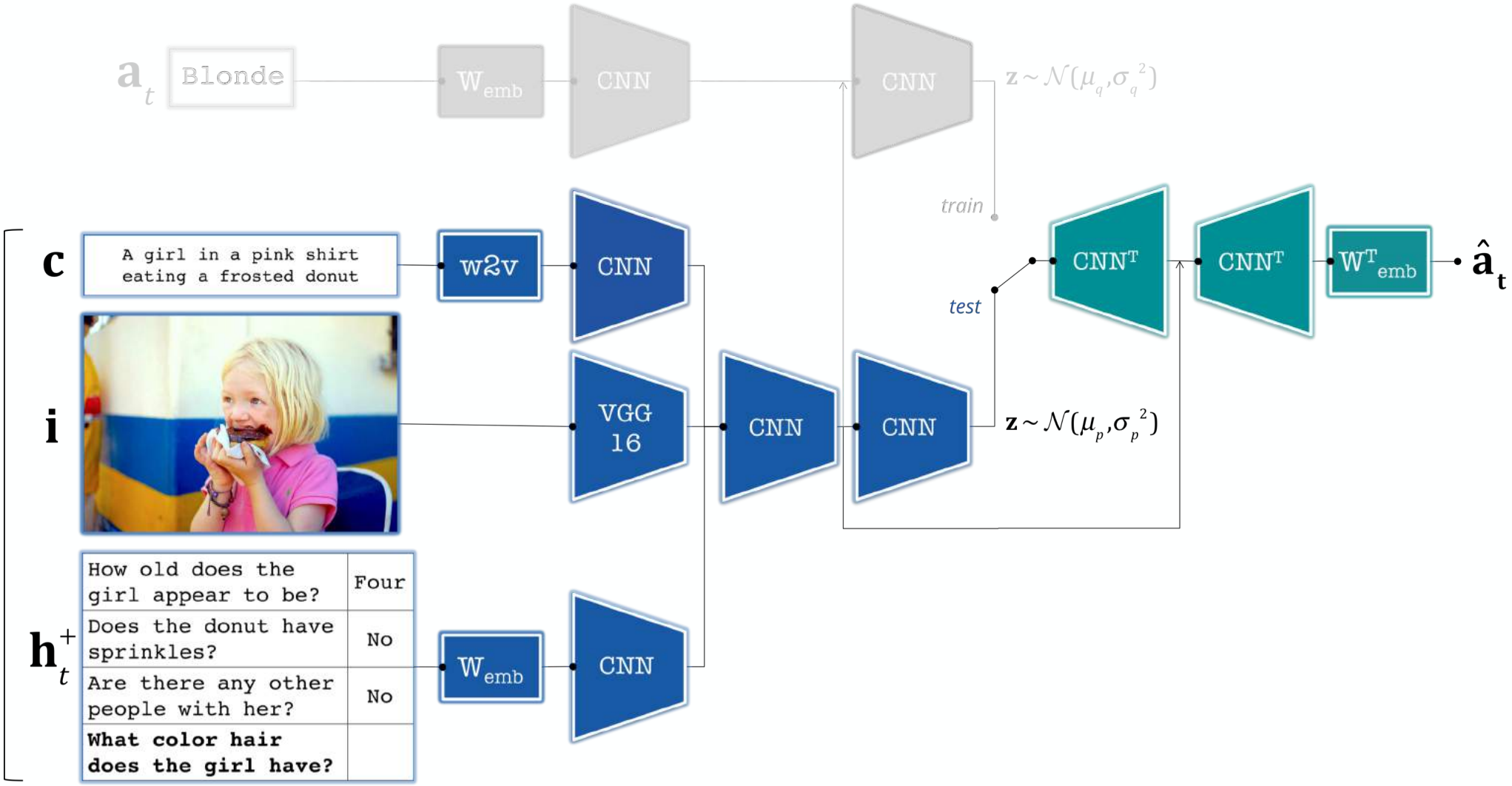
$$\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)}[\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)]$$
$$- \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+))$$



$$\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \geq \mathbb{E}_{q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)}[\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)]$$
$$- \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+))$$

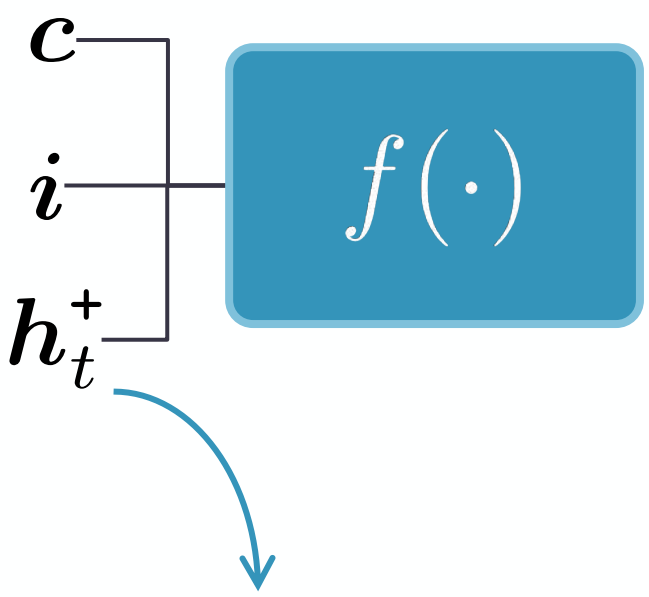


$$\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)}[\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)]$$
$$- \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+))$$



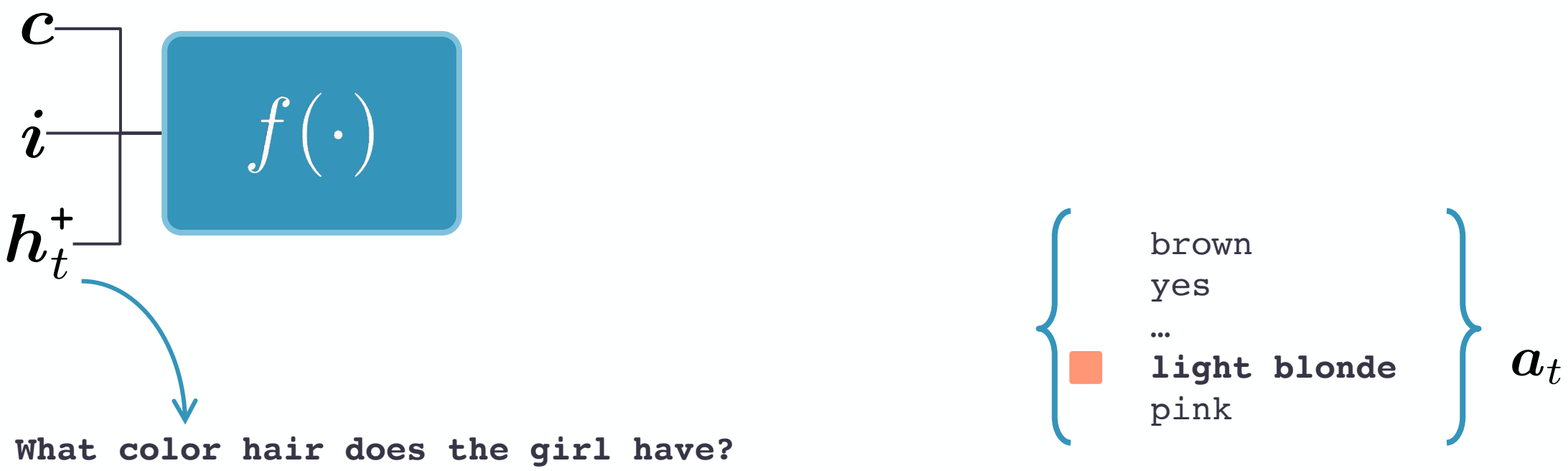
$$\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)}[\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)]$$
$$- \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+))$$

IVD EVALUATION

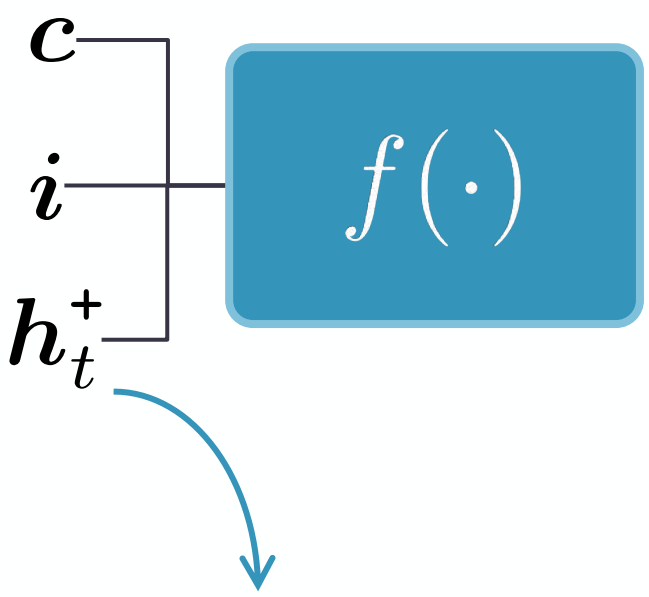


What color hair does the girl have?

IVD EVALUATION



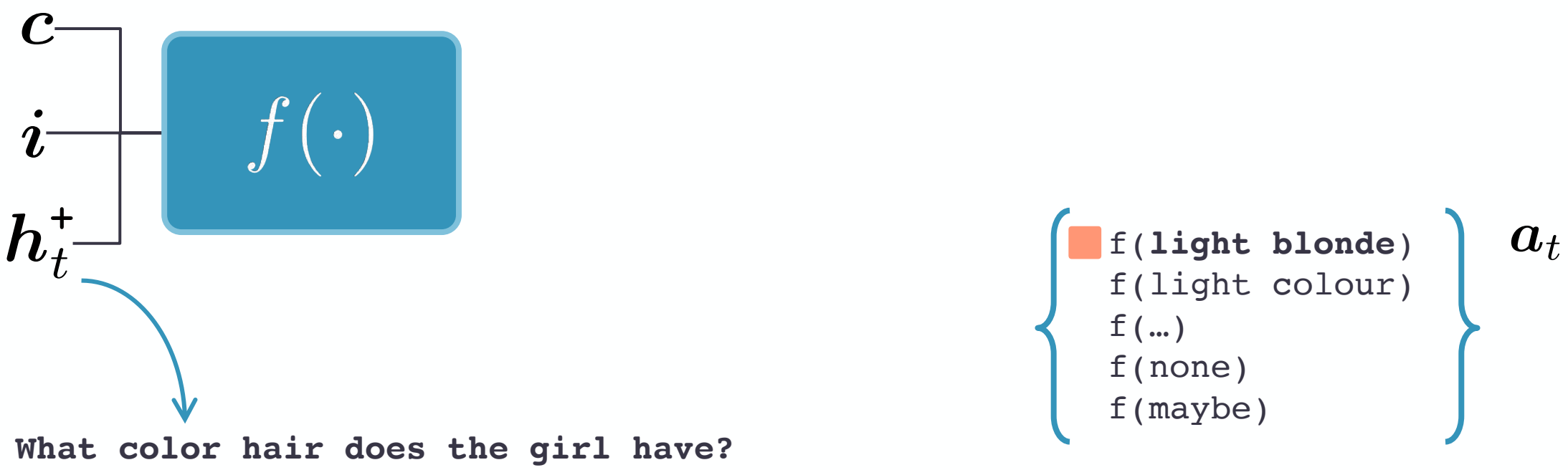
IVD EVALUATION



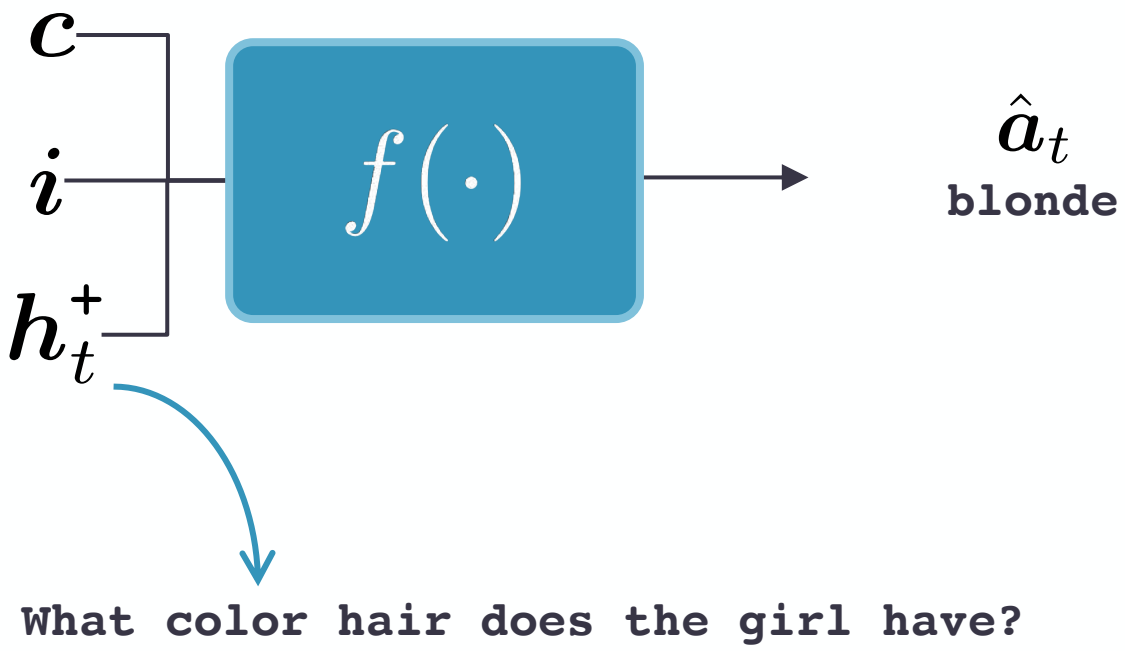
What color hair does the girl have?

$$\left\{\begin{array}{l} \texttt{f(brown)} \\ \texttt{f(yes)} \\ \texttt{f(…)} \\ \texttt{f(}\textbf{light blonde}\texttt{)} \\ \texttt{f(pink)} \end{array}\right\} a_t$$

IVD EVALUATION



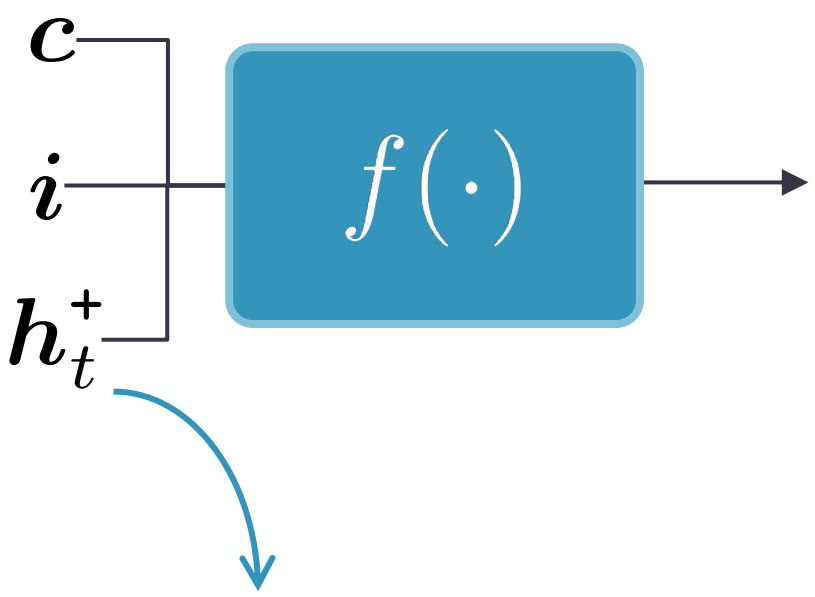
IVD EVALUATION



f(**light blonde**)
f(light colour)
f(…)
f(none)
f(maybe)

a_t

IVD EVALUATION



What color hair does the girl have?

