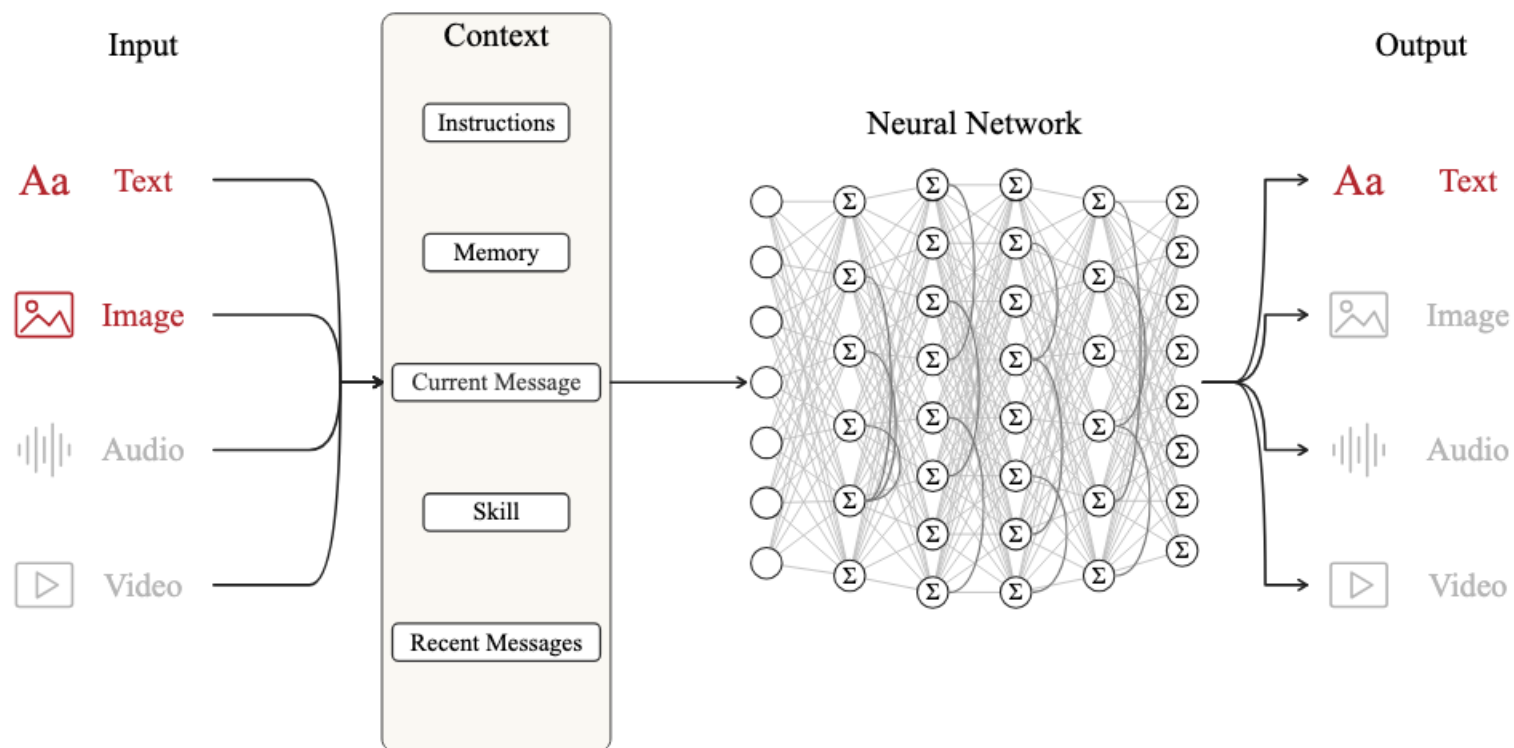


L-5



Example 1

Aa +  → Aa

Answer a question about a picture

Example 2

 +  → 

Answer a spoken question about a picture

Example 3

 + Aa → Aa

Summarize a clip

Example 4

 → Aa

Transcribe speech

Example 5

Aa → 

Generate a picture

Example 6

Aa → 

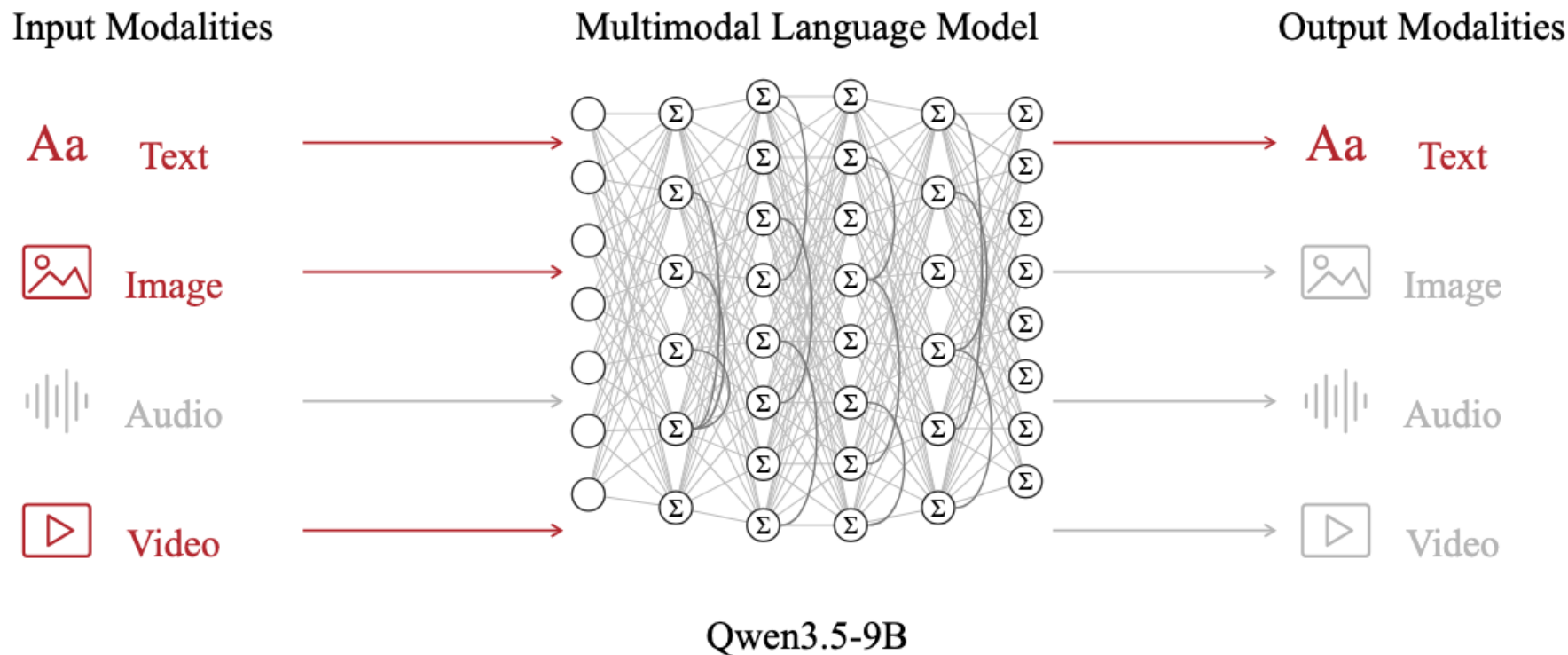
Generate speech

Beyond Text: Learning Across Modalities



DEMO 1 · L4 Exit Ticket Results

Models Support Different Modalities



Images and video can inform text answers; video input does not include native speech understanding

Qwen3.5-9B

Parakeet Unified
English 0.6B

DeepSeek V4 Pro
1.6T · 49B active

MagpieTTS v2607
364M

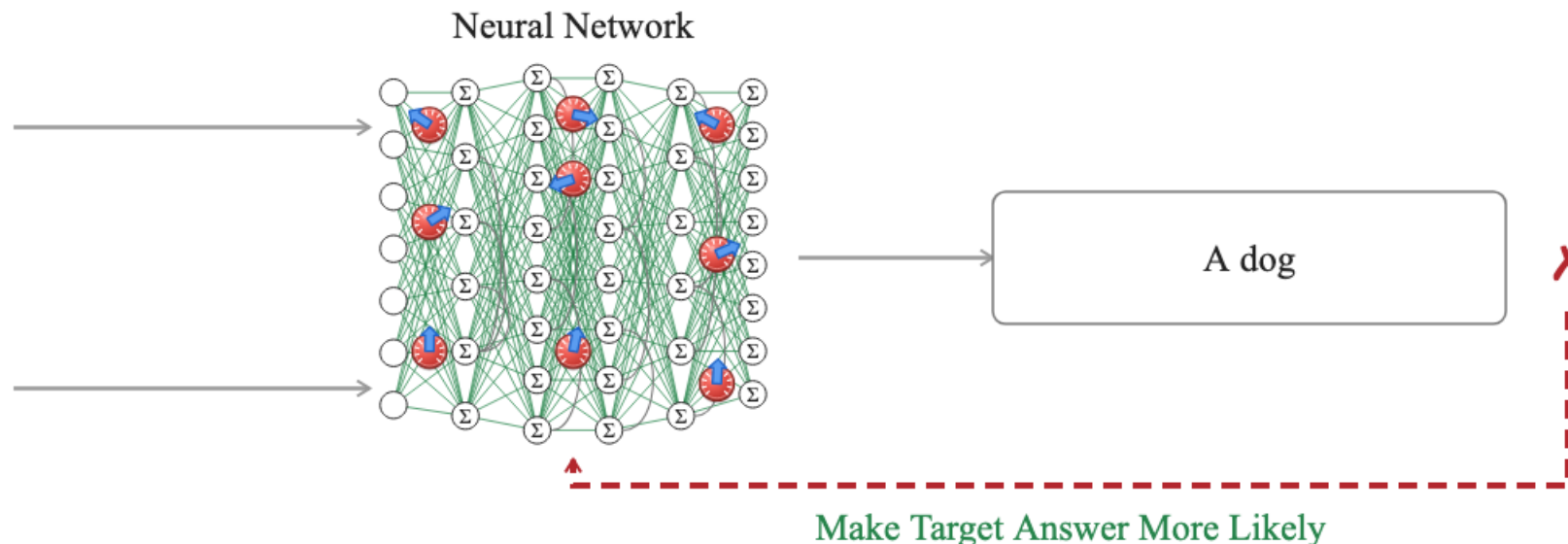
FLUX.2 klein 4B

FLUX 3 Video
Size undisclosed

Training Connects Modalities



What animal is this?



Training Data



Q: What animal?
A: A cat



Q: What animals?
A: Dogs



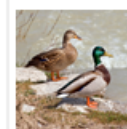
Q: What animals?
A: Horses



Q: Describe the ears
A: Long ears



Q: What colours?
A: Brown and white



Q: What birds?
A: Ducks



Q: How many wheels?
A: Two wheels



Q: What colour?
A: Silver



Q: What fruit?
A: Bananas



Q: What fruit?
A: Apples



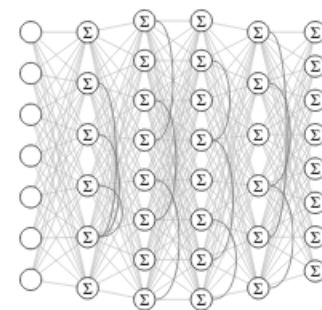
Q: Describe the neck
A: A long neck

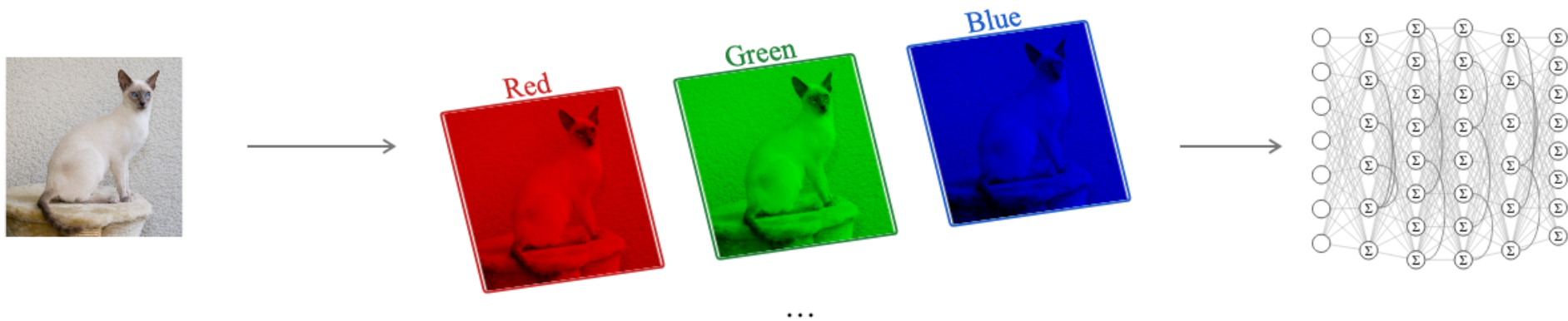


Q: What animal?
A: An elephant



?

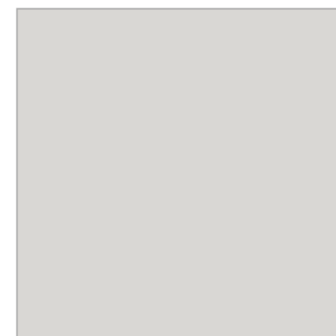




Default image · Move to explore



Move over the image, then inspect a pixel



Red 217 Green 215 Blue 212

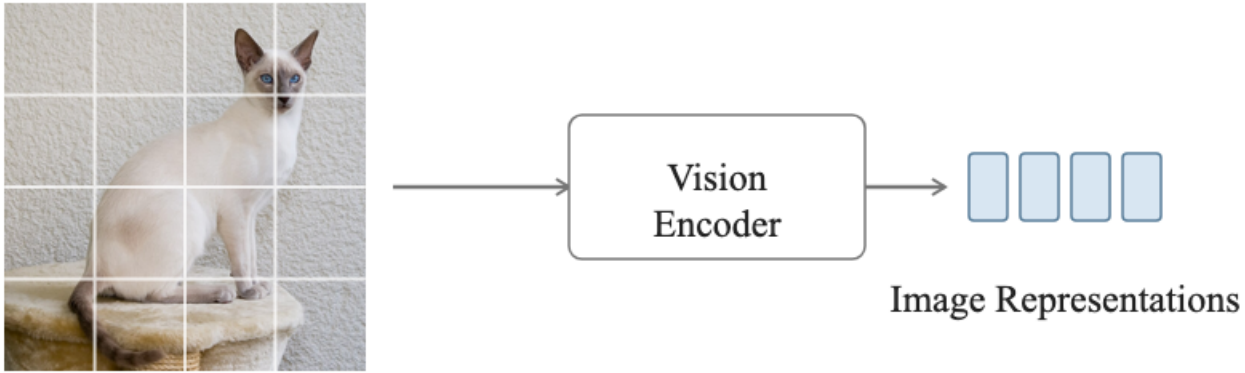
Each value ranges from 0 to 255

From Pixels to Context



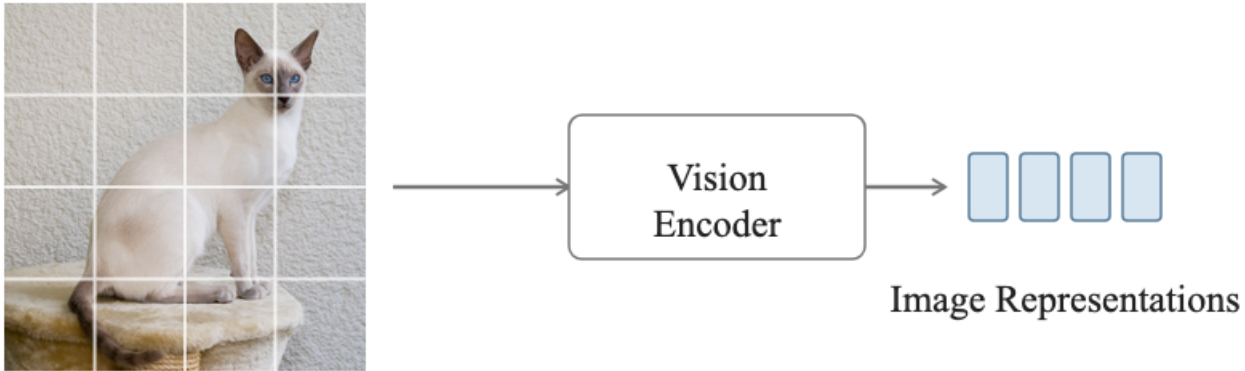
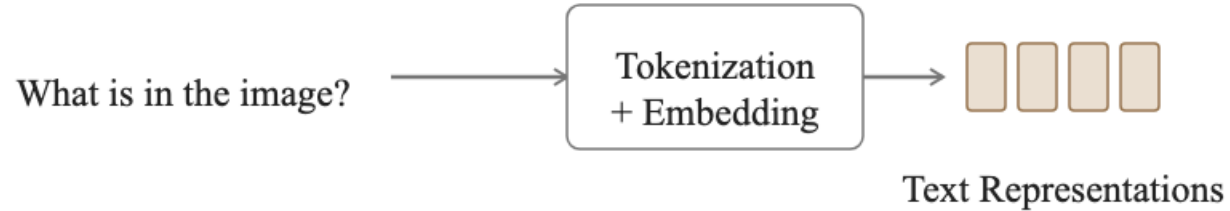
How do colour values become useful to the model?

