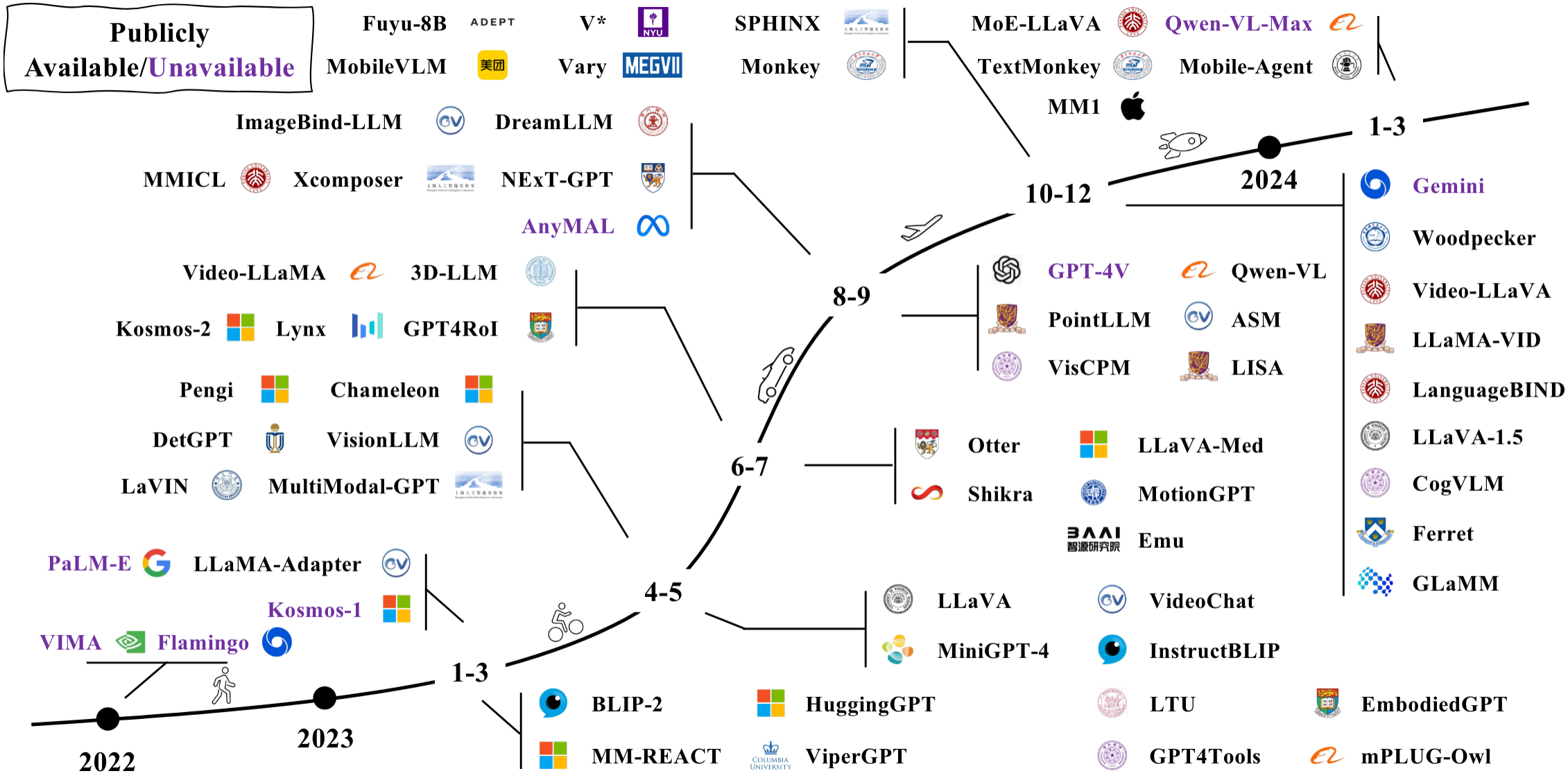
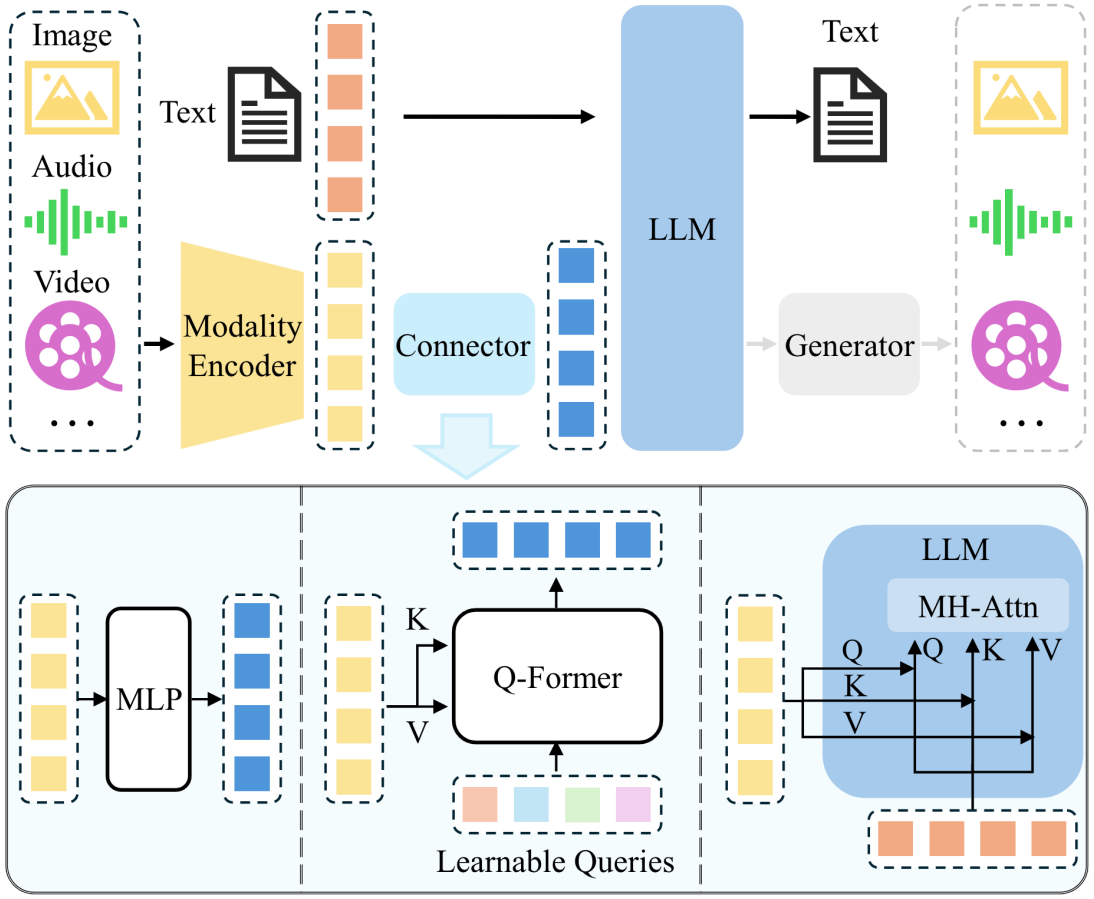
Multimodal Large Language Model