$\hat{a}_t$

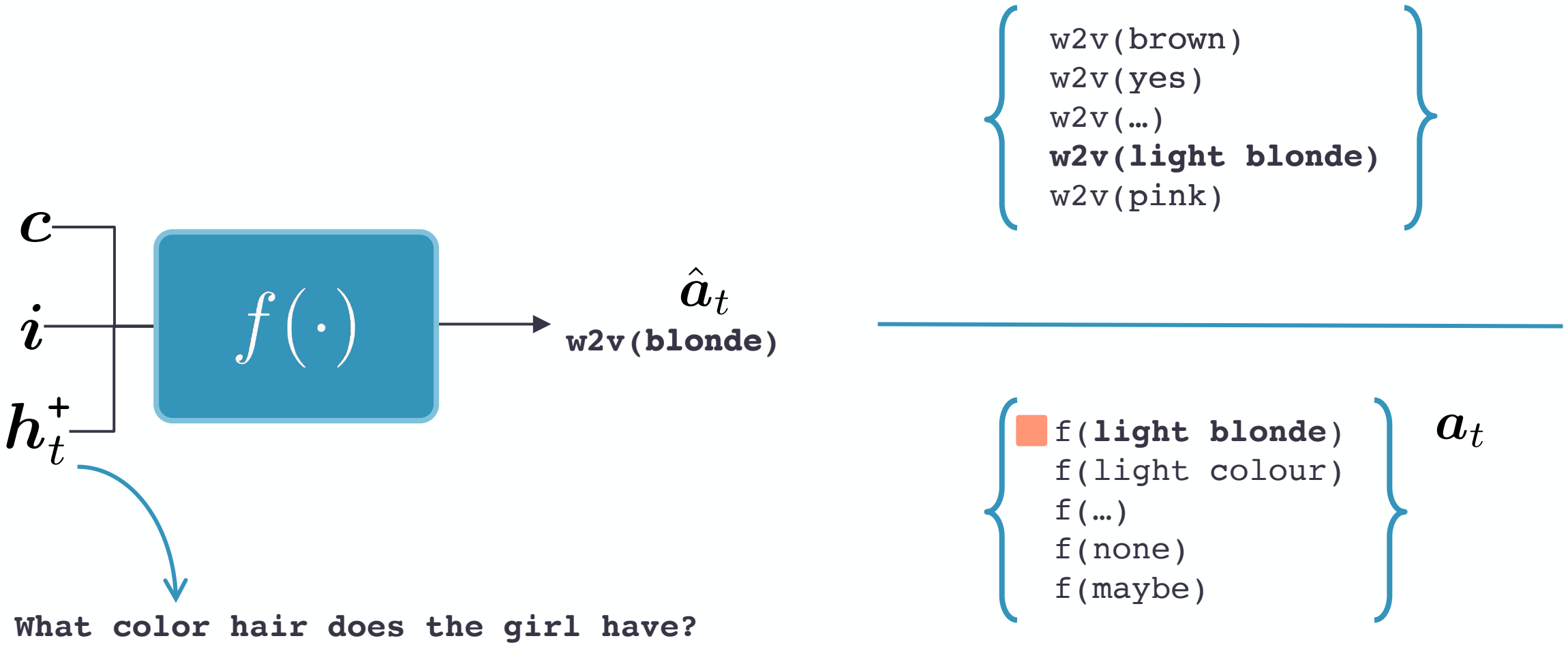
blonde

brown
yes
…
light blonde
pink

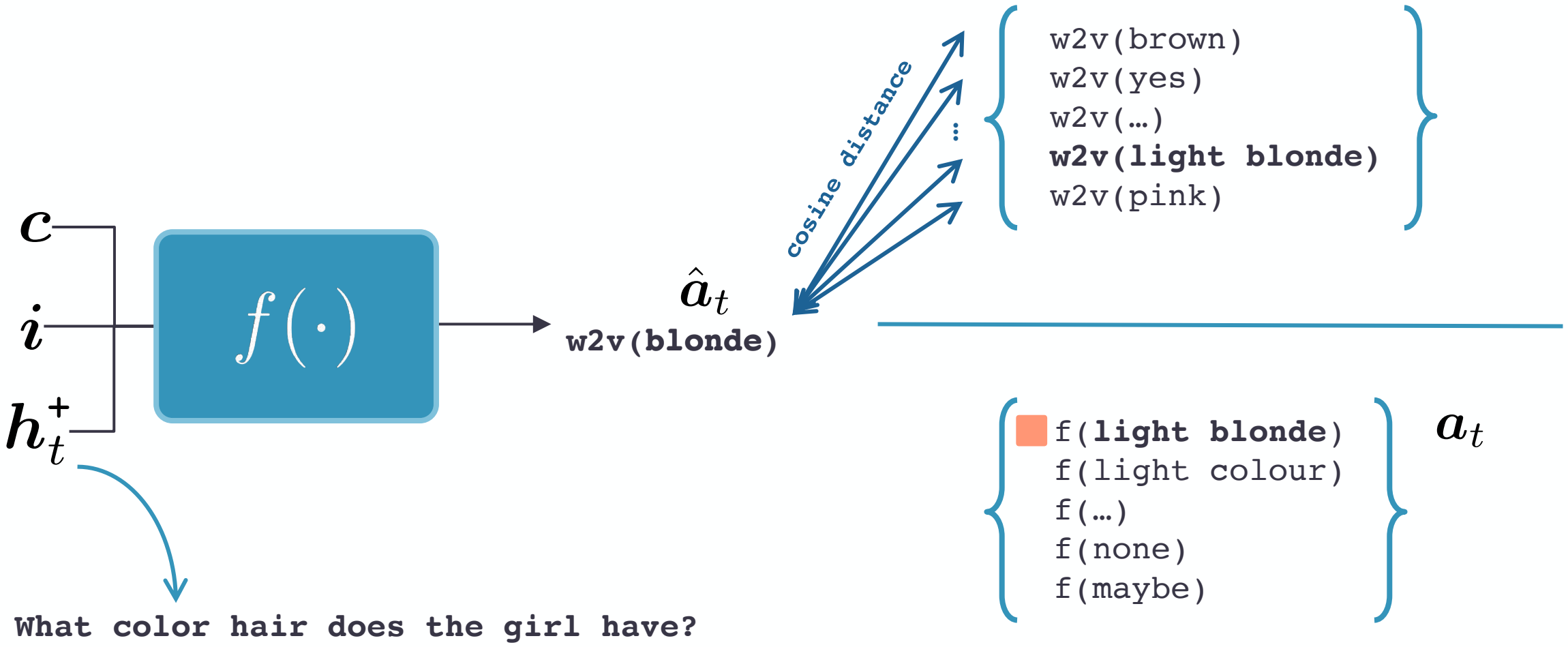
f(**light blonde**)
f(light colour)
f(…)
f(none)
f(maybe)

a_t

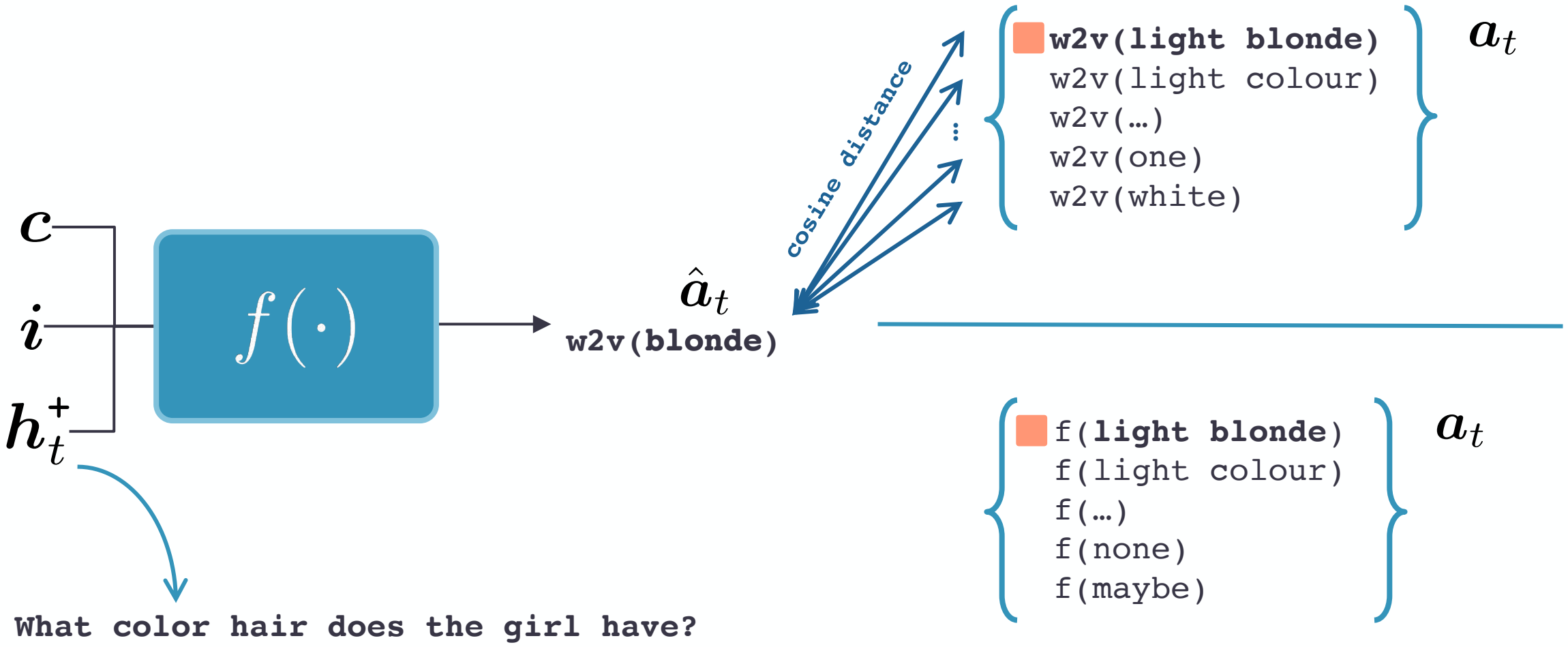
IVD EVALUATION



IVD EVALUATION



IVD EVALUATION



IVD EVALUATION (I)

Rank with…		Mean Rank
model	RL-QAbot [1]	21.13
	MN-QIH-G [2]	**17.06**
	Ours	20.38
w2v cosine	MN-QIH-G [2]	31.31
	Ours	**25.65**

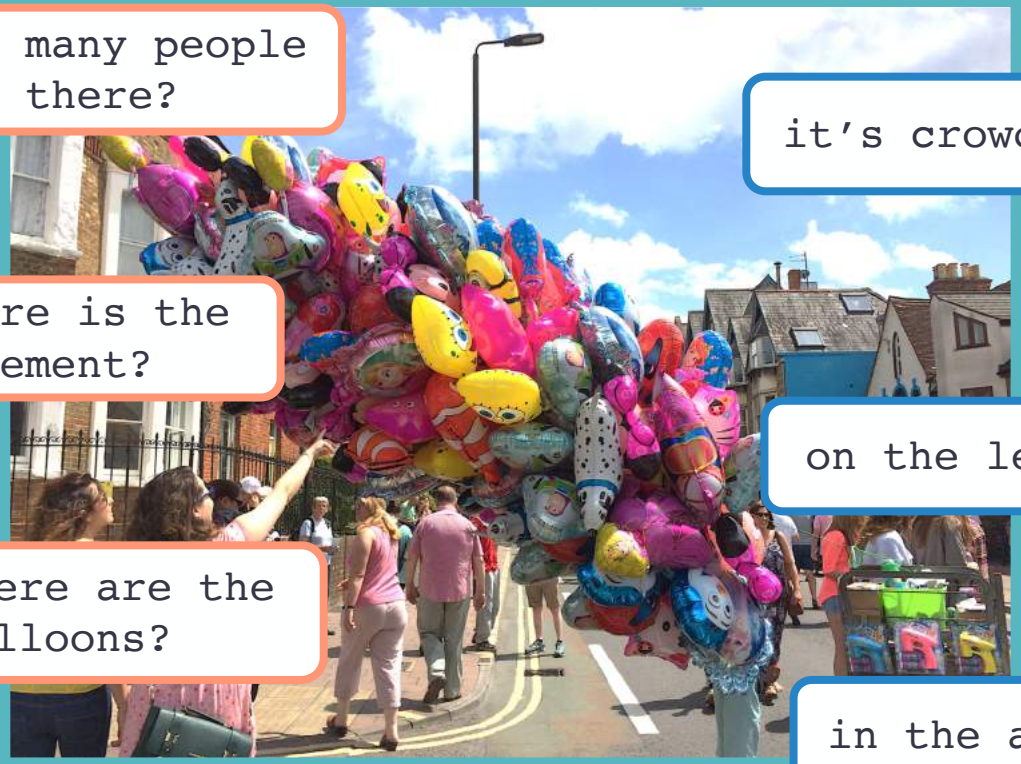
[1] Das et al. Learning cooperative visual dialog agents with deep reinforcement learning. In ICCV, 2017
[2] Das et al. Visual Dialog. In CVPR, 2017

IVD EVALUATION (II)

How old is the man?	He looks thirty	Twenties	Thirtys	I say teenager
What race is the man?	White	White	I can't see	Caucasian
Is he wearing a shirt?	Yes	Yes	No	Yes
What color?	Grey	Grey	White	Black
What gender is the baby?	I	Boy	Male	I
What is the baby wearing?	Shirt	T-shirt has pants	Shirt and shirt	Looks like white
What color is the remote?	White	Silver	It is white and black	White
Is the chair wood?	No	No	No it's a chair	No
What color is the chair?	Brown	White has checkered	Gray	Brown with texture

A man sitting in a chair holding a baby who is chewing on a remote

ONE-WAY VISUAL DIALOGUE (1VD)



How many people are there?

it's crowded

Where is the pavement?

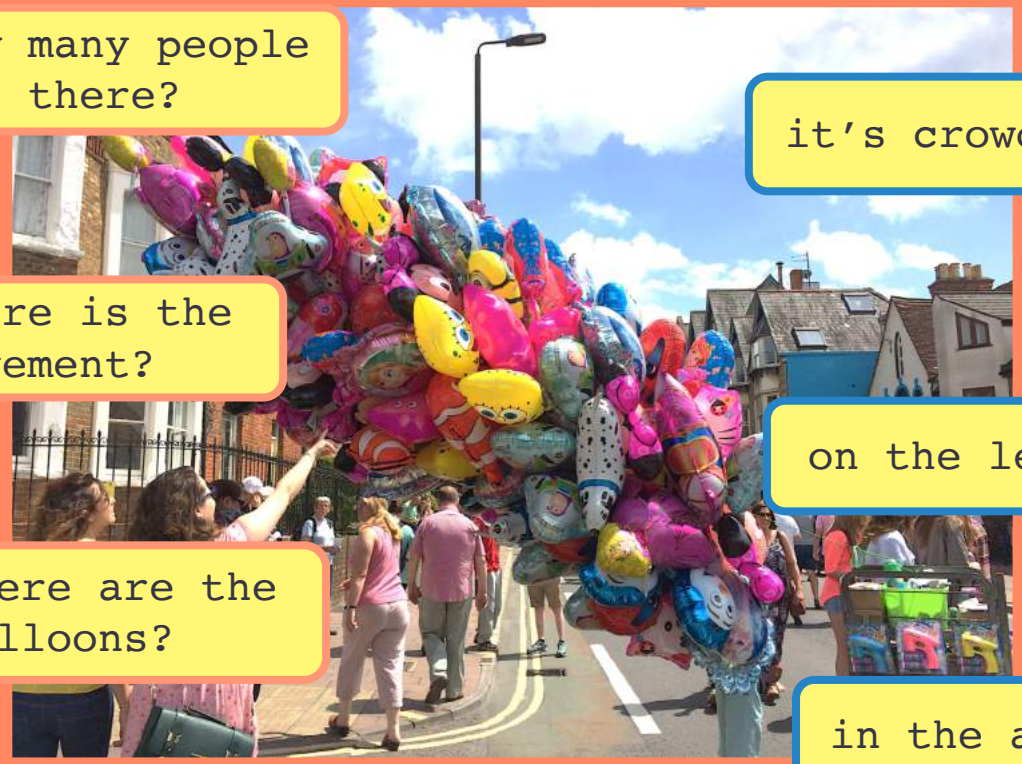
on the left

Where are the balloons?

in the air

A street with people and balloons

TWO-WAY VISUAL DIALOGUE (2VD)



How many people are there?

it's crowded

Where is the pavement?

on the left

Where are the balloons?

in the air

A street with people and balloons

CONDITIONAL VAE FOR IVD

image $\boldsymbol{i}$

A girl in a pink shirt eating a frosted donut

caption $\boldsymbol{c}$

How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	

dialogue history $\boldsymbol{h}_t^+ = \{\boldsymbol{h}_{t-1}, \boldsymbol{q}_t\}$

$$\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)$$

Blonde

CONDITIONAL VAE FOR 2VD

image $\boldsymbol{i}$

A girl in a pink shirt eating a frosted donut

caption $\boldsymbol{c}$

How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	Blonde

dialogue $\boldsymbol{d}_{1:T}$

$$\log p_\theta(\boldsymbol{d}_{1:T} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})$$

CONDITIONAL VAE FOR 2VD

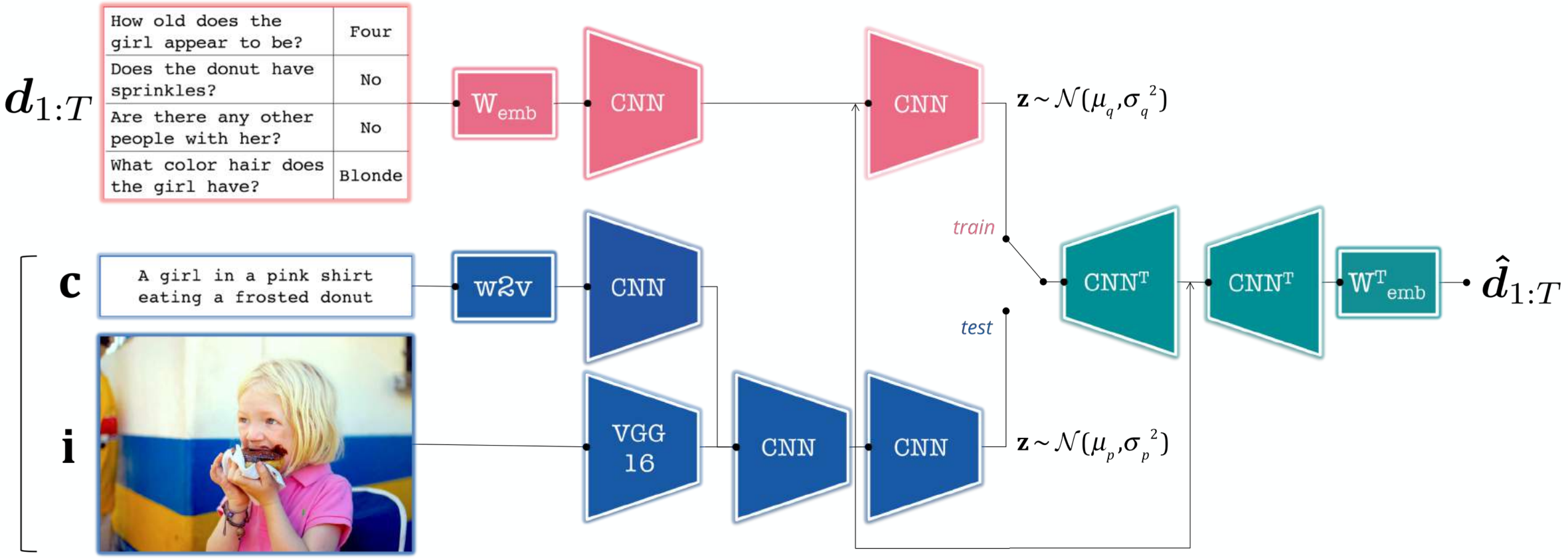
image $\boldsymbol{i}$

A girl in a pink shirt eating a frosted donut

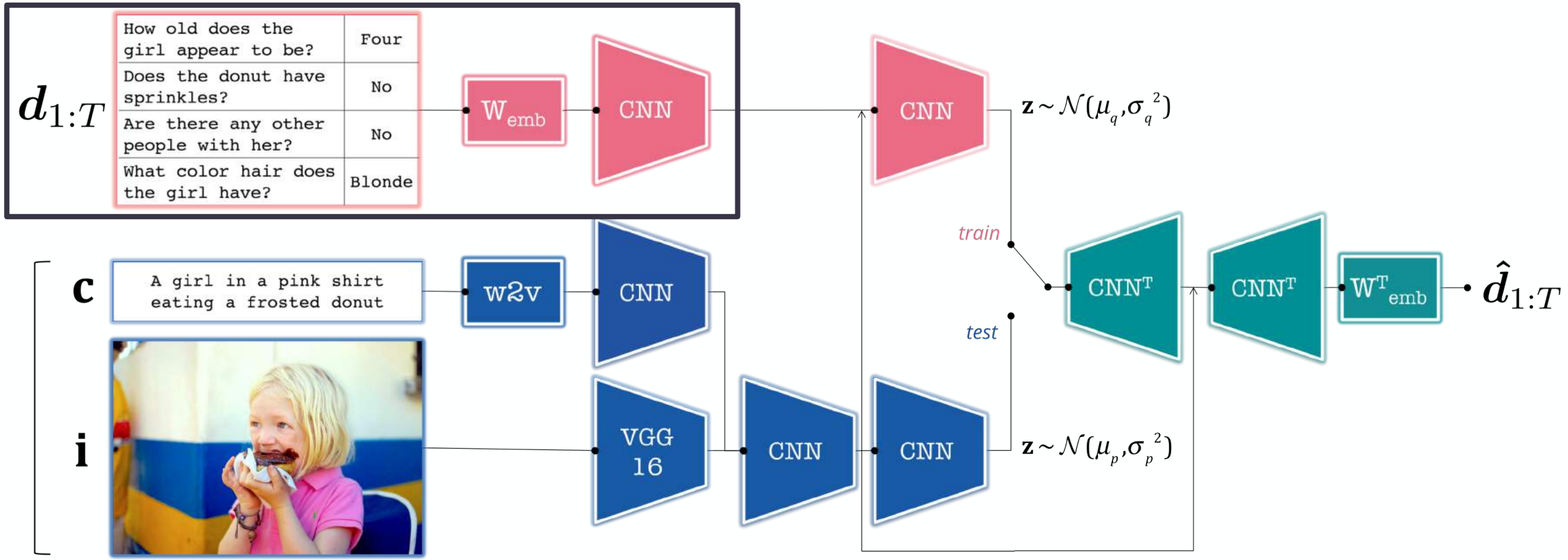
caption $\boldsymbol{c}$

dialogue history $\boldsymbol{h} = \emptyset$

$$\log p_\theta(\boldsymbol{d}_{1:T} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z} \mid \boldsymbol{d}_{1:T}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})}[\log p_\theta(\boldsymbol{d}_{1:T} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})] - \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{d}_{1:T}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}))$$



$$\log p_\theta(\boldsymbol{d}_{1:T} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z} \mid \boldsymbol{d}_{1:T}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})}[\log p_\theta(\boldsymbol{d}_{1:T} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})]$$
$$- \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{d}_{1:T}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}))$$



$$\log p_\theta(\boldsymbol{d}_{1:T} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z} \mid \boldsymbol{d}_{1:T}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})}[\log p_\theta(\boldsymbol{d}_{1:T} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})]$$
$$- \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{d}_{1:T}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}))$$

