Paper Reading
Minghao Liu
2023/07/31

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li Dongxu Li Silvio Savarese Steven Hoi
Salesforce Research

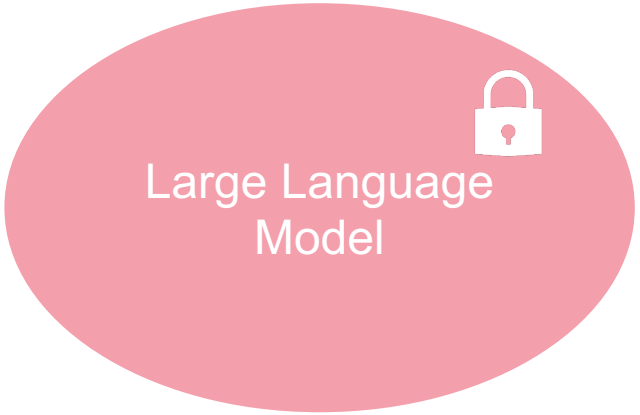
arXiv:2301.12597

Motivation

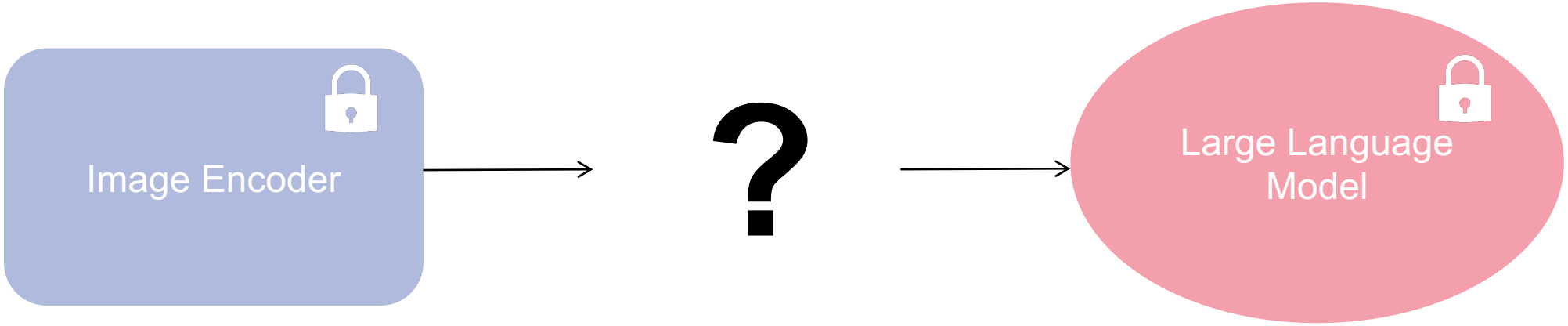
Image Encoder

Large Language Model

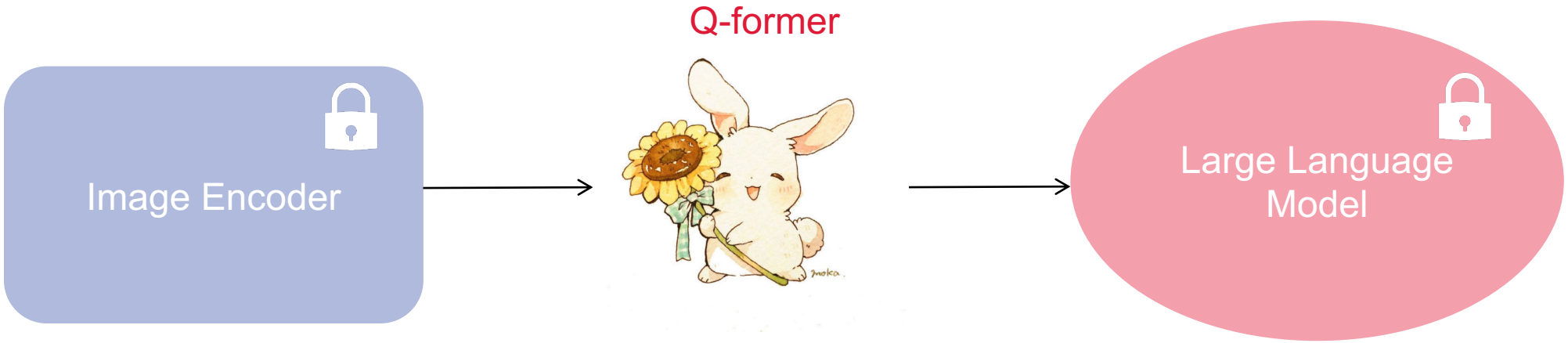
Motivation



Motivation



Motivation



Q-former

Large Language Model

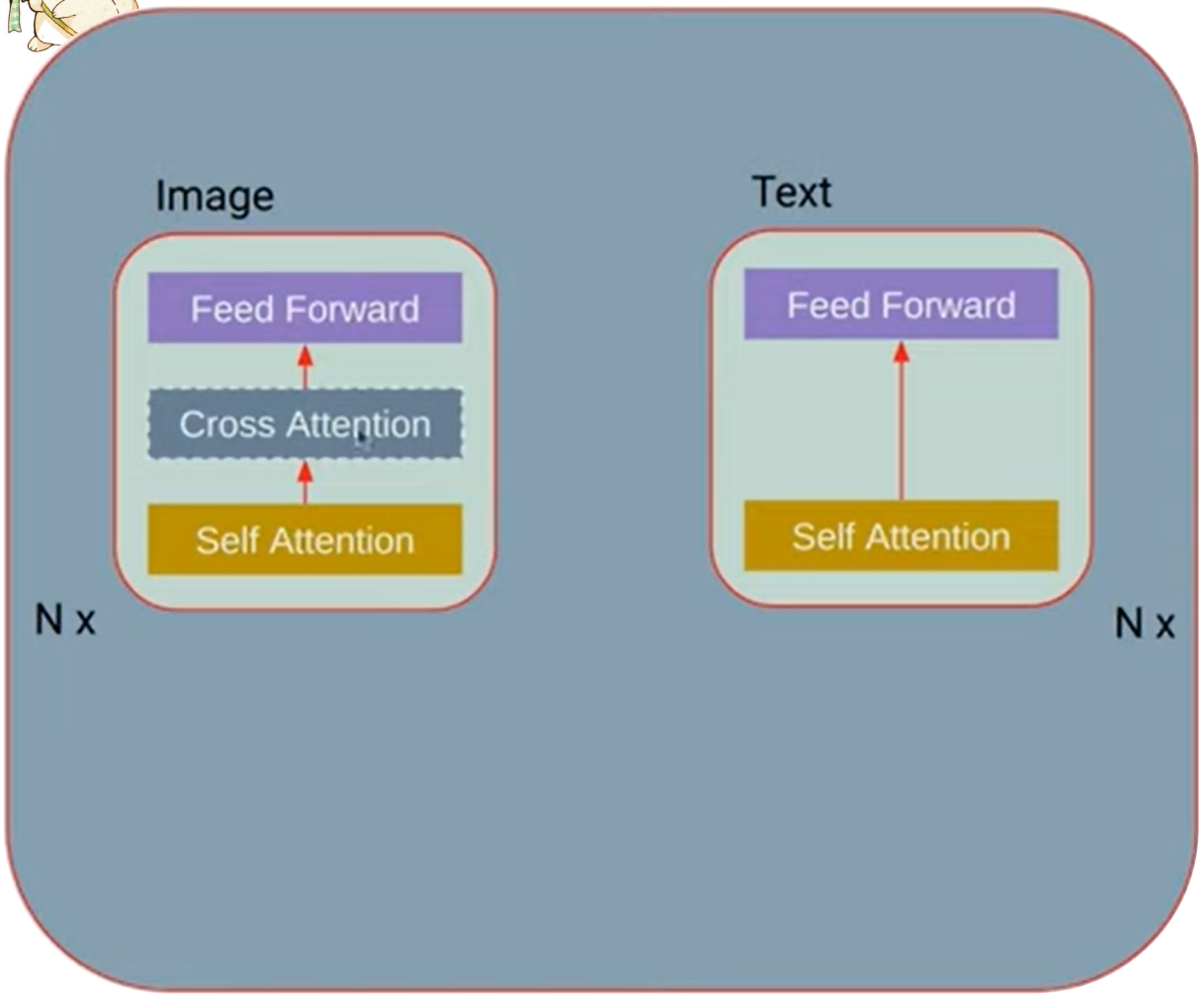


Image Encoder

Reference : https://www.youtube.com/watch?v=k0DAtZCCI1w