Yongxin Wang

06/28/2024


Publicly Available/Unavailable
2022
2023
2024
1-3
4-5
6-7
8-9
10-12
1-3
PaLM-E
LLaMA-Adapter
Kosmos-1
VIMA
Flamingo
Pengi
Chameleon
DetGPT
VisionLLM
LaVIN
MultiModal-GPT
Video-LLaMA
3D-LLM
Kosmos-2
Lynx
GPT4RoI
ImageBind-LLM
DreamLLM
MMICL
Xcomposer
NExT-GPT
AnyMAL
Fuyu-8B
V*
MobileVLM
Vary
SPHINX
Monkey
MoE-LLaVA
Qwen-VL-Max
TextMonkey
Mobile-Agent
MM1
BLIP-2
HuggingGPT
MM-REACT
ViperGPT
LLaVA
VideoChat
MiniGPT-4
InstructBLIP
LTU
EmbodiedGPT
GPT4Tools
mPLUG-Owl
Otter
LLaVA-Med
Shikra
MotionGPT
Emu
GPT-4V
Qwen-VL
PointLLM
ASM
VisCPM
LISA
Gemini
Woodpecker
Video-LLaVA
LLaMA-VID
LanguageBIND
LLaVA-1.5
CogVLM
Ferret
GLaMM