From Pixels to Context



The vision encoder produces learned image representations

From Pixels to Context



Text has its own familiar preparation step

From Pixels to Context

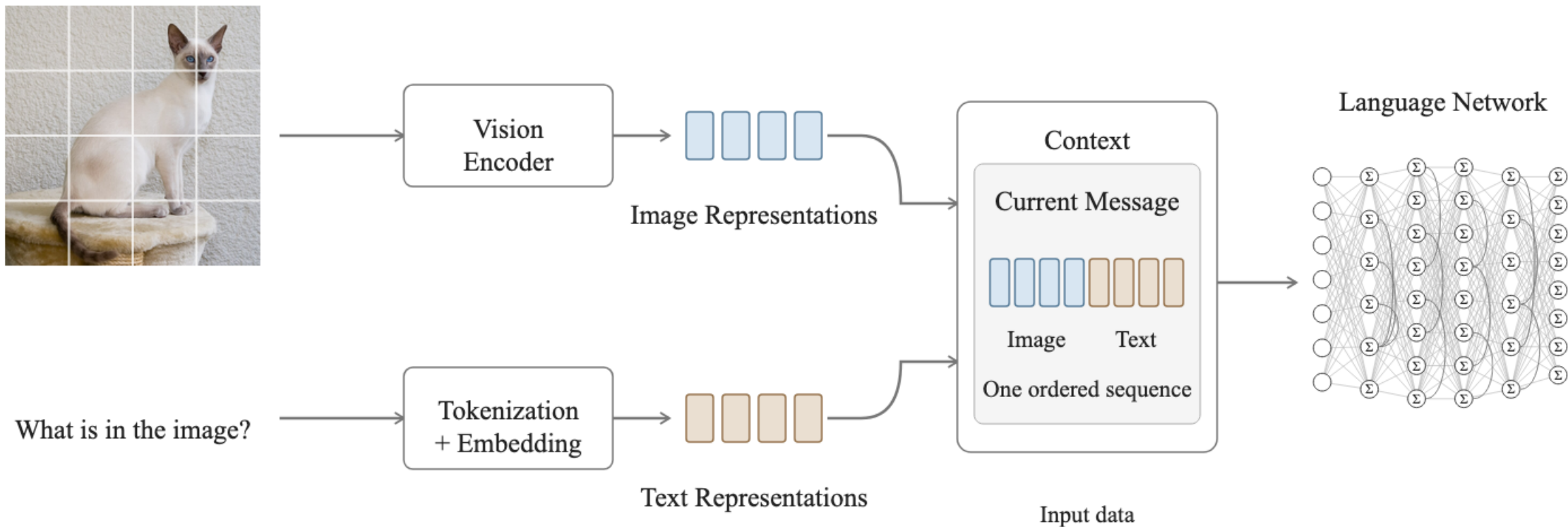
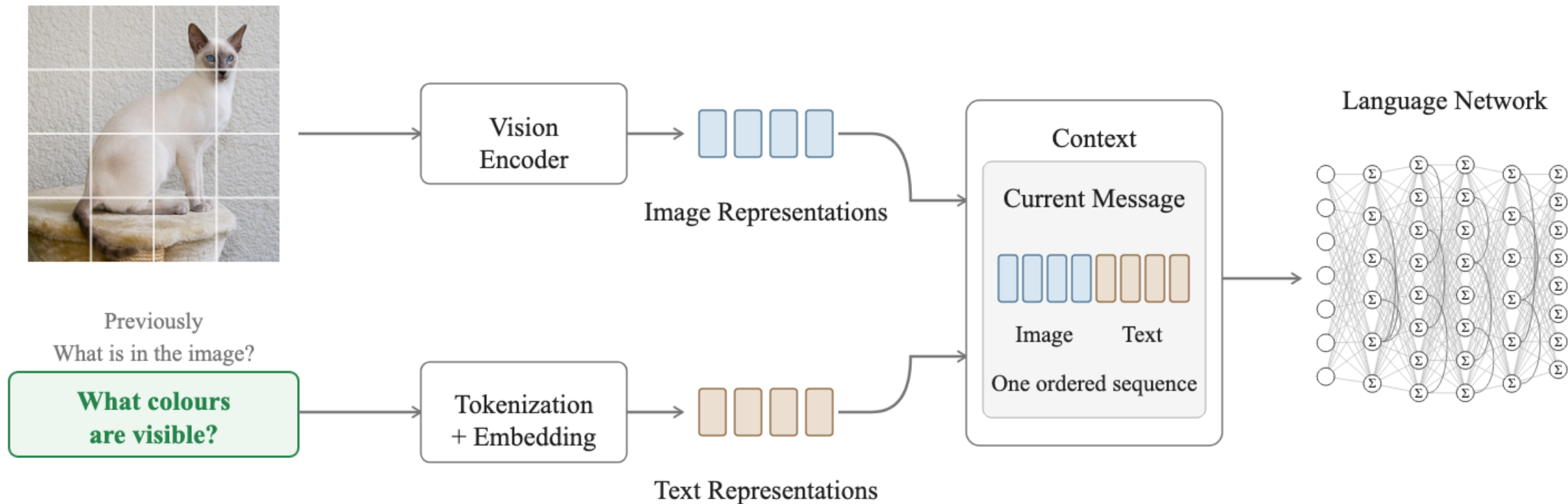


Image and text representations combine in context

Inference Uses Fixed Weights

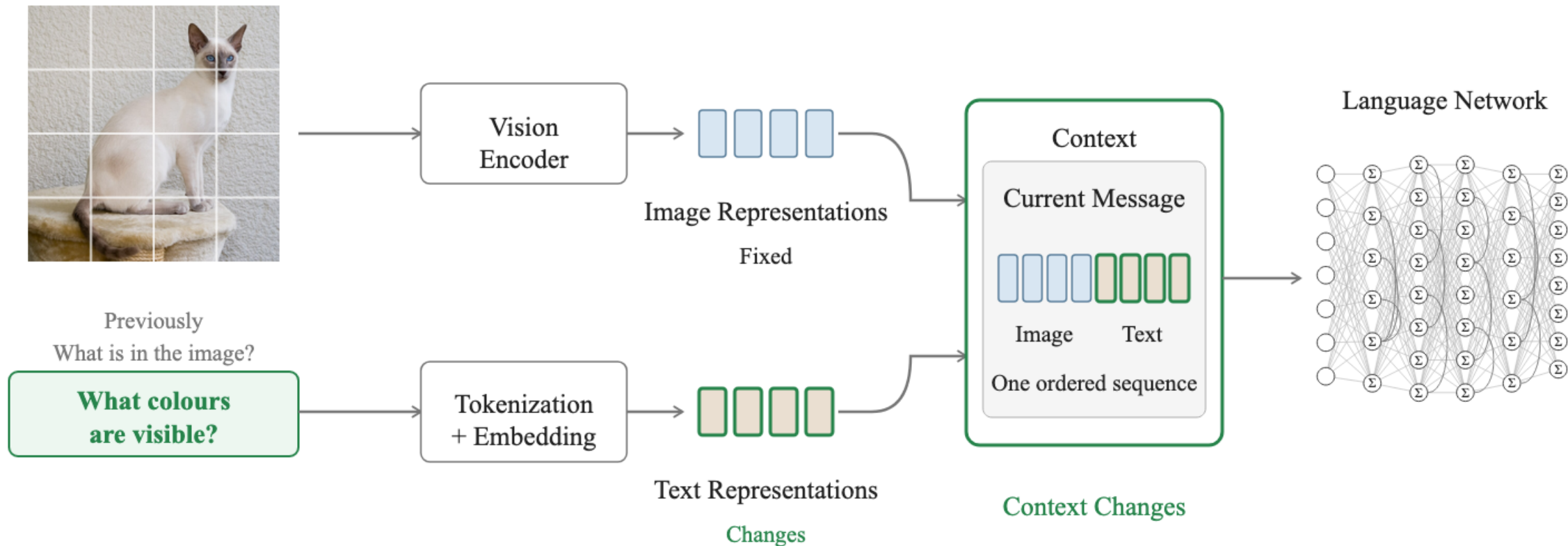
Same image, new question



Which changes: image representations, text representations, or weights?

Inference Uses Fixed Weights

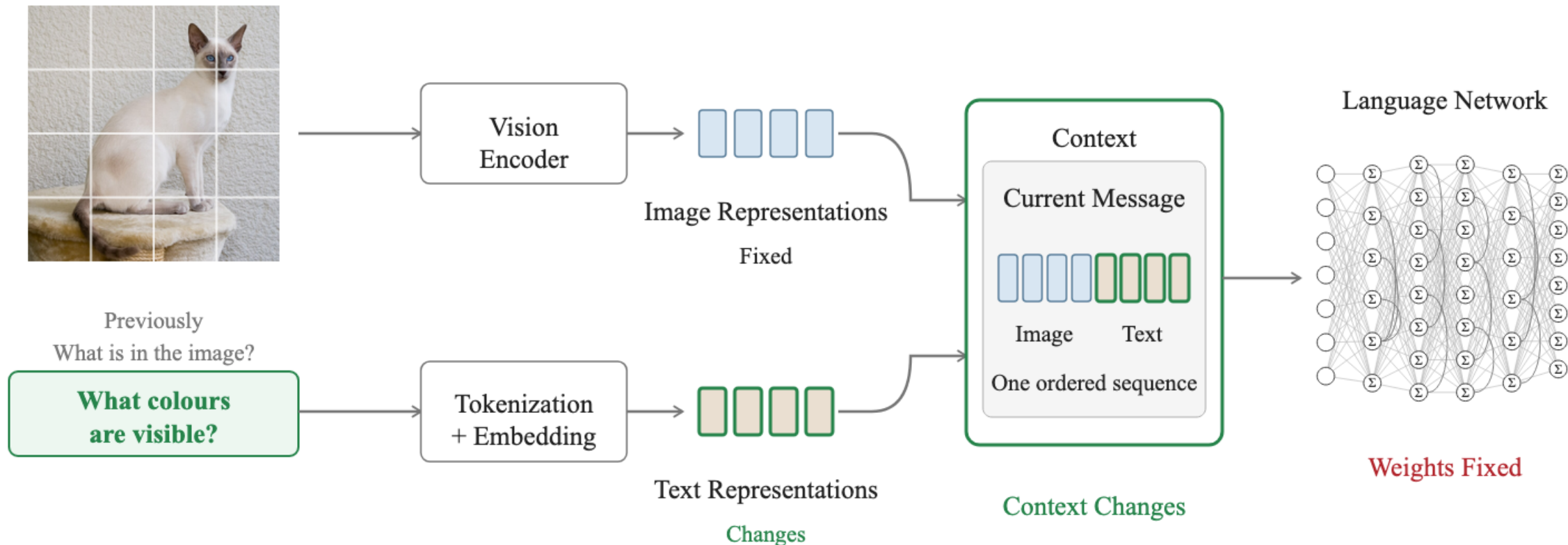
Same image, new question



New text representations change the context

Inference Uses Fixed Weights

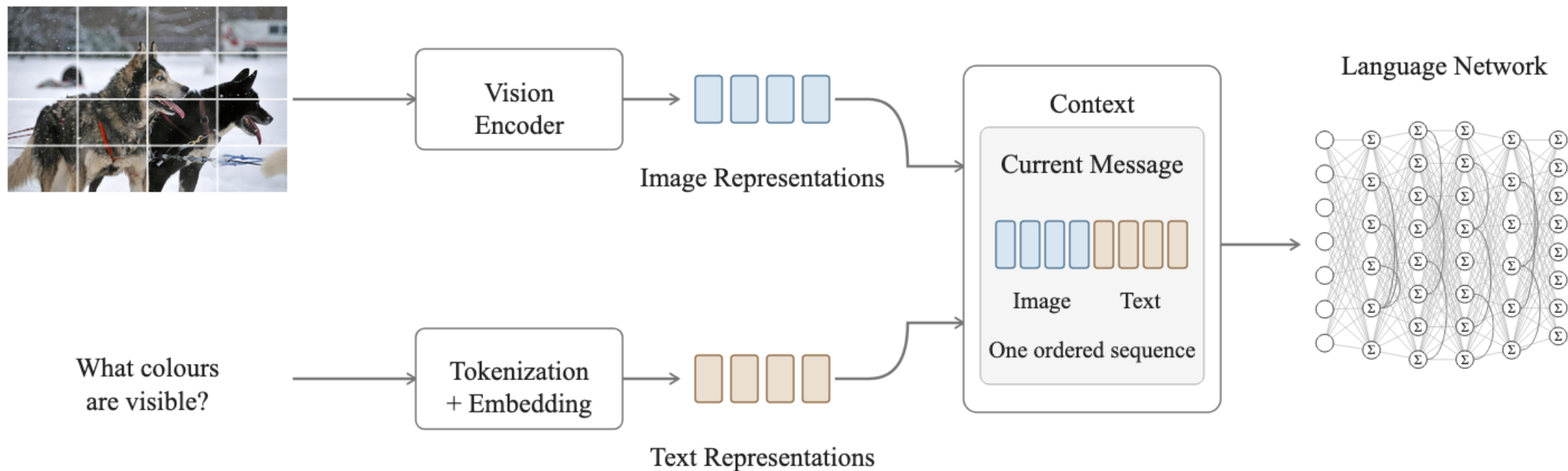
Same image, new question



We change the input, not the trained weights

Inference Uses Fixed Weights

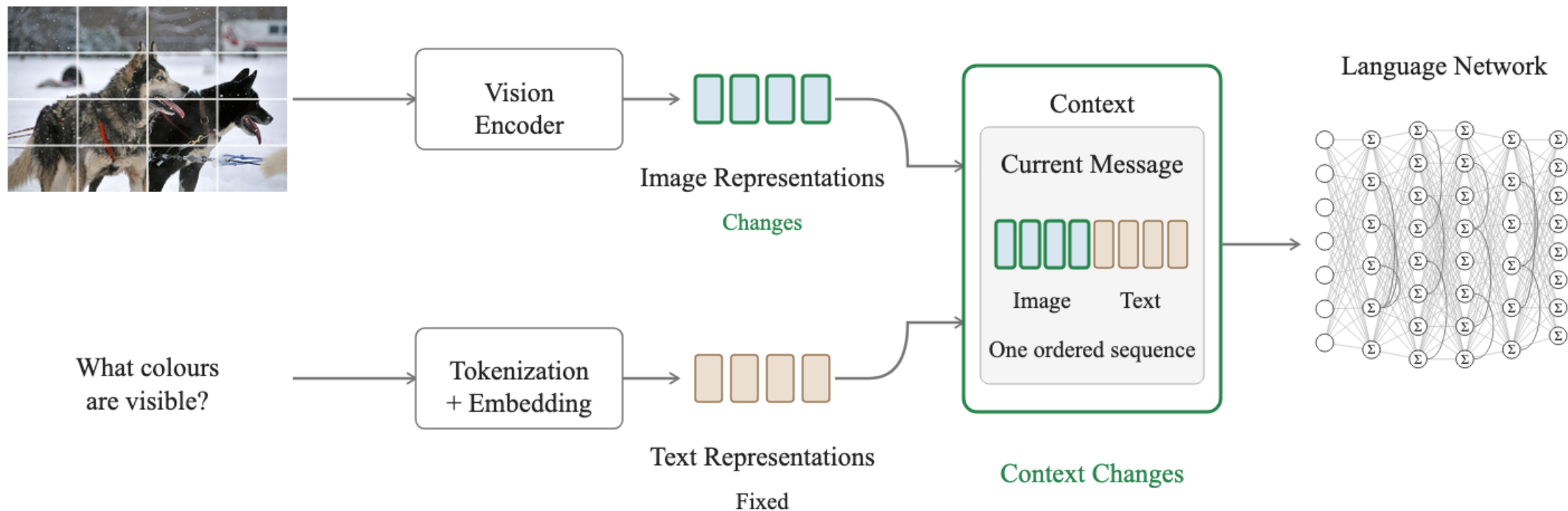
Same question, new image



Which changes: image representations, text representations, or weights?