Q-former

Large Language Model

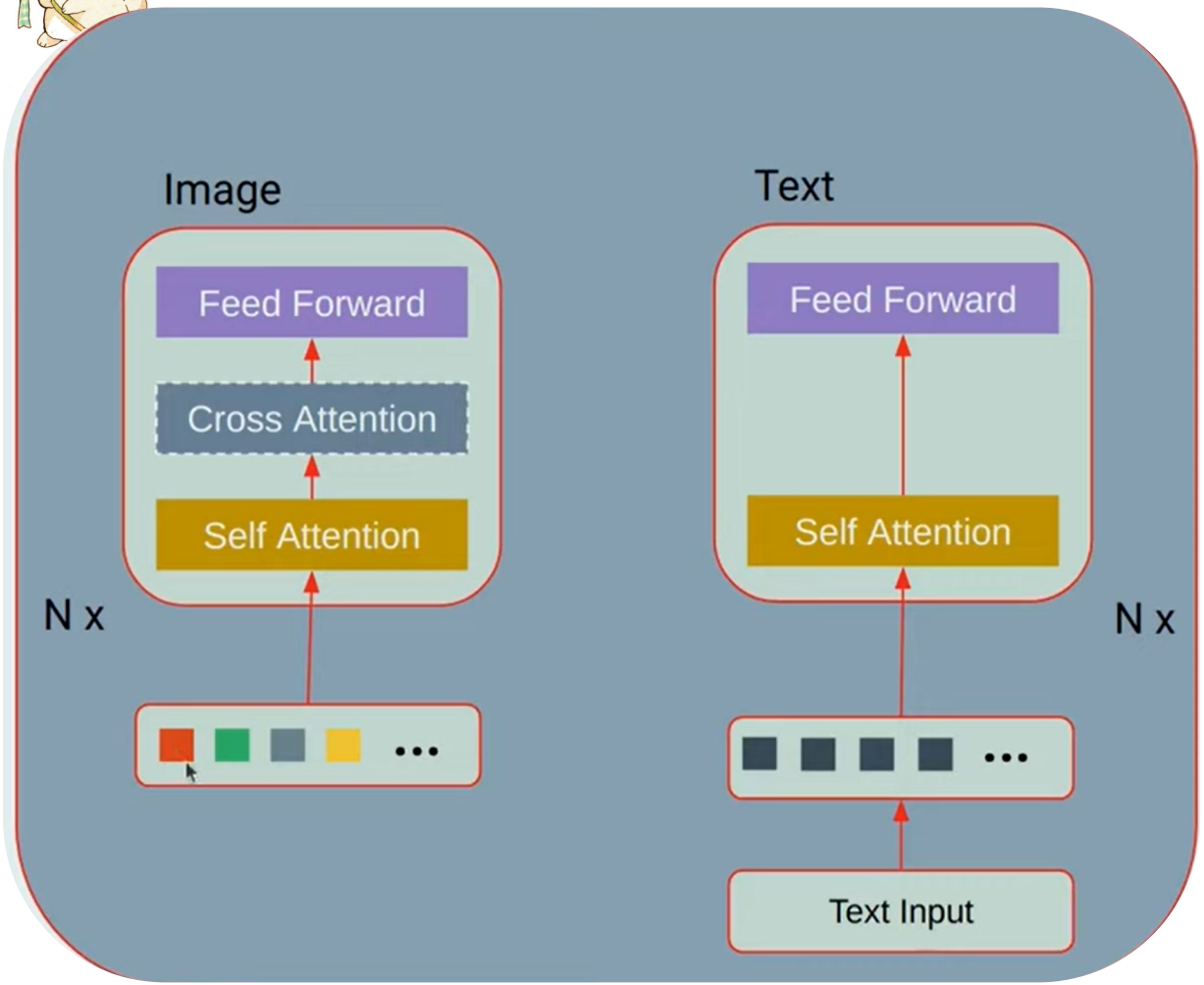


Image Encoder

Reference : https://www.youtube.com/watch?v=k0DAtZCCI1w

Q-former

Large Language Model

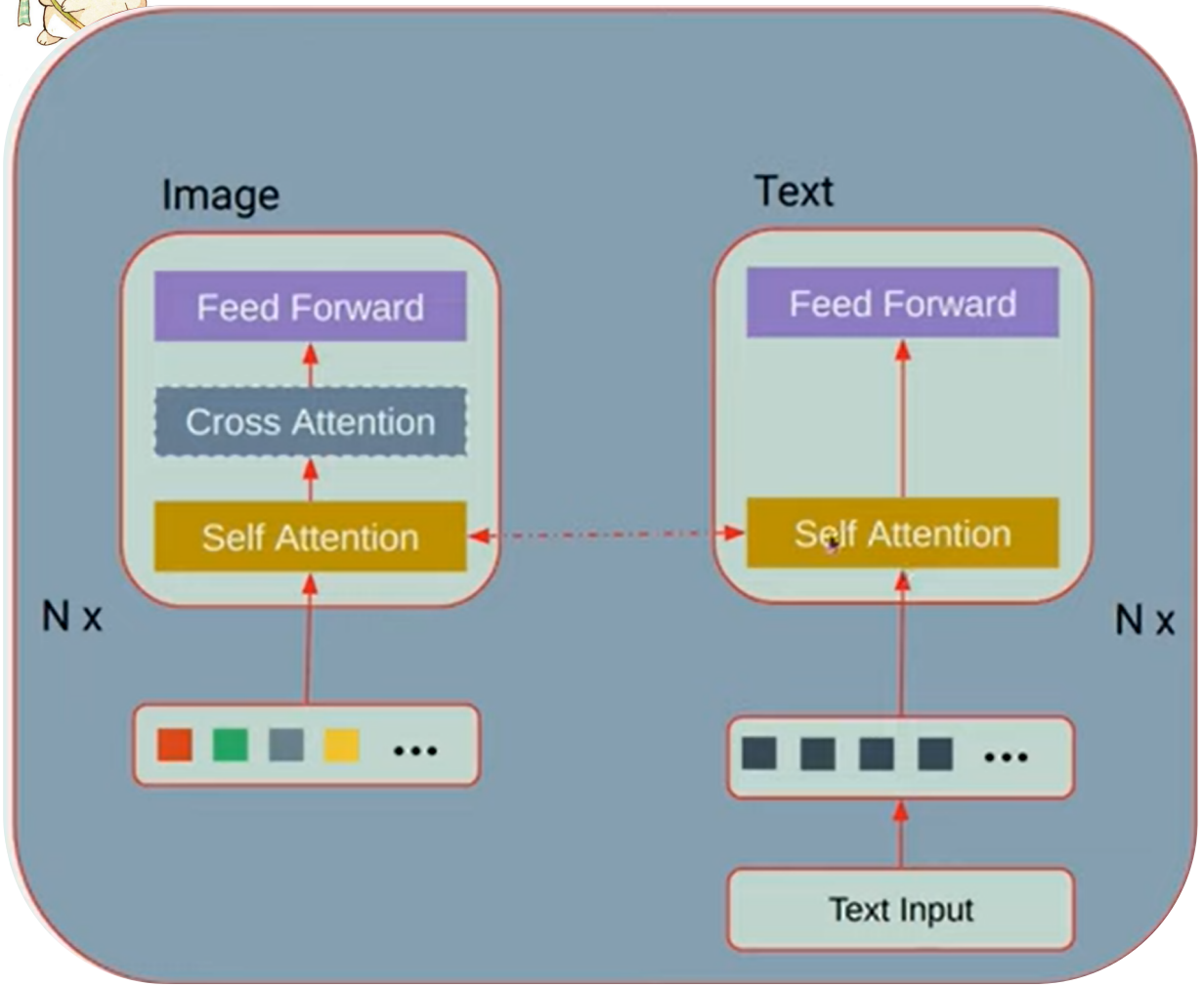


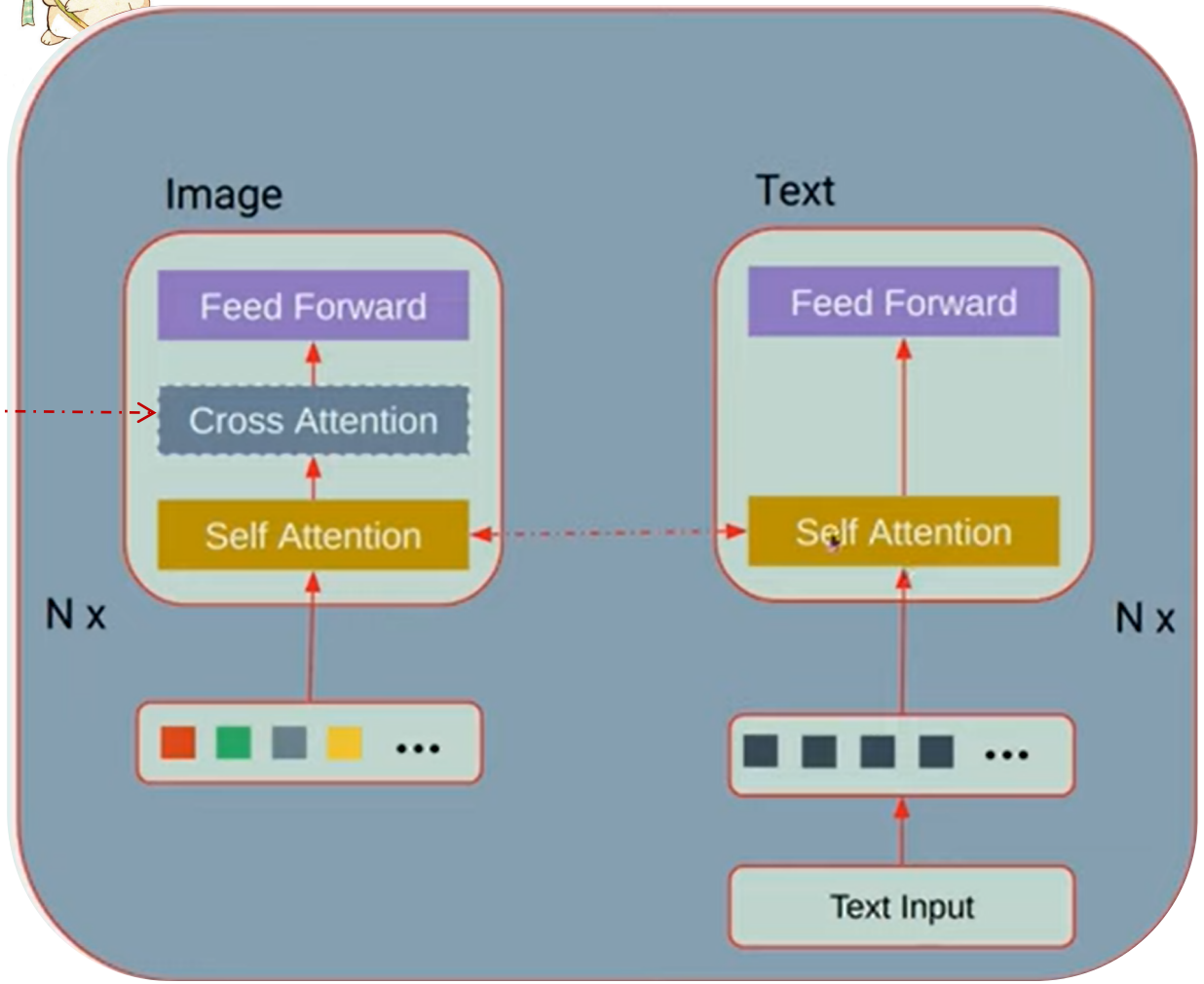
Image Encoder

Reference :
https://www.youtube.com/watch?v=k0DAtZCCI1w

Q-former

Large Language Model

Image Encoder

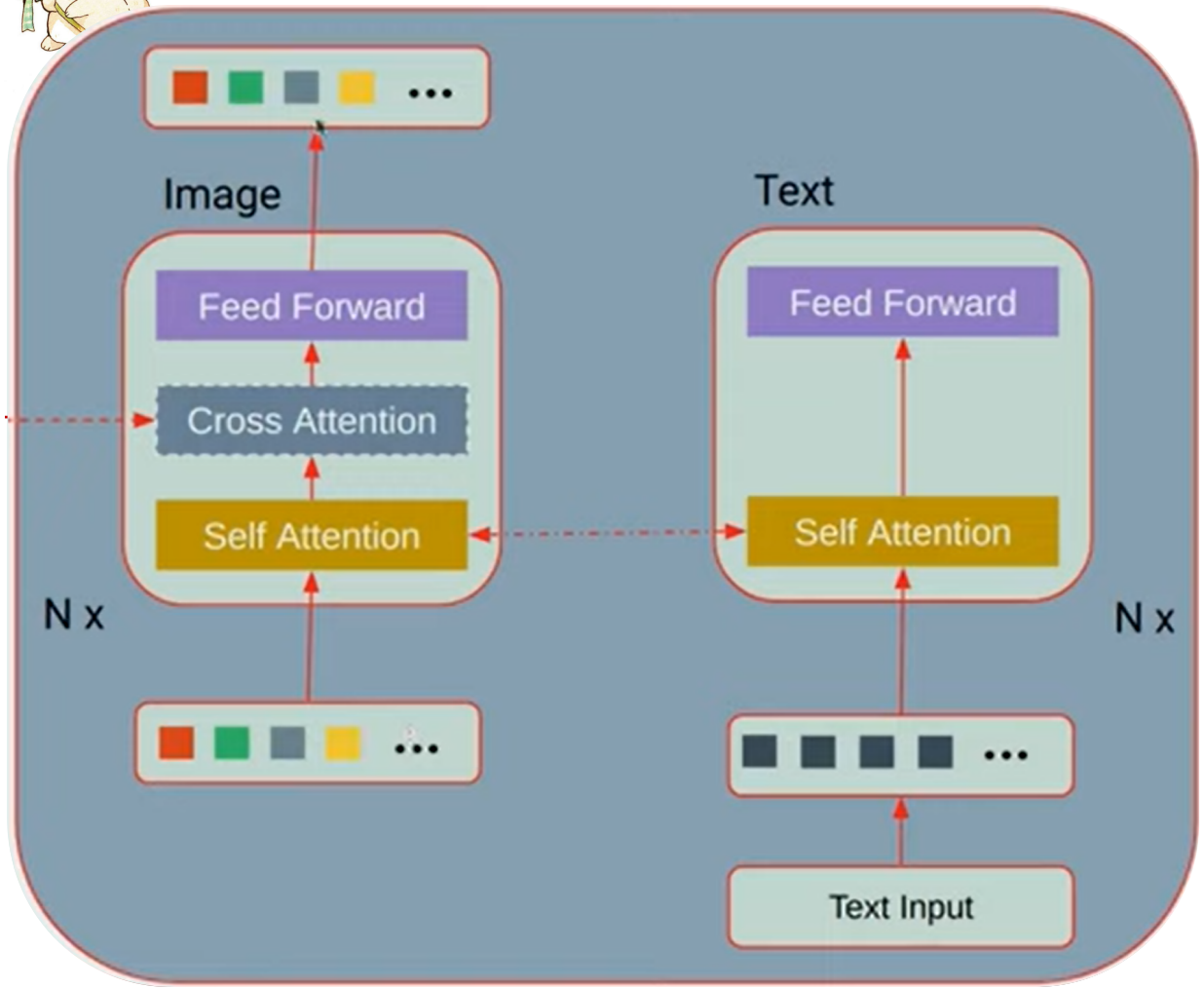


Reference :
https://www.youtube.com/watch?v=k0DAtZCCI1w