Architecture



LLaVA-series

BLIP2, InstructBLIP-series

Flamingo, LLaMA-Adapter

Q-Former

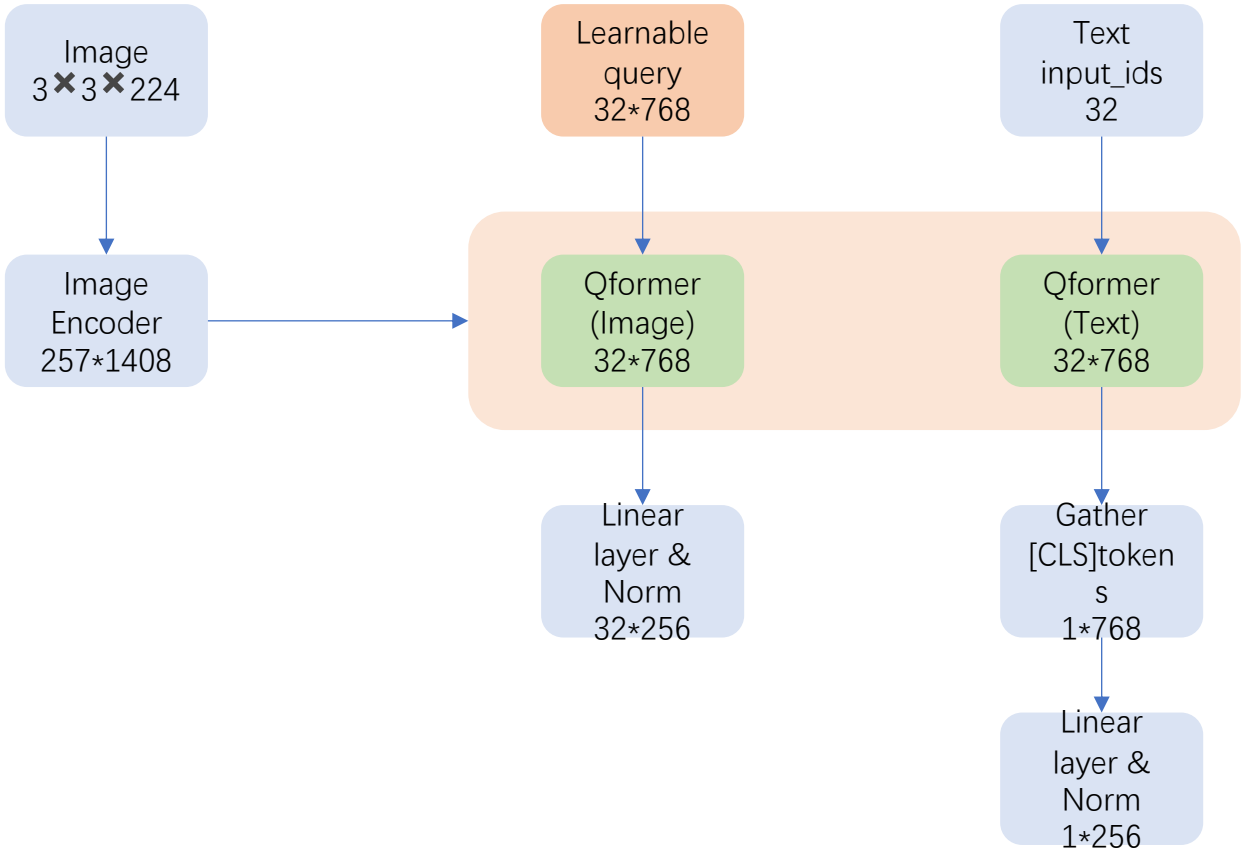


Image
3✖3✖224
Learnable query
32*768
Text input_ids
32
Image Encoder
257*1408
Qformer (Image)
32*768
Qformer (Text)
32*768
Linear layer & Norm
32*256
Gather [CLS]token s
1*768
Linear layer & Norm
1*256