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	Blonde

Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

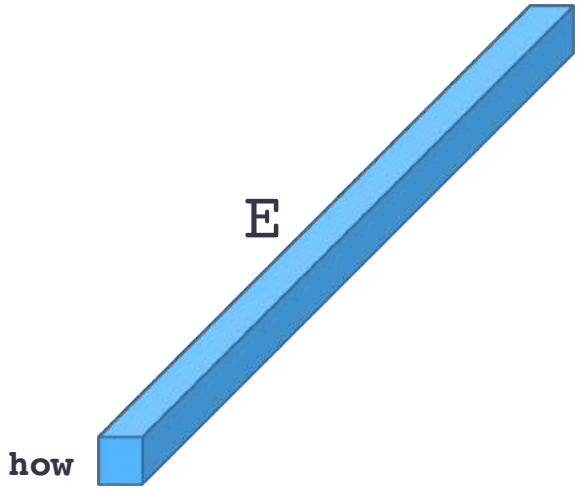
‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde

Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

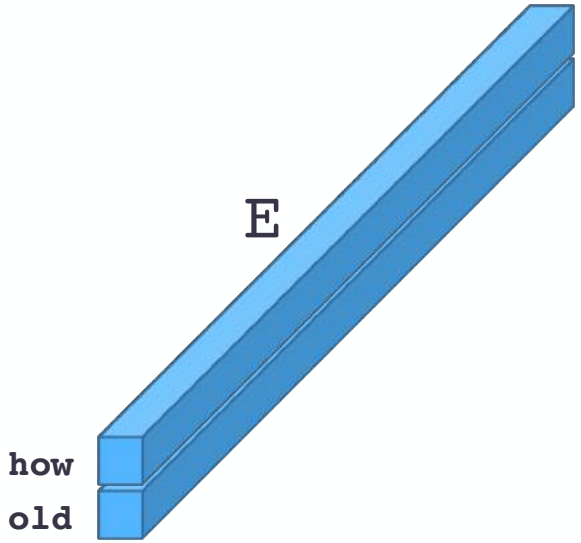
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

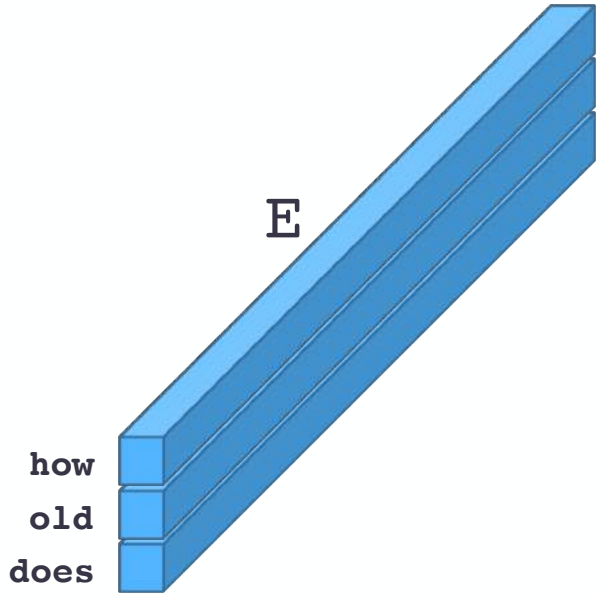
How $\boxed{\text{old}}$ does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

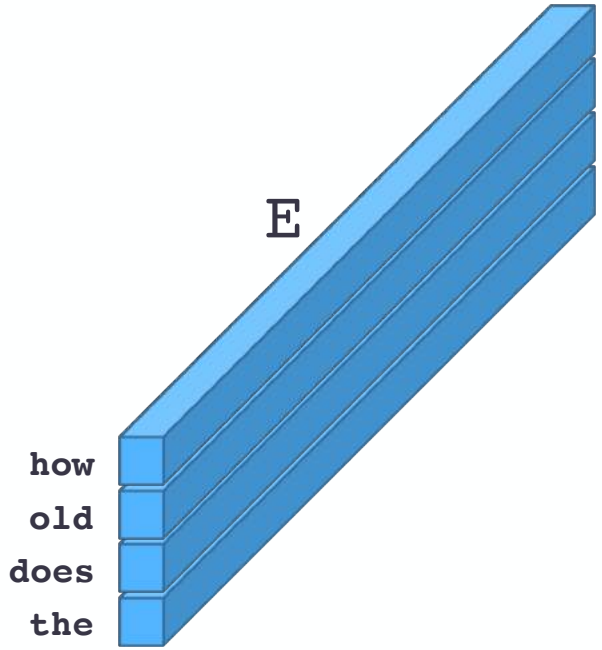
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

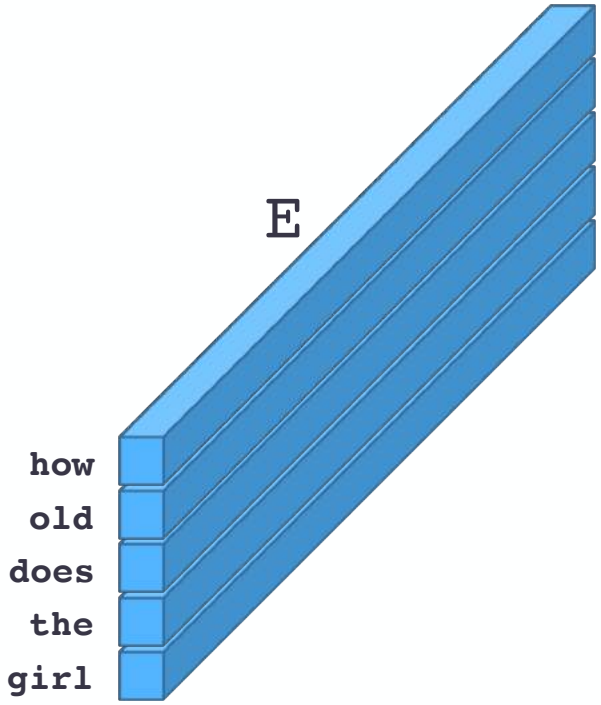
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

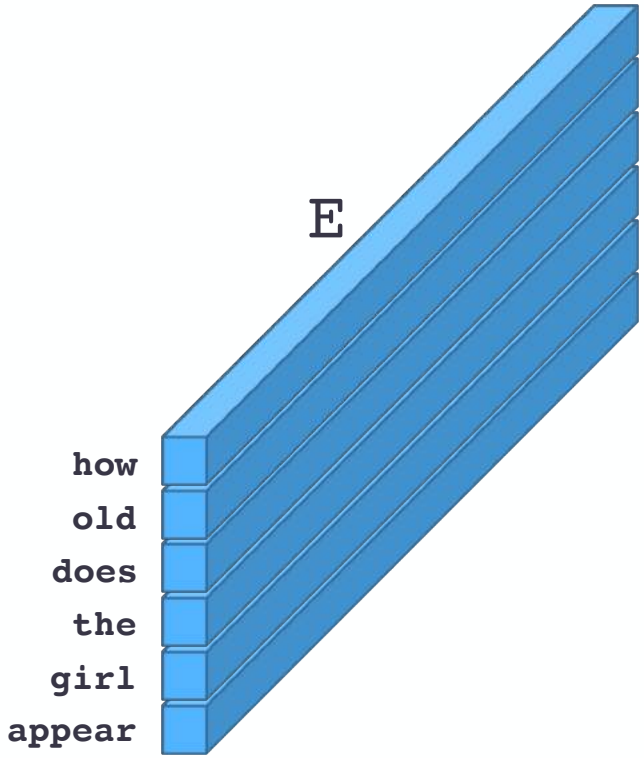
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

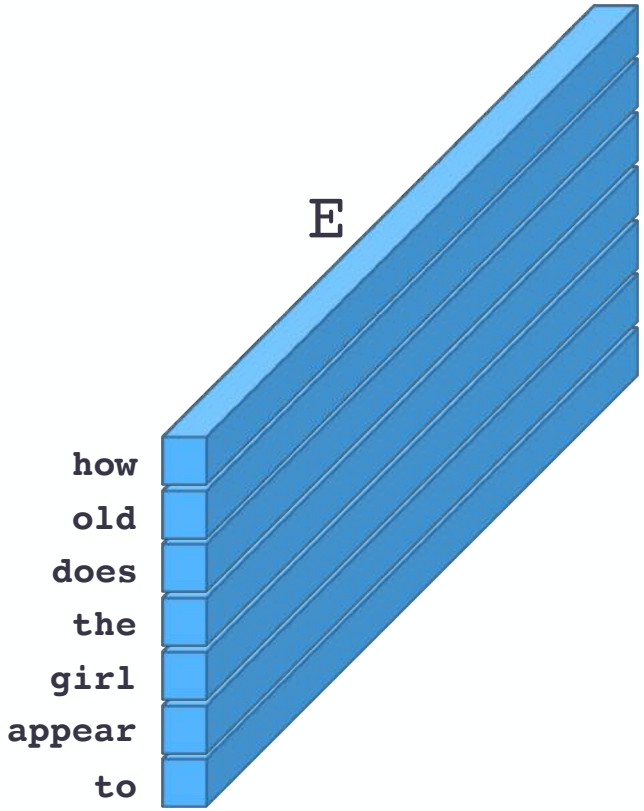
How old does the girl $\boxed{\text{appear}}$ to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

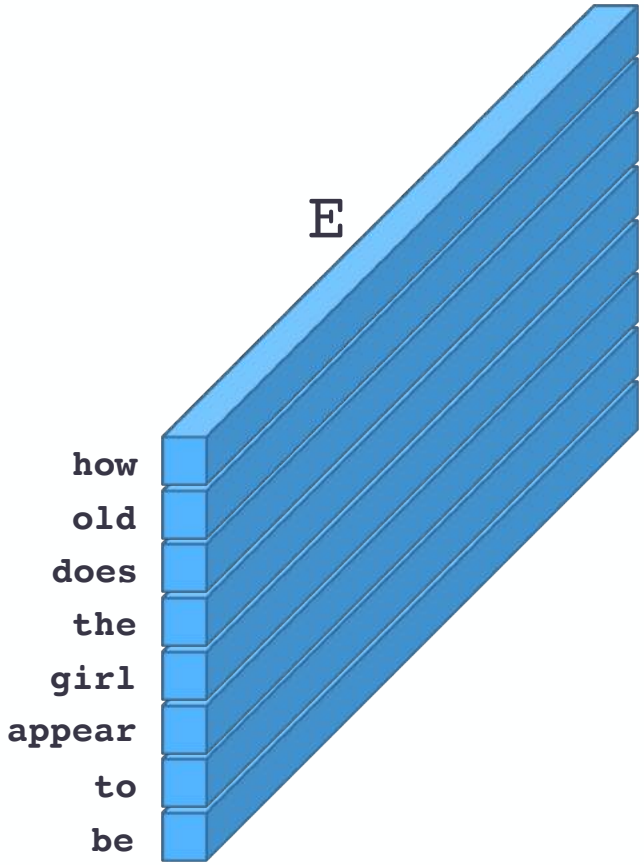
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

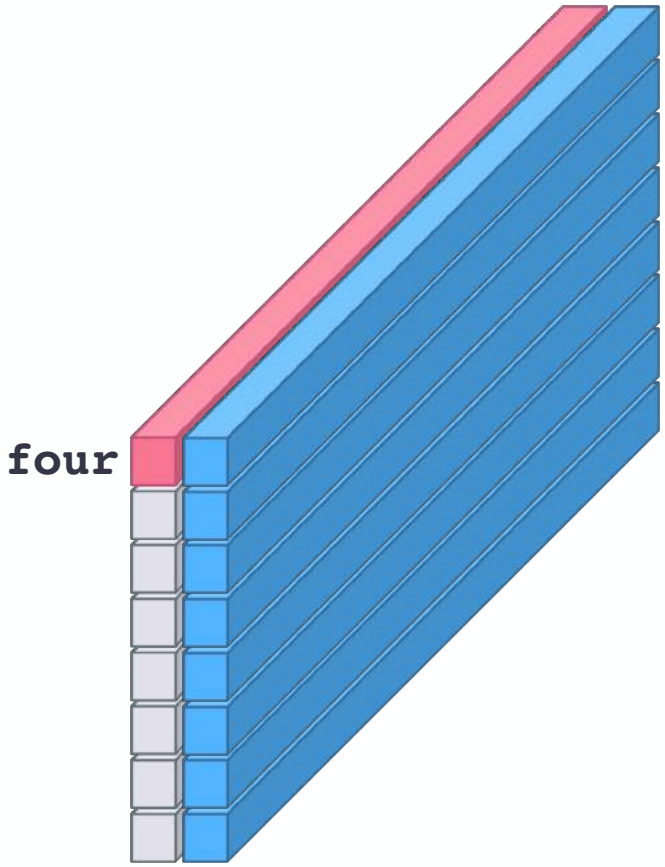
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

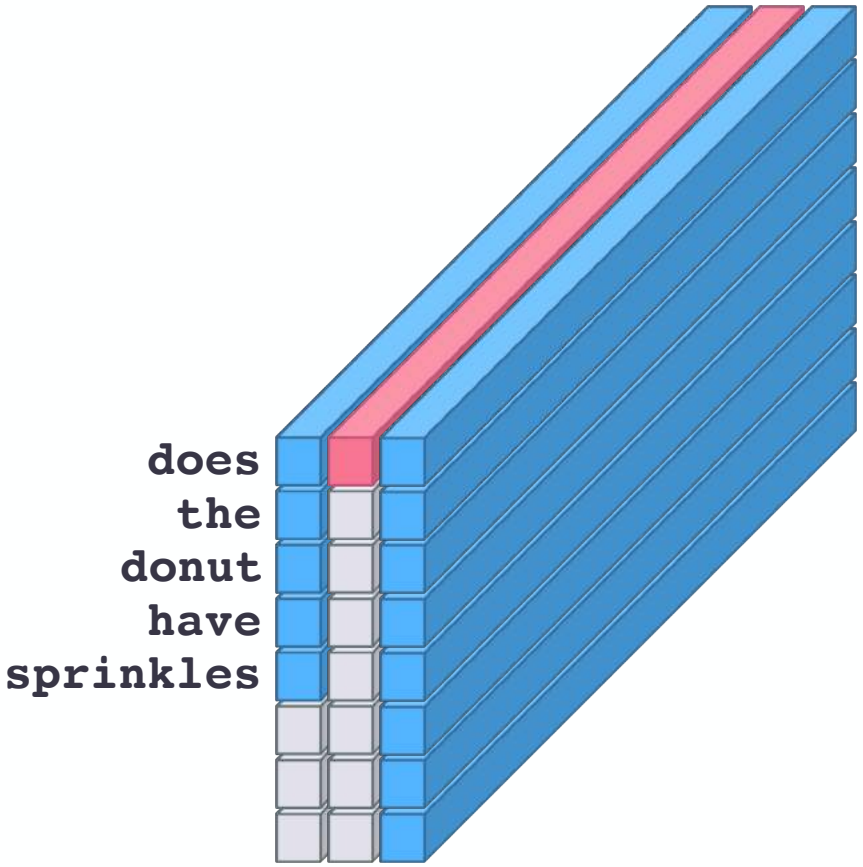
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

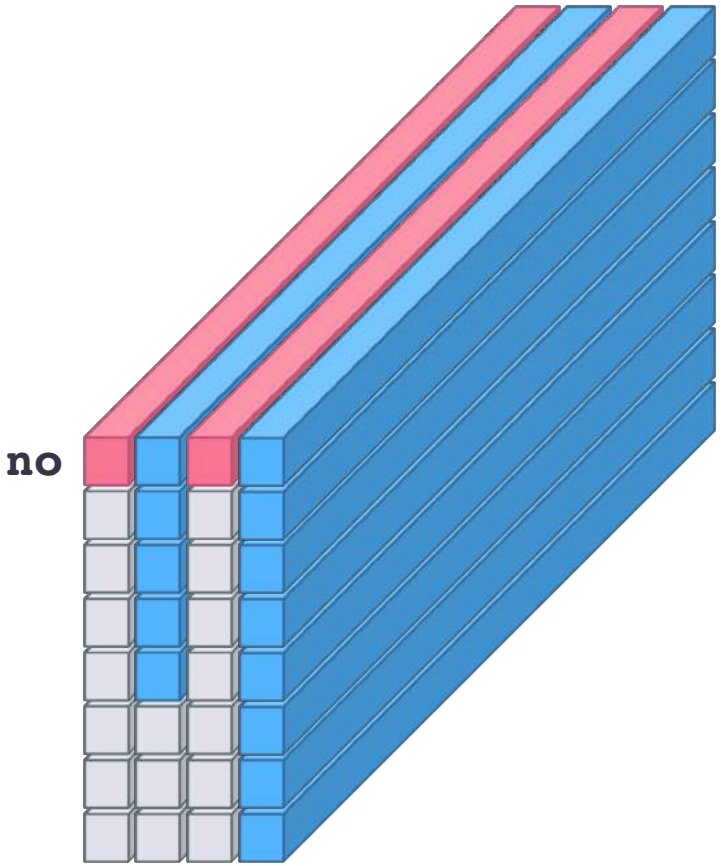
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

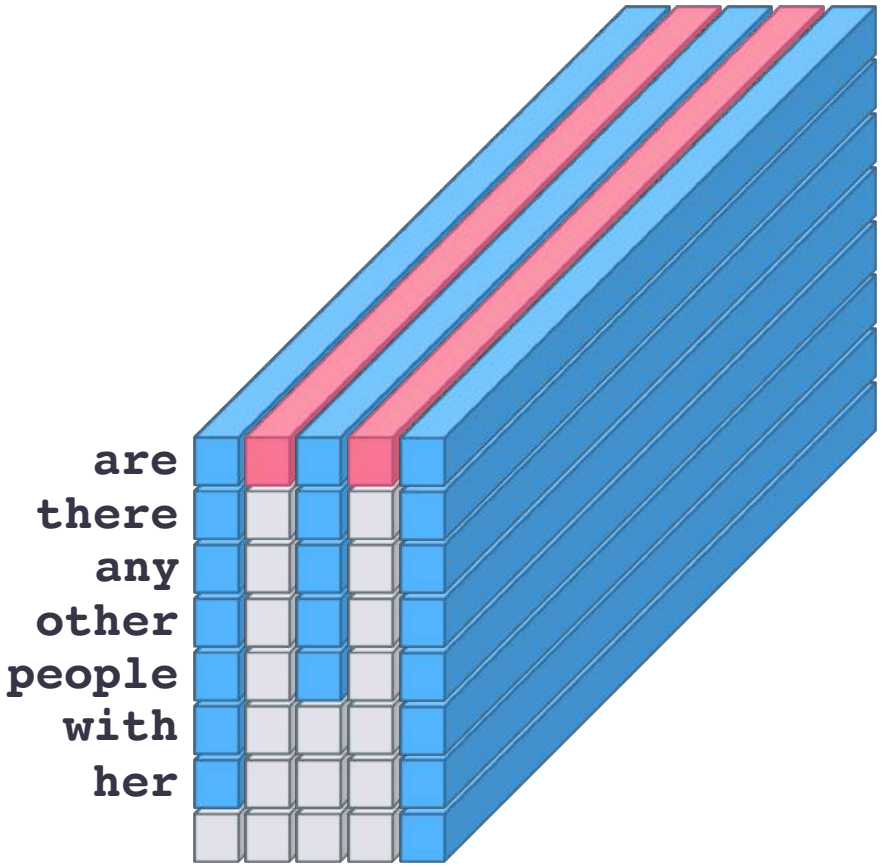
How old does the girl appear to be?
Four
Does the donut have sprinkles?
→ No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