Q-former

Large Language Model

Image Encoder

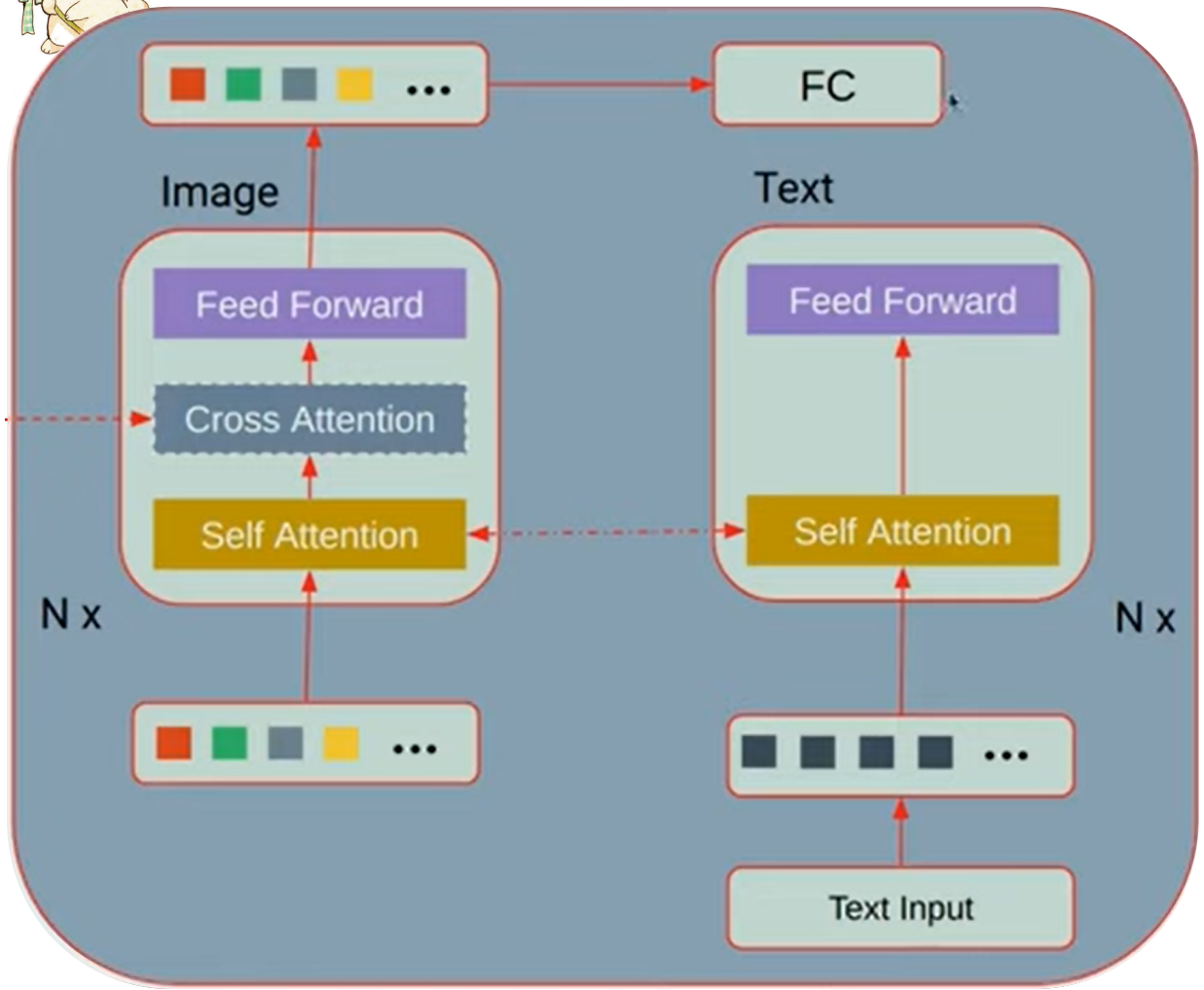


Reference : https://www.youtube.com/watch?v=k0DAtZCCI1w

Q-former

Large Language Model

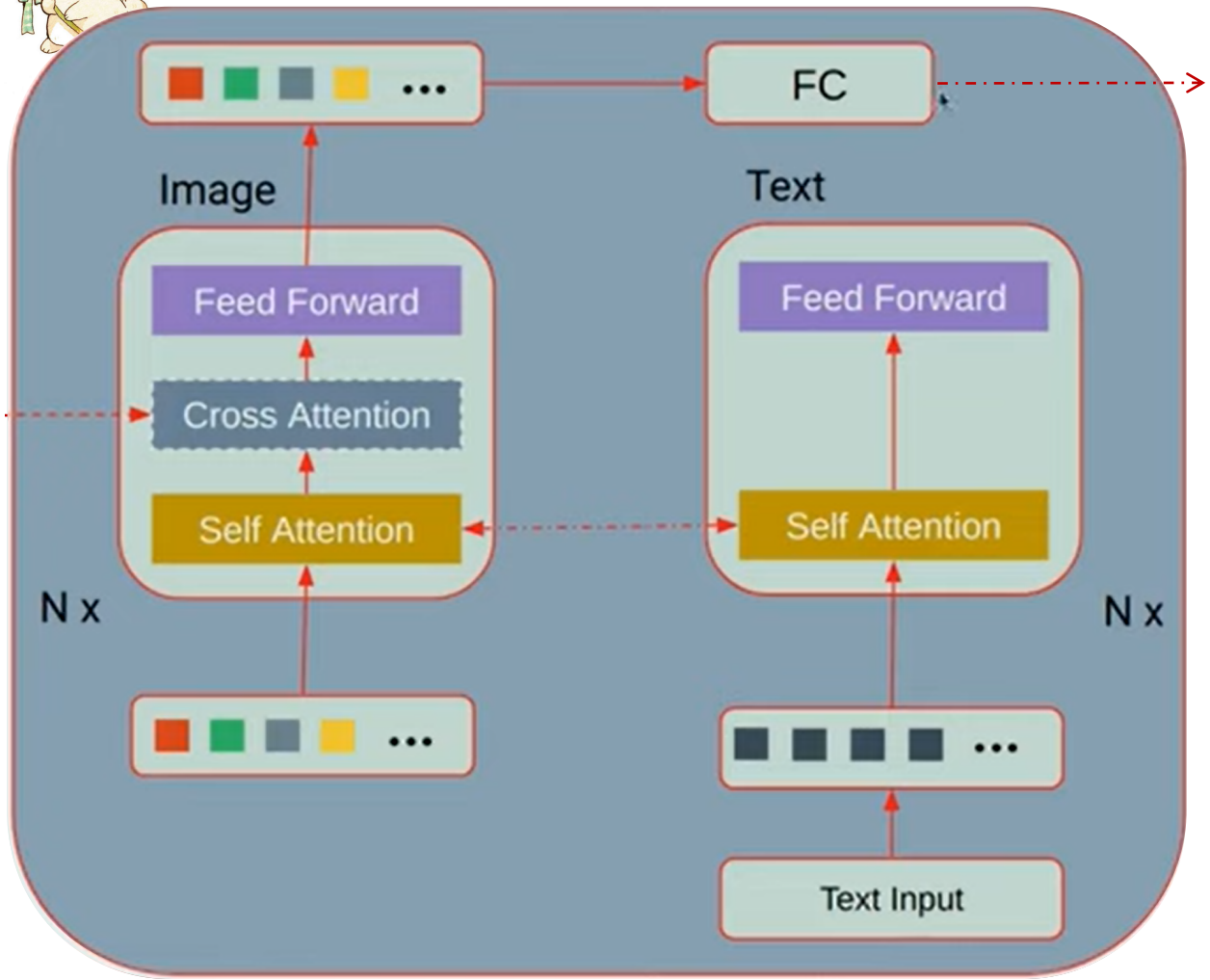
Image Encoder



Reference : https://www.youtube.com/watch?v=k0DAtZCCI1w

Q-former

Image Encoder

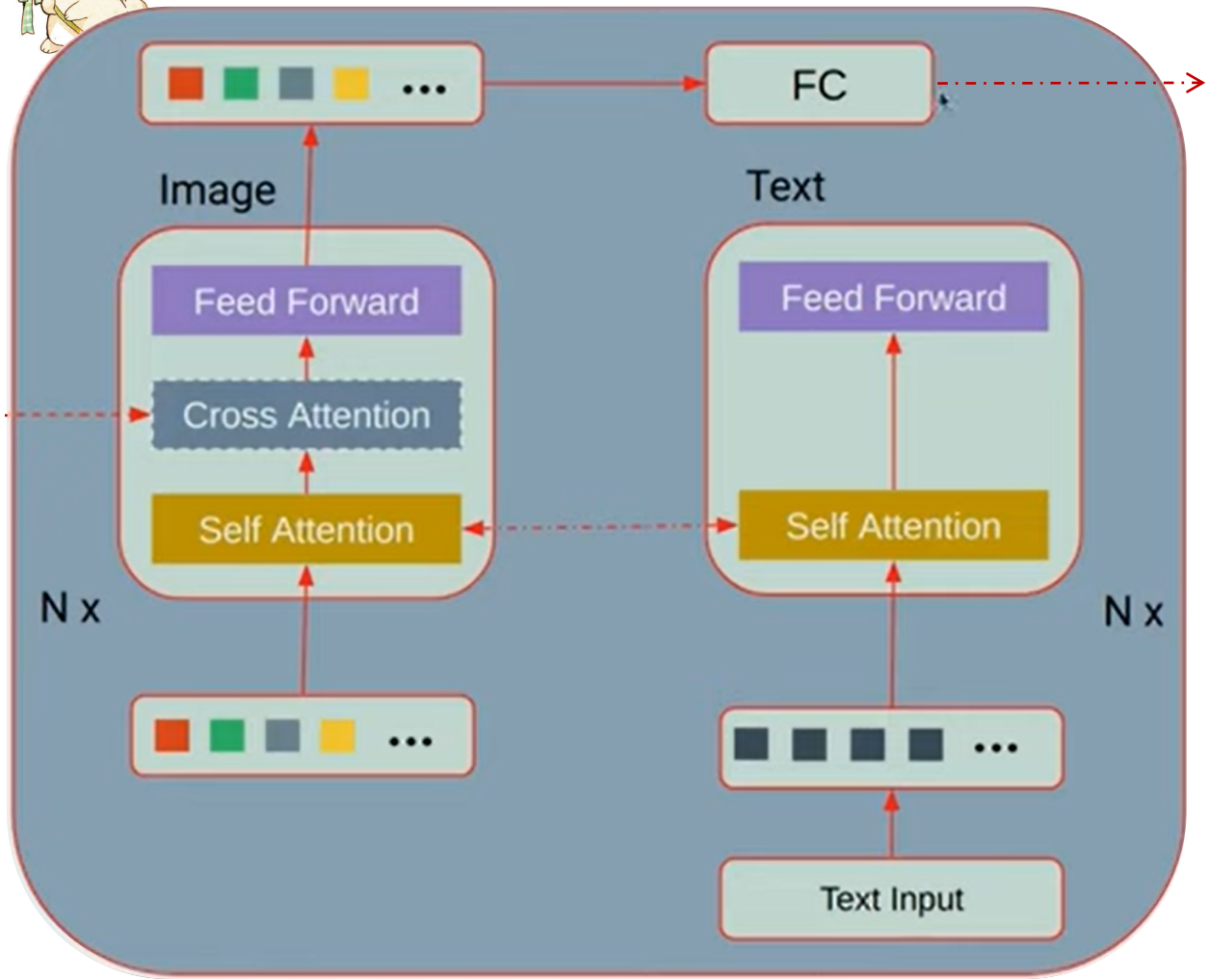


Large Language Model

Reference :
https://www.youtube.com/watch?v=k0DAtZCCI1w

Q-former

Image Encoder

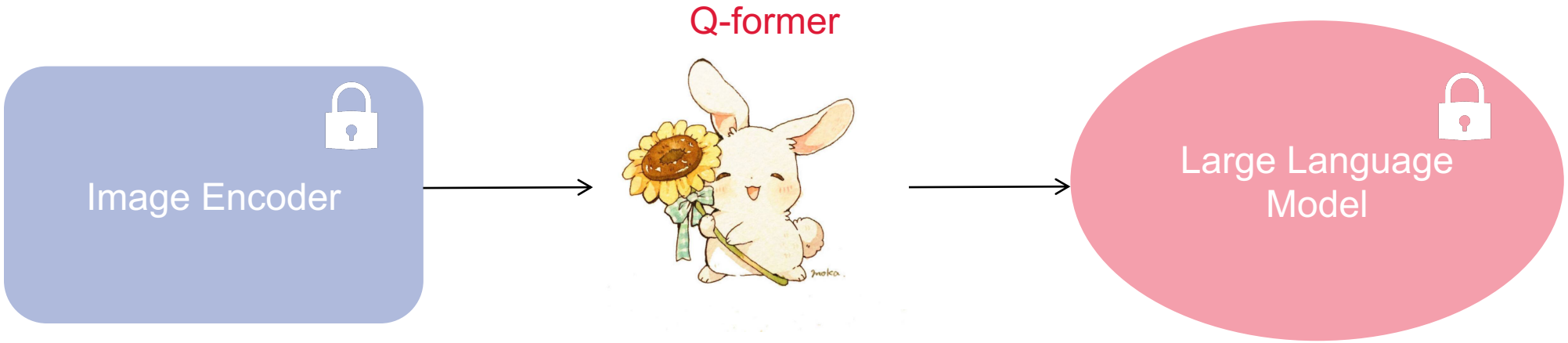