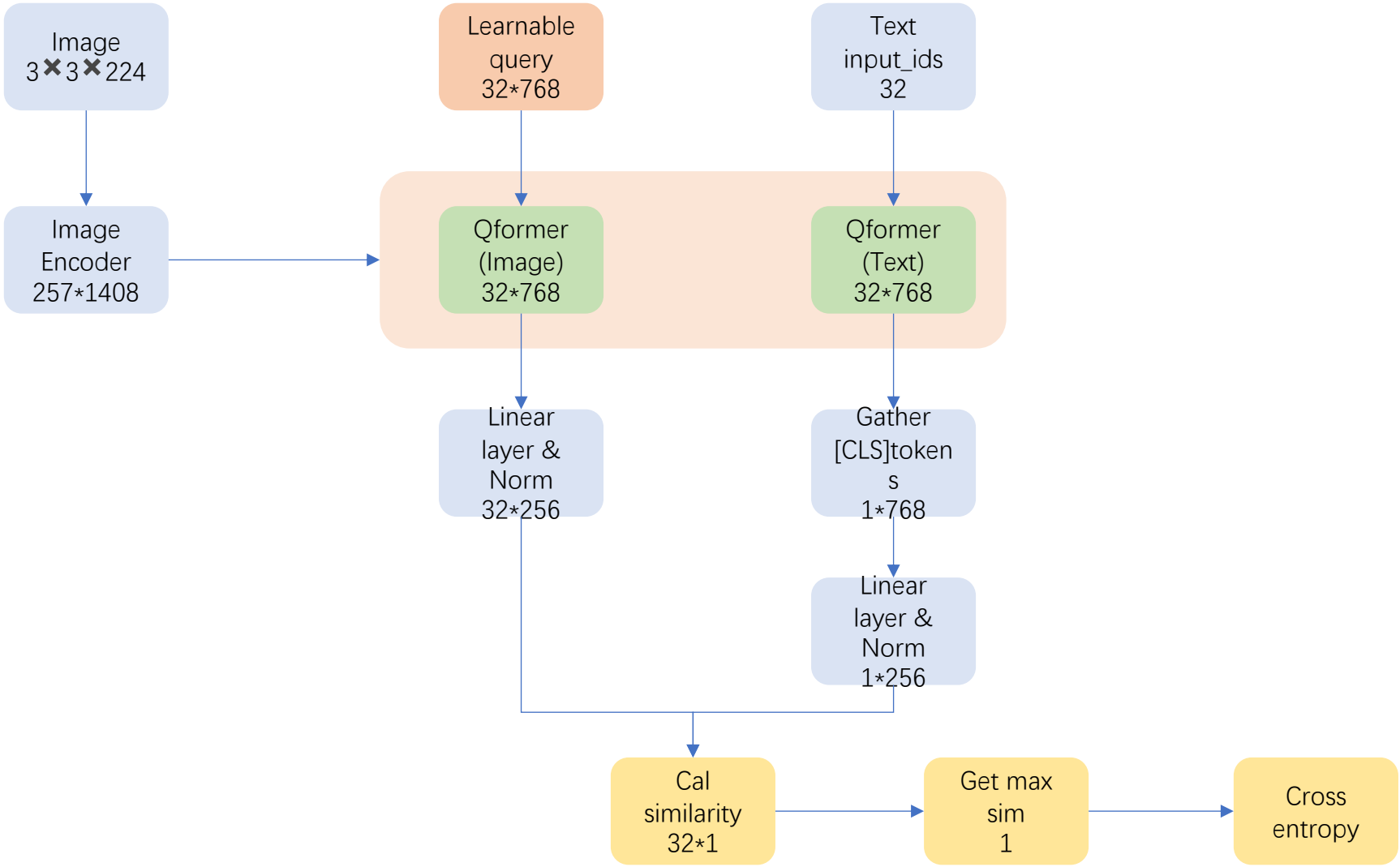
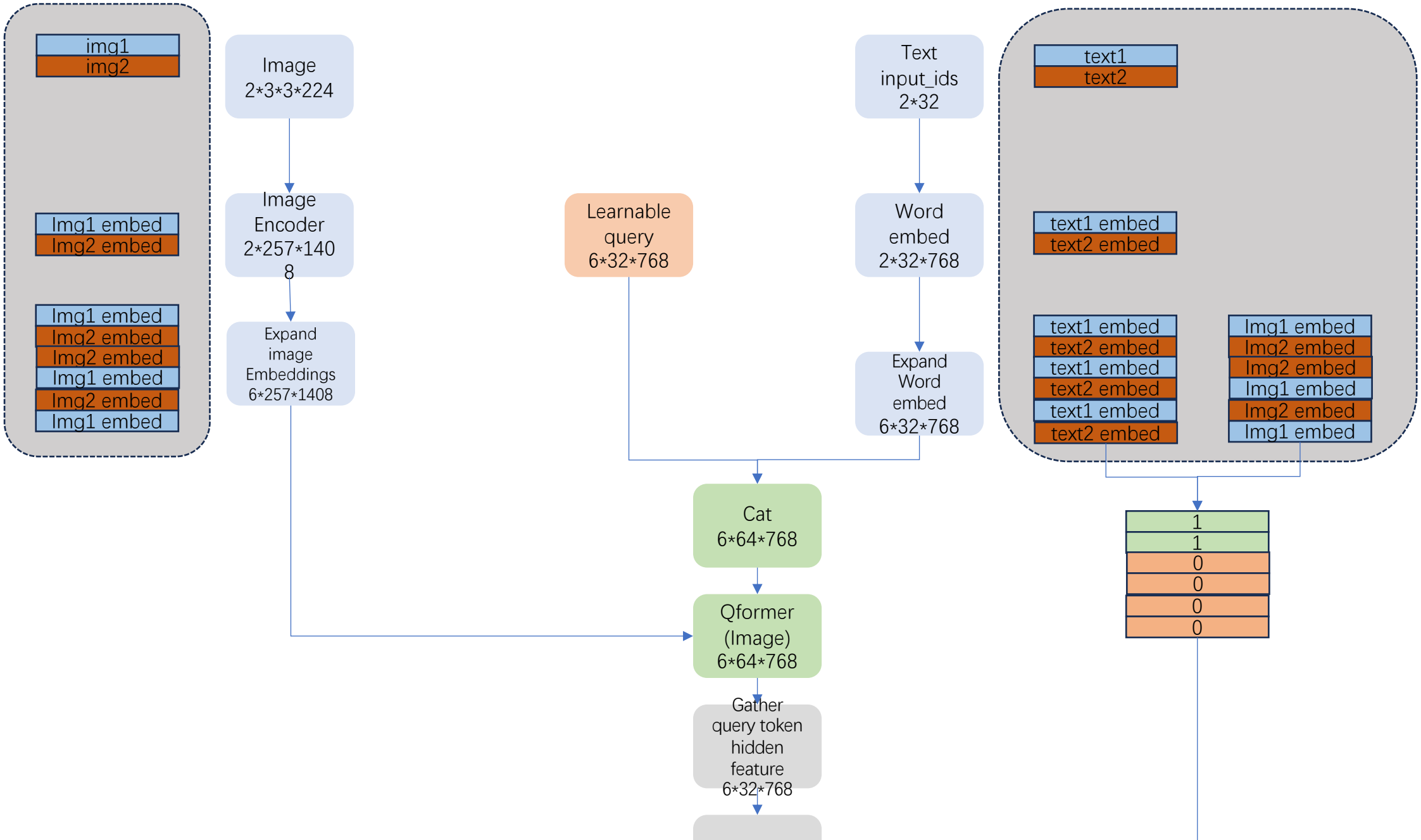