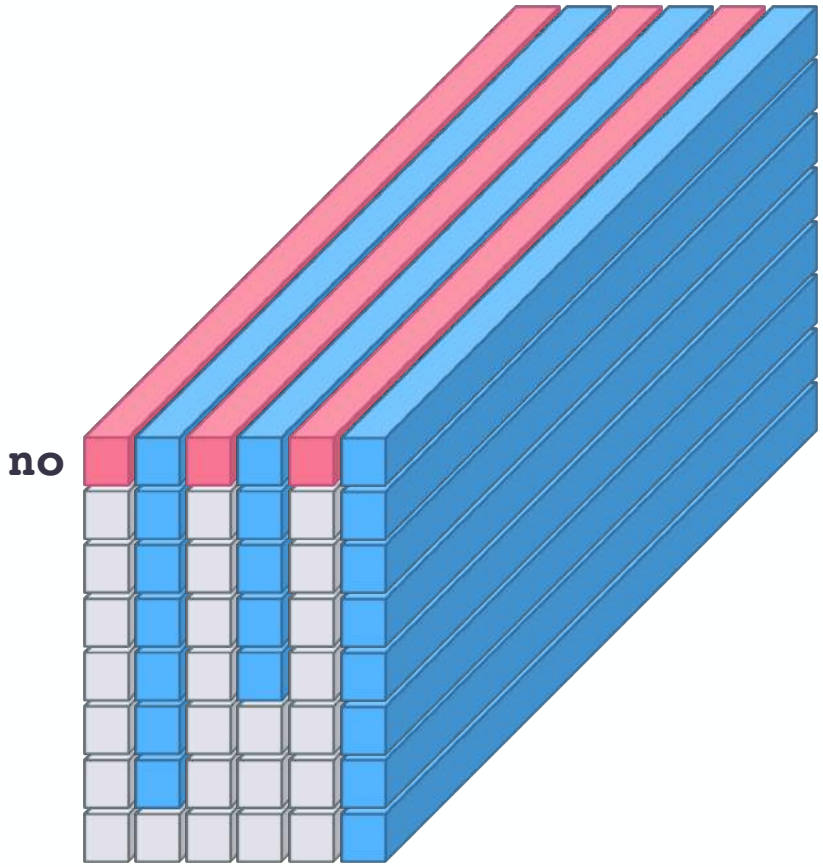
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
→ Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

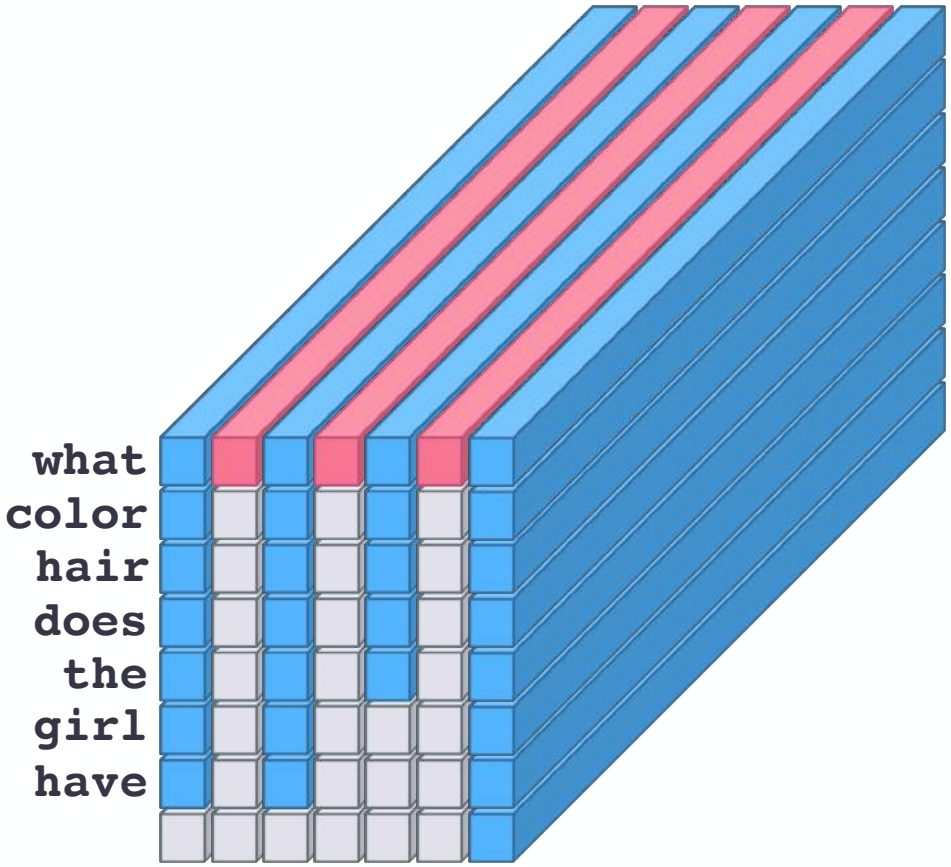
How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
→ No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

How old does the girl appear to be?

Four

Does the donut have sprinkles?

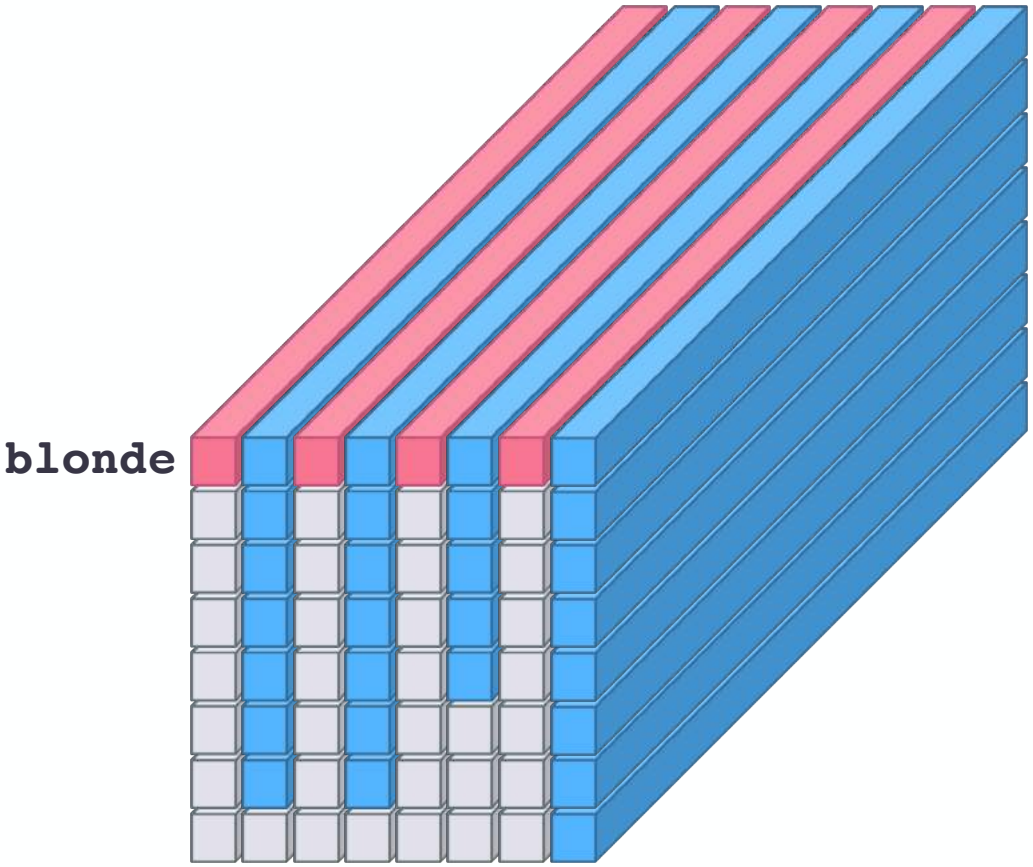
No

Are there any other people with her?

No

What color hair does the girl have?

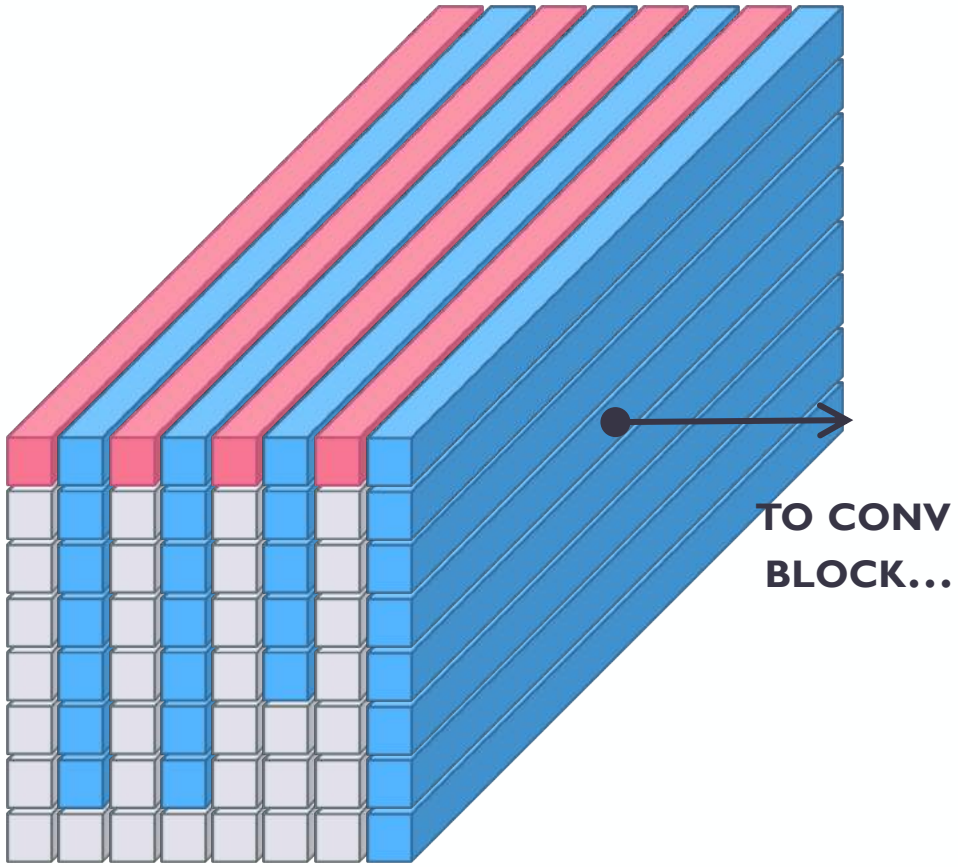
→ Blonde



Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

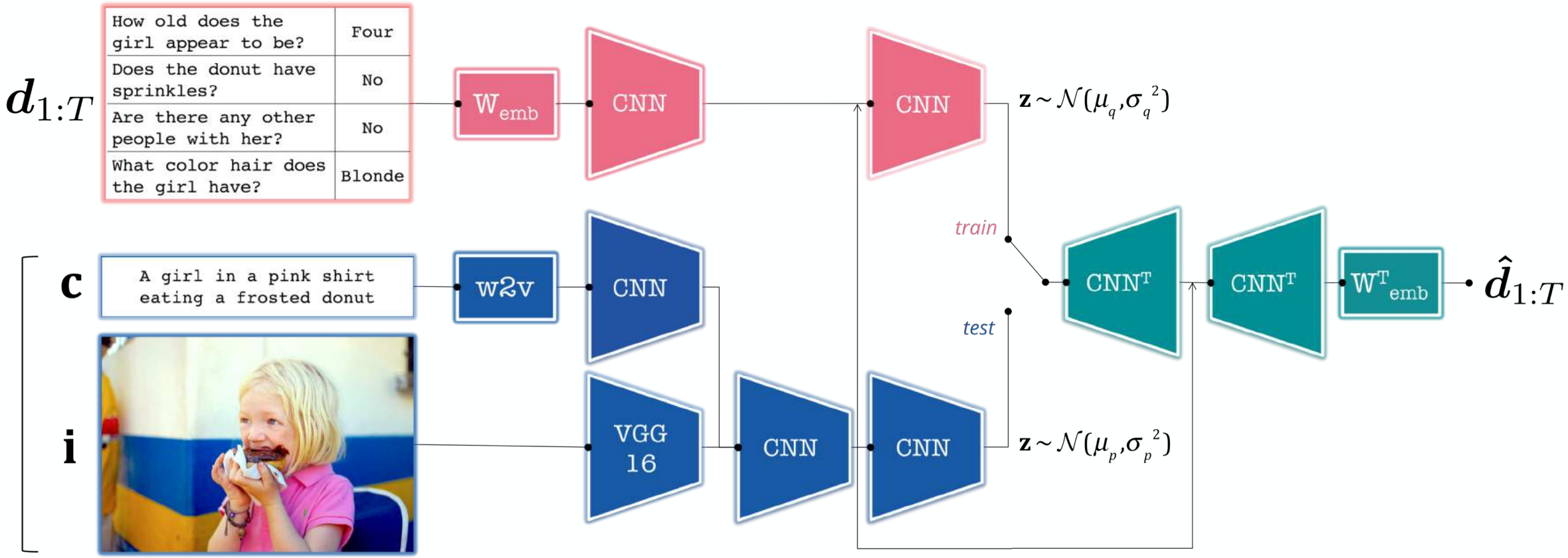
‘COLOURING’ DIALOGUE WITH CONVOLUTIONS

How old does the girl appear to be?
Four
Does the donut have sprinkles?
No
Are there any other people with her?
No
What color hair does the girl have?
→ Blonde



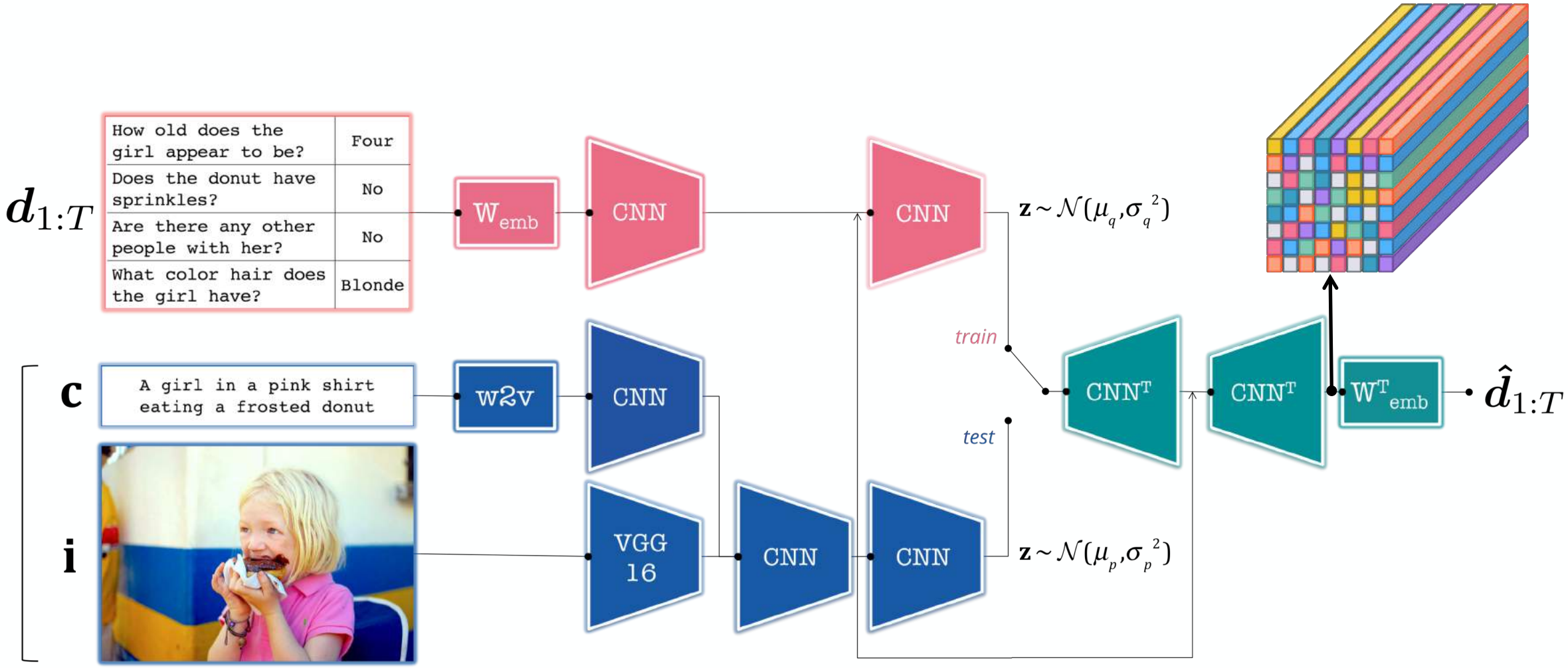
Convolutions for language: Kalchbrenner et al. (2014), Pham et al. (2016), Bai, et al. (2018)

EXPLICIT SEQUENCE WITH AN AUTOREGRESSIVE DECODER



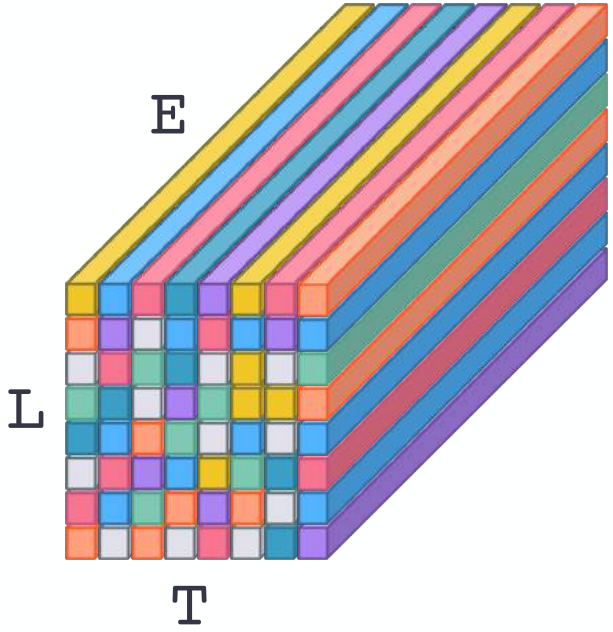
van den Oord et al. (2016)