Large Language Model

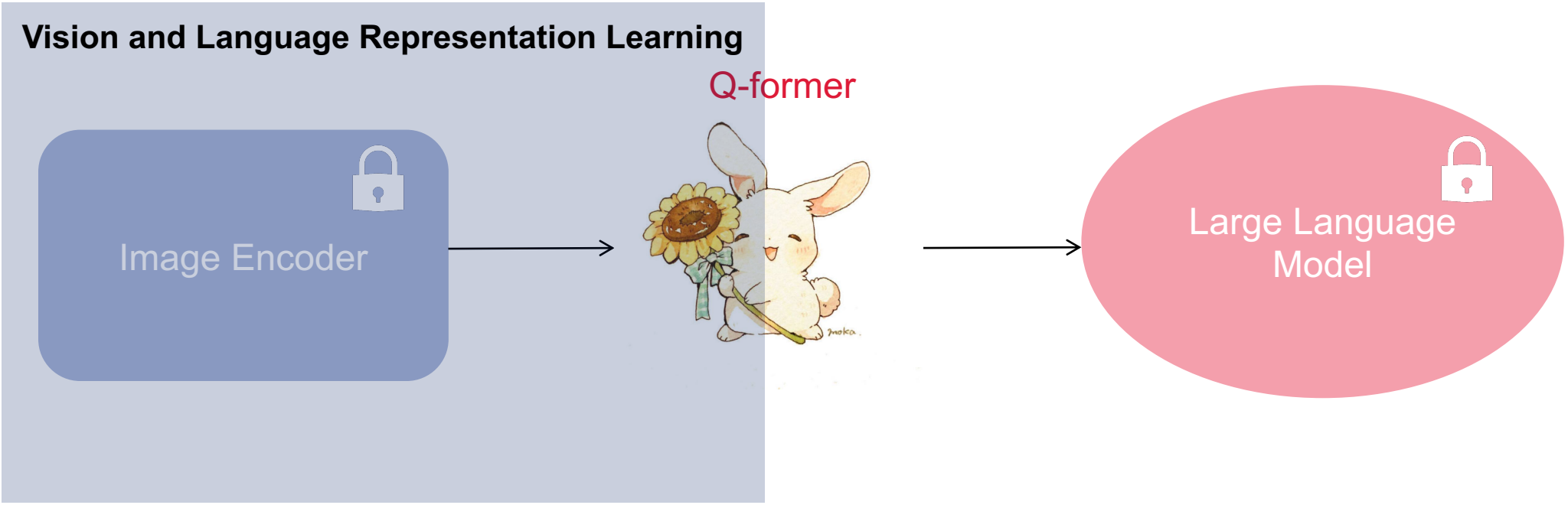
Text Output

Reference :
https://www.youtube.com/watch?v=k0DAtZCCI1w

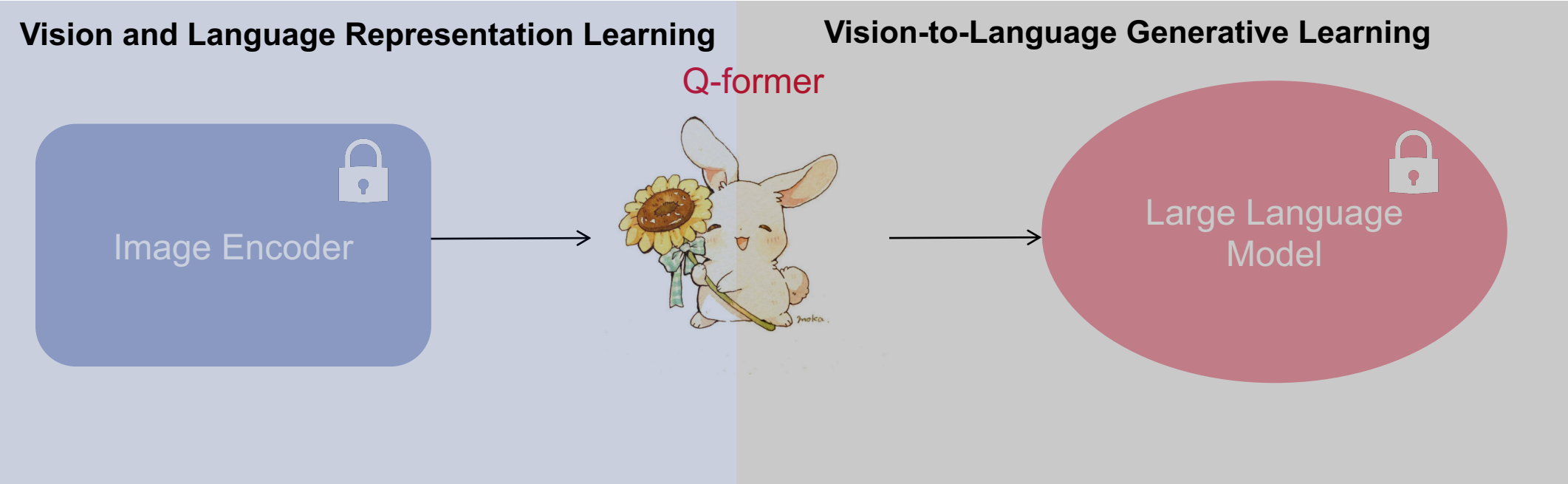
Training



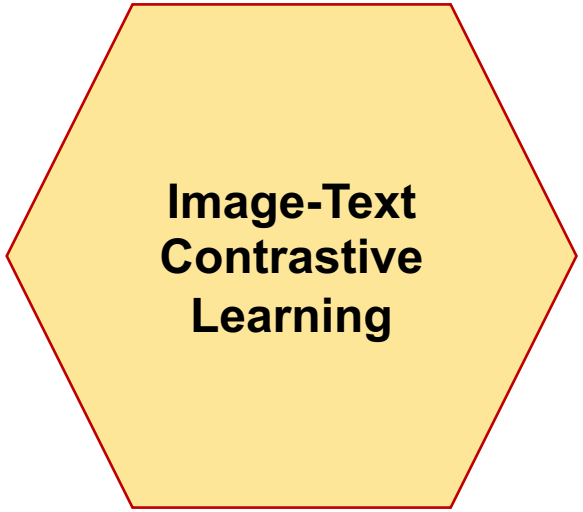
Training


Vision and Language Representation Learning
Q-former
Image Encoder
Large Language Model

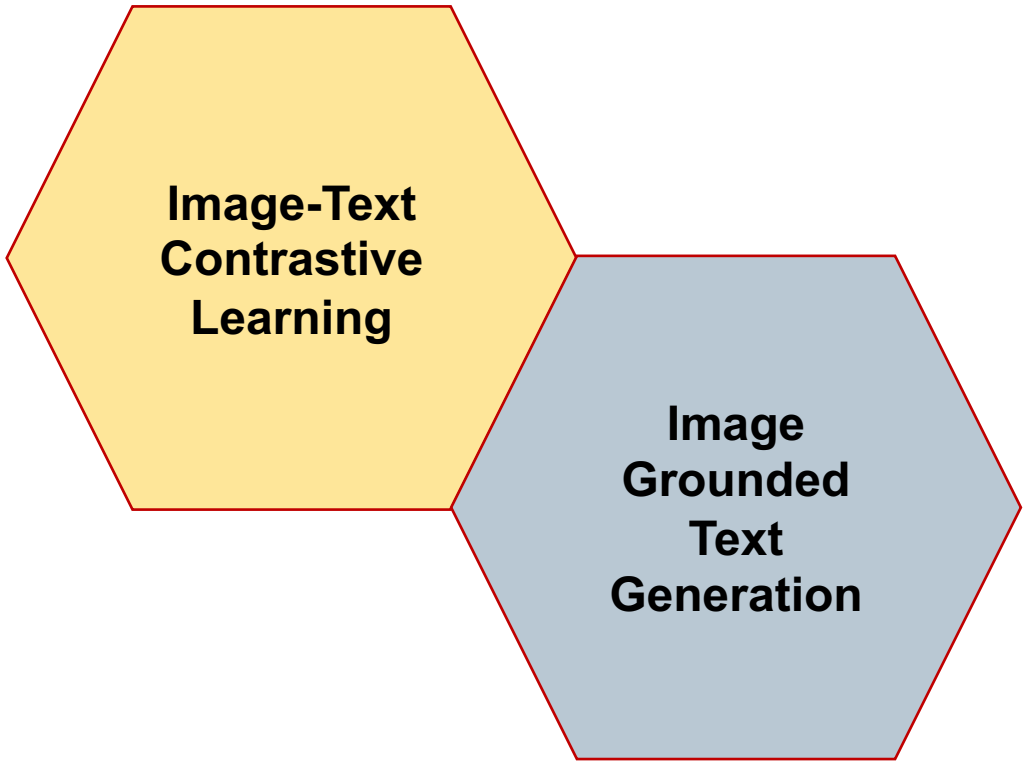
Training


Vision and Language Representation Learning
Vision-to-Language Generative Learning
Q-former
Image Encoder
Large Language Model