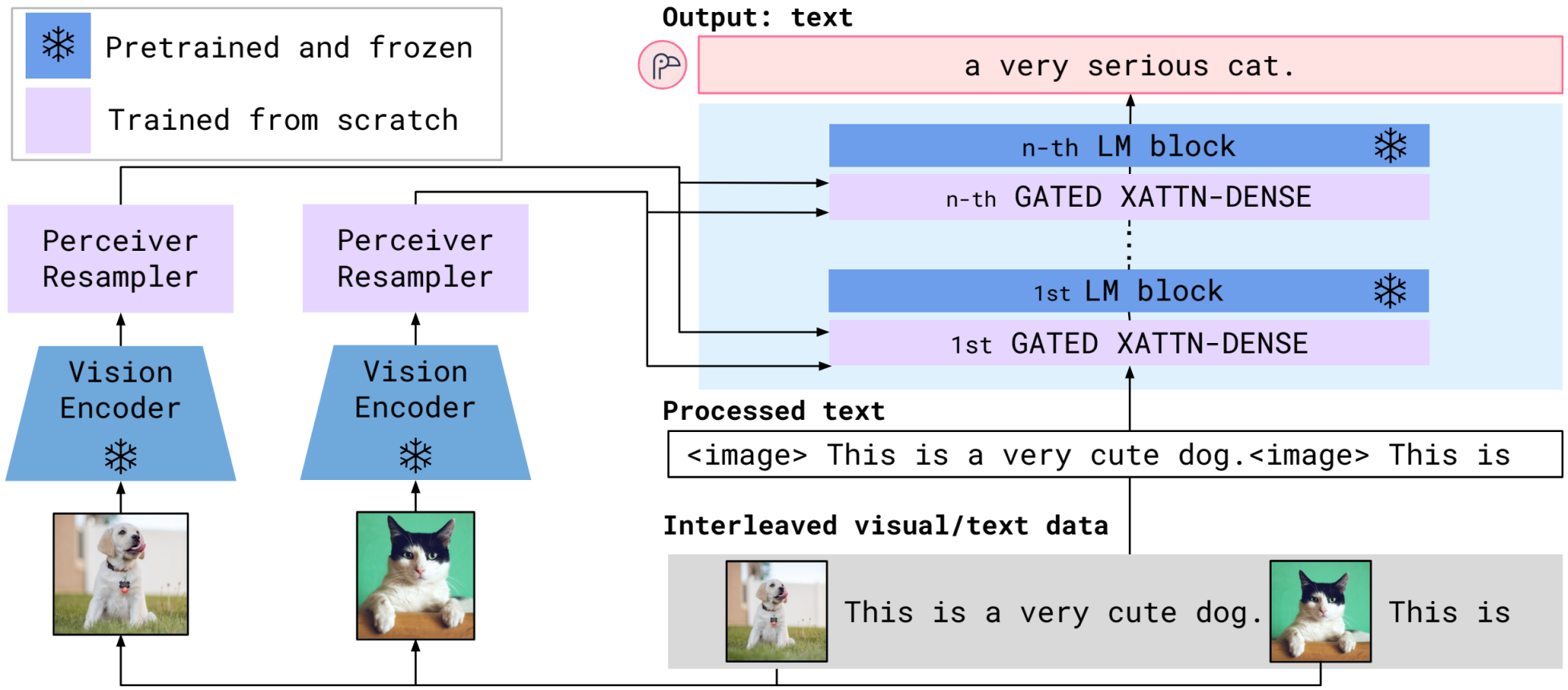
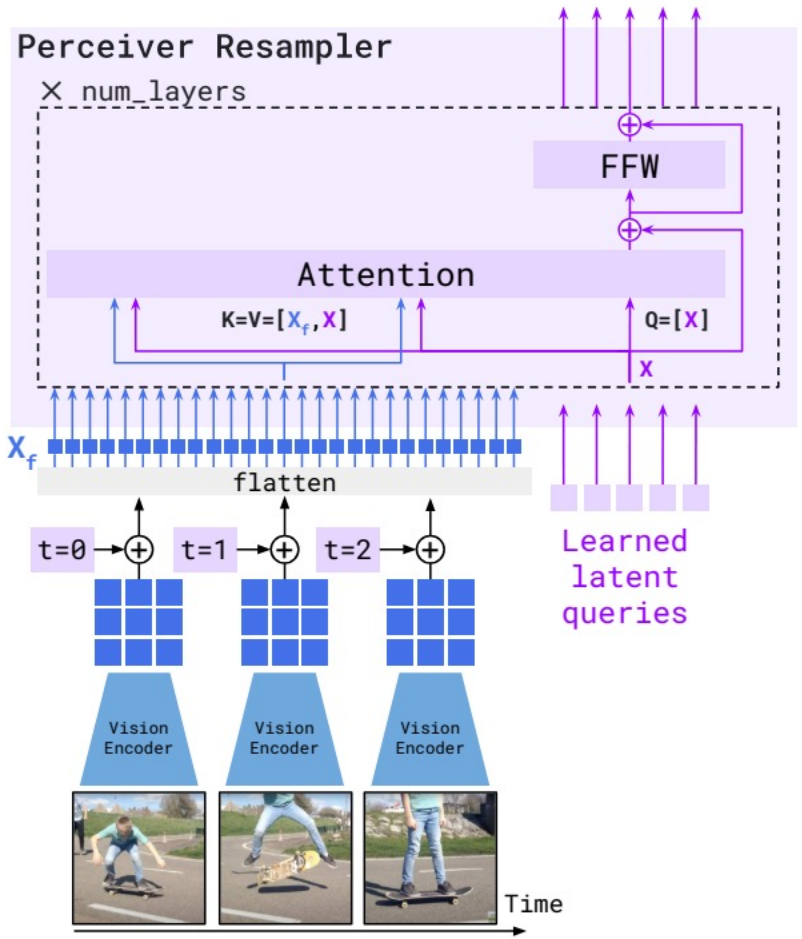