Inference Uses Fixed Weights

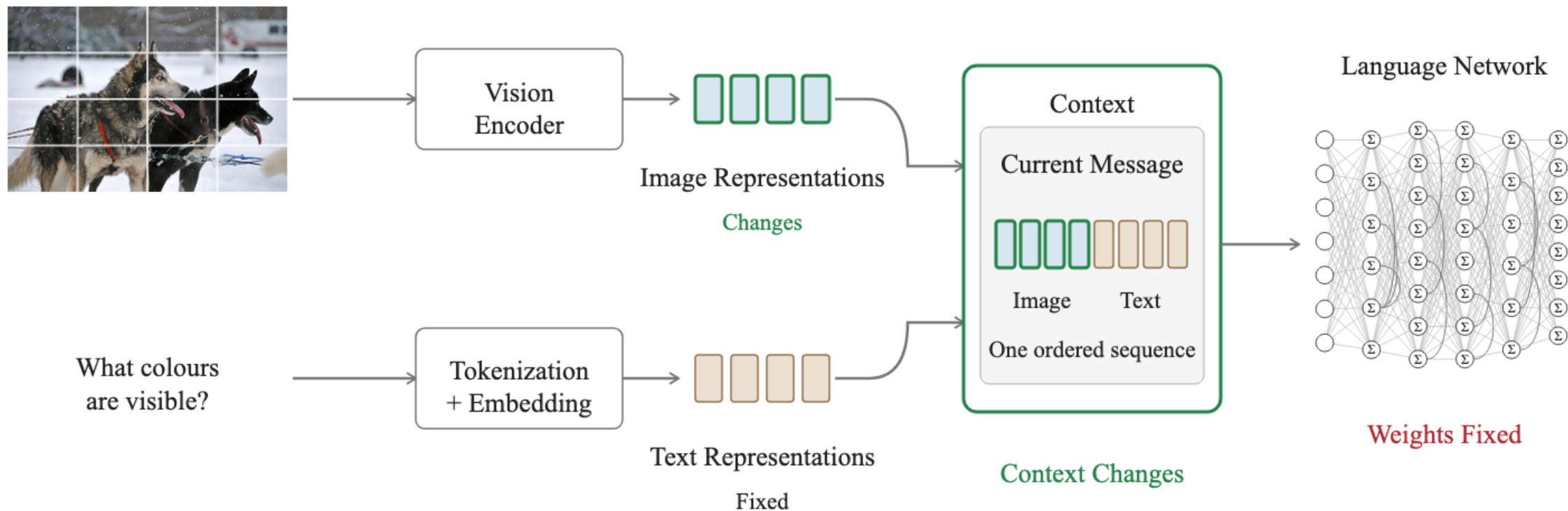
Same question, new image



New image representations change the context

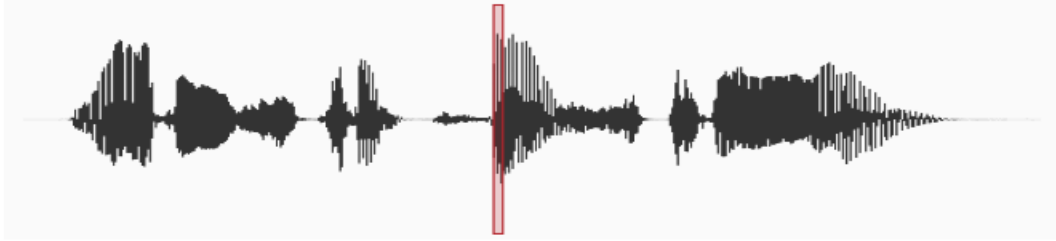
Inference Uses Fixed Weights

Same question, new image



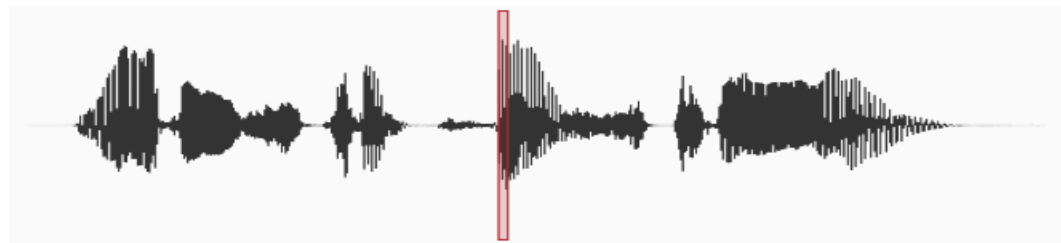
We change the input, not the trained weights

Speech Is Also Numerical Input

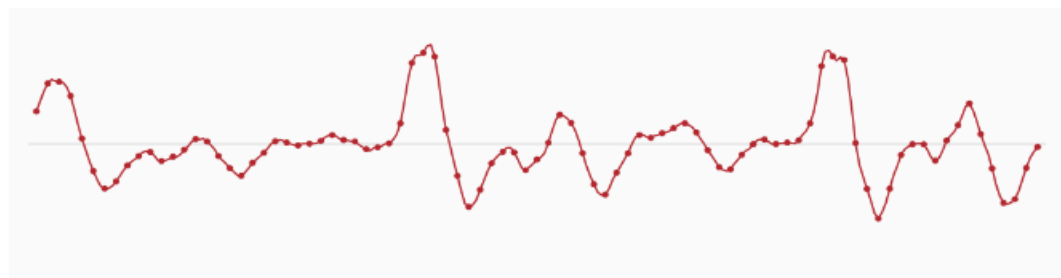


Default recording · Saved transcript

Speech Is Also Numerical Input



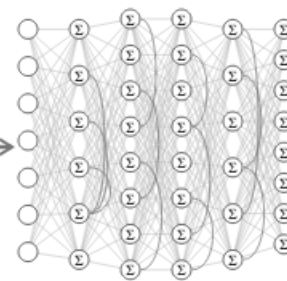
Default recording · Saved transcript



Each dot is one audio sample

Audio
Encoding

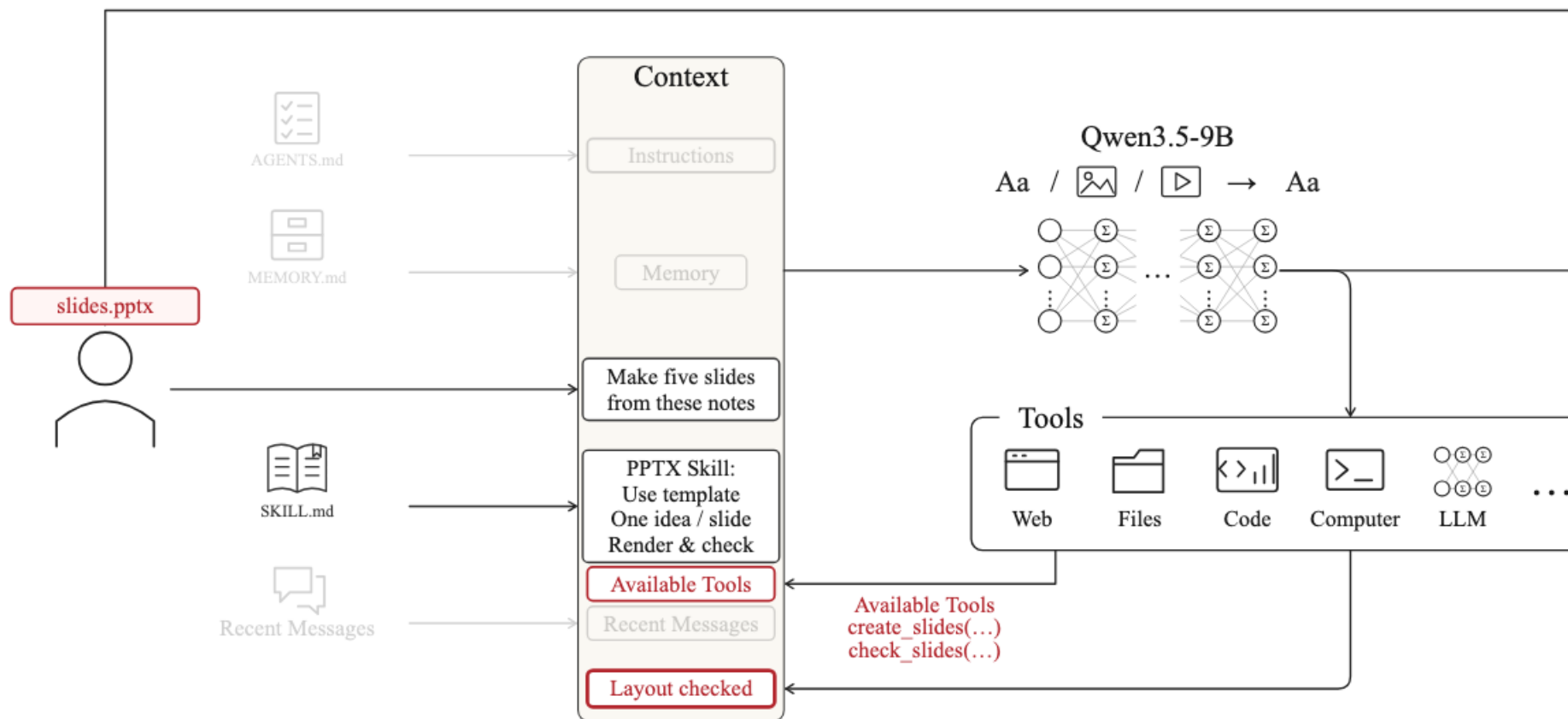
Neural Network



Parakeet Unified
English 0.6B

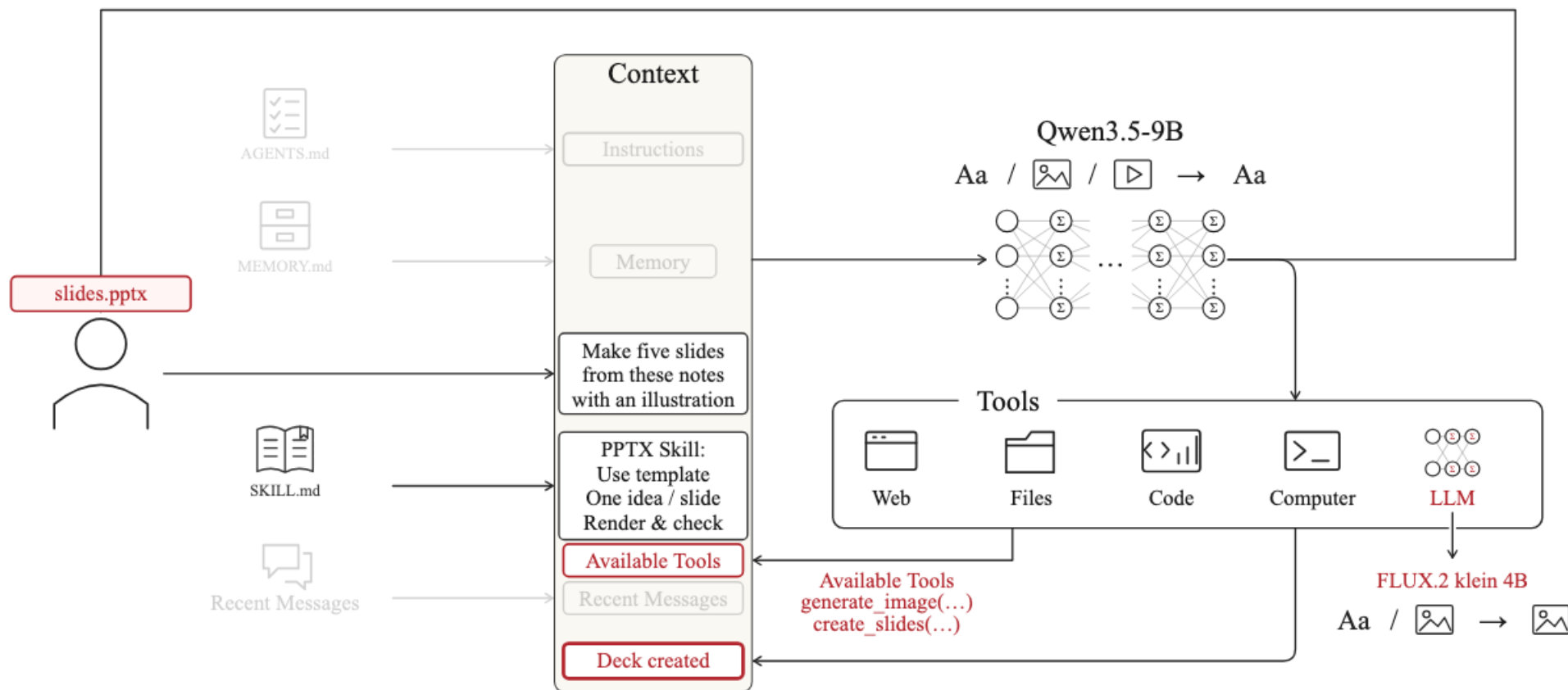
What is two plus two?