Vision and Language Representation Learning



Vision and Language Representation Learning



Vision and Language Representation Learning

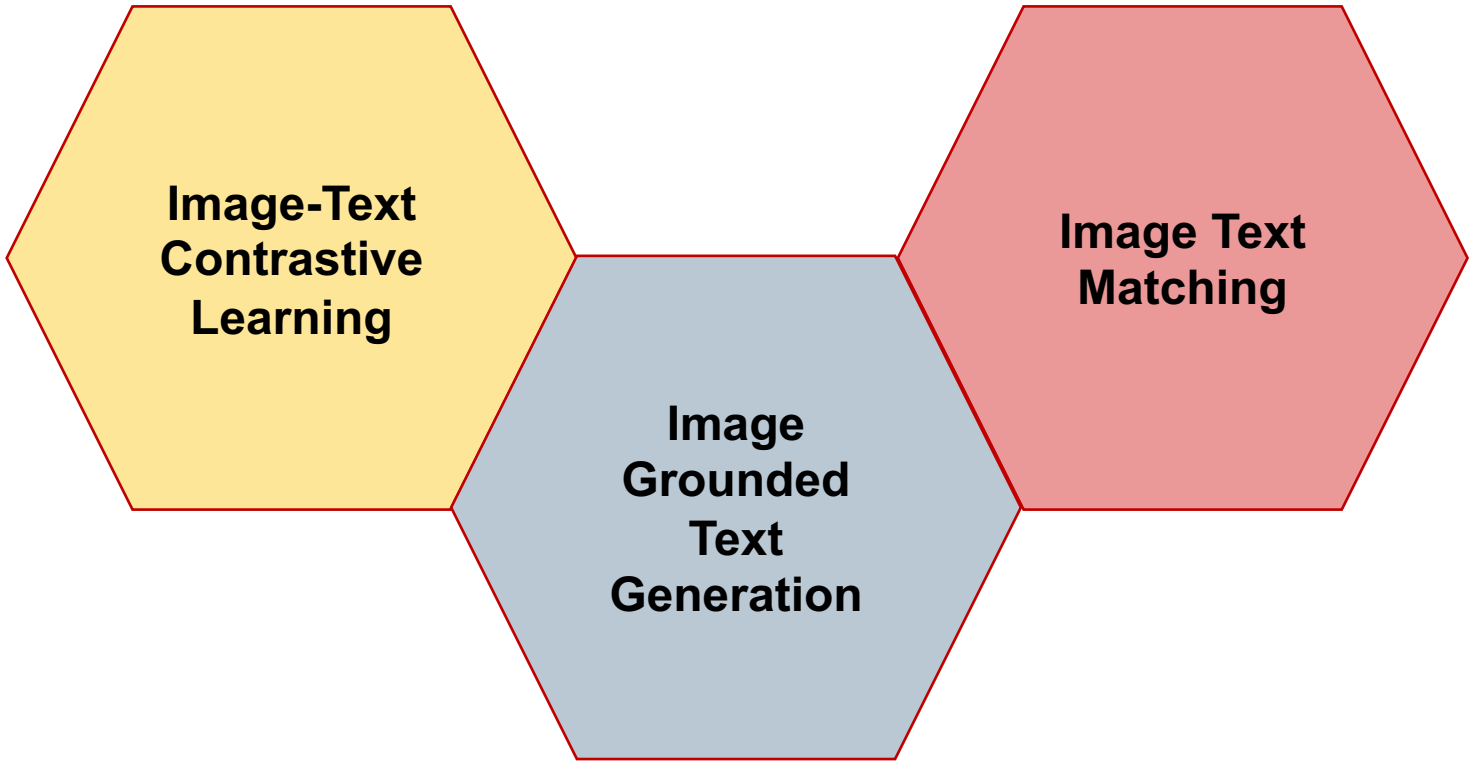


Image-Text Contrastive Learning

Image Encoder

Image-Text
Contrastive Learning

Image

Feed Forward

Cross Attention

Self Attention

N x

Text

Feed Forward

Self Attention

N x

Text Input

An eval AI overlord

Image-Text Contrastive Learning

Image Encoder

Image-Text
Contrastive Learning

Image

Feed Forward

Cross Attention

Self Attention

N x

Text

Feed Forward

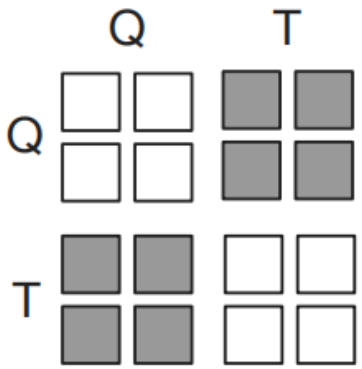
Self Attention

N x

Text Input

An eval AI overlord

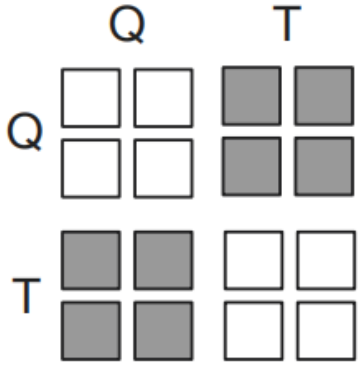
Uni-modal Self-Attention Mask



Uni-modal
Self-Attention Mask

Image-Text
Contrastive Learning

Uni-modal Self-Attention Mask



Uni-modal
Self-Attention Mask

Image-Text
Contrastive Learning

[CLS]

Image Grounded Text Generation

Image Encoder

Image Grounded Text Generation

Image

Feed Forward

Cross Attention

Self Attention

N x

Text

Feed Forward

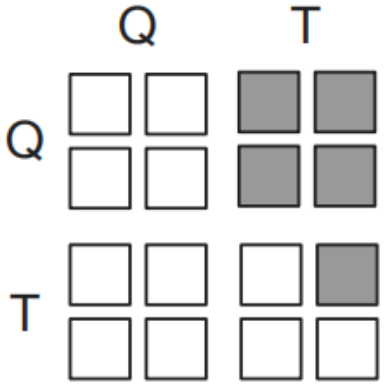
Self Attention

N x

Text Input

An eval AI overlord

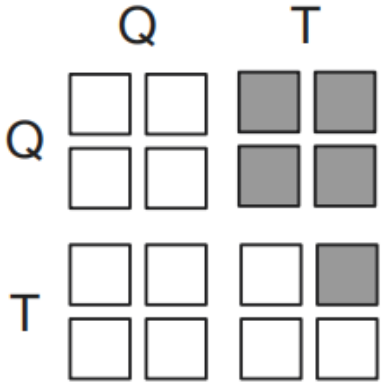
Multi-modal Causal Self-Attention Mask



Multi-modal Causal Self-Attention Mask

Image-Grounded Text Generation

Multi-modal Causal Self-Attention Mask



Multi-modal Causal Self-Attention Mask

Image-Grounded Text Generation

[DEC]

Image Text Matching

Image Encoder

Image Text Matching

Image

Feed Forward

Cross Attention

Self Attention

N x

Text

Feed Forward

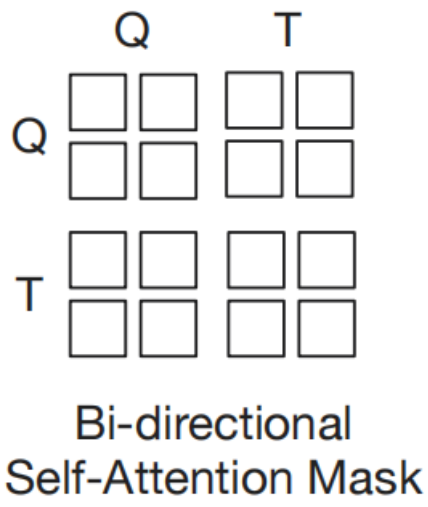
Self Attention

N x

Text Input

An eval AI overlord

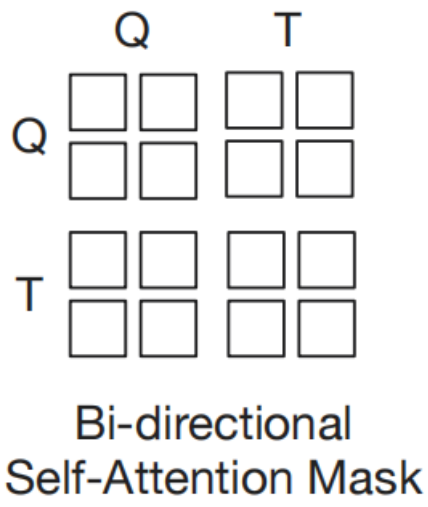
Bi-directional Self-Attention Mask



Bi-directional
Self-Attention Mask

Image-Text
Matching

Bi-directional Self-Attention Mask

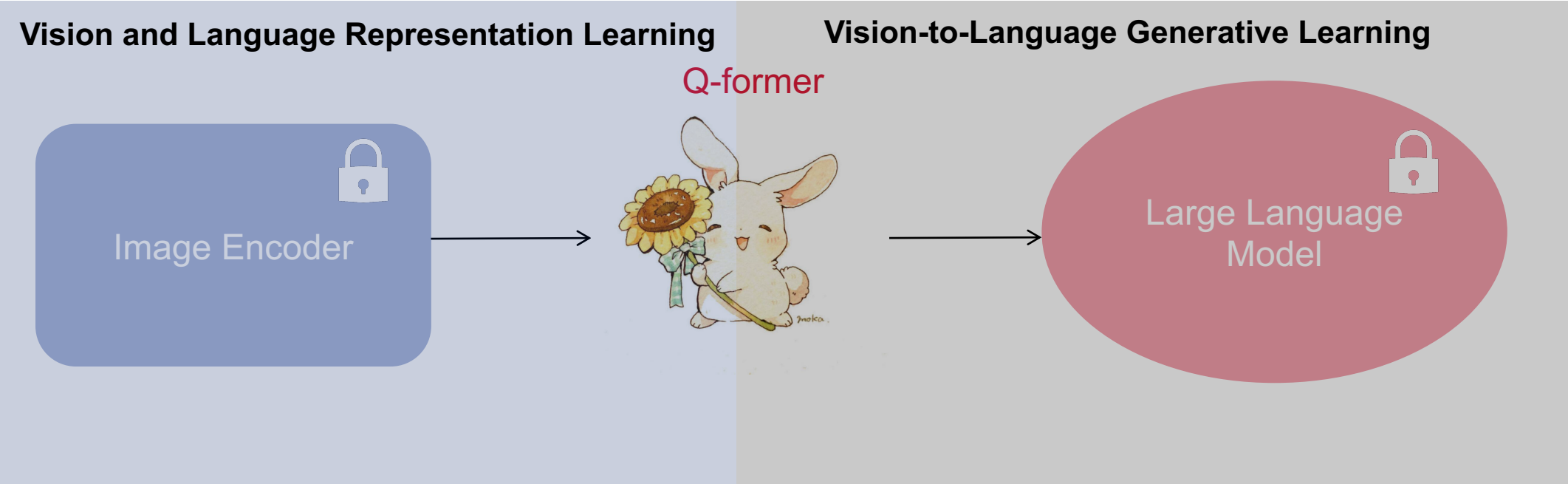


Bi-directional
Self-Attention Mask

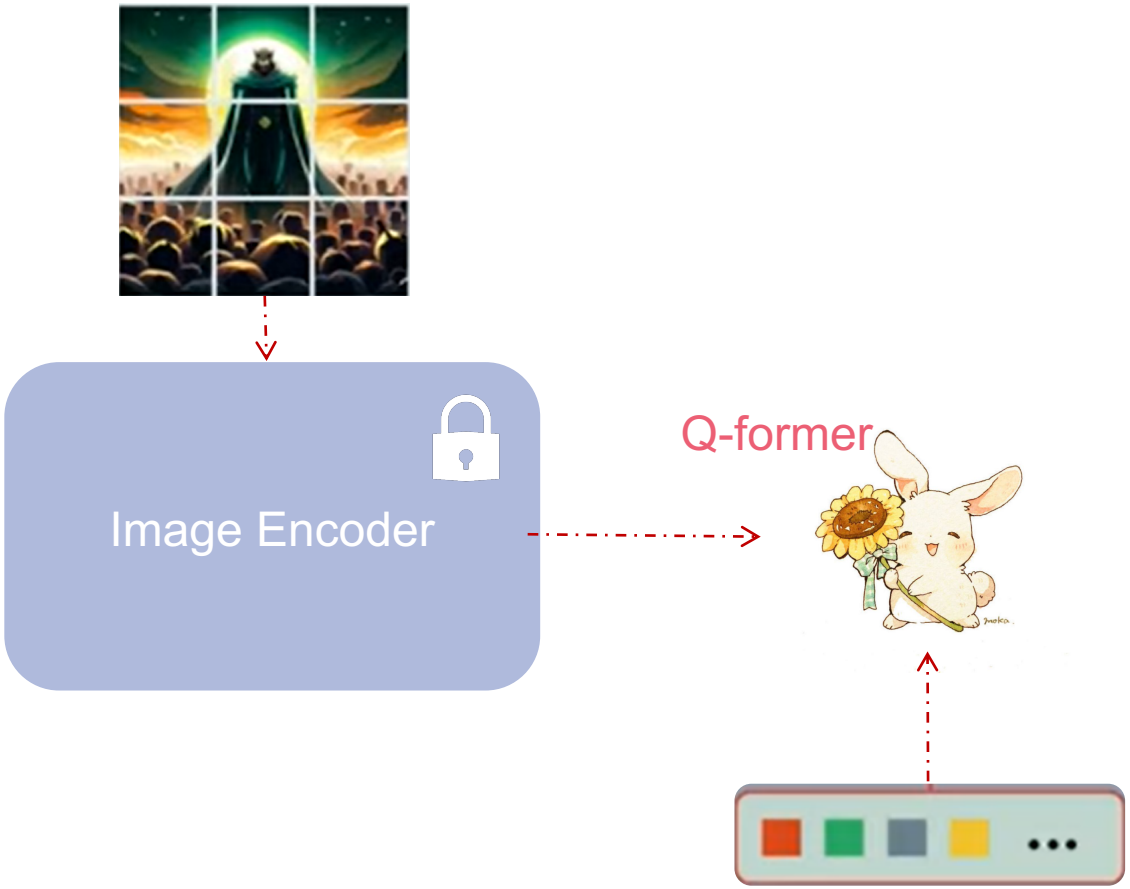
Image-Text
Matching

[CLS]