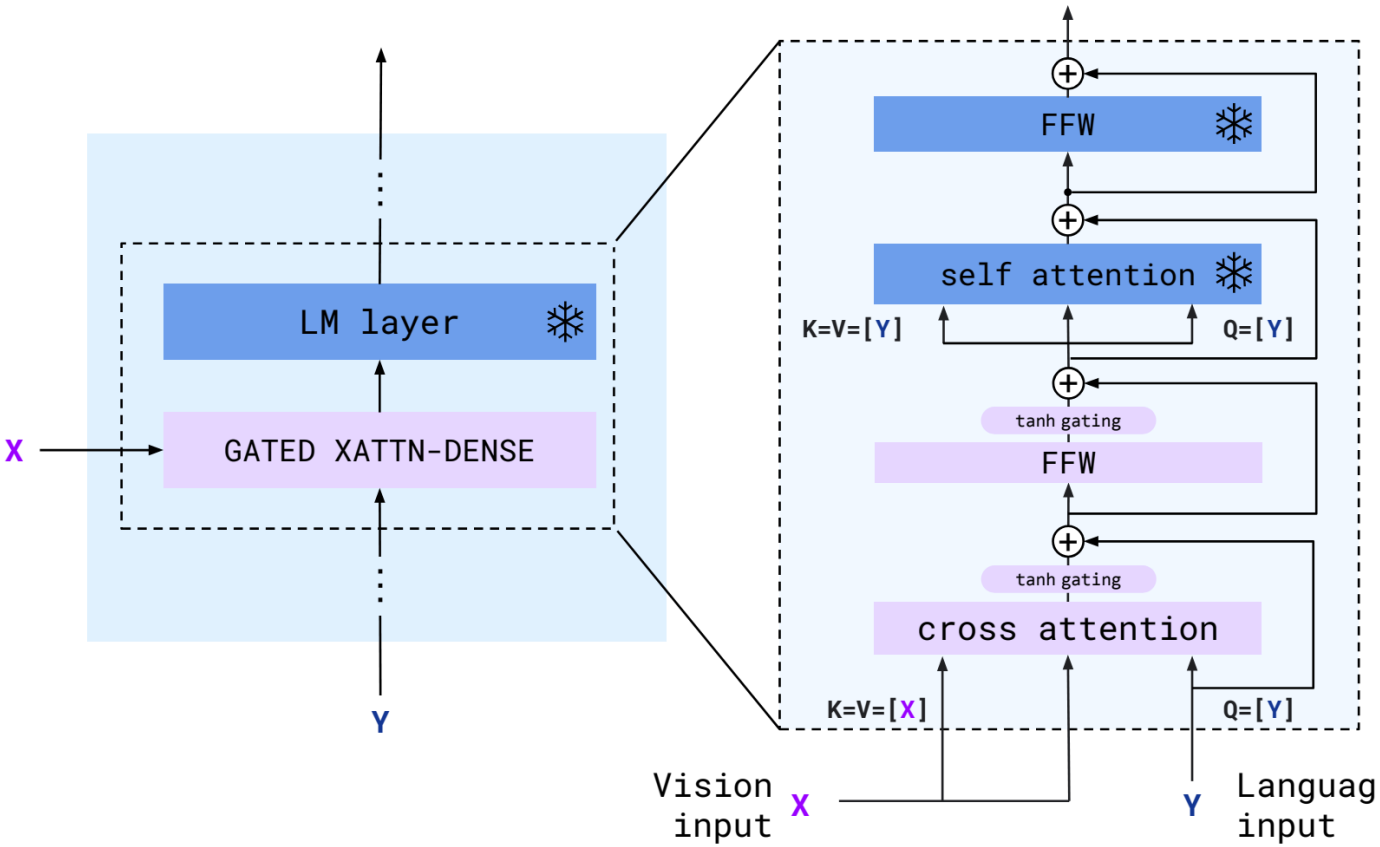
Image
3✖3✖224
Learnable
query
32*768
Text
input_ids
32
Image
Encoder
257*1408
Qformer
(Image)
32*768
Qformer
(Text)
32*768
Linear
layer &
Norm
32*256
Gather
[CLS]token
s
1*768
Linear
layer &
Norm
1*256
Cal
similarity
32*1
Get max
sim
1
Cross
entropy