Revisiting L4: Calling Tools



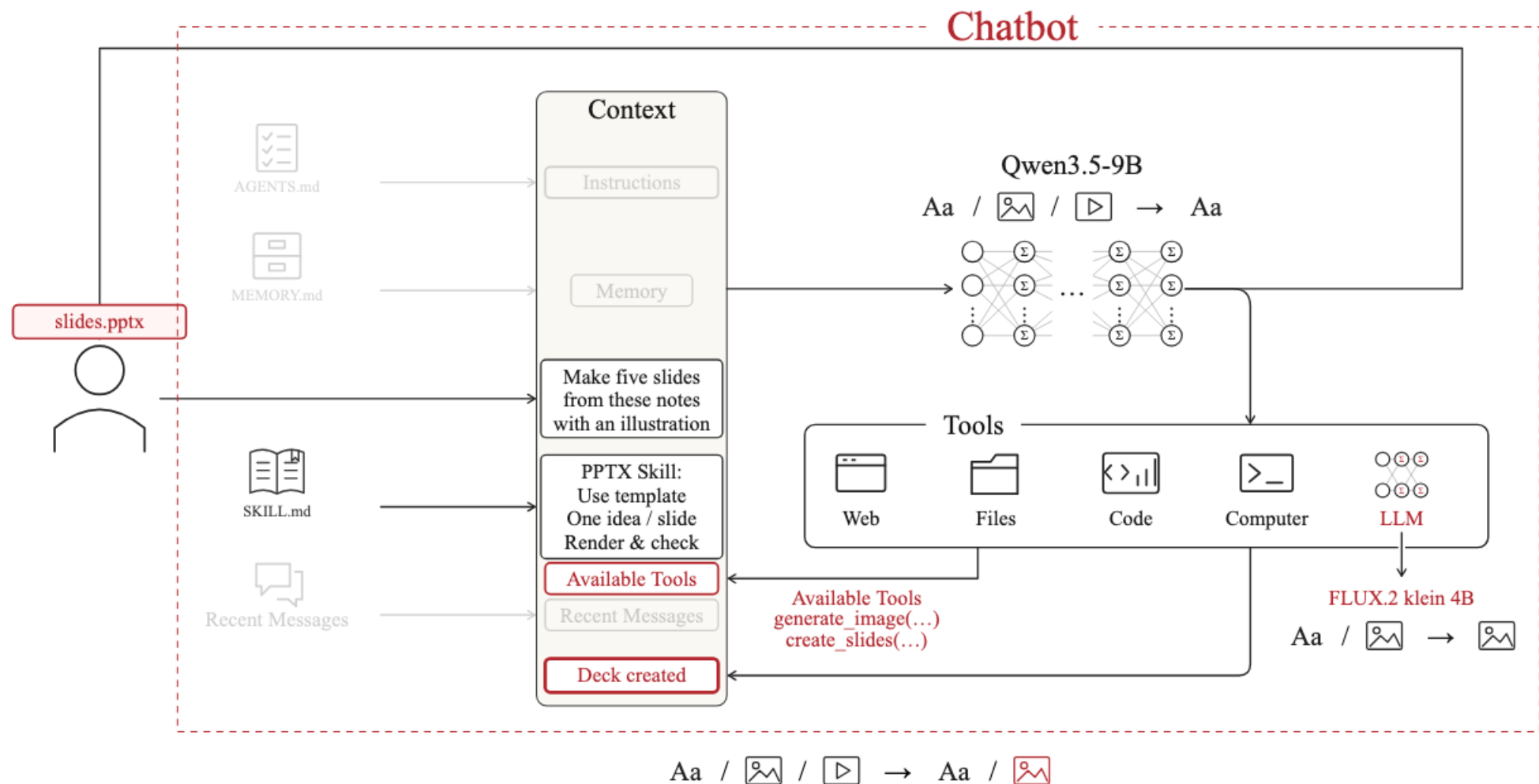
A tool turns text instructions into a PPTX file

A Specialist Model as a Tool



The chatbot application can offer more capabilities than its central language model has by itself

A Specialist Model as a Tool



The chatbot application can offer more capabilities than its central language model has by itself

Changing an Image

Cat



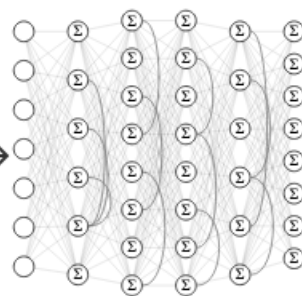
Choose an image and describe a change

Put the cat in a spacesuit

Generate

FLUX.2 klein 4B

 + Aa → 



Generated image appears here

The Challenge: Generating a Plausible World



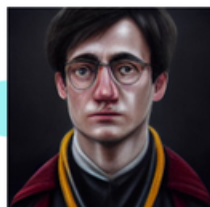
What makes these cats look wrong?

The Challenge: Generating a Plausible World

“A hyper-realistic image of Harry Potter”



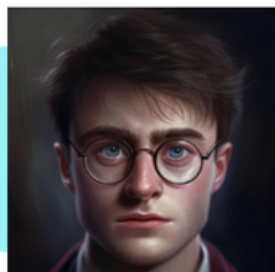
V1, February 2022



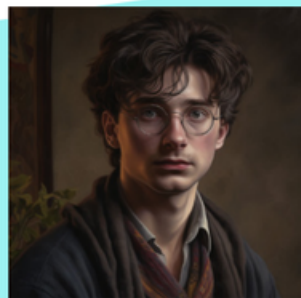
V2, April 2022



V3, July 2022



V4, November 2022



V5, March 2023



V5.1, March 2023



V5.2, June 2023



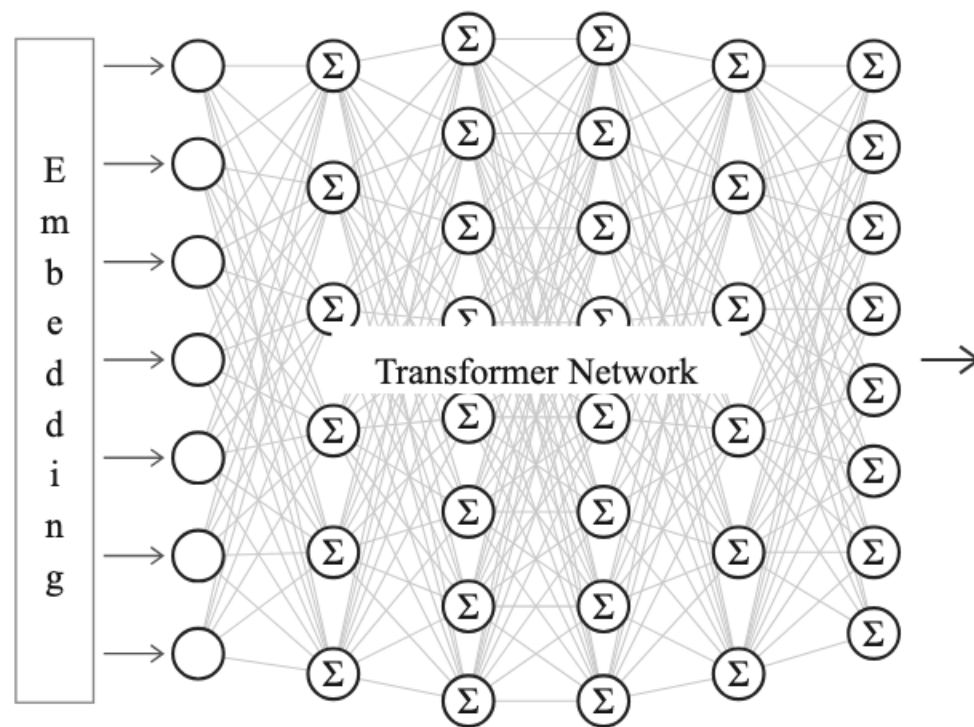
V6, December 2023

What has improved beyond sharper pixels?

Historical Midjourney Examples

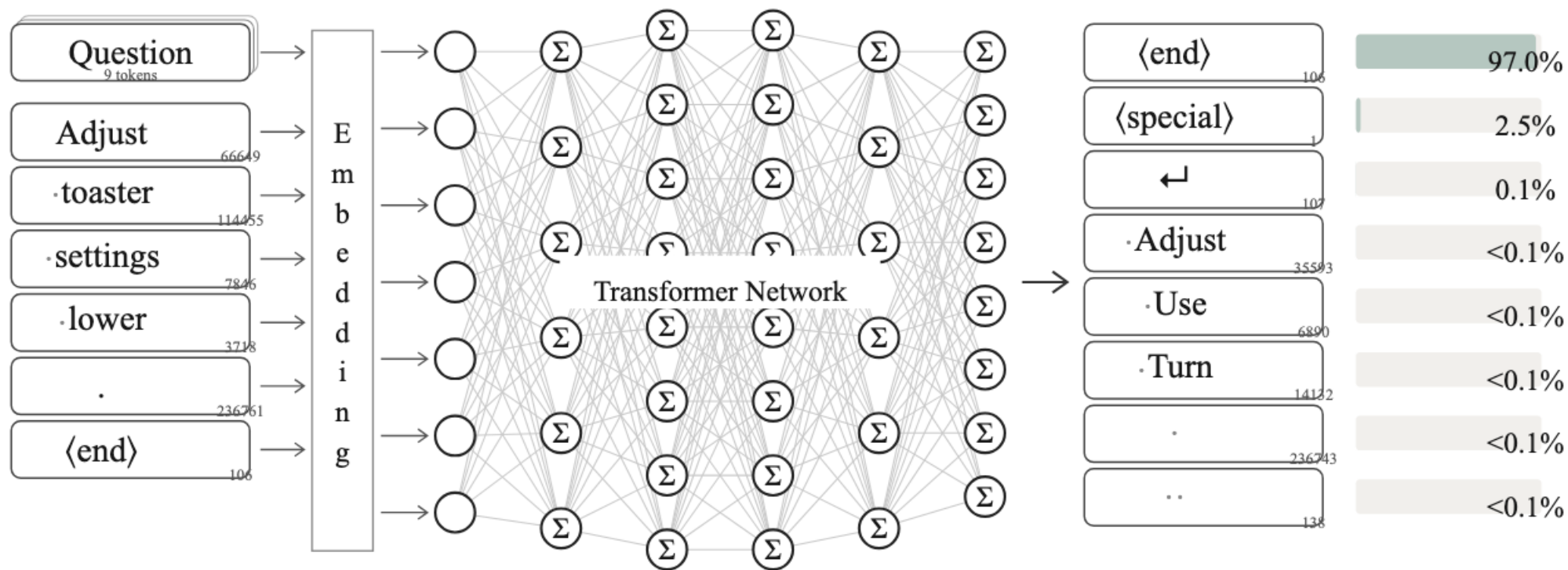
Revisiting L3: Generating Text

How 3910	·can 740	·I 564	·stop 4721	·my 1041	·toast 40233	·from 699	·burning 18830	? 236881
-------------	-------------	-----------	---------------	-------------	-----------------	--------------	-------------------	-------------



Revisiting L3: Generating Text

How ·can ·I ·stop ·my ·toast ·from ·burning ?



Generating an Image from Text

“A person dancing outside a restaurant”

Text adds a token. How can we generate an image?



Image

Generating an Image from Text

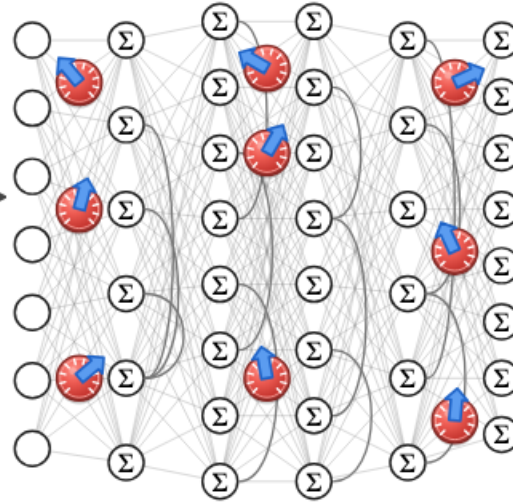
“A person dancing outside a restaurant”