EXPLICIT SEQUENCE WITH AN AUTOREGRESSIVE DECODER



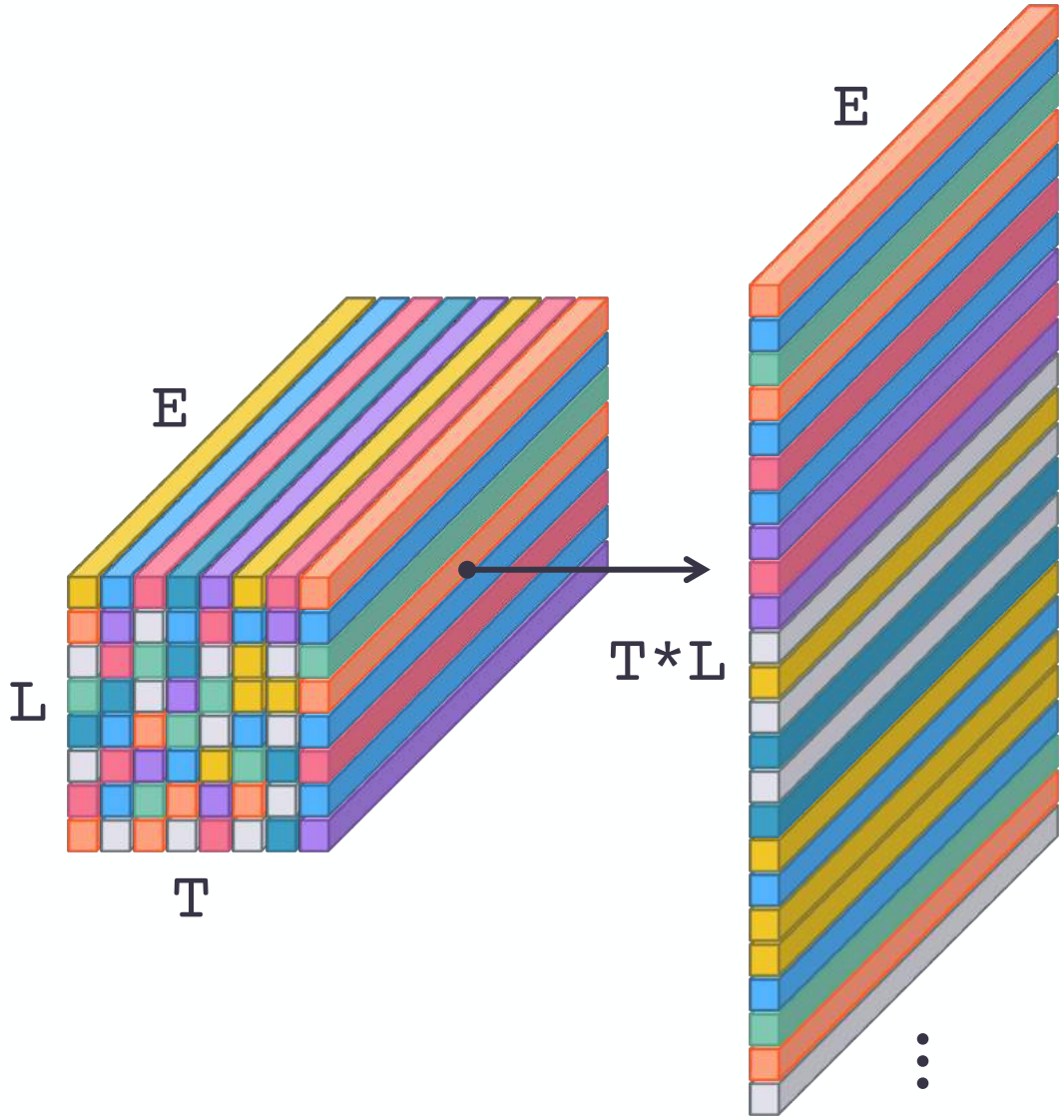
van den Oord et al. (2016)

AN AUTO-REGRESSIVE (AR) LAYER



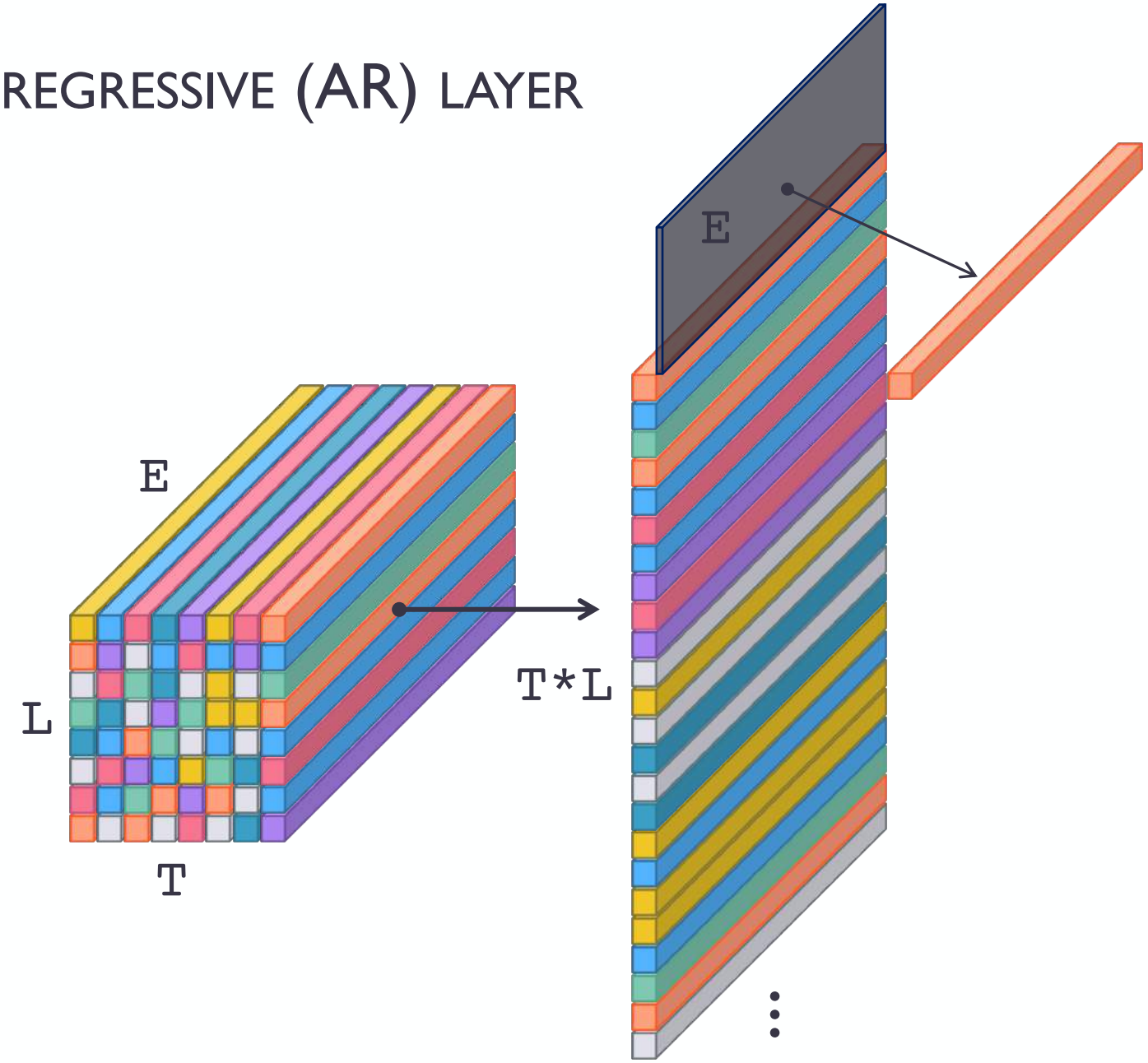
van den Oord et al. (2016)

AN AUTO-REGRESSIVE (AR) LAYER



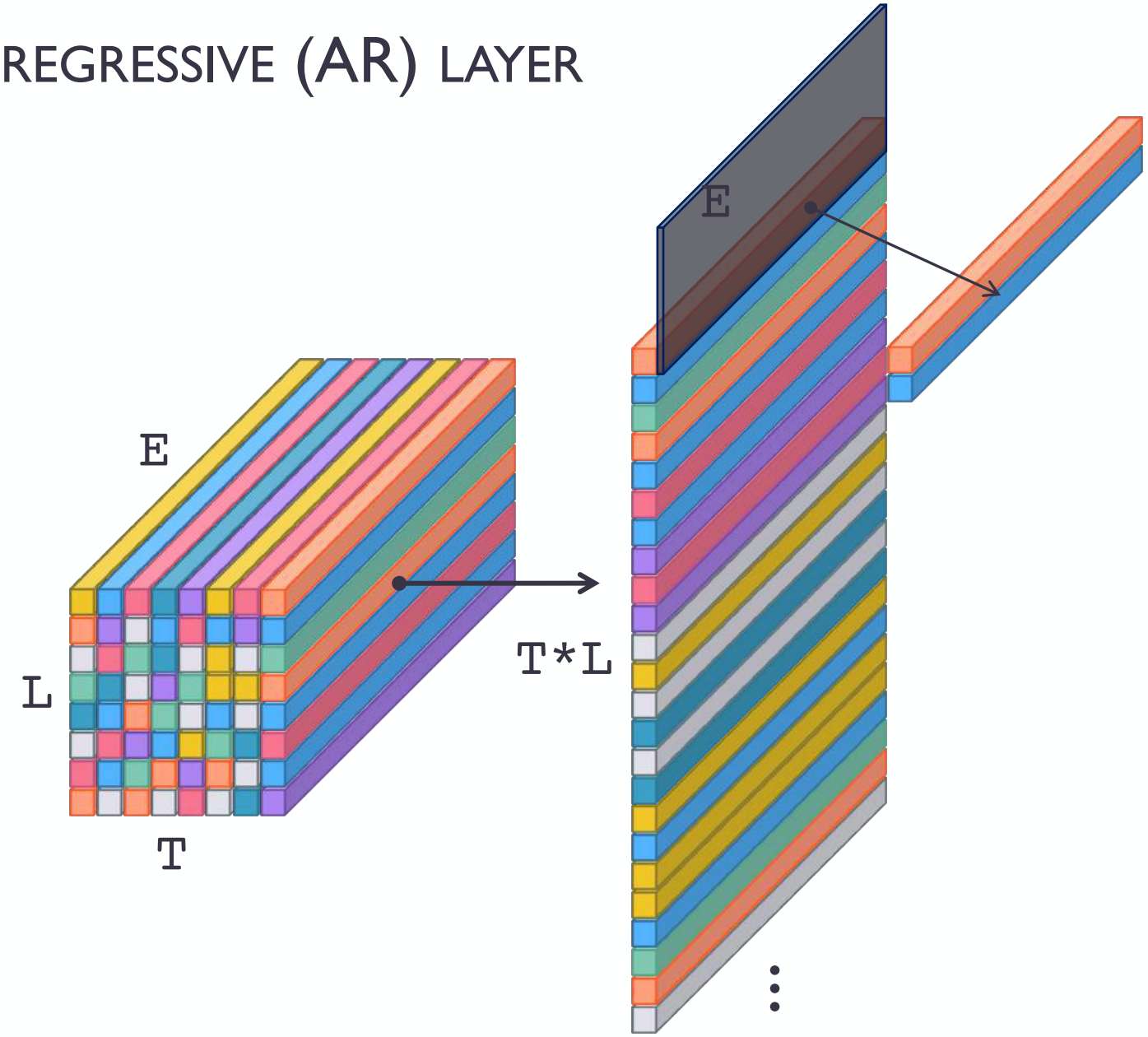
van den Oord et al. (2016)

AN AUTO-REGRESSIVE (AR) LAYER



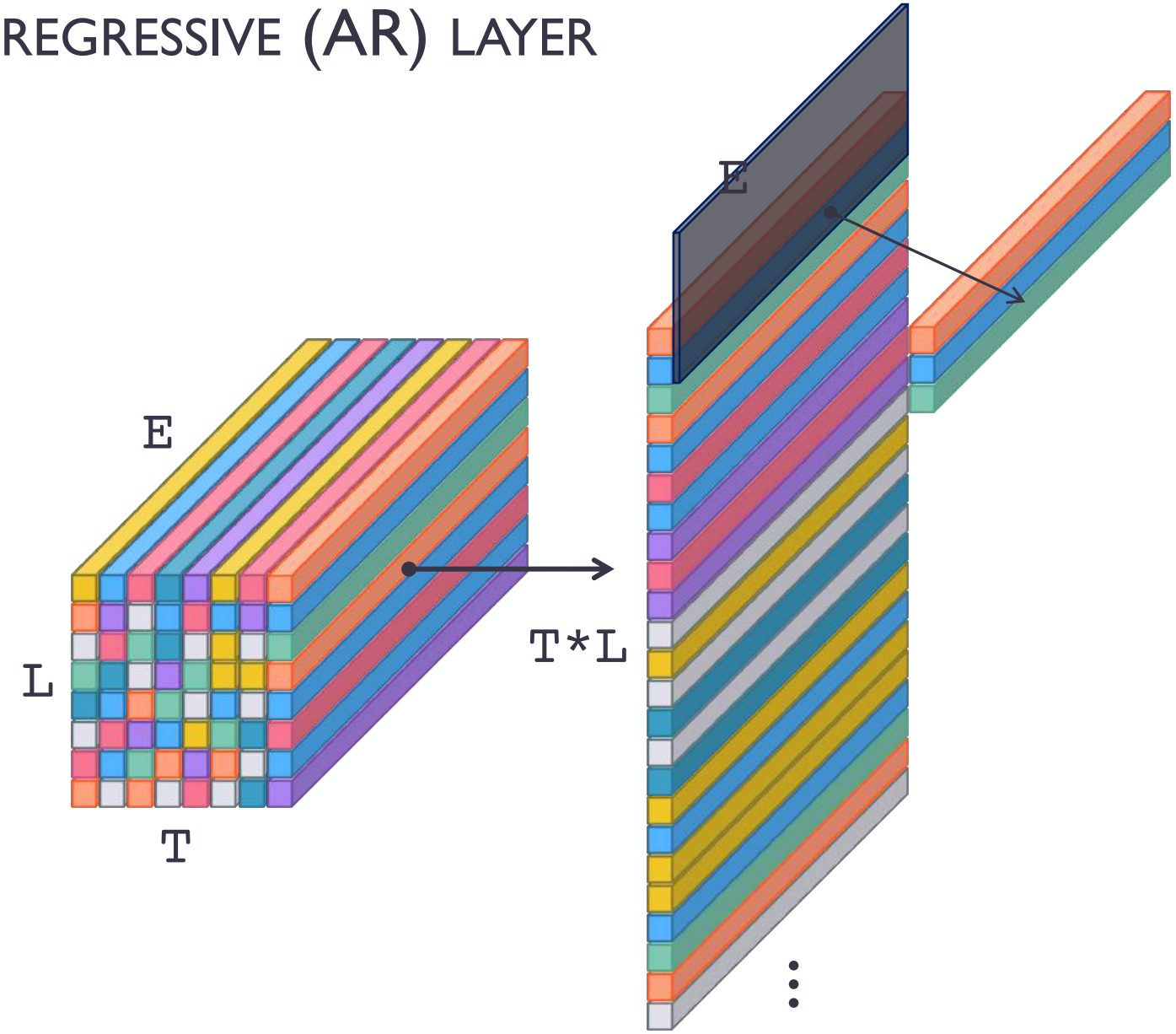
van den Oord et al. (2016)

AN AUTO-REGRESSIVE (AR) LAYER



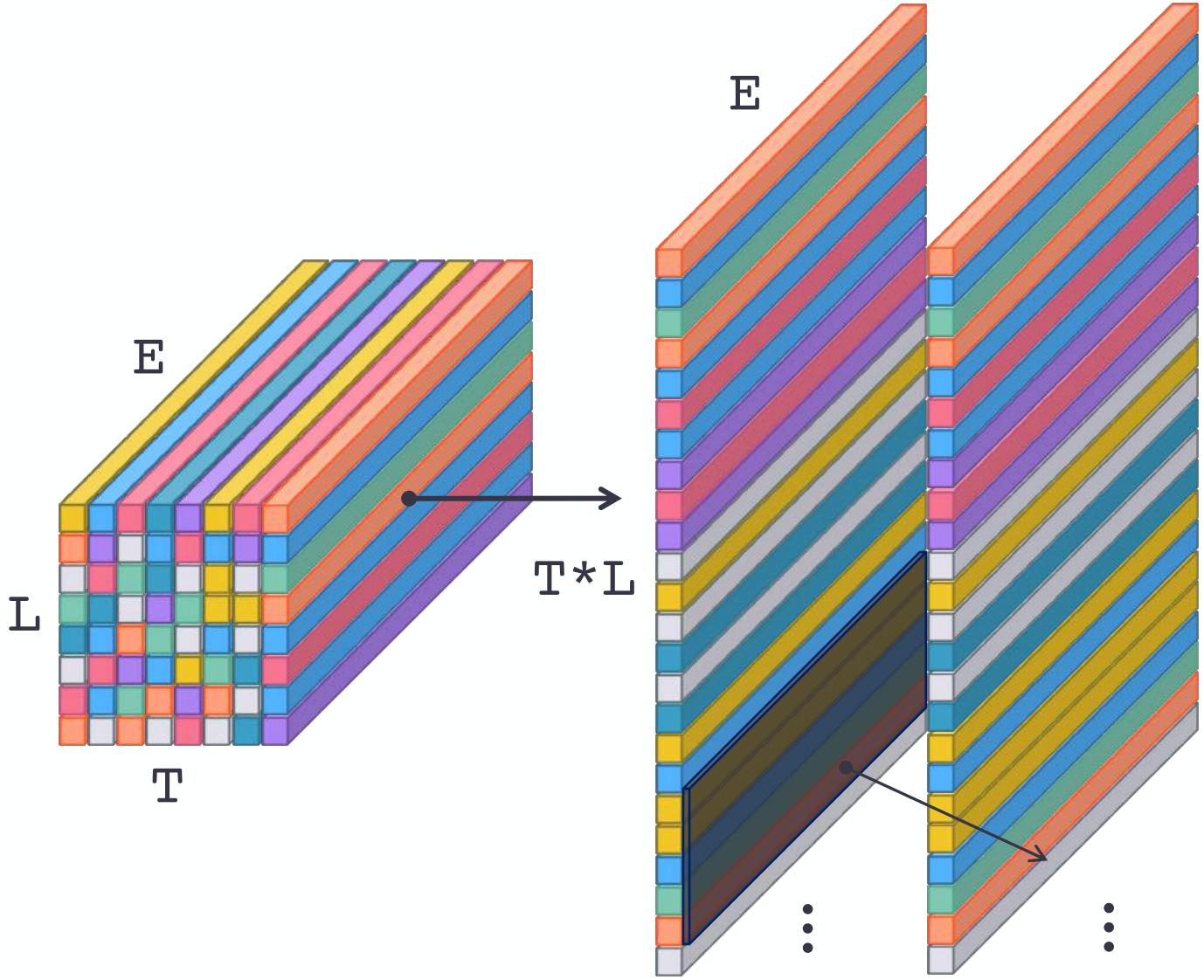
van den Oord et al. (2016)