img1
img2
Image
2*3*3*224
Image
Encoder
2*257*140
8
Expand
image
Embeddings
6*257*1408
Img1 embed
Img2 embed
Img1 embed
Img2 embed
Img2 embed
Img1 embed
Img2 embed
Img1 embed
Learnable
query
6*32*768
Text
input_ids
2*32
Word
embed
2*32*768
Expand
Word
embed
6*32*768
Cat
6*64*768
Qformer
(Image)
6*64*768
Gather
query token
hidden
feature
6*32*768
text1
text2
text1 embed
text2 embed
text1 embed
text2 embed
text1 embed
text2 embed
Img1 embed
Img2 embed
Img2 embed
Img1 embed
Img2 embed
Img1 embed
1
1
0
0
0
0

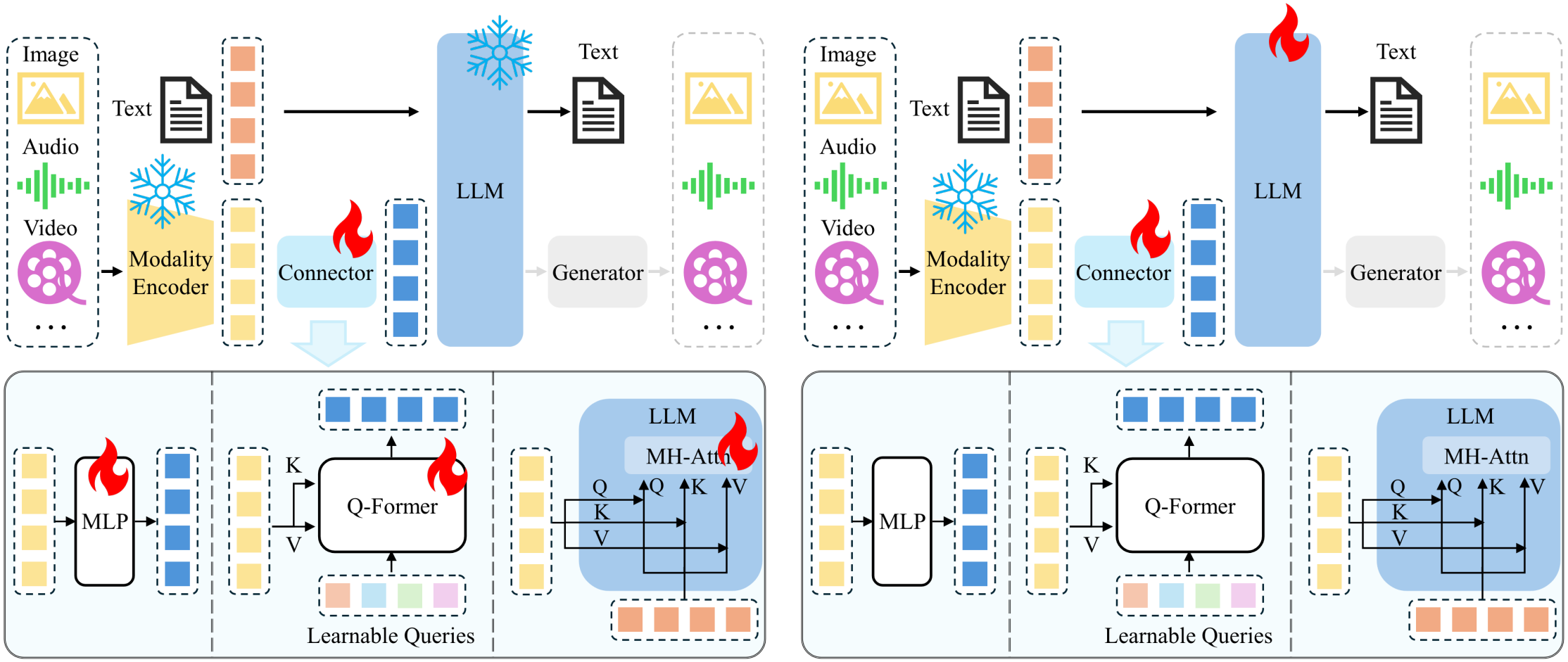
Flamingo


Pretrained and frozen
Trained from scratch
Output: text
a very serious cat.
n-th LM block
n-th GATED XATTN-DENSE
1st LM block
1st GATED XATTN-DENSE
Perceiver Resampler
Perceiver Resampler
Vision Encoder
Vision Encoder
Processed text
<image> This is a very cute dog.<image> This is
Interleaved visual/text data
This is a very cute dog.
This is