Neural Network



Starting Noise



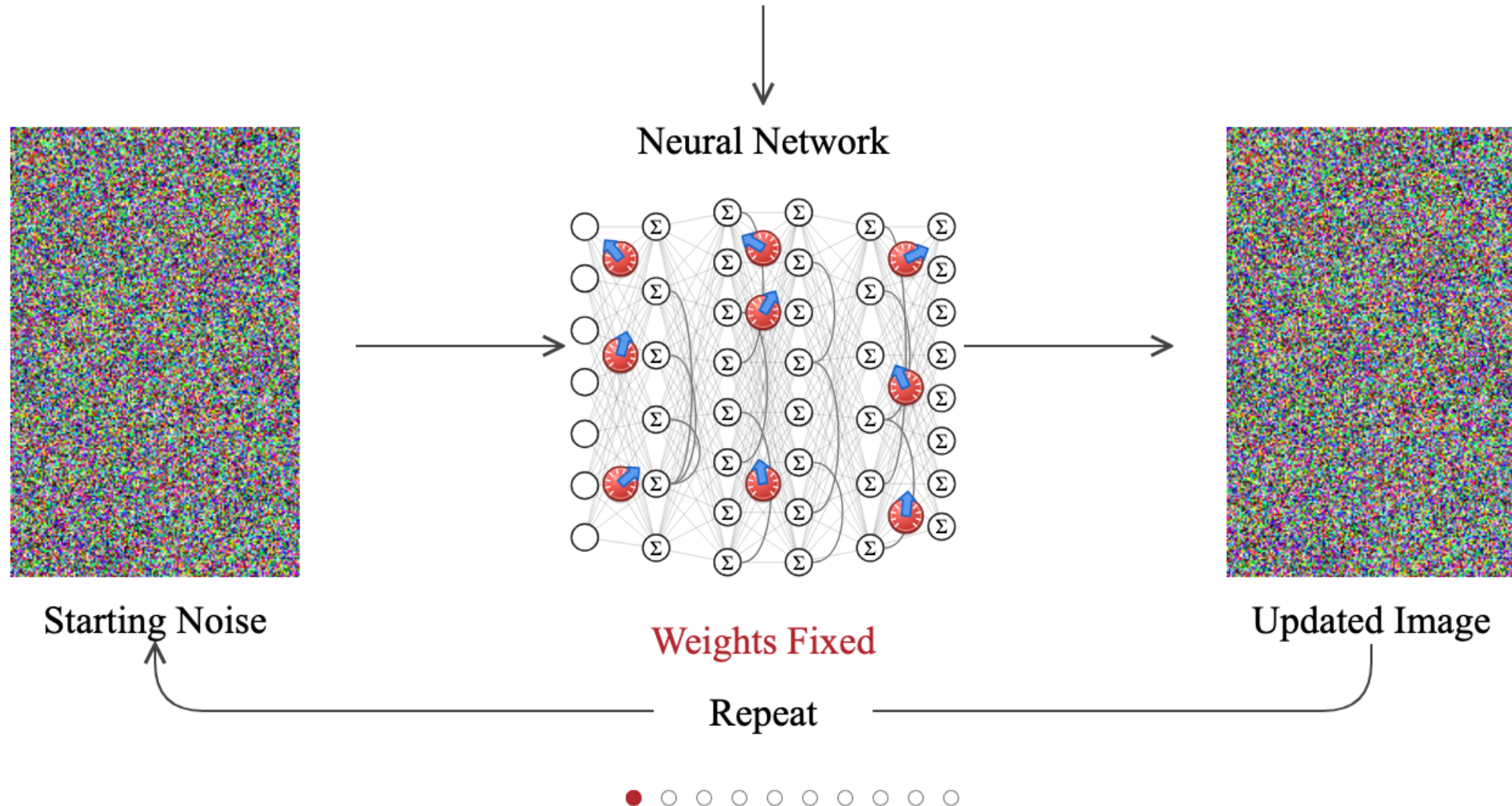
Weights Fixed



Next Image

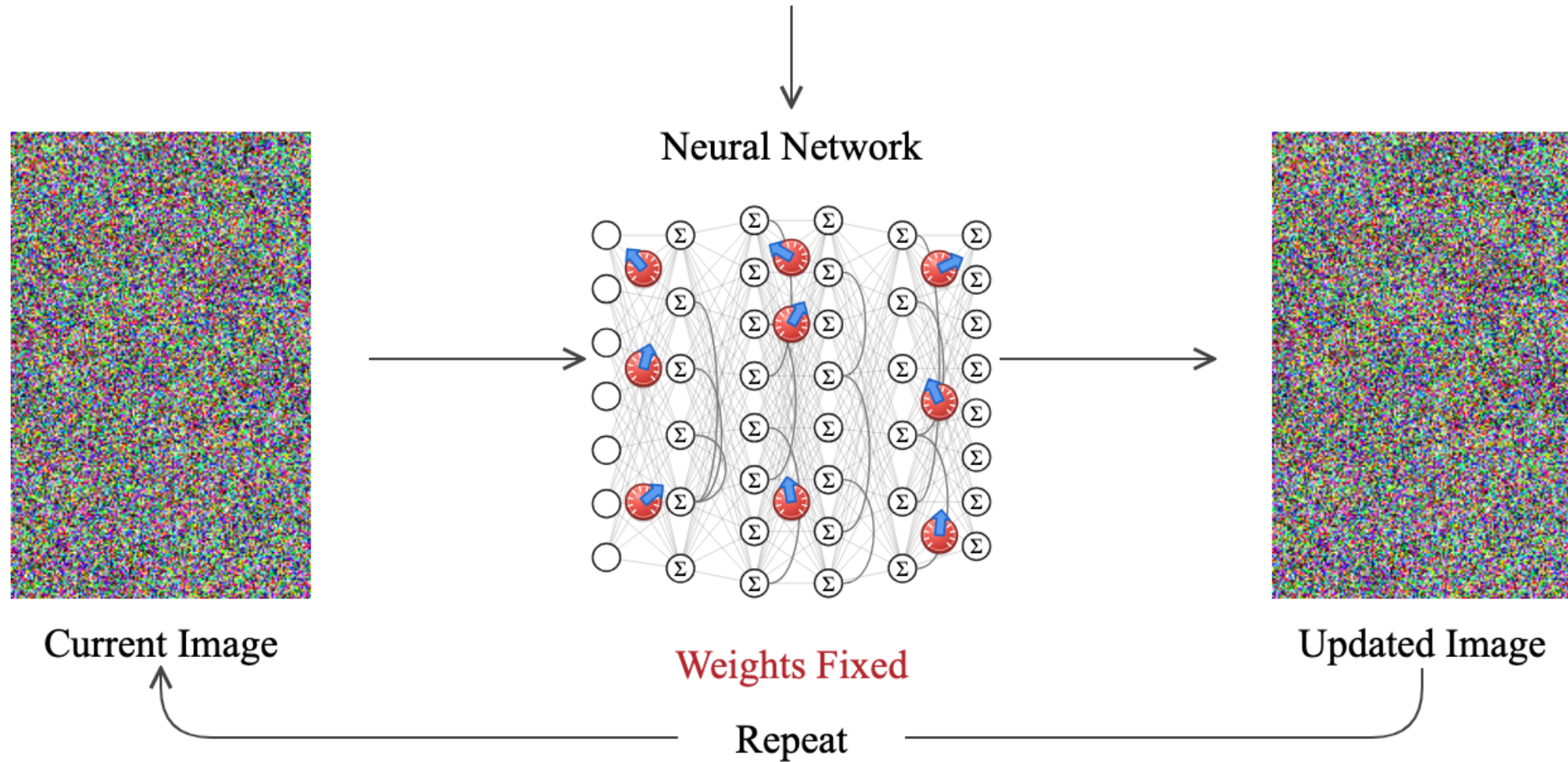
Generating an Image from Text

“A person dancing outside a restaurant”



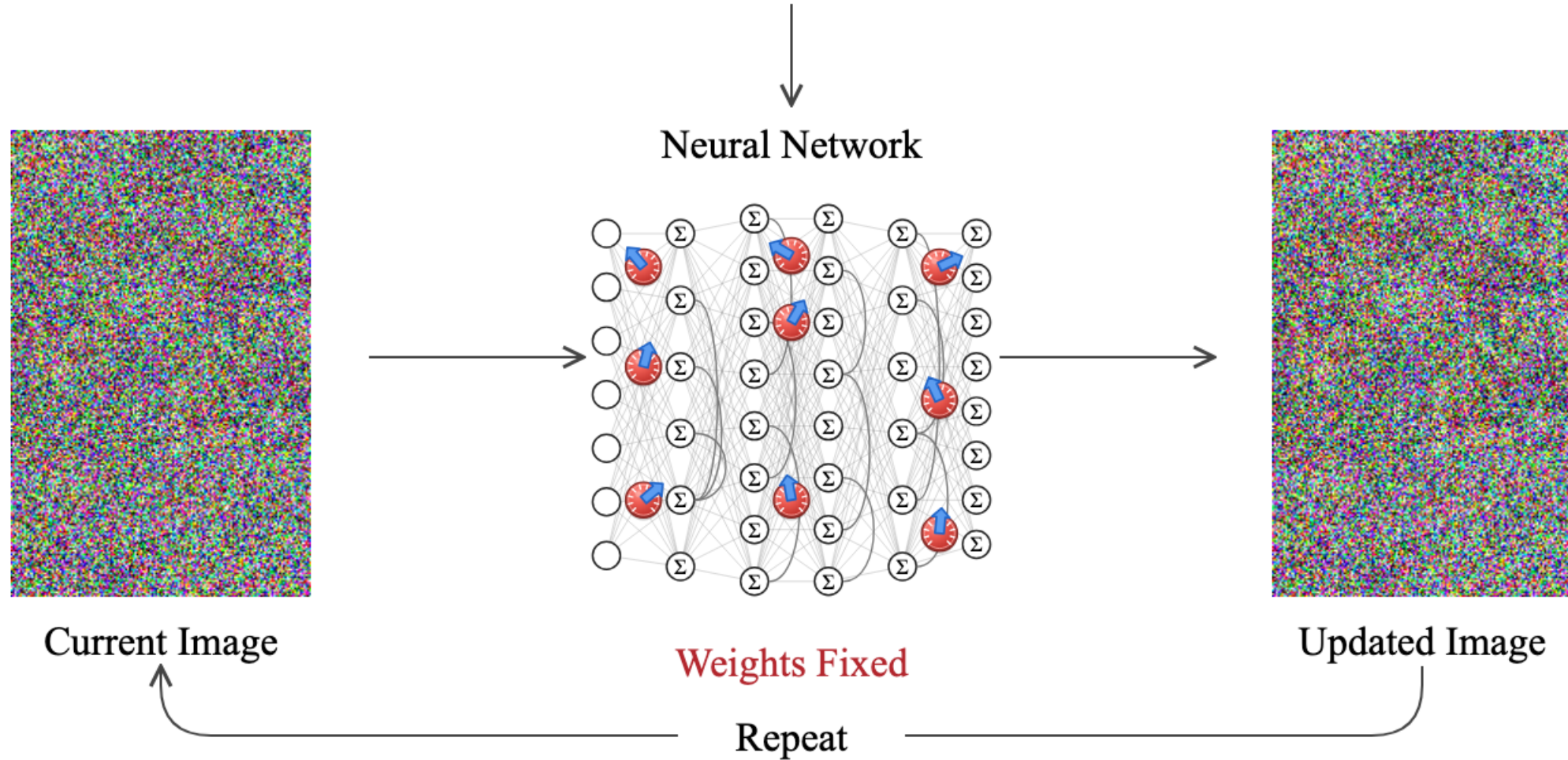
Generating an Image from Text

“A person dancing outside a restaurant”



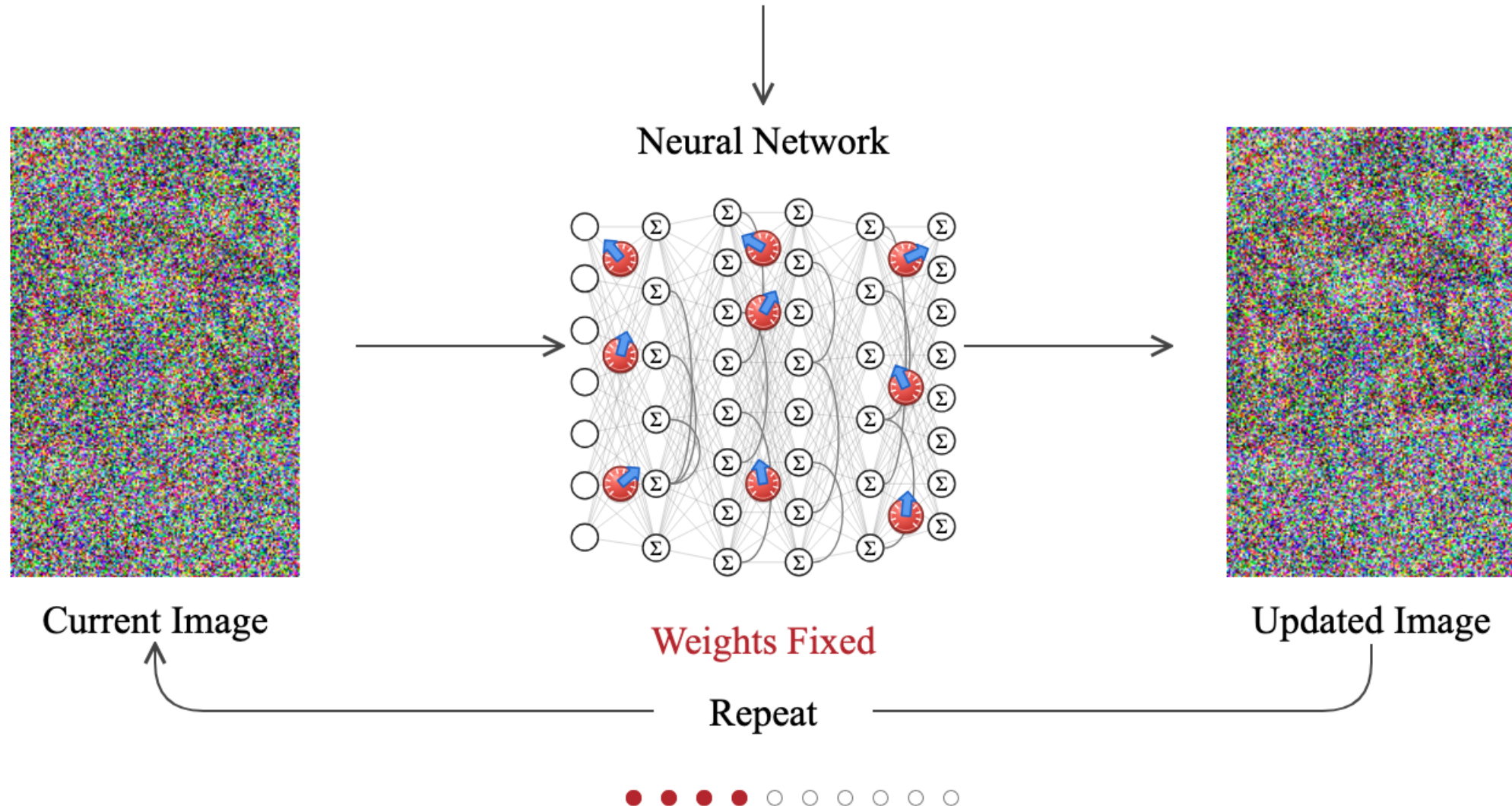
Generating an Image from Text

“A person dancing outside a restaurant”



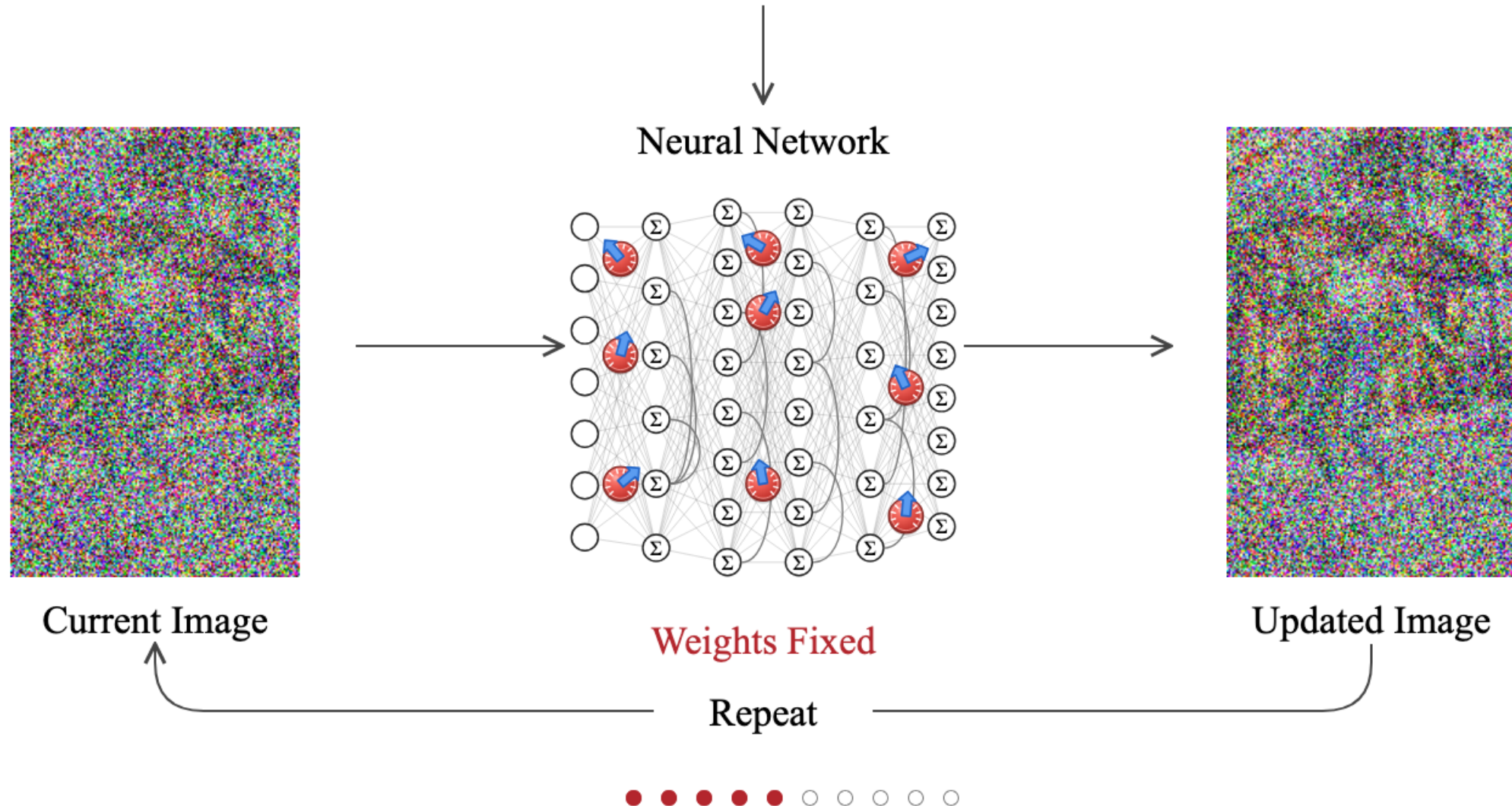
Generating an Image from Text

“A person dancing outside a restaurant”