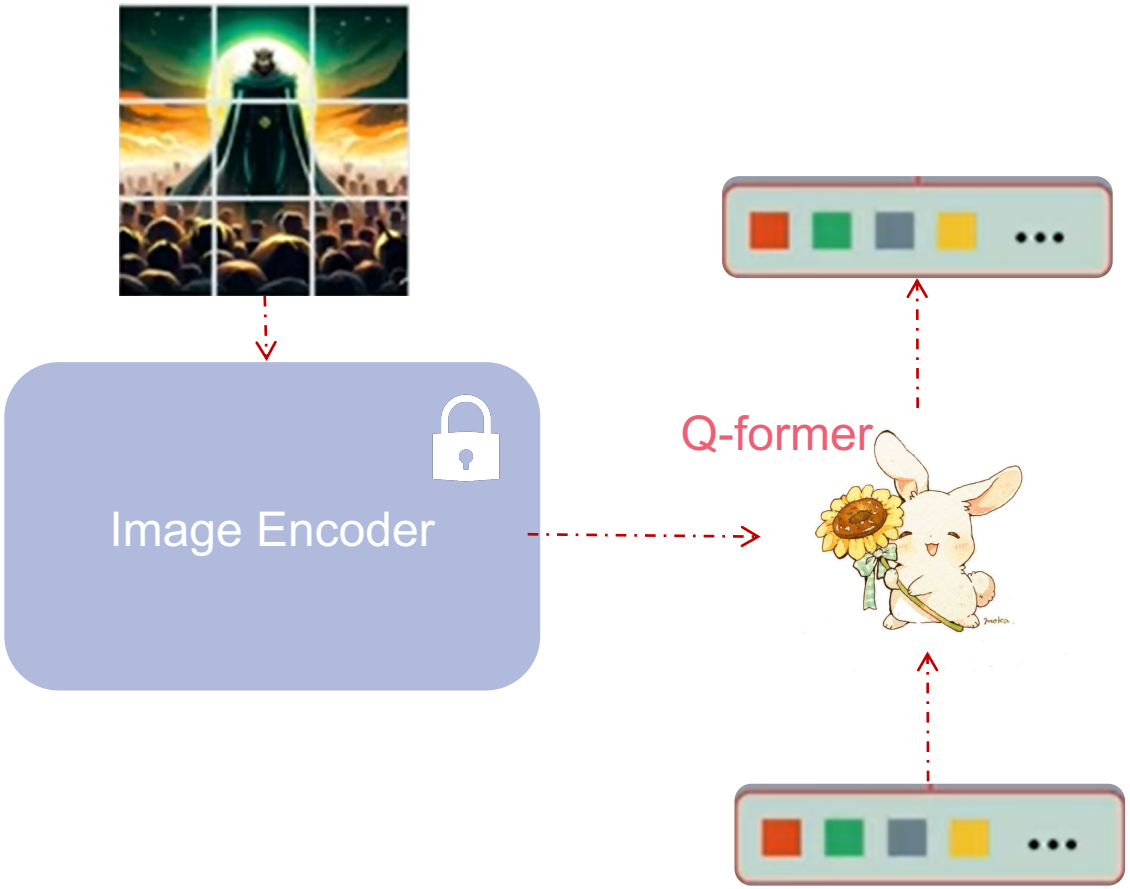
Training


Vision and Language Representation Learning
Vision-to-Language Generative Learning
Q-former
Image Encoder
Large Language Model