Perceiver Resampler
× num_layers
FFW
Attention
K=V=[X_f,X]
Q=[X]
X
X_f
flatten
t=0
t=1
t=2
Learned latent queries
Vision Encoder
Vision Encoder
Vision Encoder
Time


LM layer
GATED XATTN-DENSE
X
Y
FFW
self attention
K=V=[Y]
Q=[Y]
tanh gating
FFW
tanh gating
cross attention
K=V=[X]
Q=[Y]
Vision input
X
Y
Languag input

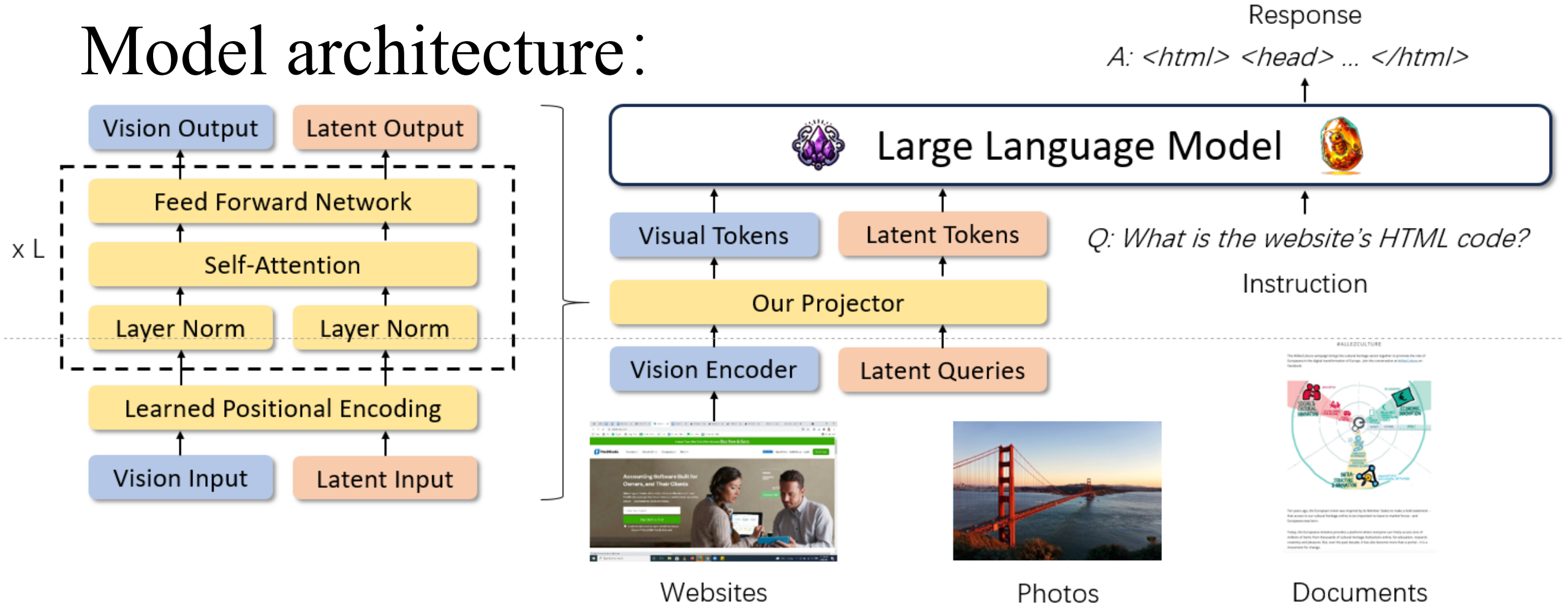
Training Details



- image embeddings for perception and reasoning should be different
  - For example, reasoning often requires multiple perceptual image tokens
  - However, existing frameworks use only global tokens (like Flamingo and BLIP-2) or only local tokens (like LLaVA)
  - considering to use both global and local tokens as image embeddings
    - **Global** tokens from both image and text
    - **Local** tokens from the image

- Explicitly modeling relation tokens as inputs of LLM
- Inspired on scene graph domain, relation tokens extracted from object tokens can bring rich information
  - Like Flamingo, we add new modules to model object tokens and relation tokens
  - New learnable object and relation tokens are given as inputs of LLM

Model architecture:



- Aligning text and visual embeddings using image-text pairs.
- Refining the model's instruction-following capability using multi-modal tasks and natural language processing data.