AN AUTO-REGRESSIVE (AR) LAYER



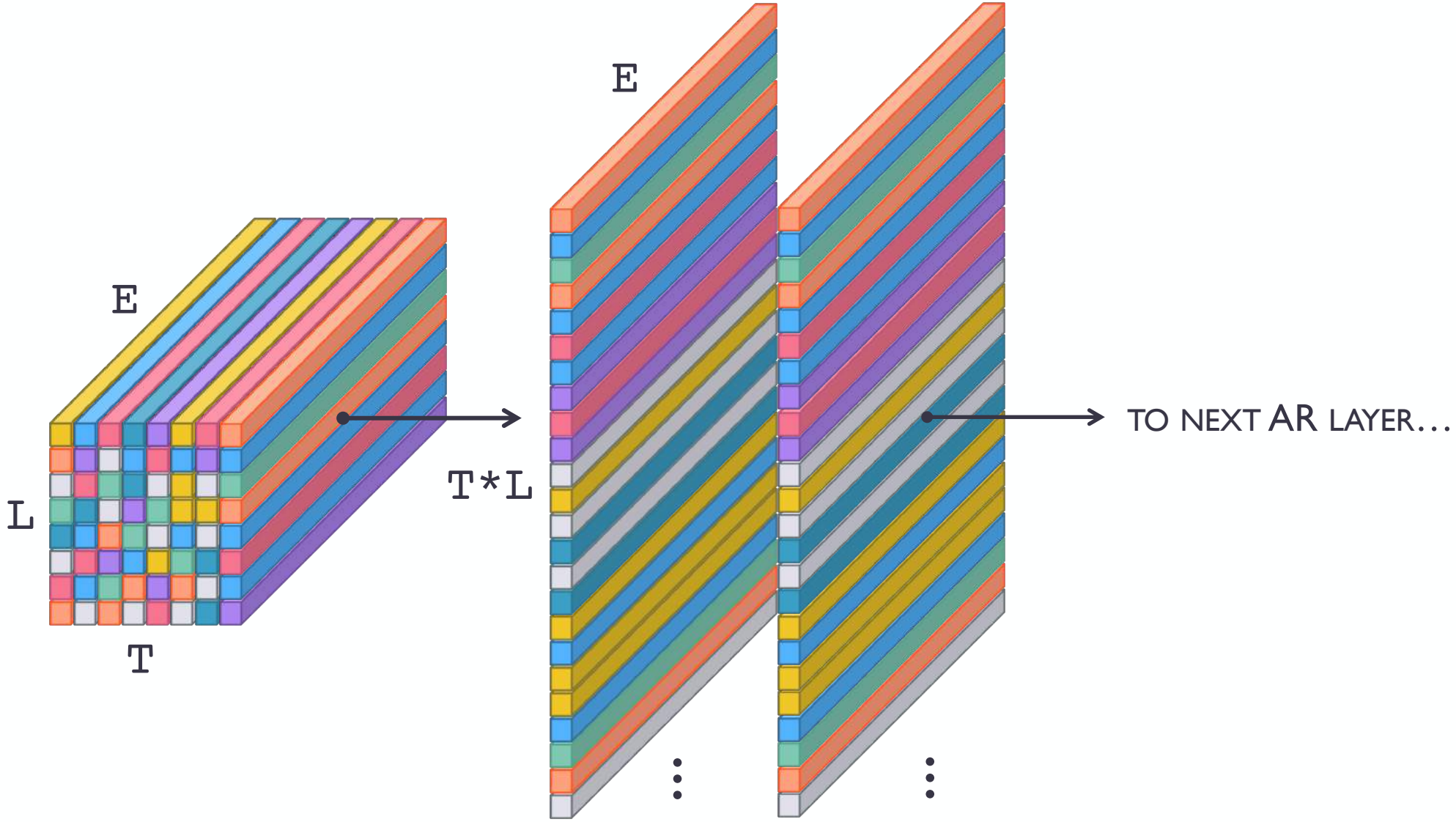
van den Oord et al. (2016)

AN AUTO-REGRESSIVE (AR) LAYER



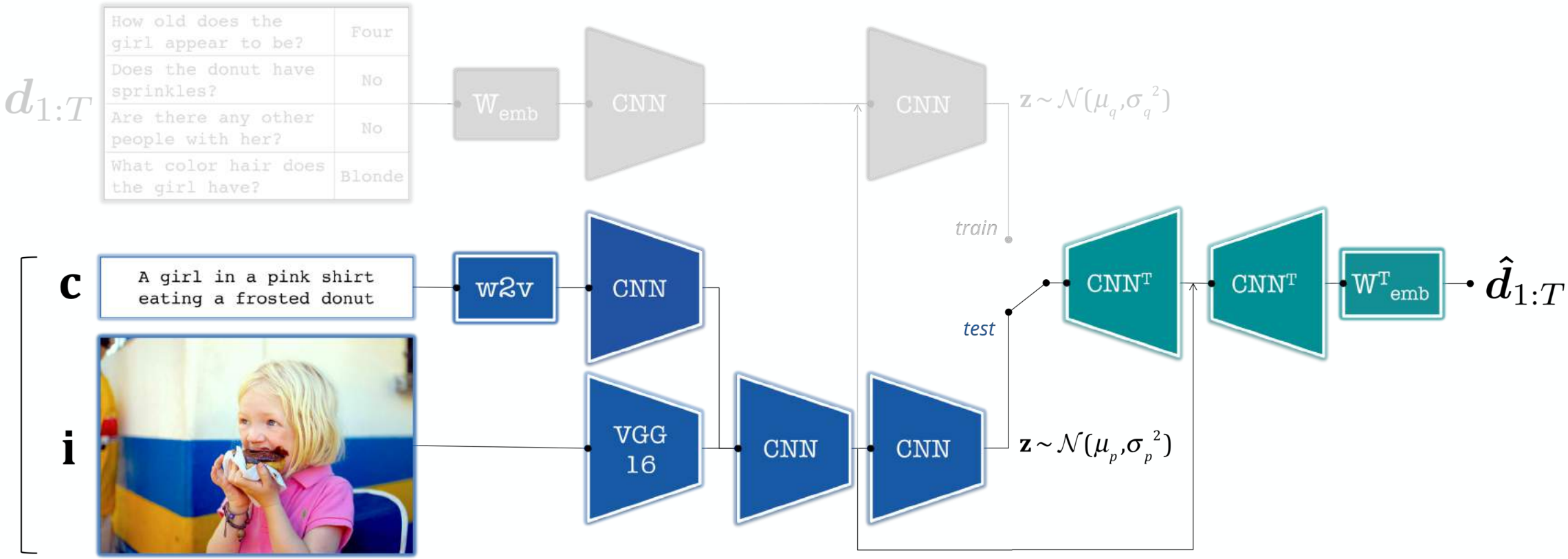
van den Oord et al. (2016)

AN AUTO-REGRESSIVE (AR) LAYER



van den Oord et al. (2016)

2VD EVALUATION



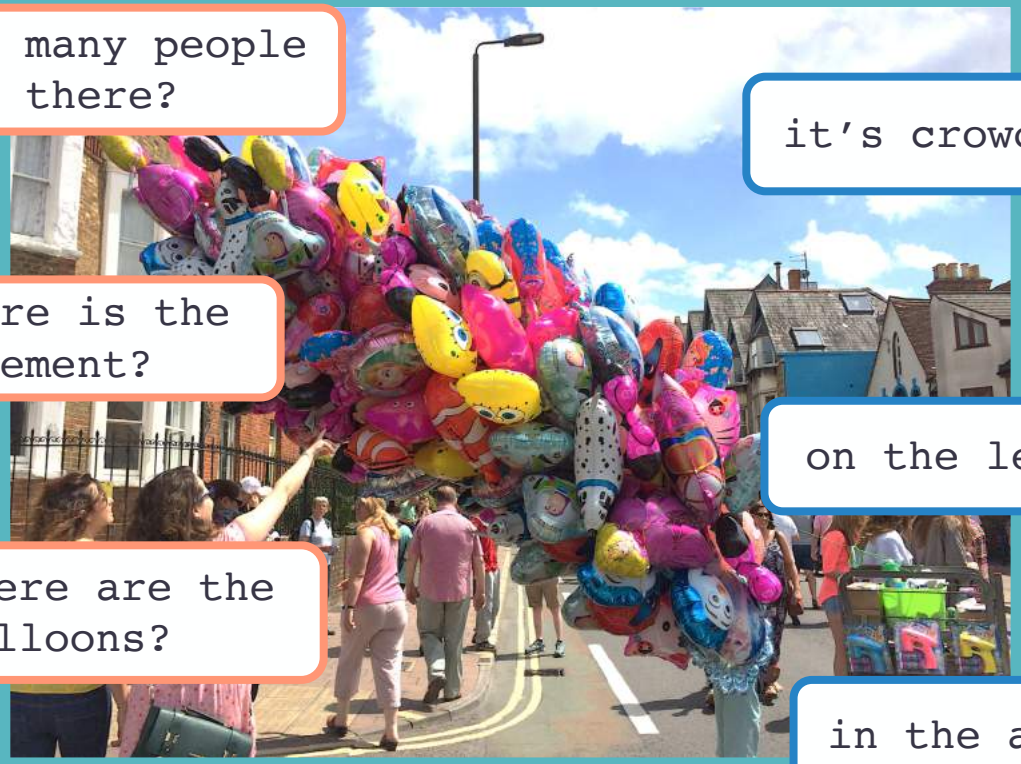
2VD EVALUATION (I)

	History?	NLL (⇩)
Ours	∅	31.18
	✓	25.40
$\text{Ours}_{\text{AR}}8$	∅	28.81
	✓	26.60
$\text{Ours}_{\text{AR}}10$	∅	28.49
	✓	**24.93**

∅ - generate whole dialogue block

✓ - iteratively condition on previously *generated* history

ONE-WAY VISUAL DIALOGUE (1VD)



How many people are there?

it's crowded

Where is the pavement?

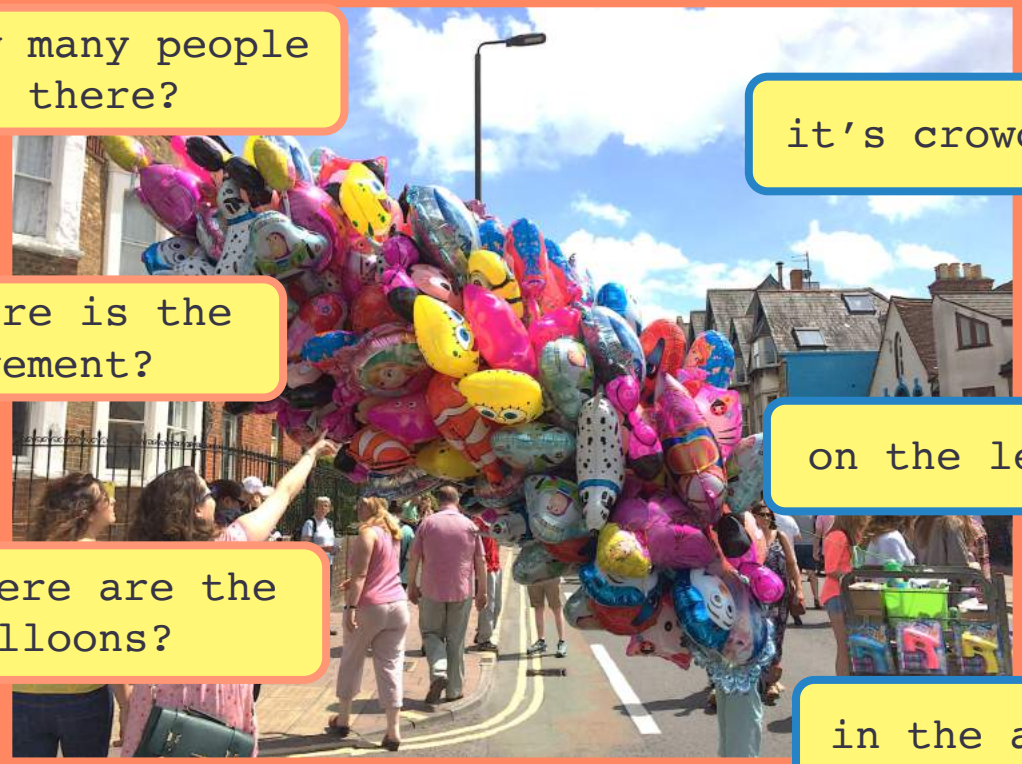
on the left

Where are the balloons?

in the air

A street with people and balloons

TWO-WAY VISUAL DIALOGUE (2VD)



How many people are there?

it's crowded

Where is the pavement?

on the left

Where are the balloons?

in the air

A street with people and balloons

LATENT SPACE DISPERSION ($\text{SIM}_{\circlearrowleft}$)

ground-truth dialogue $\boldsymbol{d}_{1:T}$

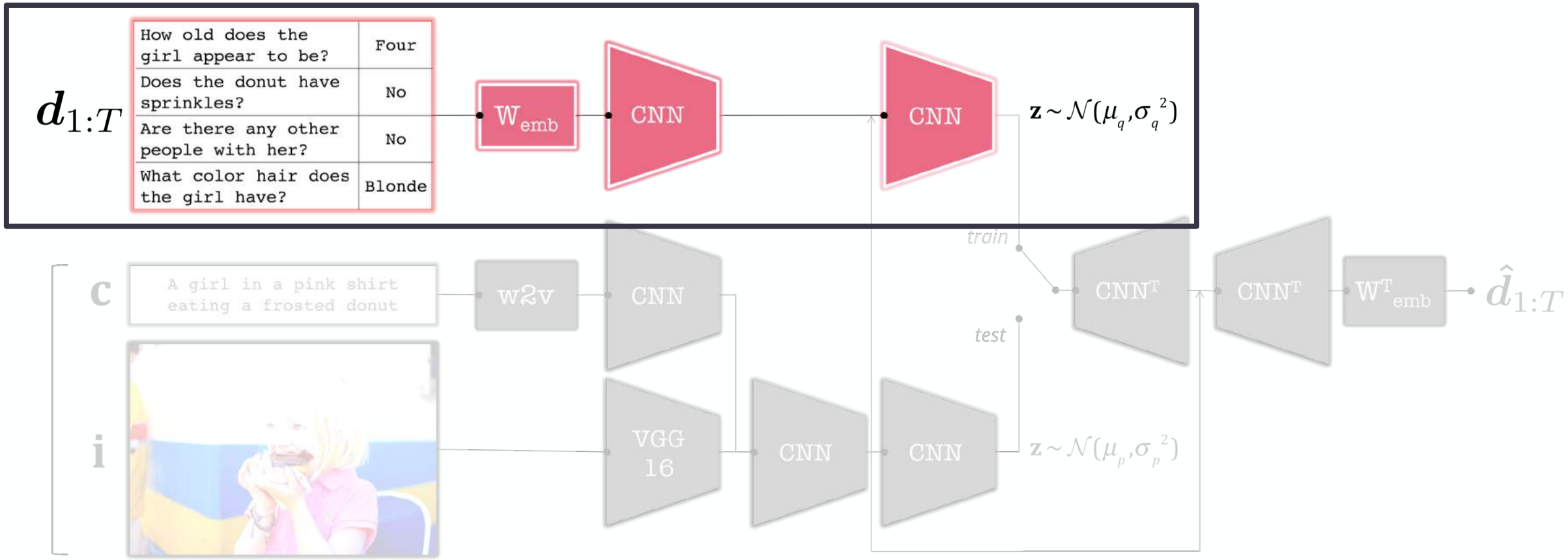
How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	Blonde

generated dialogue $\hat{\boldsymbol{d}}_{1:T}$

What is the girl wearing?	A pink shirt
Is she eating?	Yes
Is she young or old?	Young
What type of donut is it?	Chocolate

$\boldsymbol{z}$

LATENT SPACE DISPERSION (SIM↺)

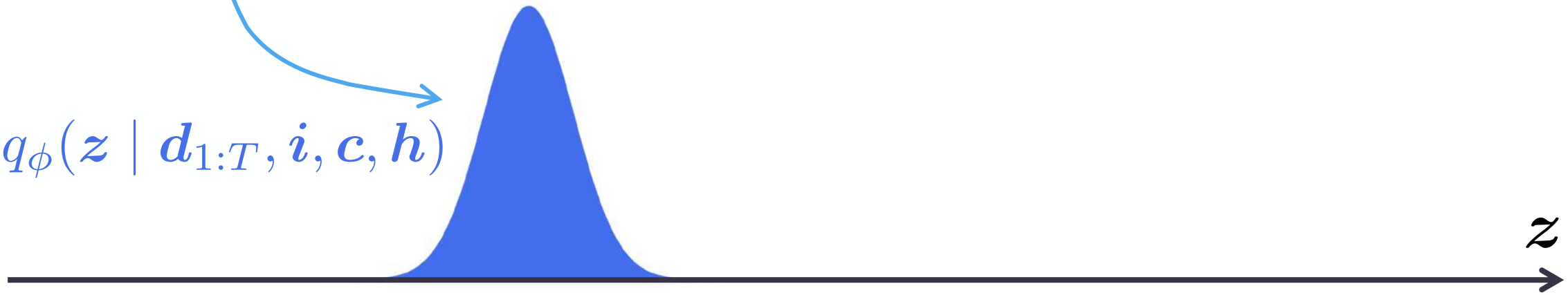


LATENT SPACE DISPERSION (SIM$_{\circlearrowleft}$)

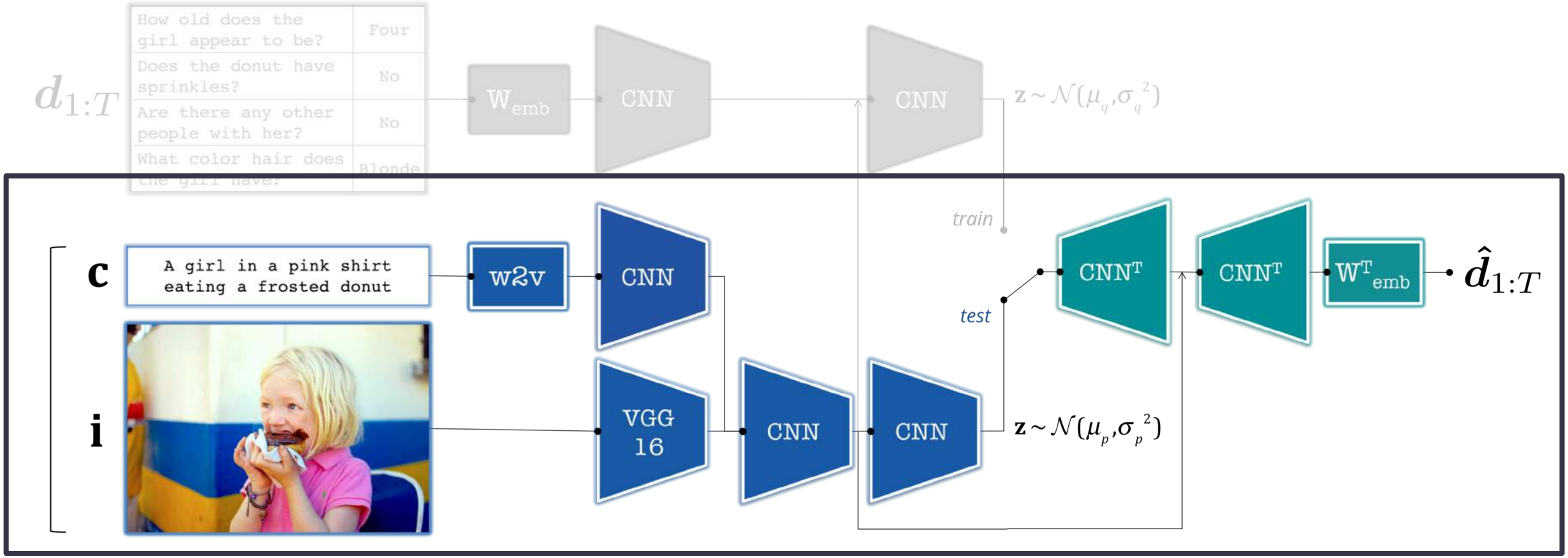
ground-truth dialogue $\boldsymbol{d}_{1:T}$