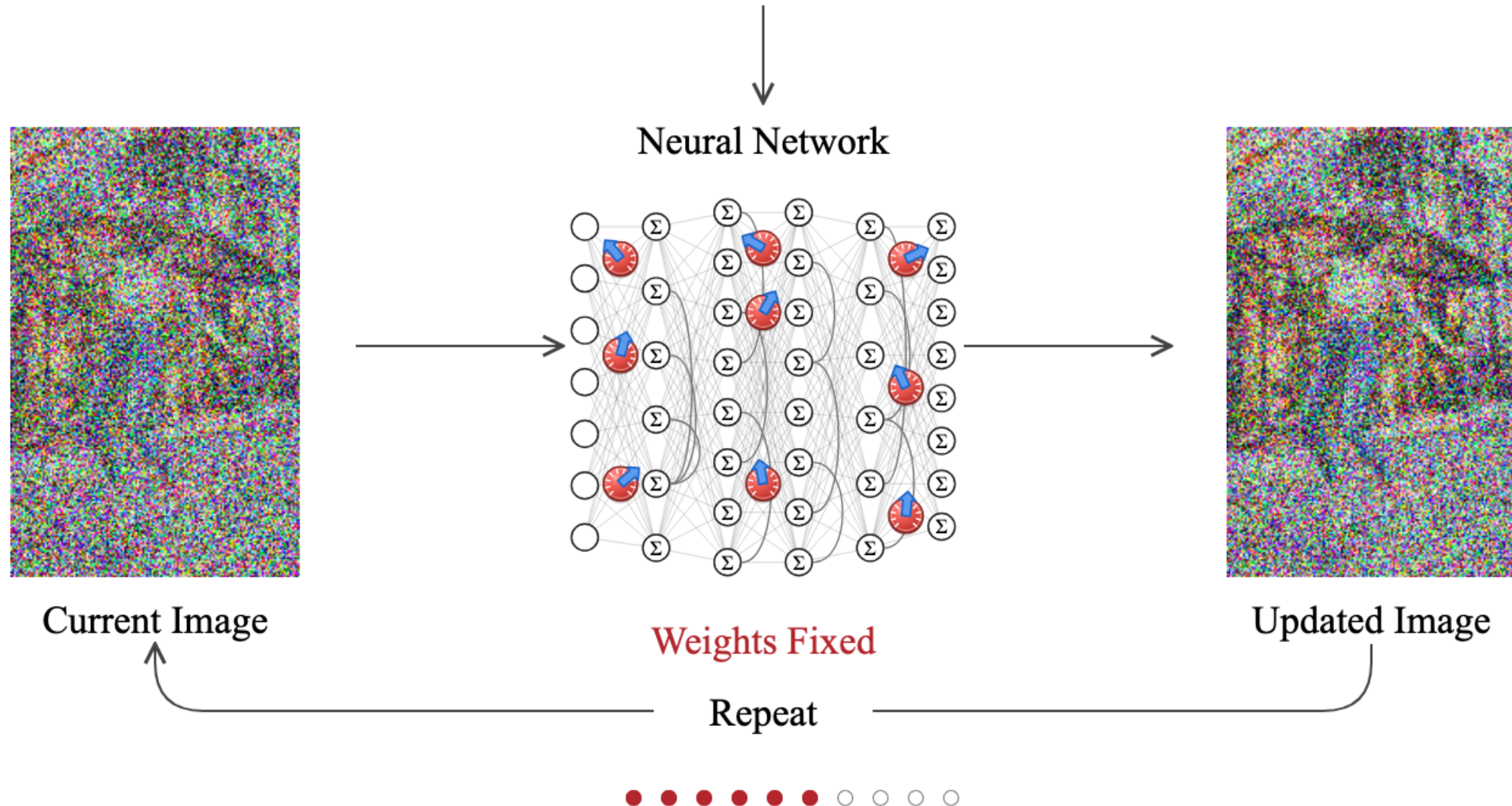
Generating an Image from Text

“A person dancing outside a restaurant”



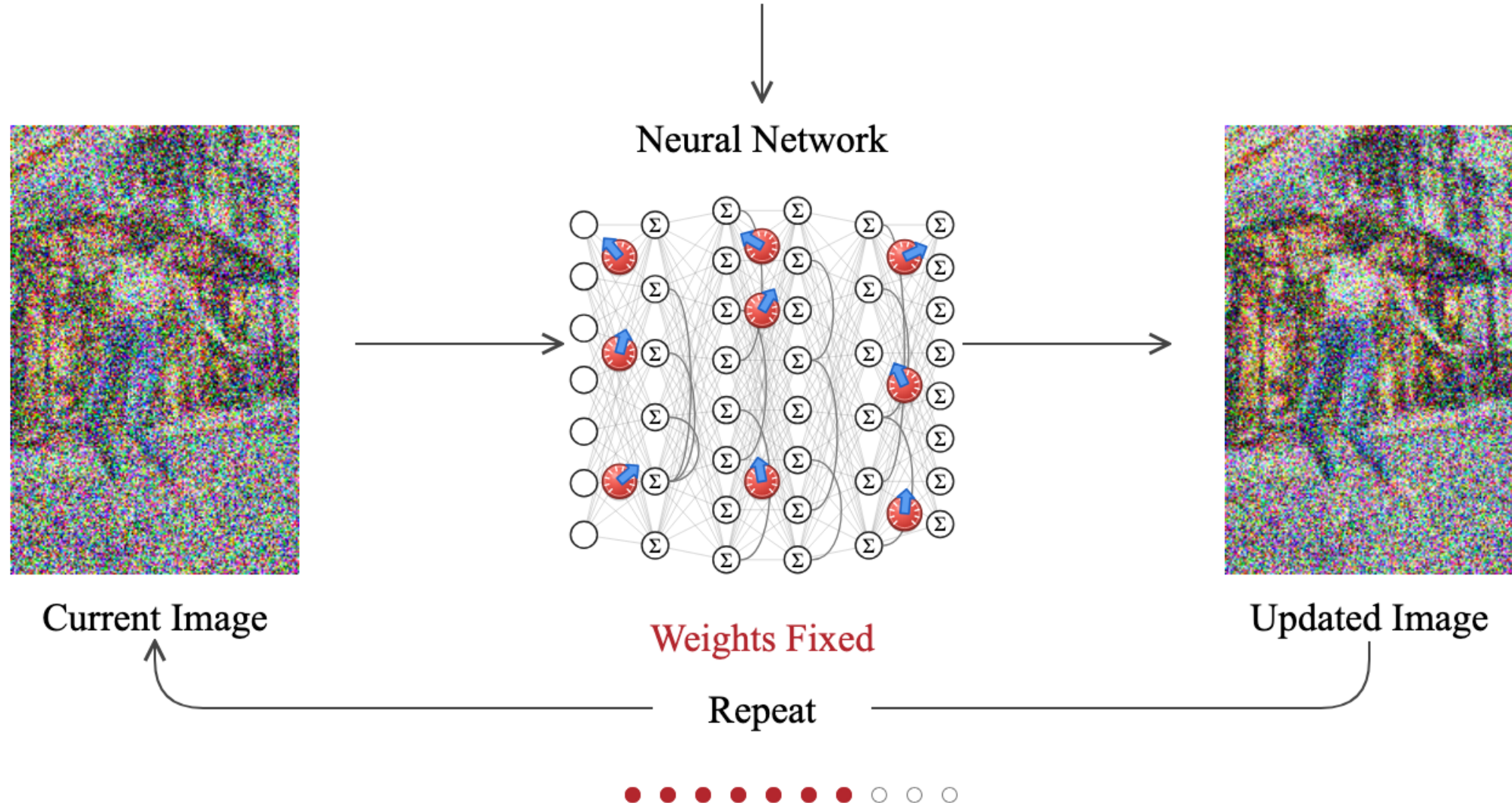
Generating an Image from Text

“A person dancing outside a restaurant”



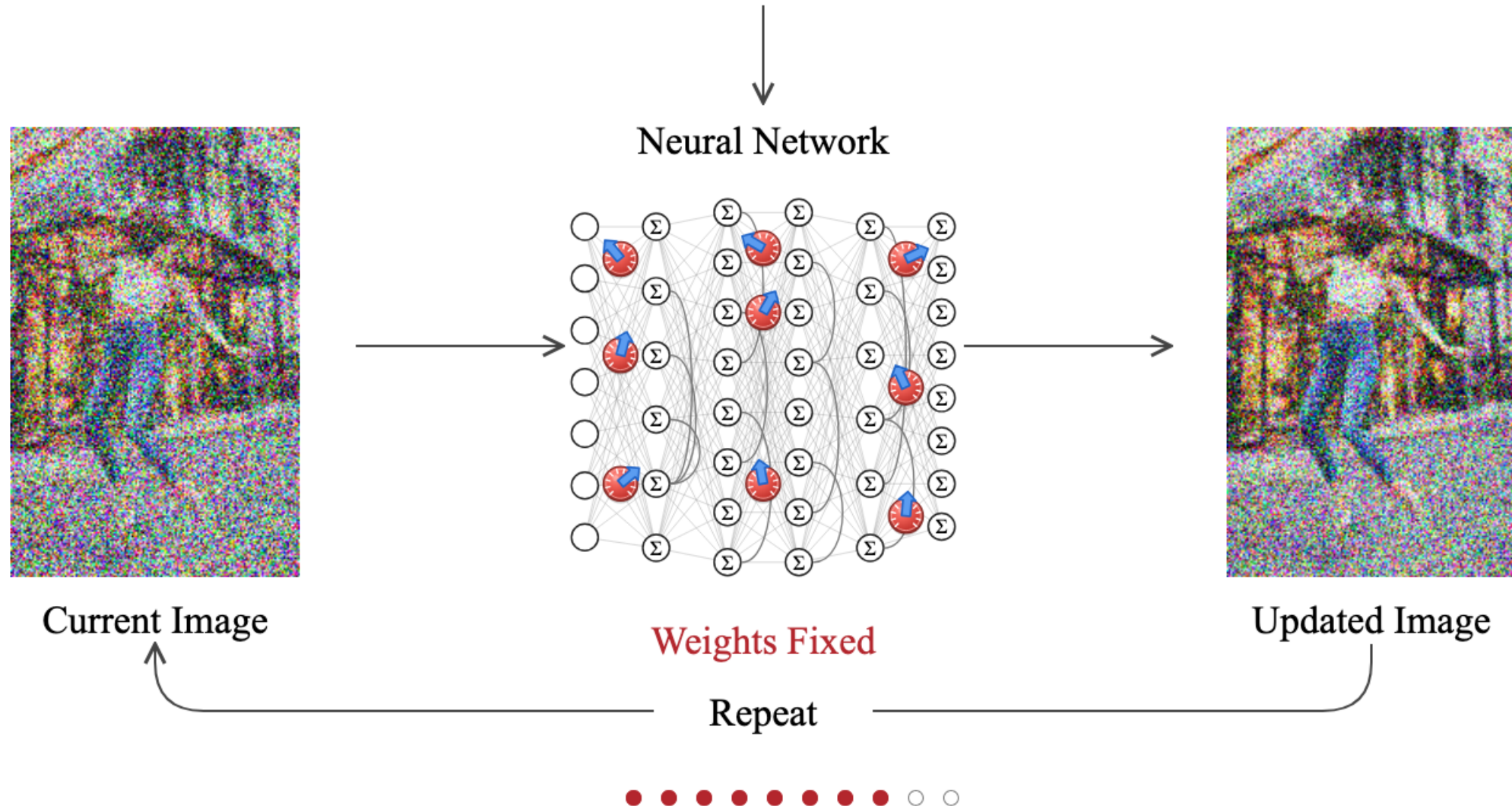
Generating an Image from Text

“A person dancing outside a restaurant”



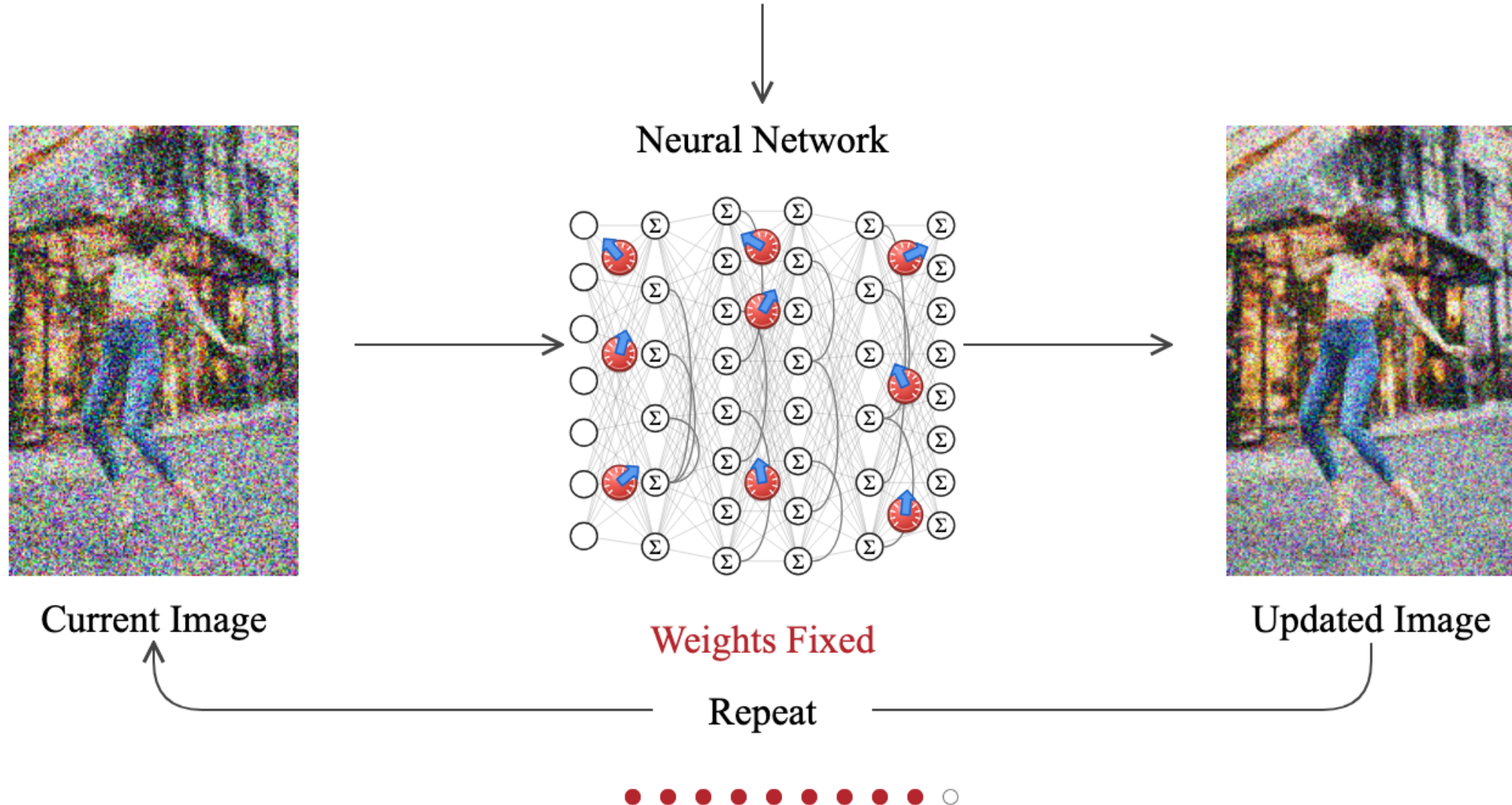
Generating an Image from Text

“A person dancing outside a restaurant”



Generating an Image from Text

“A person dancing outside a restaurant”



Generating an Image from Text

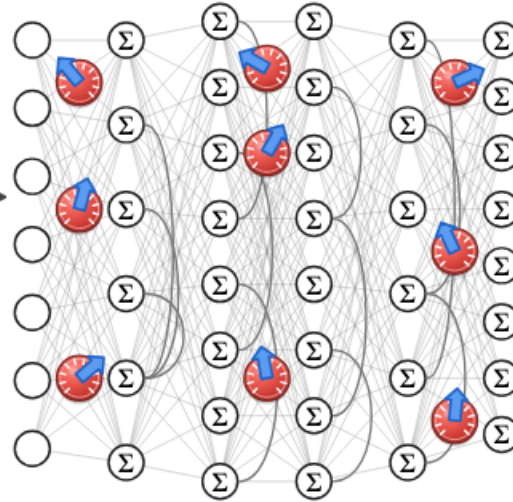
“A person dancing outside a restaurant”



Neural Network



Current Image



Weights Fixed

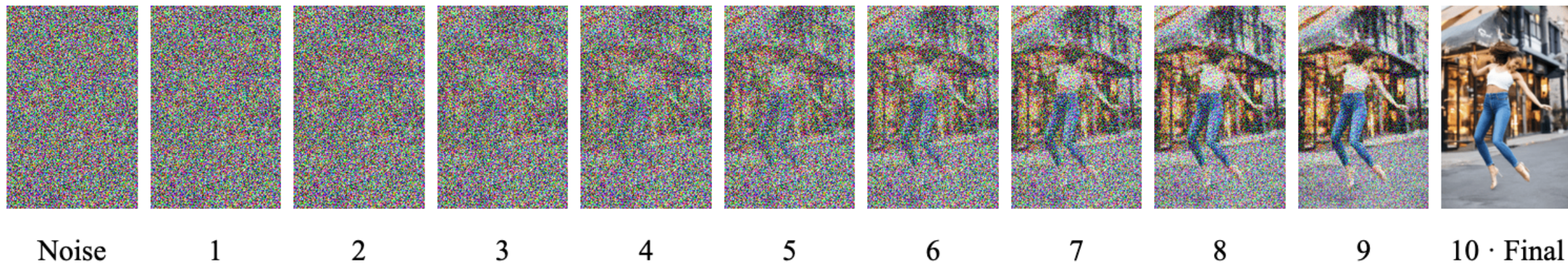


Final Image



Generating an Image from Text

“A person dancing outside a restaurant”



How does it learn to do this?