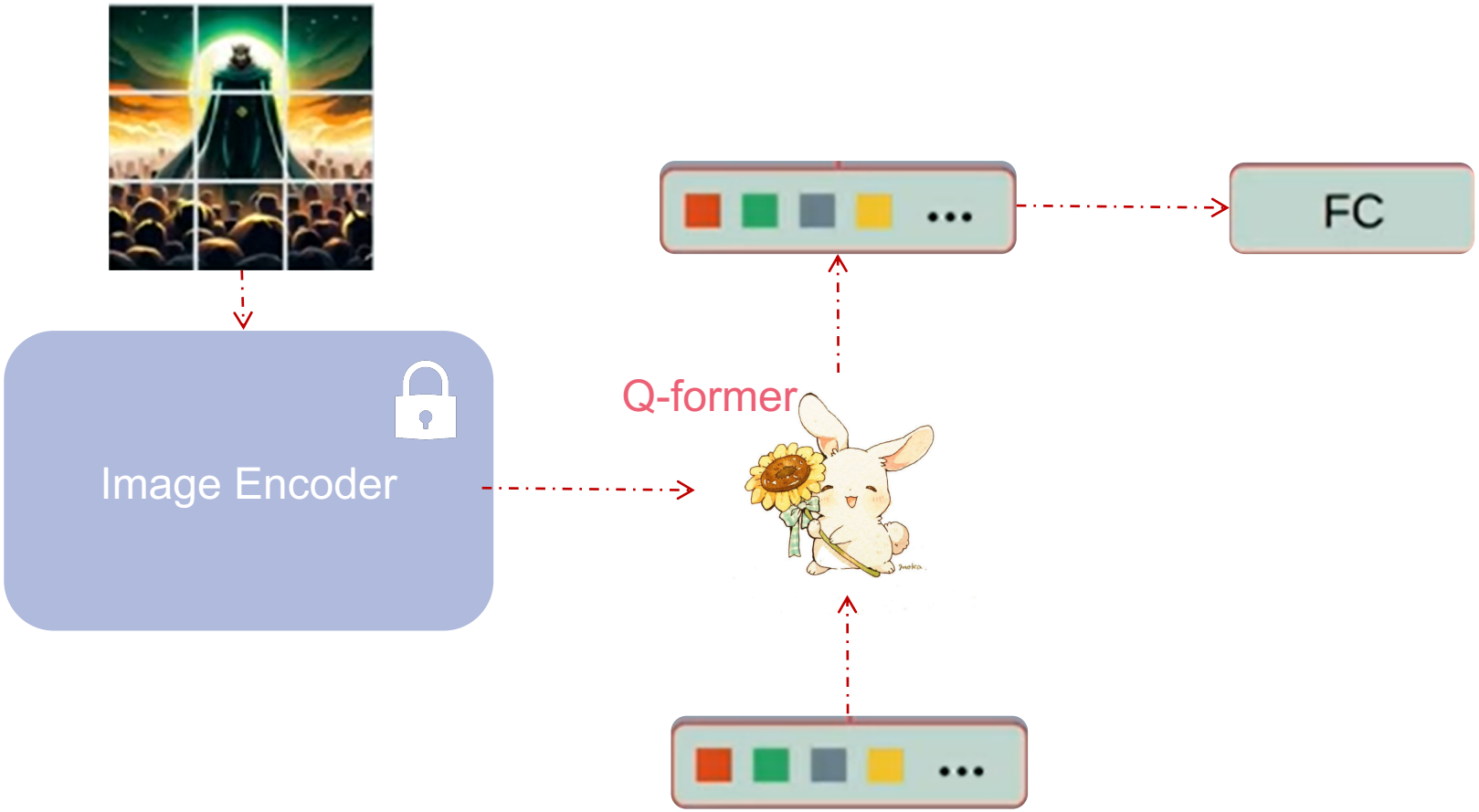
Vision-to-Language Generative Learning



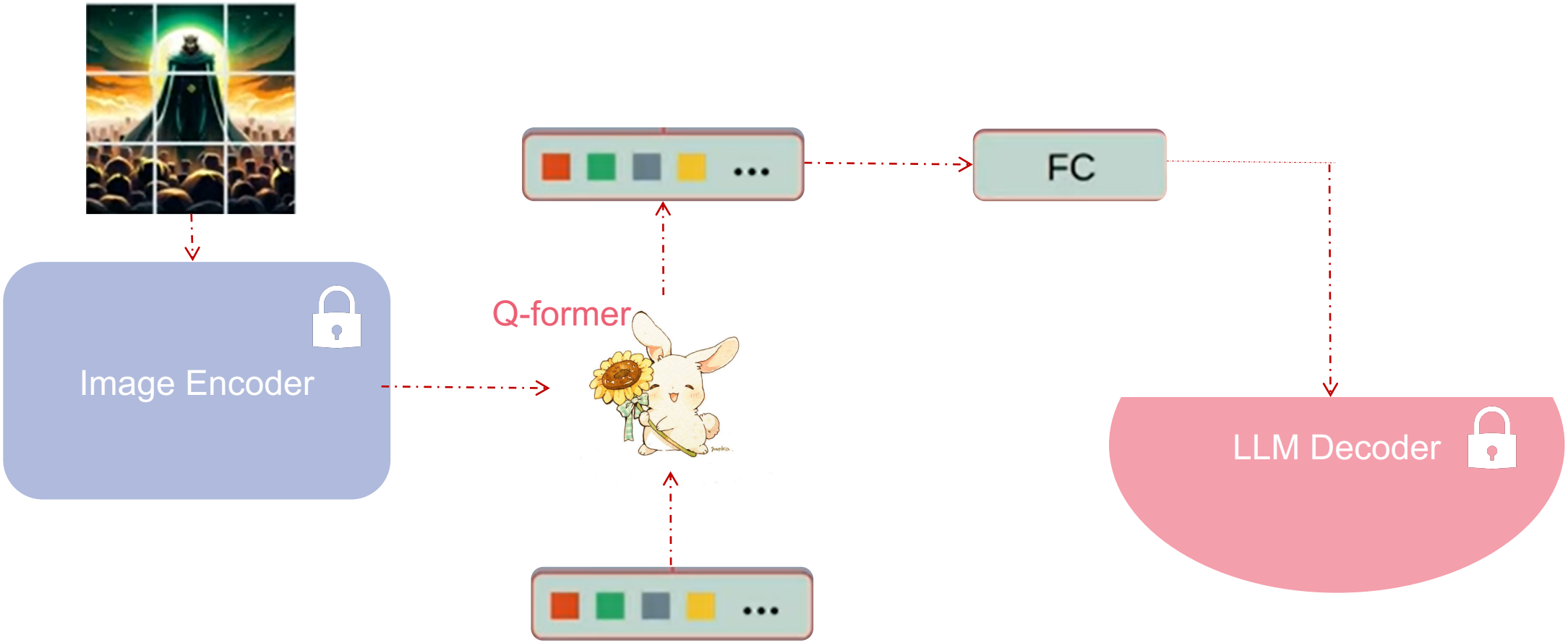
Vision-to-Language Generative Learning



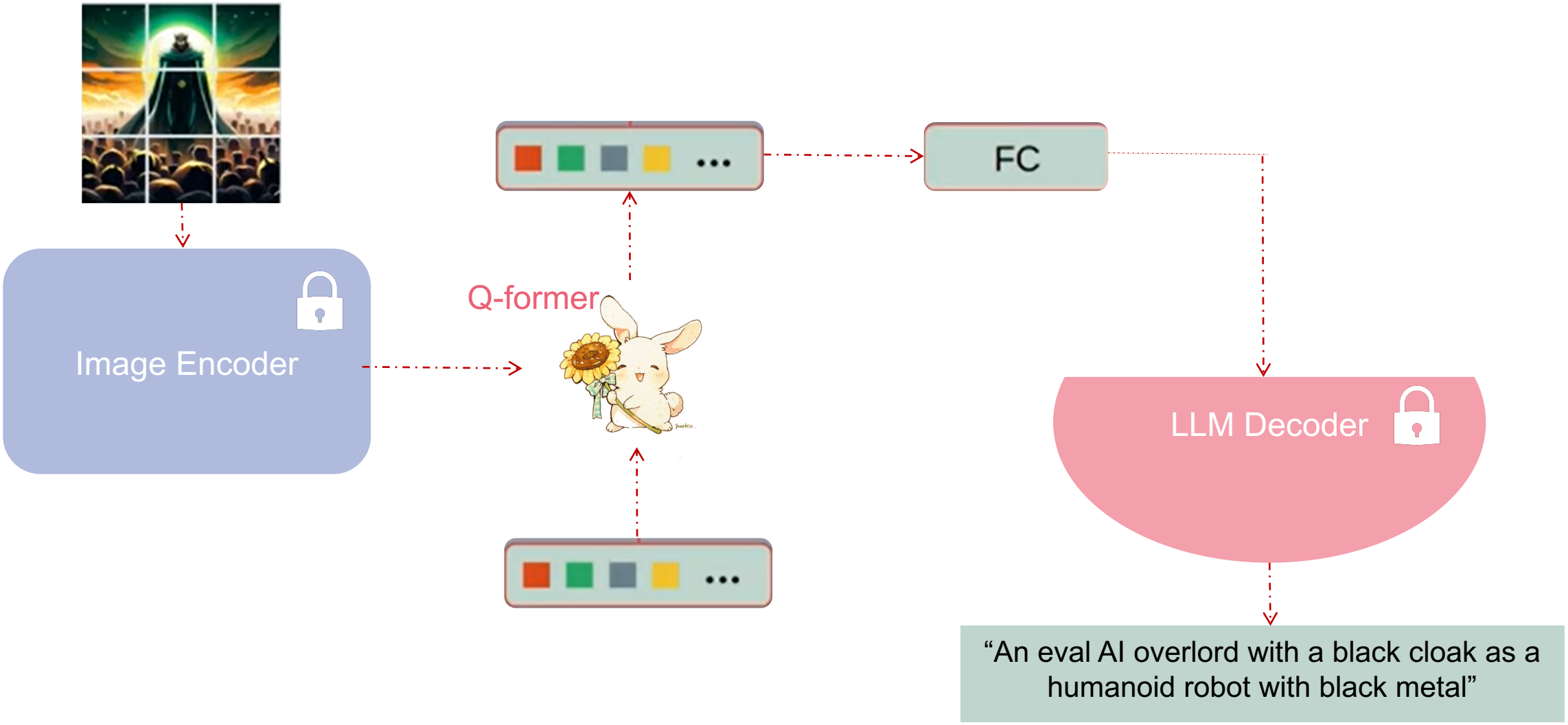
Vision-to-Language Generative Learning



Vision-to-Language Generative Learning

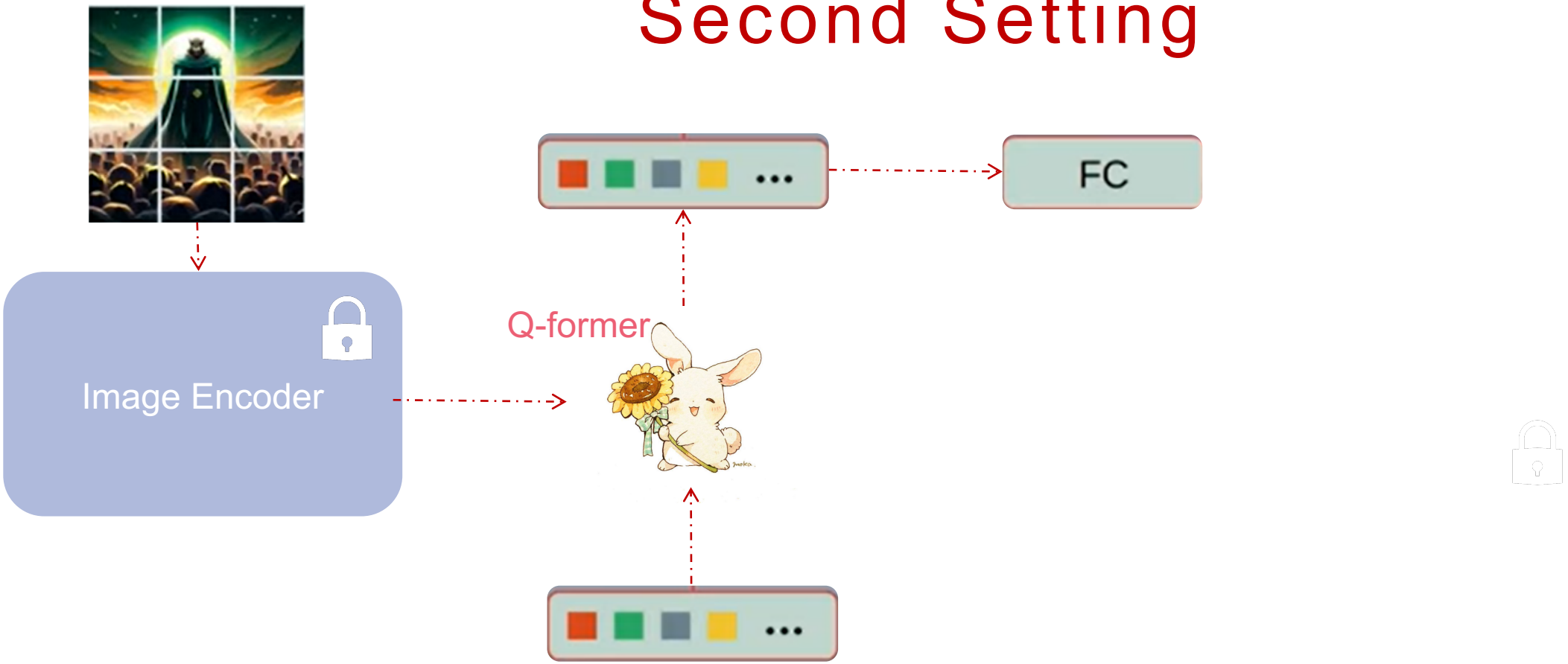


Vision-to-Language Generative Learning



Vision-to-Language Generative Learning

Second Setting

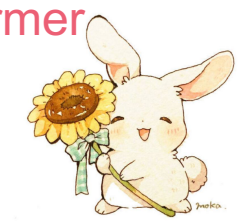


Vision-to-Language Generative Learning

Second Setting

Image Encoder

Q-former

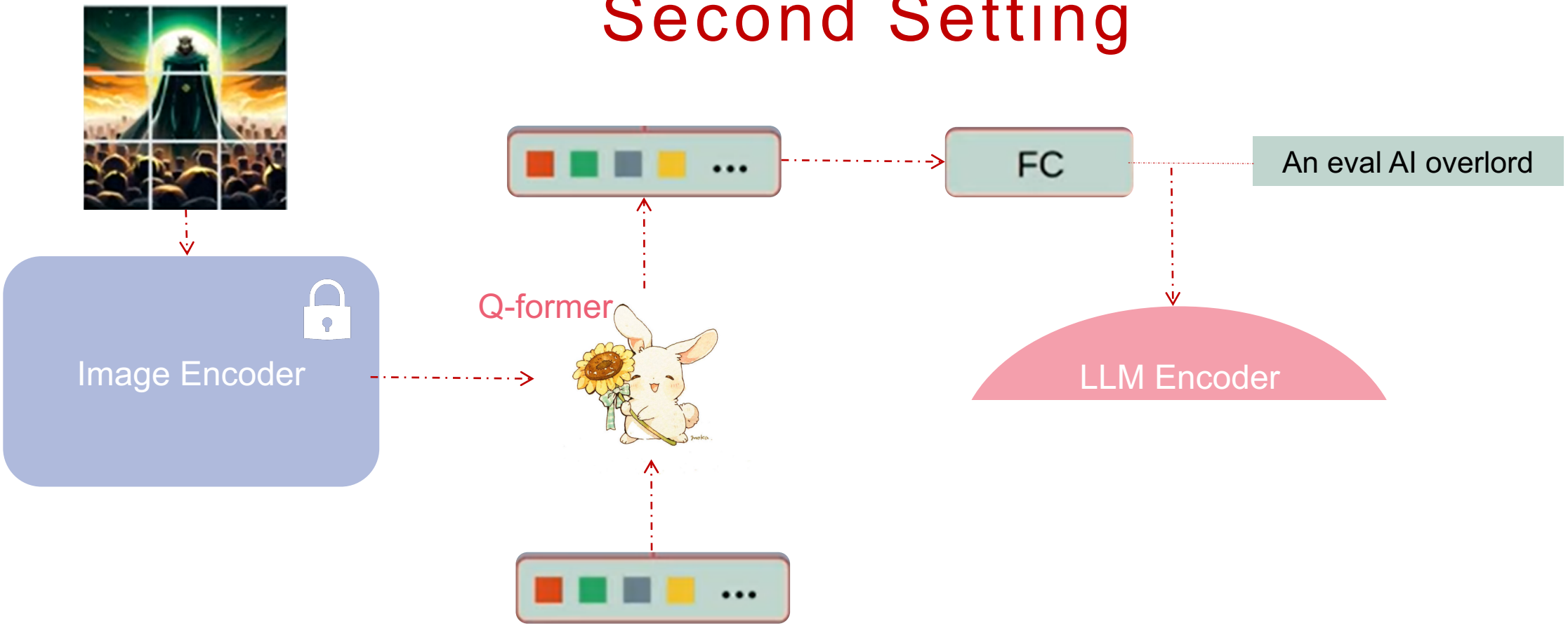


FC

An eval AI overlord

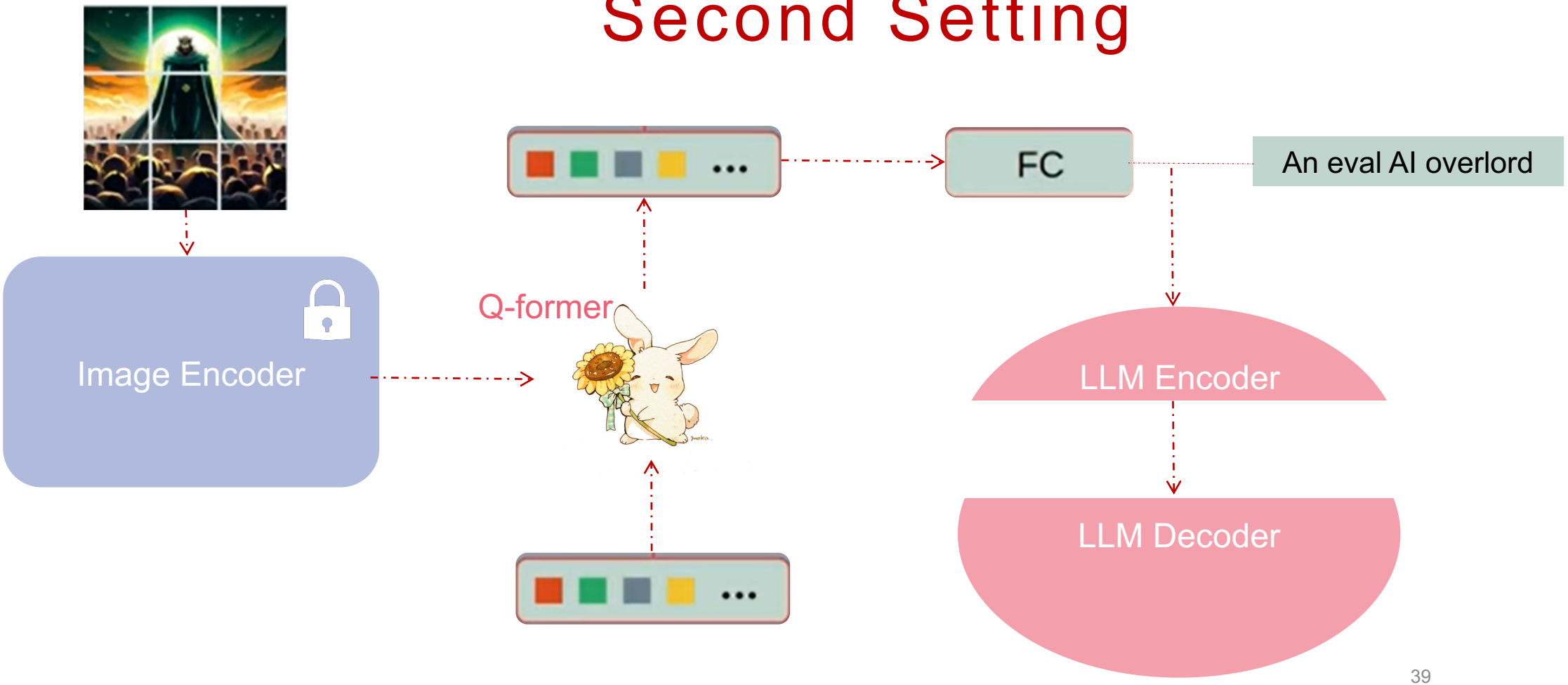
Vision-to-Language Generative Learning

Second Setting



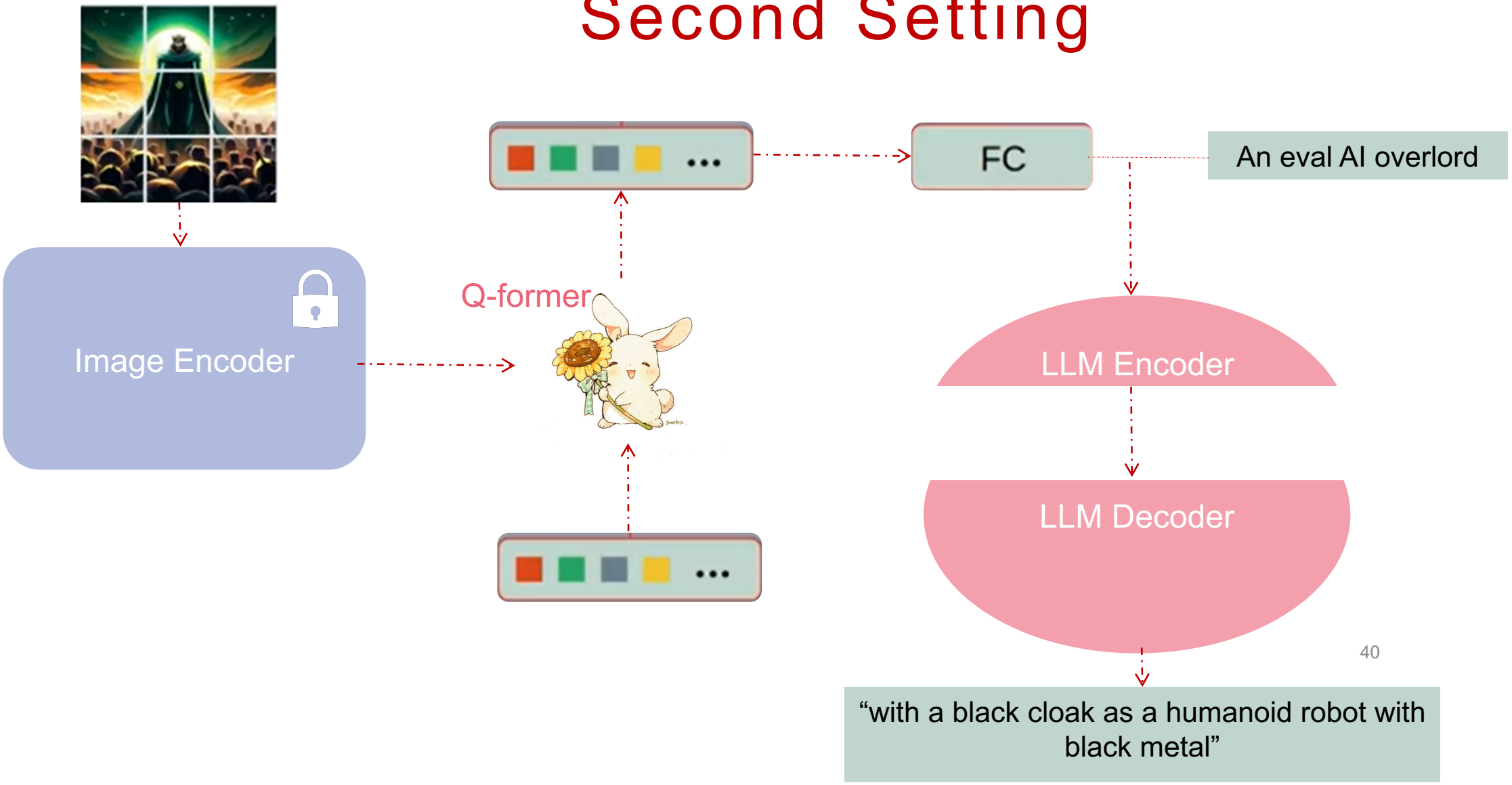
Vision-to-Language Generative Learning

Second Setting



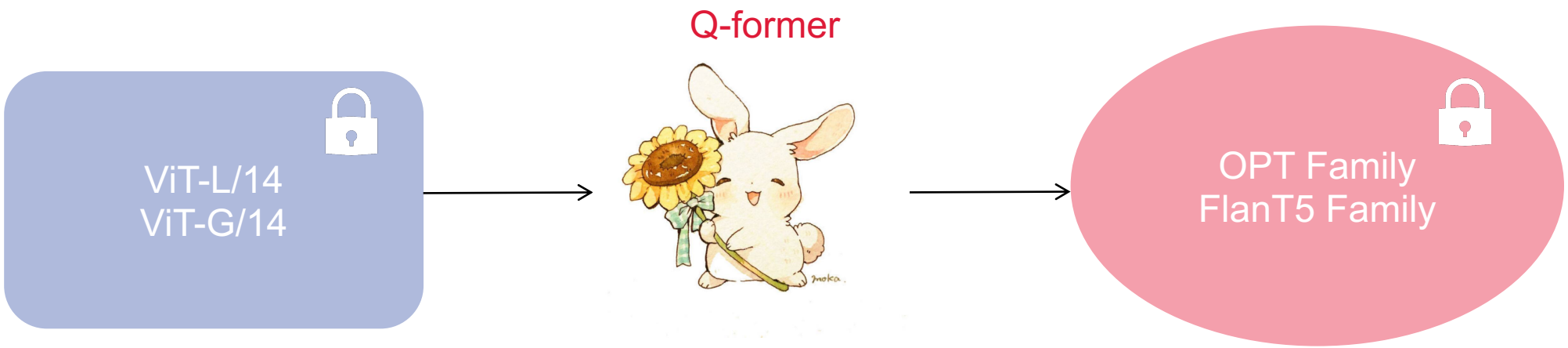
Vision-to-Language Generative Learning

Second Setting



Experiment-Architecture

Data: 6 datasets (129M), some from web.



Experiment-VQA-Zero-Shot

Models	#Trainable Params	#Total Params	VQAv2		OK-VQA	GQA
			val	test-dev	test	test-dev
VL-T5$_{\text{no-vqa}}$	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	**50.6**	-
BLIP-2 ViT-L OPT$_{2.7B}$	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT$_{2.7B}$	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT$_{6.7B}$	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5$_{XL}$	103M	3.4B	62.6	62.3	39.4	44.4
BLIP-2 ViT-g FlanT5$_{XL}$	107M	4.1B	63.1	63.0	40.7	44.2
BLIP-2 ViT-g FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**	45.9	**44.7**

Experiment-VQA

Models	#Trainable Params	VQAv2 test-dev	VQAv2 test-std
Open-ended generation models			
ALBEF (Li et al., 2021)	314M	75.84	76.04
BLIP (Li et al., 2022)	385M	78.25	78.32
OFA (Wang et al., 2022a)	930M	82.00	82.00
Flamingo80B (Alayrac et al., 2022)	10.6B	82.00	82.10