How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	Blonde

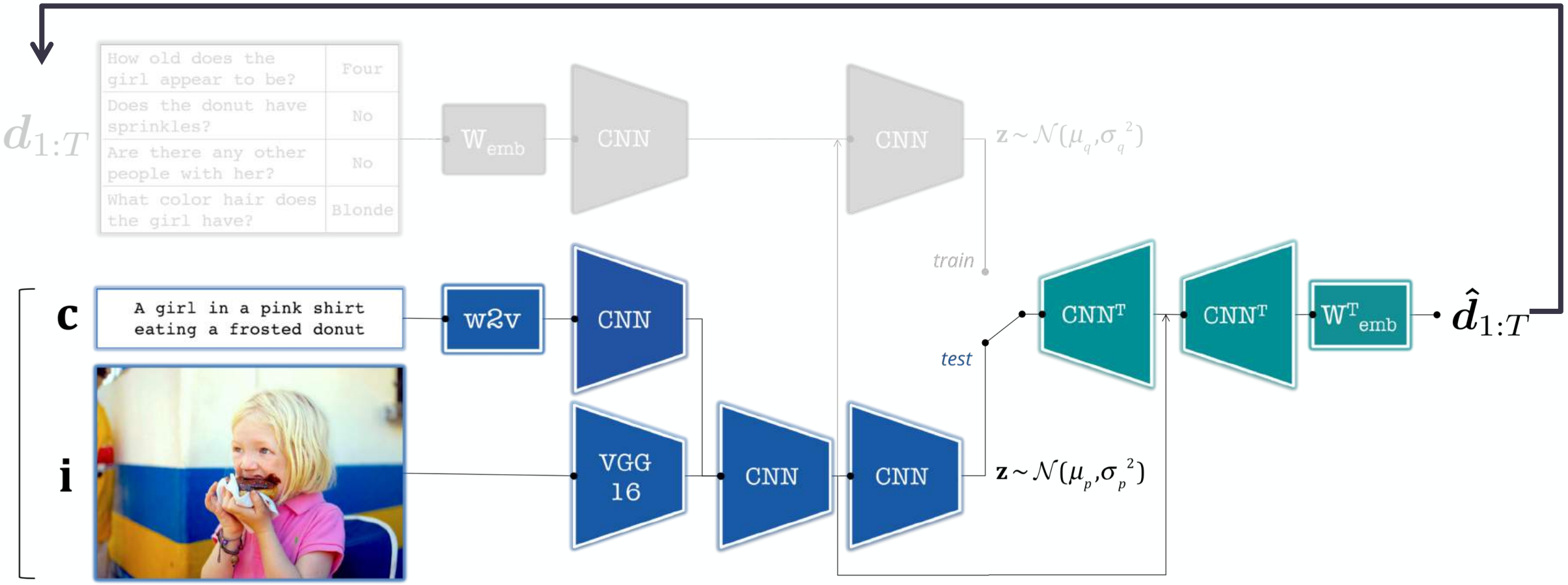
$q_\phi(\boldsymbol{z} \mid \boldsymbol{d}_{1:T}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})$



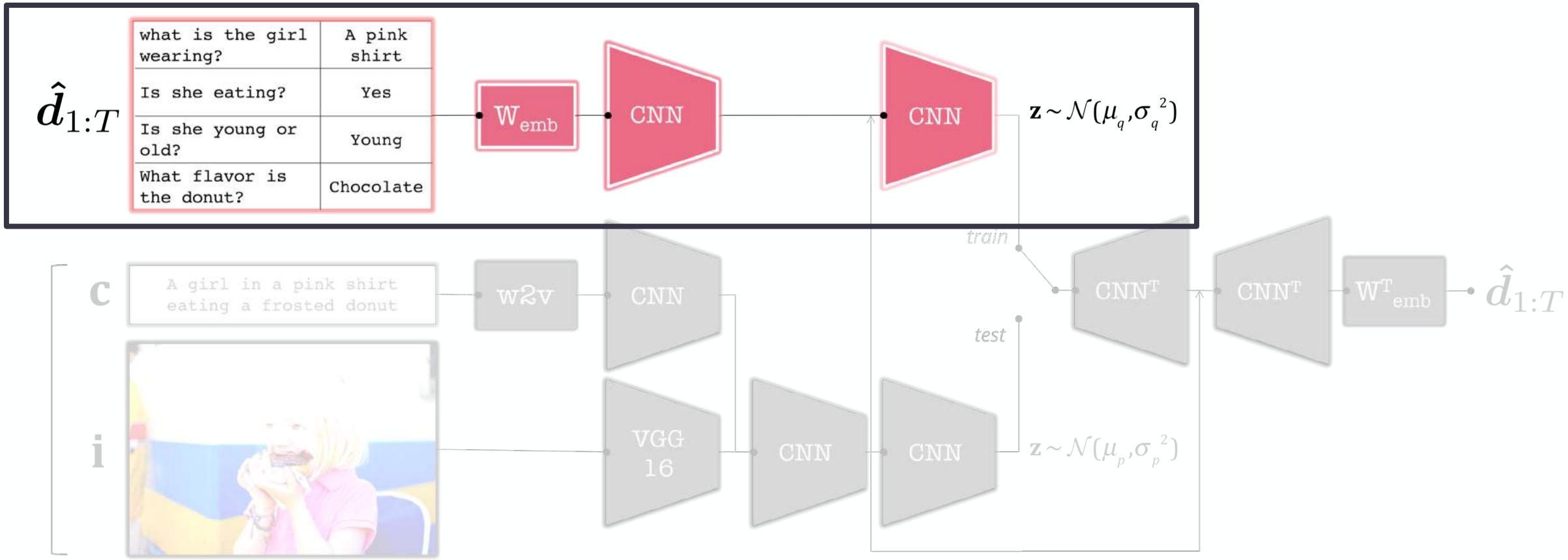
LATENT SPACE DISPERSION (SIM$_{\circlearrowleft}$)



LATENT SPACE DISPERSION ($SIM_{\circlearrowleft}$)



LATENT SPACE DISPERSION ($SIM_{\circlearrowleft}$)



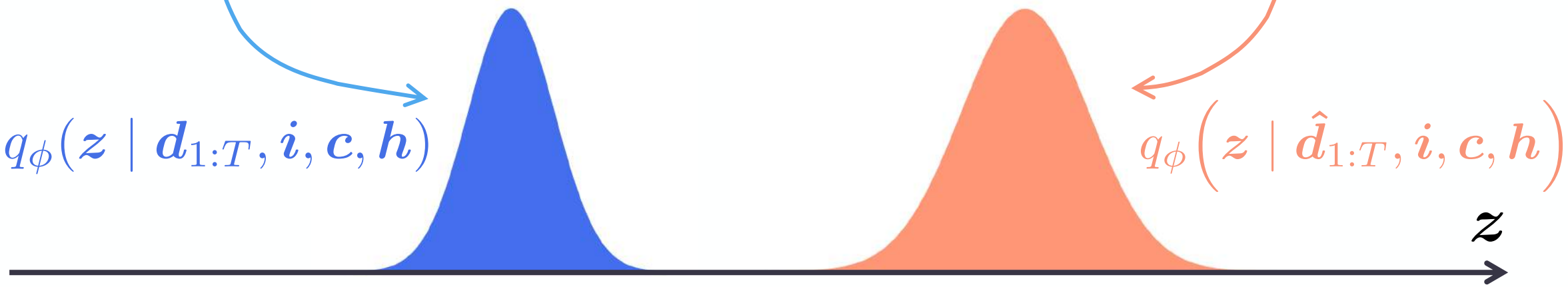
LATENT SPACE DISPERSION ($\text{SIM}_{\circlearrowleft}$)

ground-truth dialogue $\boldsymbol{d}_{1:T}$

How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	Blonde

generated dialogue $\hat{\boldsymbol{d}}_{1:T}$

What is the girl wearing?	A pink shirt
Is she eating?	Yes
Is she young or old?	Young
What type of donut is it?	Chocolate



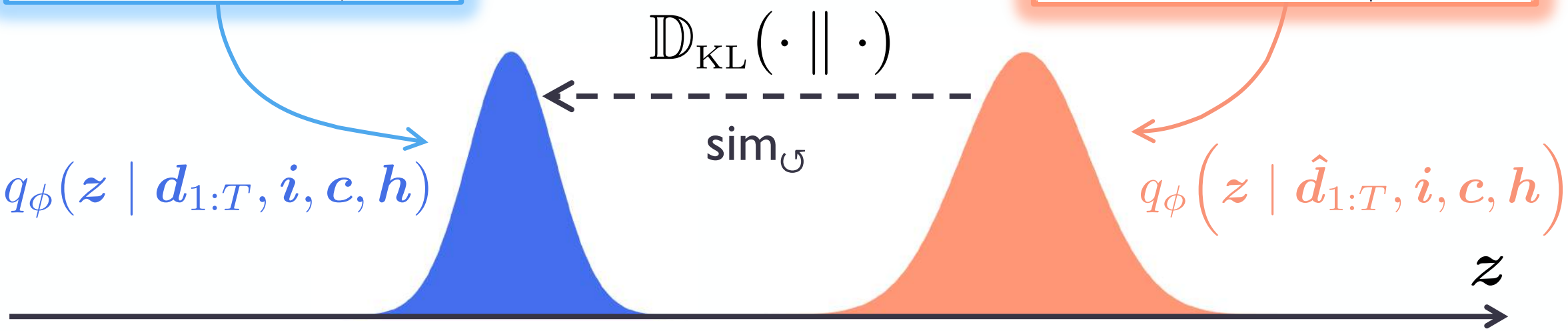
LATENT SPACE DISPERSION ($\text{SIM}_{\circlearrowleft}$)

ground-truth dialogue $\boldsymbol{d}_{1:T}$

How old does the girl appear to be?	Four
Does the donut have sprinkles?	No
Are there any other people with her?	No
What color hair does the girl have?	Blonde

generated dialogue $\hat{\boldsymbol{d}}_{1:T}$

What is the girl wearing?	A pink shirt
Is she eating?	Yes
Is she young or old?	Young
What type of donut is it?	Chocolate



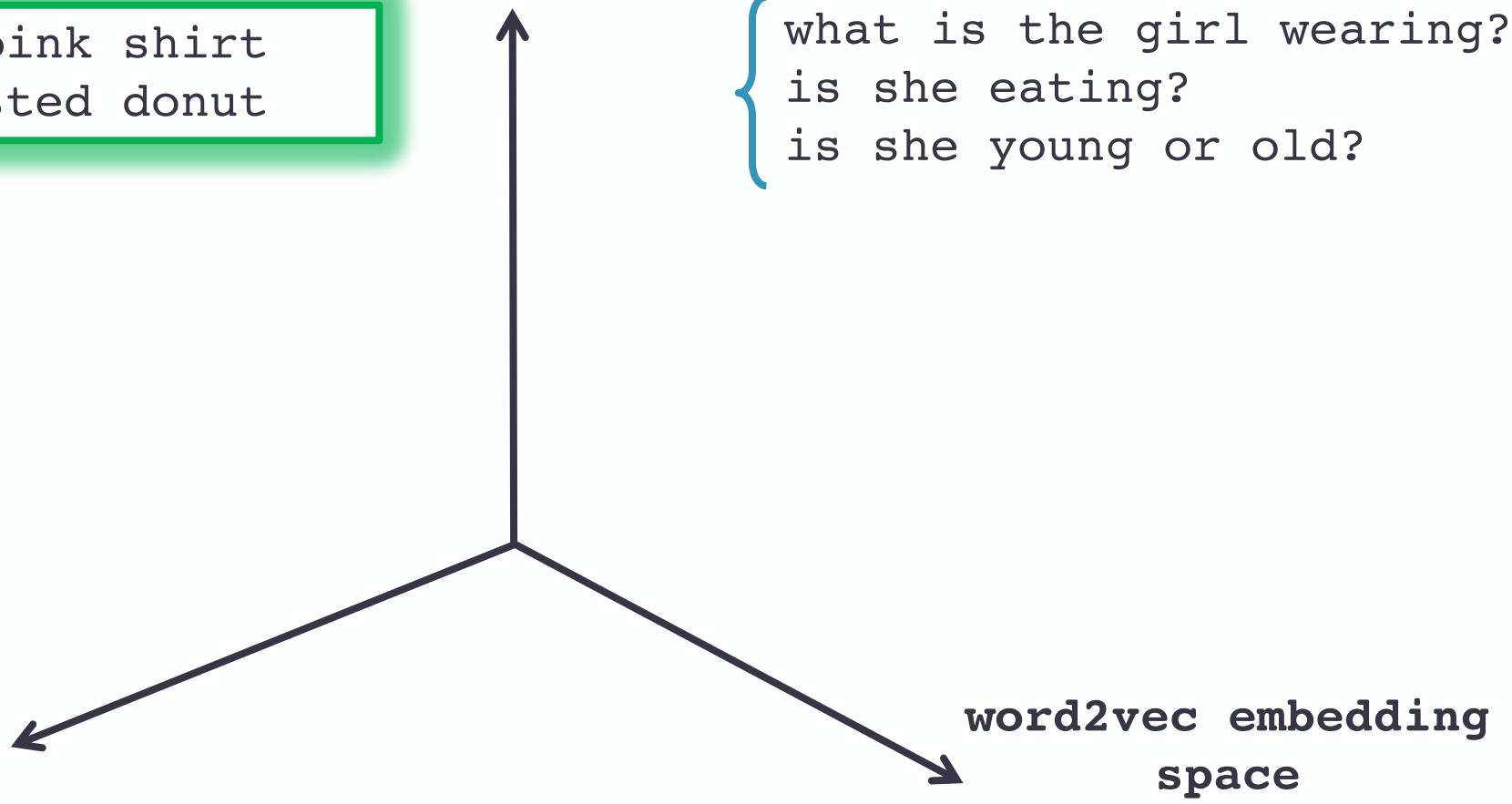
QUESTION RELEVANCY (SIM_{CQ})

caption $\boldsymbol{c}$

A girl in a pink shirt
eating a frosted donut

generated questions from $\hat{\boldsymbol{d}}_{1:T}$

what is the girl wearing?
is she eating?
is she young or old?



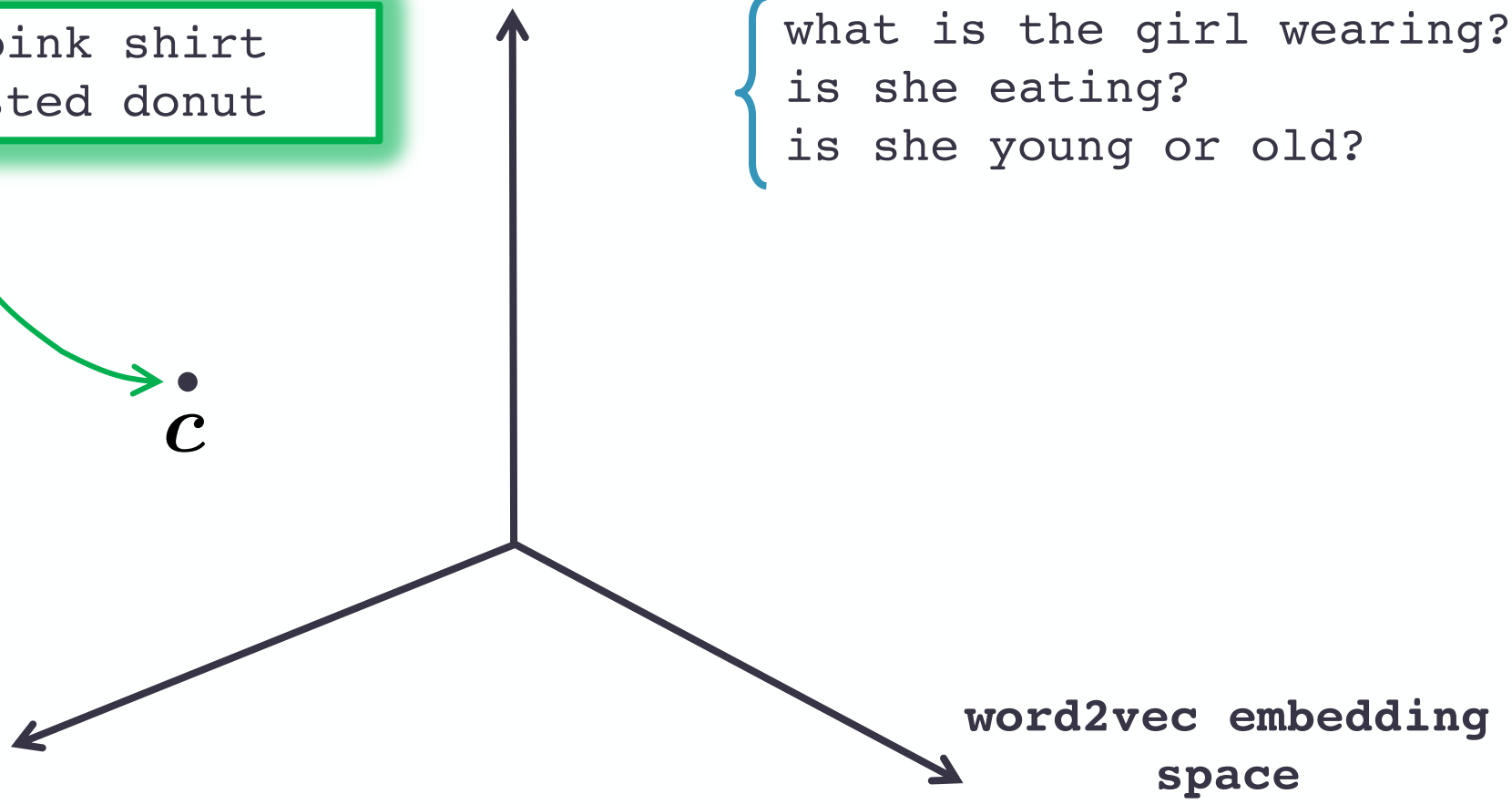
QUESTION RELEVANCY (SIM_{CQ})

caption $\boldsymbol{c}$

A girl in a pink shirt
eating a frosted donut

generated questions from $\hat{\boldsymbol{d}}_{1:T}$

what is the girl wearing?
is she eating?
is she young or old?