Create Training Examples by Adding Noise



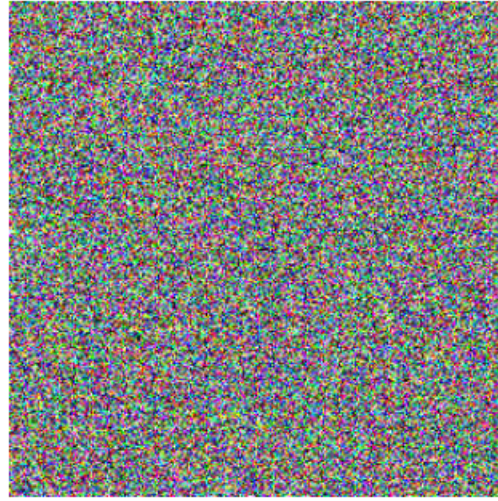
Original Image

Create Training Examples by Adding Noise



Original Image

+



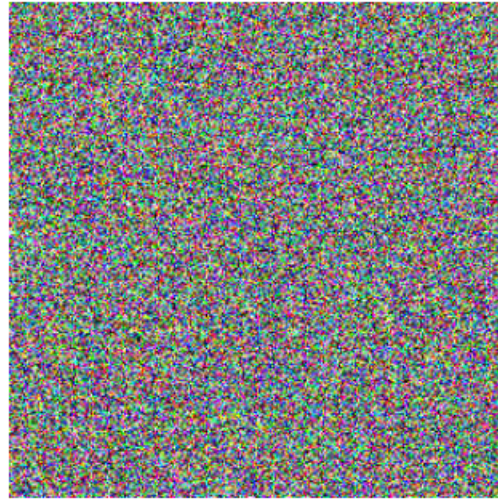
Random Noise

Create Training Examples by Adding Noise



Original Image

+



Random Noise



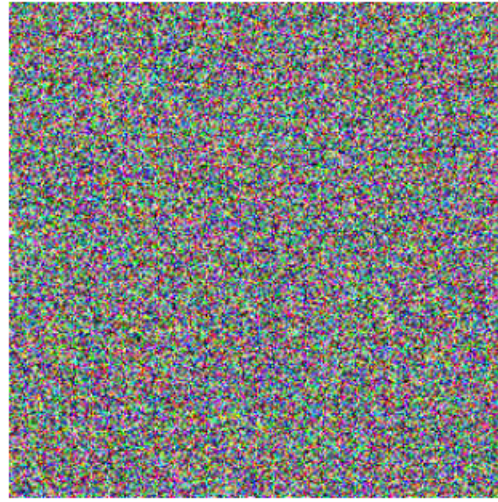
Noisy Image

Create Training Examples by Adding Noise

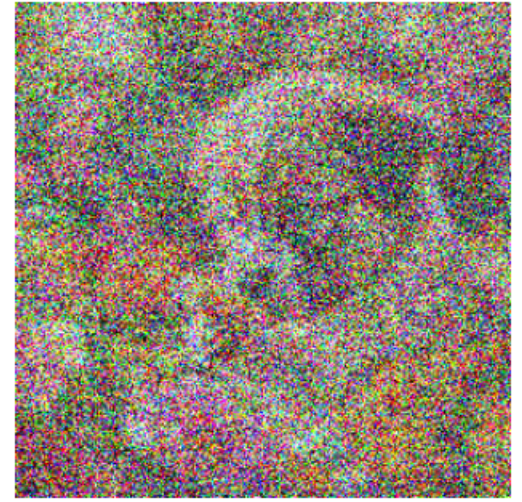


Original Image

+



Random Noise



Noisy Image

One Training Sample

Repeat With Many Images and Noise Levels

Learning to Predict the Added Noise

Create Training Examples



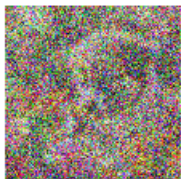
Original
Image

+



Random
Noise

→



Noisy Image

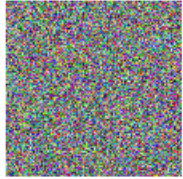
No Neural Network Needed

Learning to Predict the Added Noise

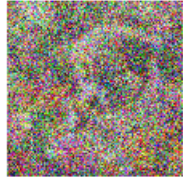
Create Training Examples



+



→

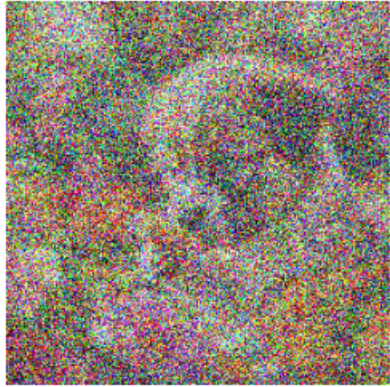


Original
Image

Random
Noise

Noisy Image

No Neural Network Needed



Noisy Image

Amount of Noise Added

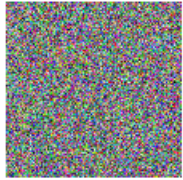
Low ———●——— High

Learning to Predict the Added Noise

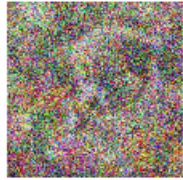
Create Training Examples



+



→

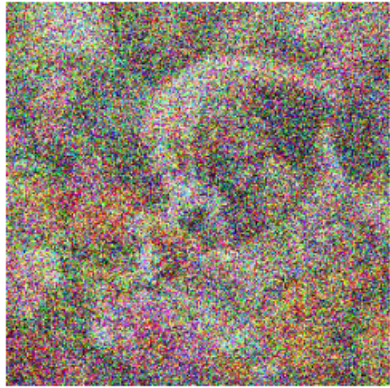


Original
Image

Random
Noise

Noisy Image

No Neural Network Needed



Noisy Image

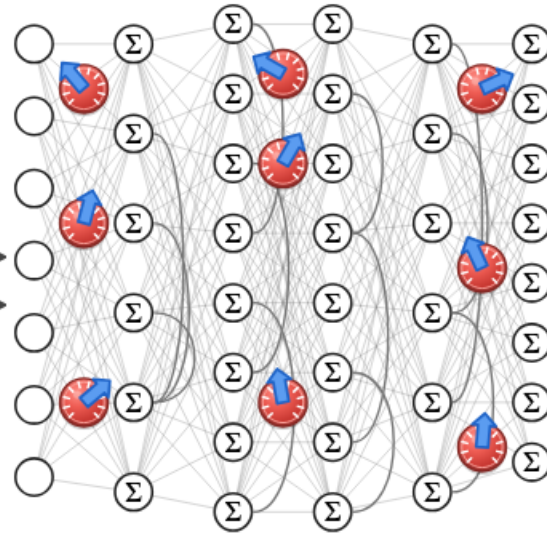
Amount of Noise Added

Low ———●——— High

“A puppy among flowers”



Neural Network

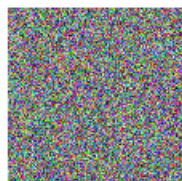


Learning to Predict the Added Noise

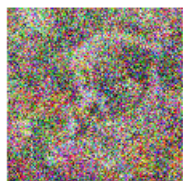
Create Training Examples



+



→



Original
Image

Random
Noise

Noisy Image

No Neural Network Needed



Noisy Image

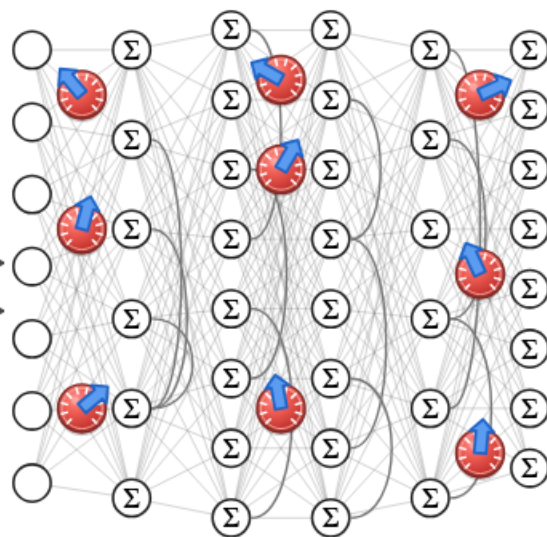
Amount of Noise Added

Low —●— High

“A puppy among flowers”



Neural Network



Estimated Added Noise

Training a Diffusion Model

“A puppy among flowers”

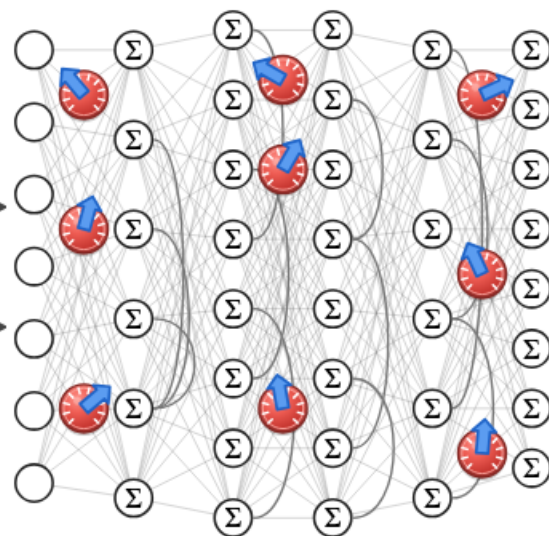
Neural Network



Noisy Image

Amount of Noise Added

Low ———●——— High



Estimated Added Noise



Known Added Noise

Training a Diffusion Model

“A puppy among flowers”

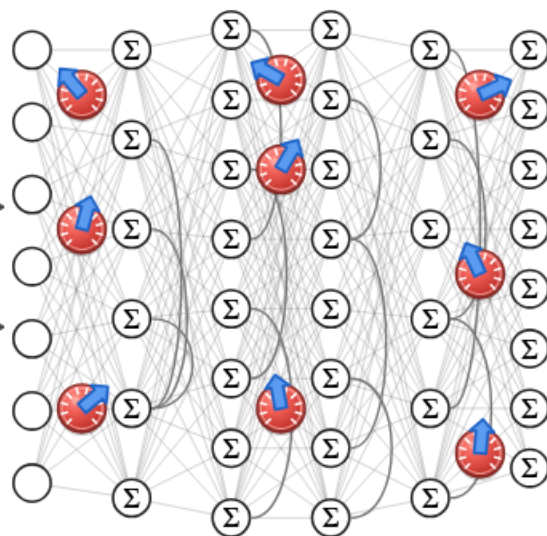
Neural Network



Noisy Image

Amount of Noise Added

Low ———●——— High



Estimated Added Noise



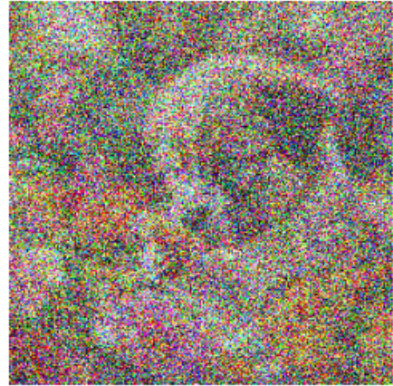
Known Added Noise

Compare

Training a Diffusion Model

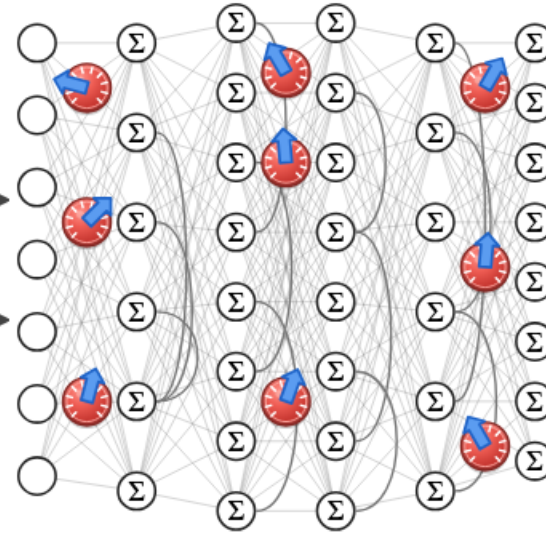
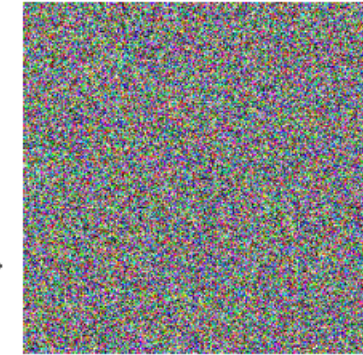
“A puppy among flowers”

Neural Network

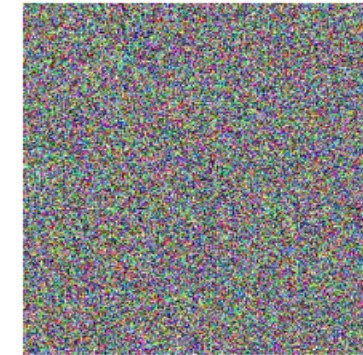


Noisy Image

Amount of Noise Added
Low ———●——— High



Estimated Added Noise

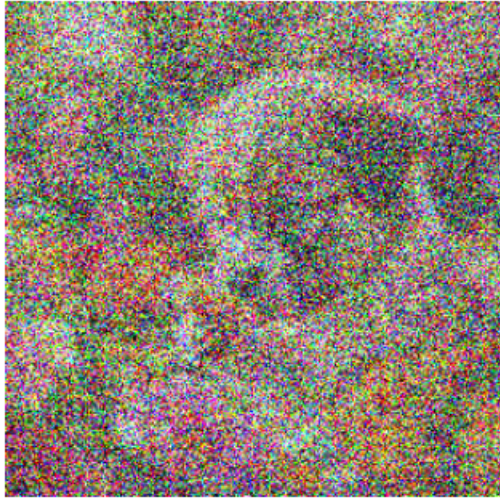


Known Added Noise

Compare

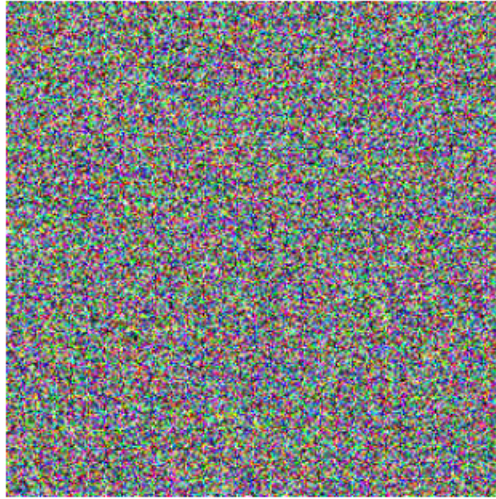
Adjust Weights to Reduce the Error

How Does Predicting Noise Help?