BLIP-2 ViT-g $\text{FlanT5}_{\text{XL}}$	1.2B	81.55	81.66
BLIP-2 ViT-g $\text{OPT}_{\text{2.7B}}$	1.2B	81.59	81.74
BLIP-2 ViT-g $\text{OPT}_{\text{6.7B}}$	1.2B	**82.19**	**82.30**
Closed-ended classification models			
VinVL	345M	76.52	76.60
SimVLM (Wang et al., 2021b)	∼1.4B	80.03	80.34
CoCa (Yu et al., 2022)	2.1B	82.30	82.30
BEIT-3 (Wang et al., 2022b)	1.9B	**84.19**	**84.03**

Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.

Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.

Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.

Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.

What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.

What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

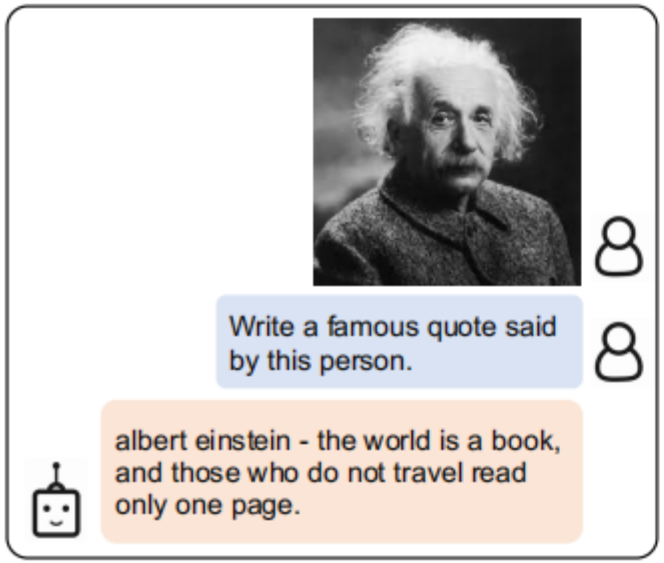
What is the first step?

Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

Experiment-Retrieval

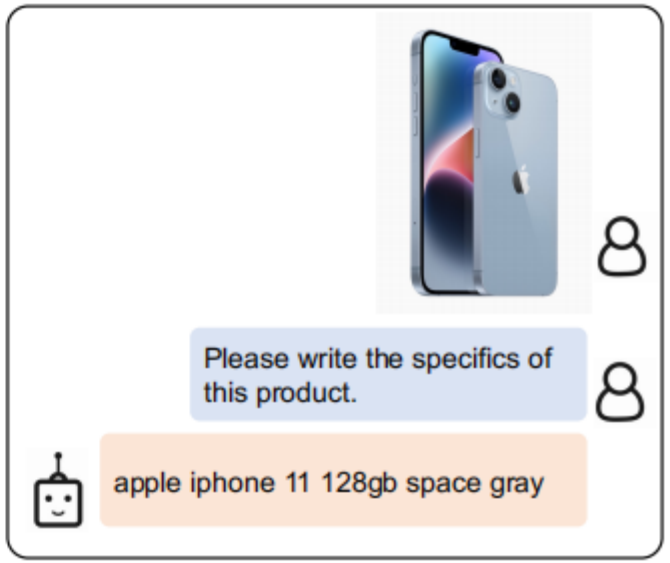
Model	#Trainable Params	Flickr30K Zero-shot (1K test set)						COCO Fine-tuned (5K test set)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Dual-encoder models													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	**100.0**	81.5	95.6	97.8	84.8	96.5	98.3	67.2	**87.7**	**92.8**
Fusion-encoder models													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
Dual encoder + Fusion encoder reranking													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	**100.0**	**100.0**	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 ViT-L	474M	96.9	**100.0**	**100.0**	88.6	97.6	**98.9**	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 ViT-g	1.2B	**97.6**	**100.0**	**100.0**	**89.7**	**98.1**	**98.9**	**85.4**	**97.0**	**98.5**	**68.3**	**87.7**	92.6

Failed Example



Inaccurate knowledge
(quote is from a different person)

Incorrect reasoning path
(should have considered weather)



Information not up-to-date
(this is iphone 14)

Thanks For Listening