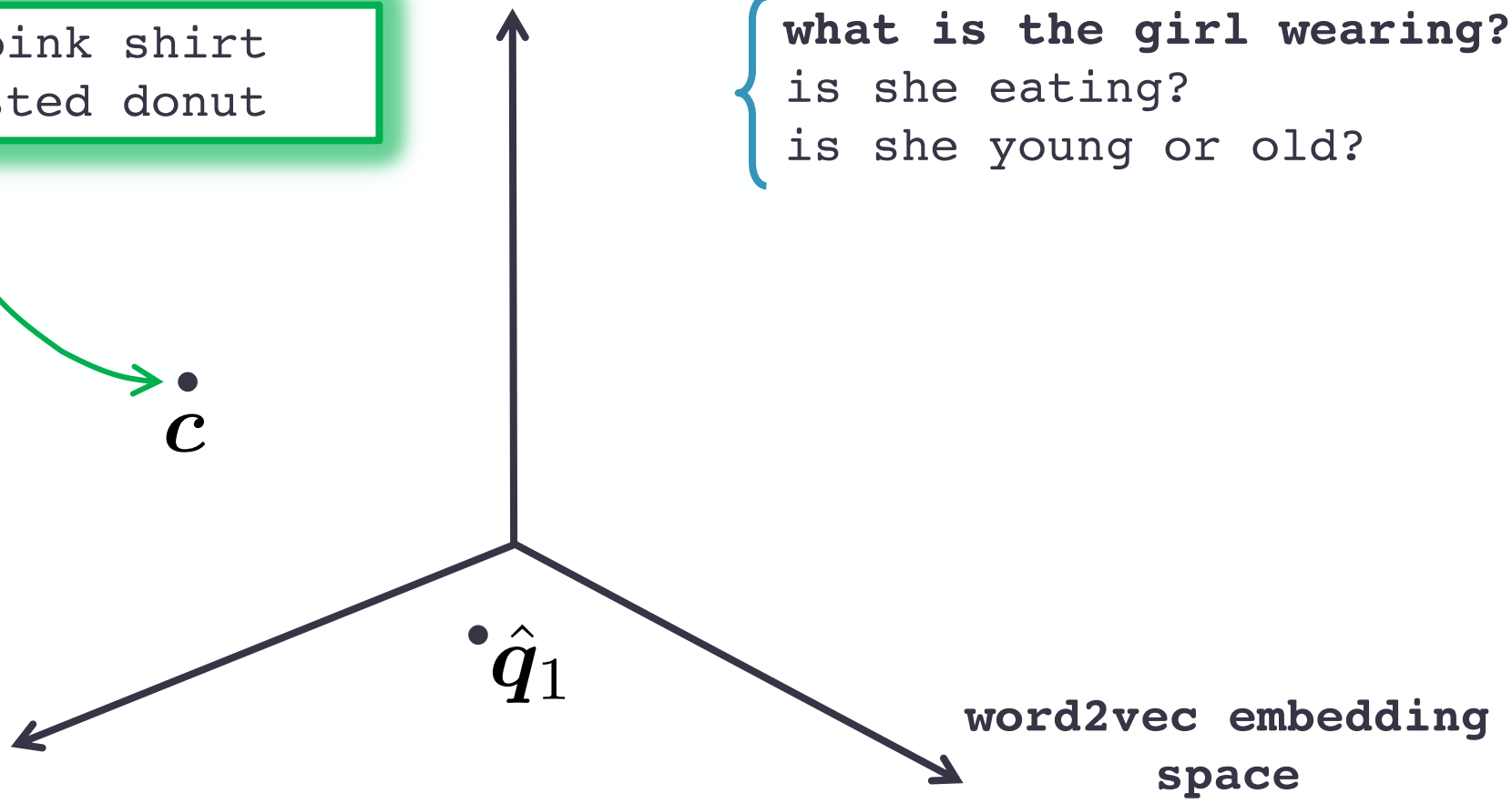
QUESTION RELEVANCY (SIM_{CQ})

caption $\boldsymbol{c}$

A girl in a pink shirt
eating a frosted donut

generated questions from $\hat{\boldsymbol{d}}_{1:T}$

what is the girl wearing?
is she eating?
is she young or old?



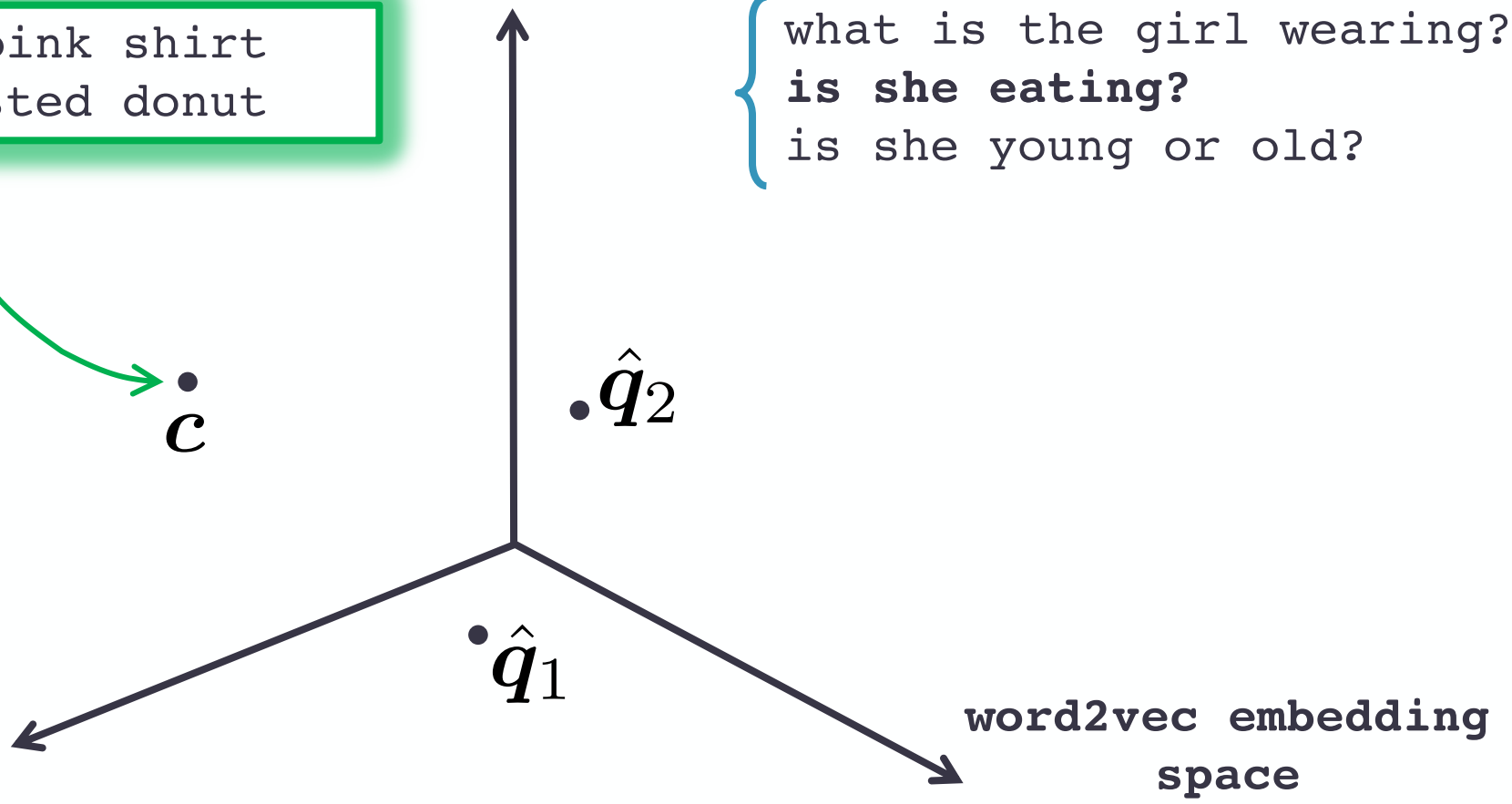
QUESTION RELEVANCY (SIM_{CQ})

caption $\boldsymbol{c}$

A girl in a pink shirt
eating a frosted donut

generated questions from $\hat{\boldsymbol{d}}_{1:T}$

what is the girl wearing?
is she eating?
is she young or old?



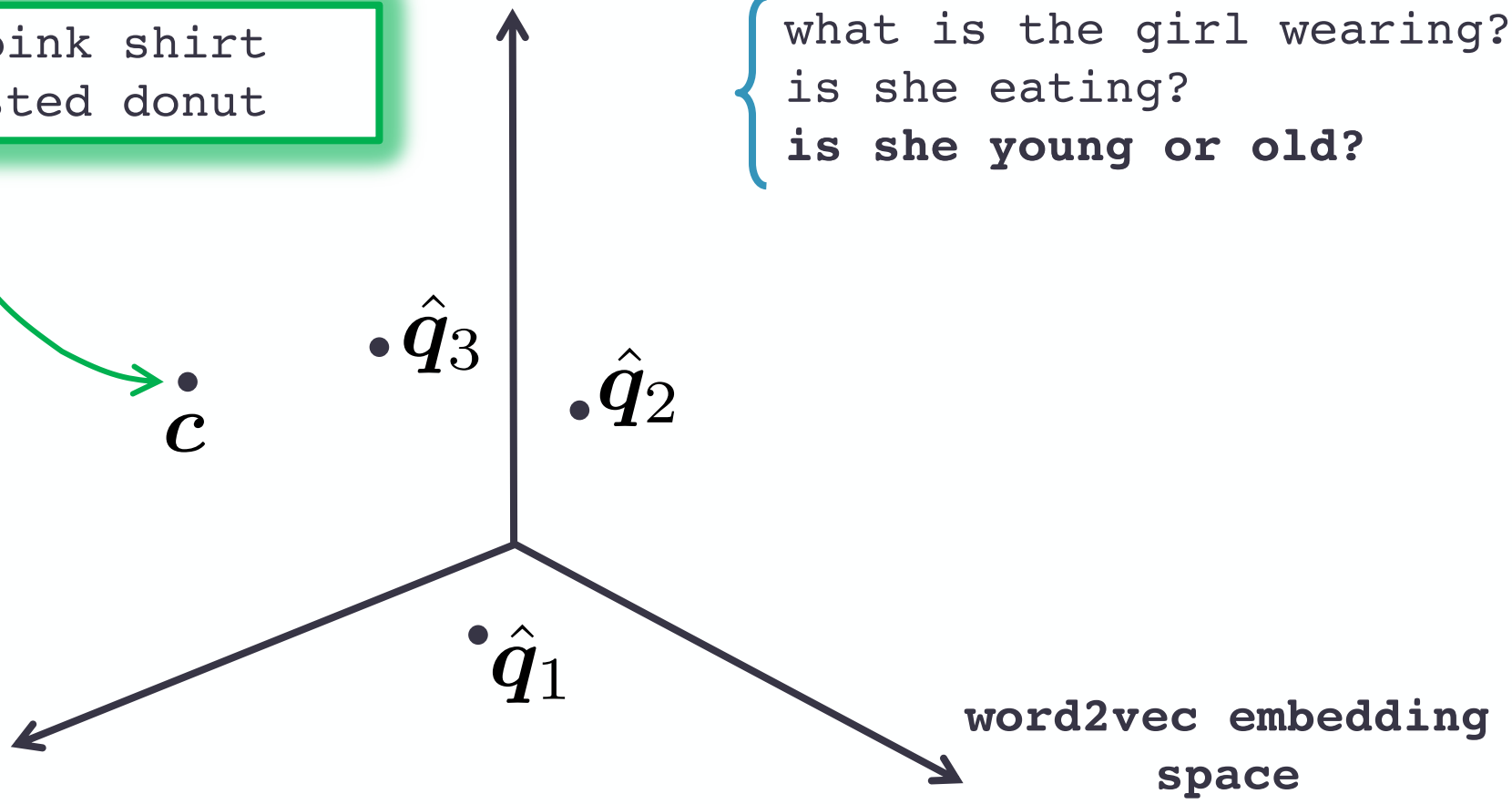
QUESTION RELEVANCY (SIM_{CQ})

caption $\boldsymbol{c}$

A girl in a pink shirt
eating a frosted donut

generated questions from $\hat{\boldsymbol{d}}_{1:T}$

what is the girl wearing?
is she eating?
is she young or old?



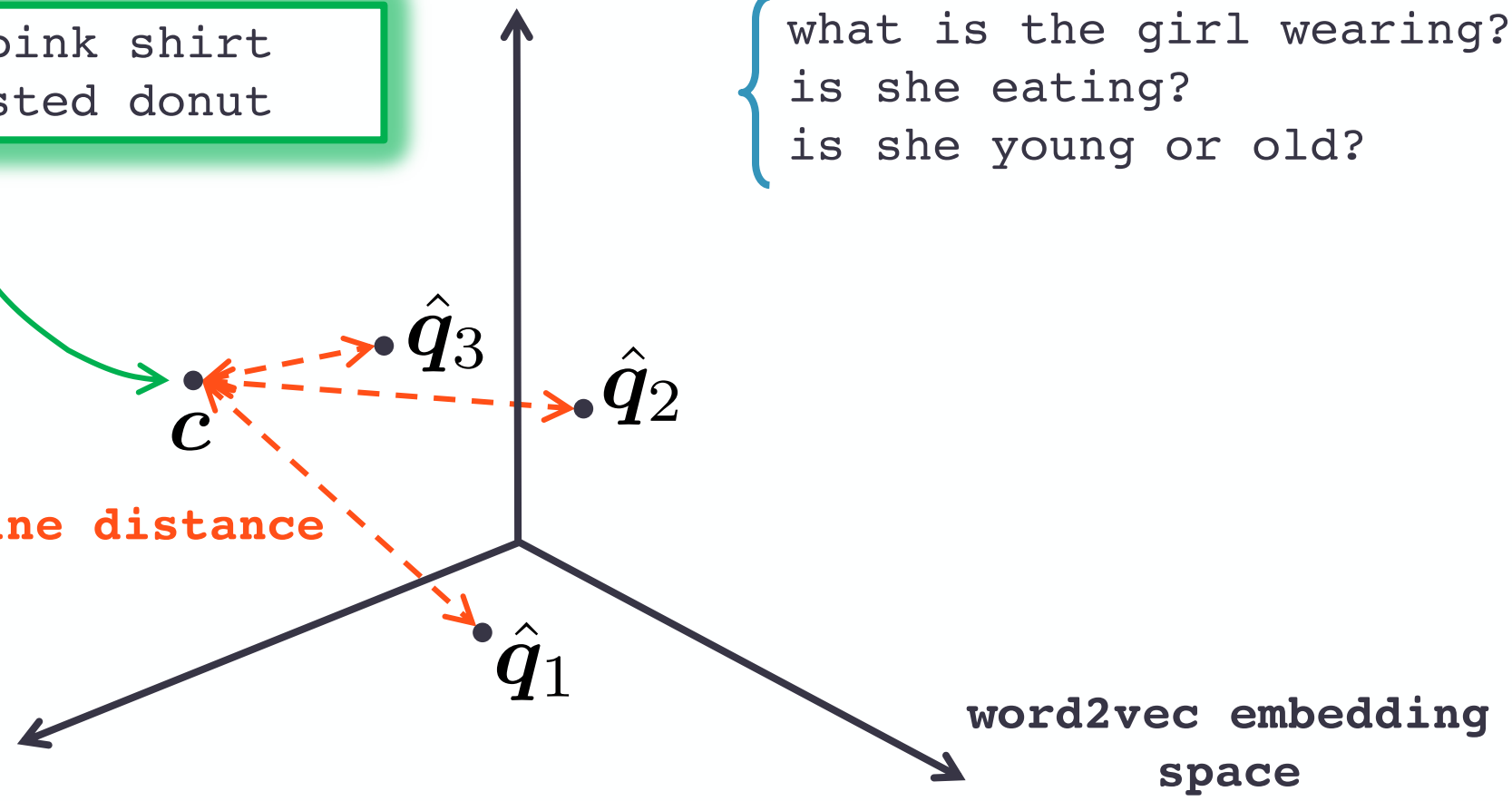
QUESTION RELEVANCY (SIM_{CQ})

caption $\boldsymbol{c}$

A girl in a pink shirt
eating a frosted donut

generated questions from $\hat{\boldsymbol{d}}_{1:T}$

what is the girl wearing?
is she eating?
is she young or old?



2VD EVALUATION (II)

	History?	NLL (⇩)	$sim_{\circlearrowleft}$ (⇩)	sim_{cq} (⇧)
Ours	∅	31.18	14.20	**0.4931**
	✓	25.40	**1.86**	0.4091
$Ours_{AR}8$	∅	28.81	31.50	0.4878
	✓	26.60	2.39	0.3884
$Ours_{AR}10$	∅	28.49	44.34	0.4927
	✓	**24.93**	2.35	0.4101

∅ - generate whole dialogue block

✓ - iteratively condition on previously *generated* history

2VD EVALUATION (III)

Man and a boy playing ball in the grass

Is the picture in color?	**Yes**
Is the photo in?	Yes
What color is the frisbee?	**White**
Can there?	No
Is the grass visible?	**Yes**
Is the sunny?	**Yes**
Is it sunny?	**Yes**
Is there other people?	**No**
What color is the frisbee?	**White**
Is the man green?	Yes

What color is the man's?	White
What is the man man?	I
How old is the man?	**I**
Can you see any ball?	**Yes**
Is there any other people?	**No**
Does the man have a?	No
Is the man wearing a hat?	**No**
Does he man have a?	No
Are there see any other people?	**No**
Can you see any sky?	**No**

CONCLUSION & FUTURE

1VD & 2VD!

✓ Generative visual dialogue

✓ Diverse questions & answers

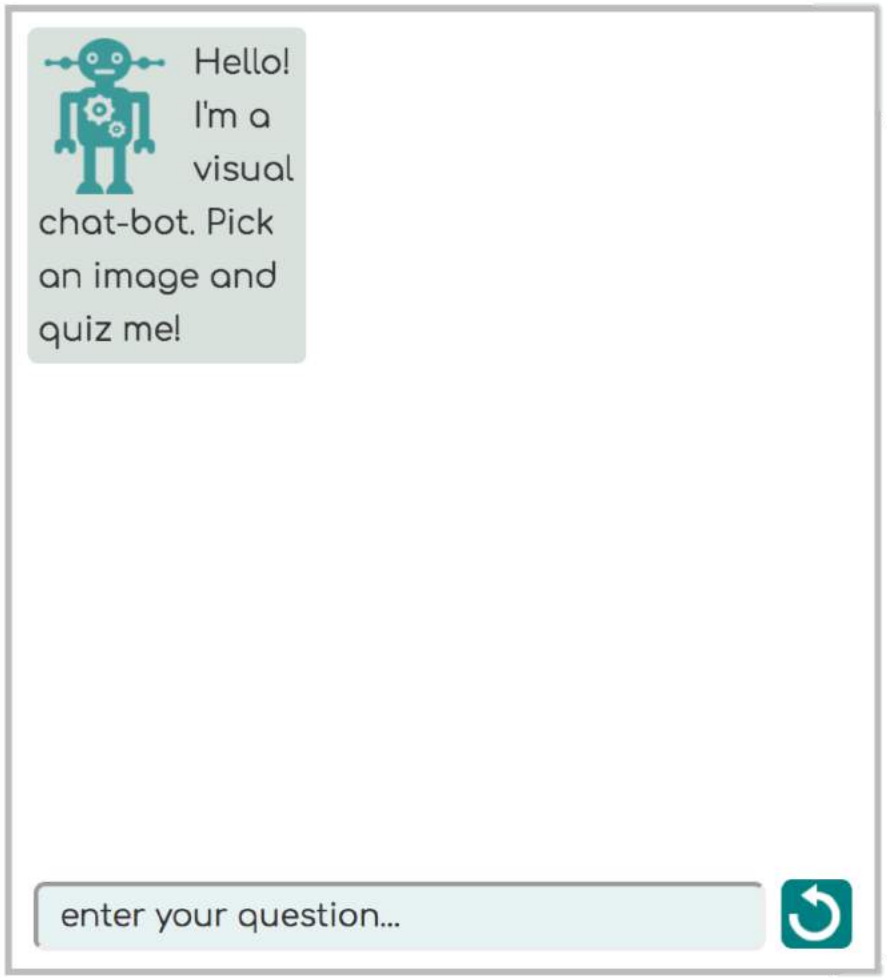
✓ Novel evaluation metrics

+ Increasing role of visual cues

+ Quantifying diversity

+ Asking informative questions

FLIPDIAL 1VD WEB-DEMO



www.robots.ox.ac.uk/~daniela/research/flipdial/1vd_demo

FLIPDIAL: FIND US AT POSTER #C18

✓Generative visual dialogue

✓Diverse questions & answers

✓Novel evaluation metrics

+ Increasing role of visual cues

+ Quantifying diversity

+ Asking informative questions

Daniela Massiceti

N. Siddharth

Puneet Dokania

Philip H.S. Torr