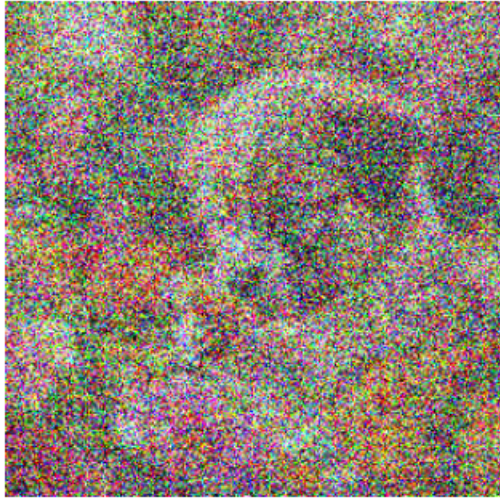
Noisy Image

—



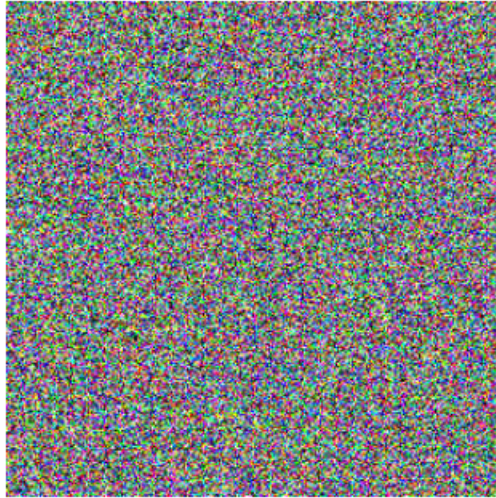
Estimated Noise

How Does Predicting Noise Help?



Noisy Image

—



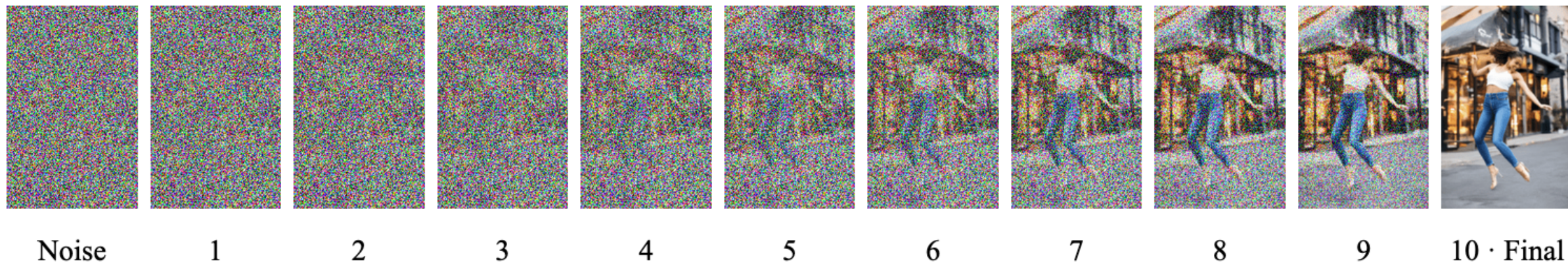
Estimated Noise



Estimate of the Clean Image

Remove a Little Noise, Then Repeat

“A person dancing outside a restaurant”



Estimate the noise → Take a small denoising step → Repeat

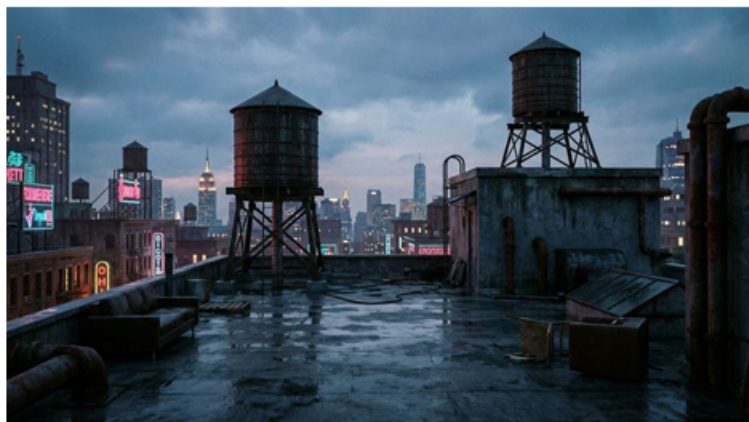
The Challenge: Not Just Realism



Boxer A



Boxer B



Rooftop

Shot 1 (2s): Wide shot, @Boxer A and @Boxer B face off in the center of the rooftop, feet apart in a boxing stance.

Shot 2 (2s): Both move in, testing each other up close: @Boxer A throws a quick punch, @Boxer B sidesteps and blocks.

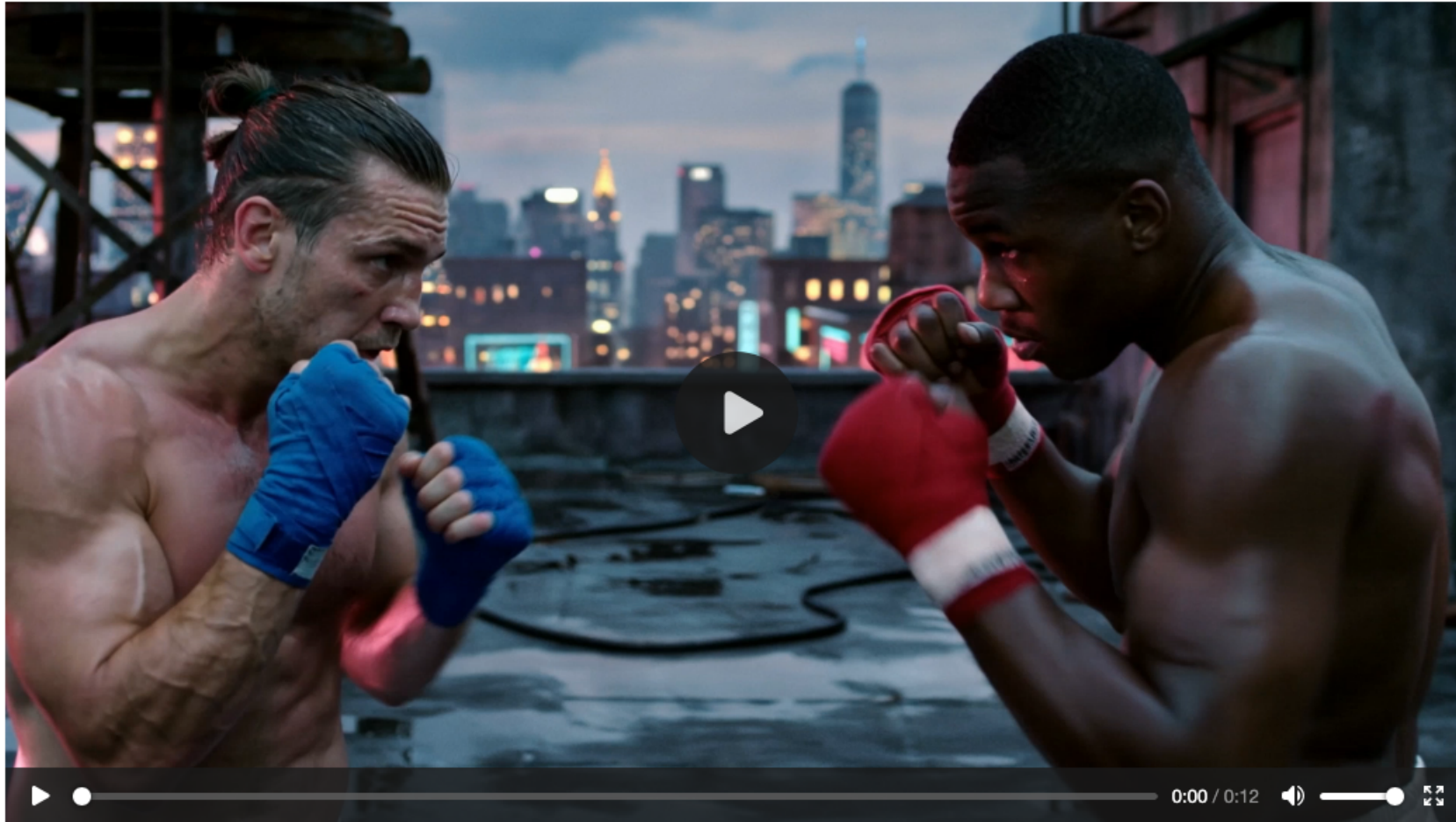
Shot 3 (3s): @Boxer A continues the attack, landing a punch on @Boxer B's head, and @Boxer B retaliates.

Shot 4 (4s): Wide shot, the two boxers continue their intense fight.

Shot 5 (2s): A bird's-eye view of the scene shows the two separated and having stopped fighting.

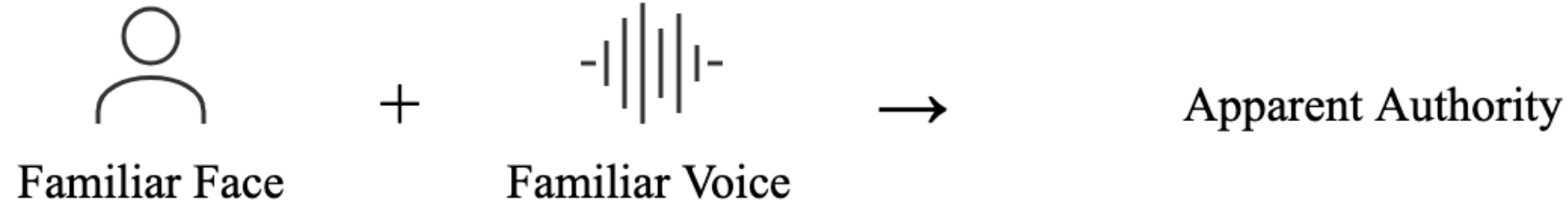
What should stay consistent when the camera changes?

The Challenge: Not Just Realism

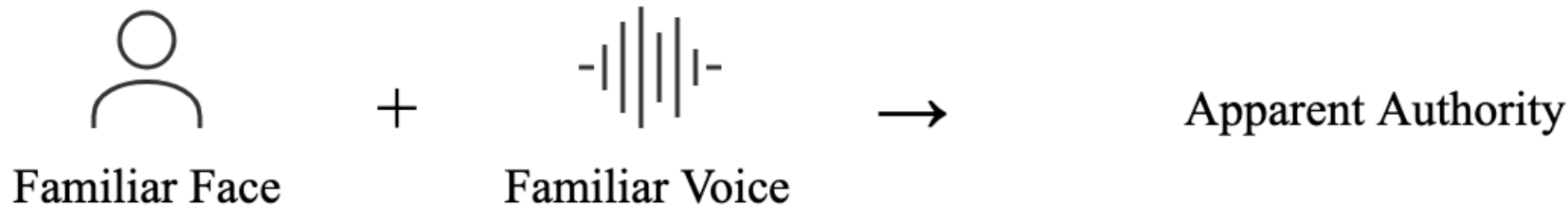


Do the people, place and action stay consistent across shots?

When Seeing and Hearing Are Not Proof



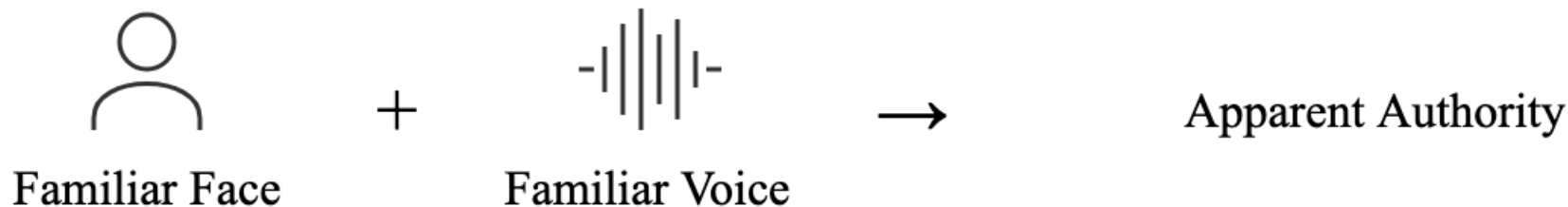
When Seeing and Hearing Are Not Proof



Hong Kong · 2024 · Deepfake Video-Conference Scam

HK\$200 Million Lost

When Seeing and Hearing Are Not Proof



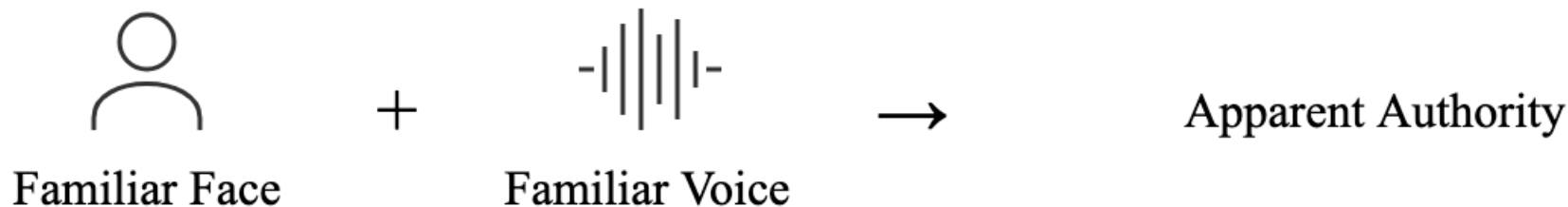
Hong Kong · 2024 · Deepfake Video-Conference Scam

HK\$200 Million Lost

What makes this different from using the same technology in a film?

Consent · Deception · Harm

When Seeing and Hearing Are Not Proof



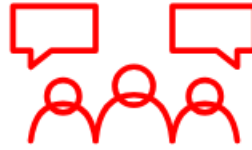
Hong Kong · 2024 · Deepfake Video-Conference Scam

HK\$200 Million Lost

What makes this different from using the same technology in a film?

Consent · Deception · Harm

Verify Consequential Requests
Through an Independently Trusted Channel



1 Ed Login
on your phone



2 Camera App



3 Scan QR



4 Submit

ACTIVITY · EXIT TICKET



