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

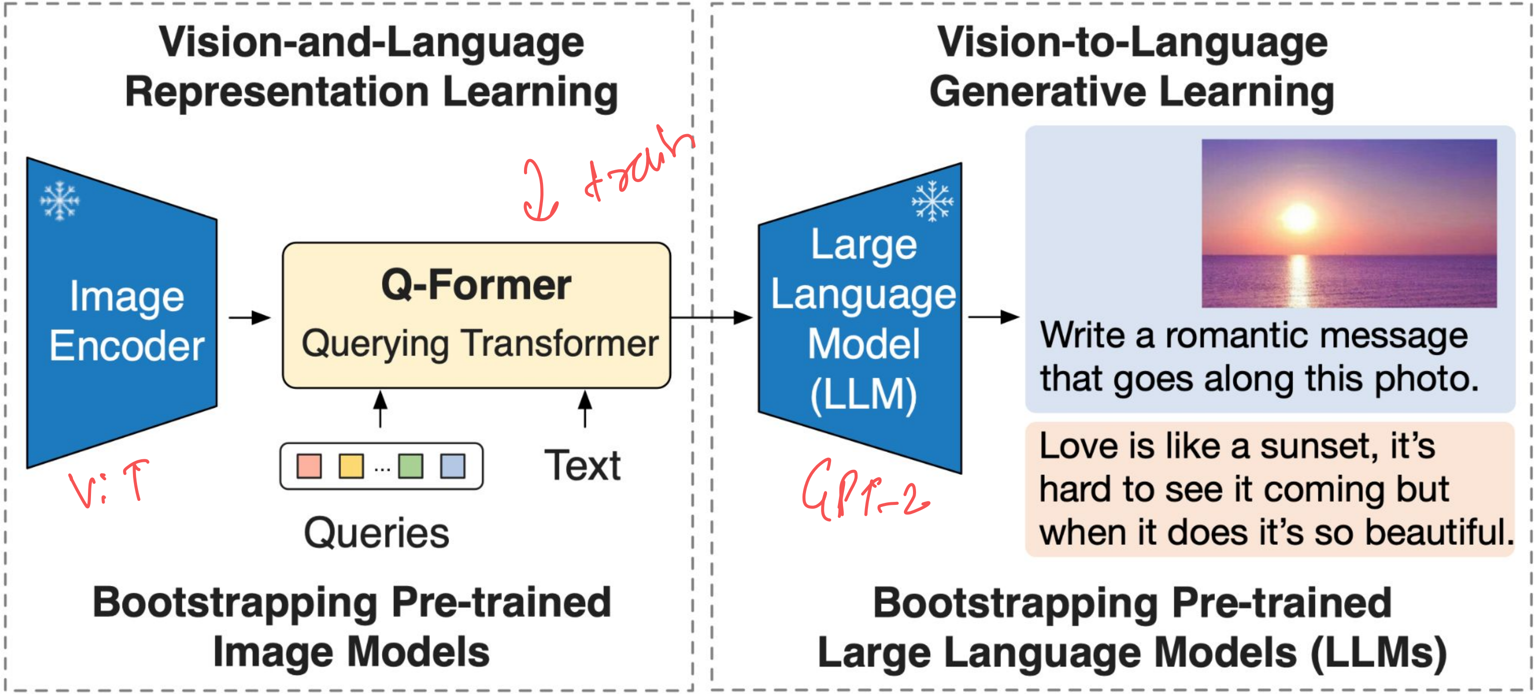
Proposes a generic and compute-efficient method by bootstrapping from off-the-shelf pre-trained vision models and language models.

- Pre-trained vision models offer high-quality visual representation.
- Pre-trained language models, in particular large language models (LLMs), offer strong language generation and zero-shot transfer abilities.

To reduce computation cost and counteract the issue of catastrophic forgetting, the unimodal pre-trained models remain frozen during the pre-training.

https://arxiv.org/pdf/2301.12597

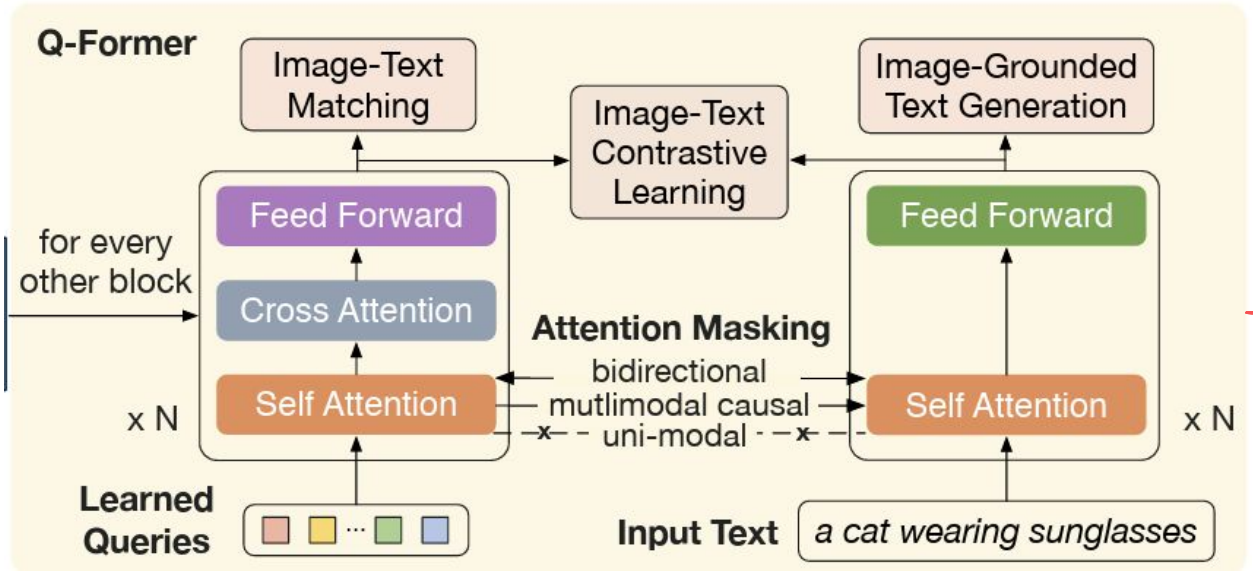
BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models


Vision-and-Language
Representation Learning
↓ train
Image
Encoder
ViT
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained
Image Models
Vision-to-Language
Generative Learning
Large
Language
Model
(LLM)
GPT-2
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained
Large Language Models (LLMs)

We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap.

https://arxiv.org/pdf/2301.12597

Querying Transformer: Architecture



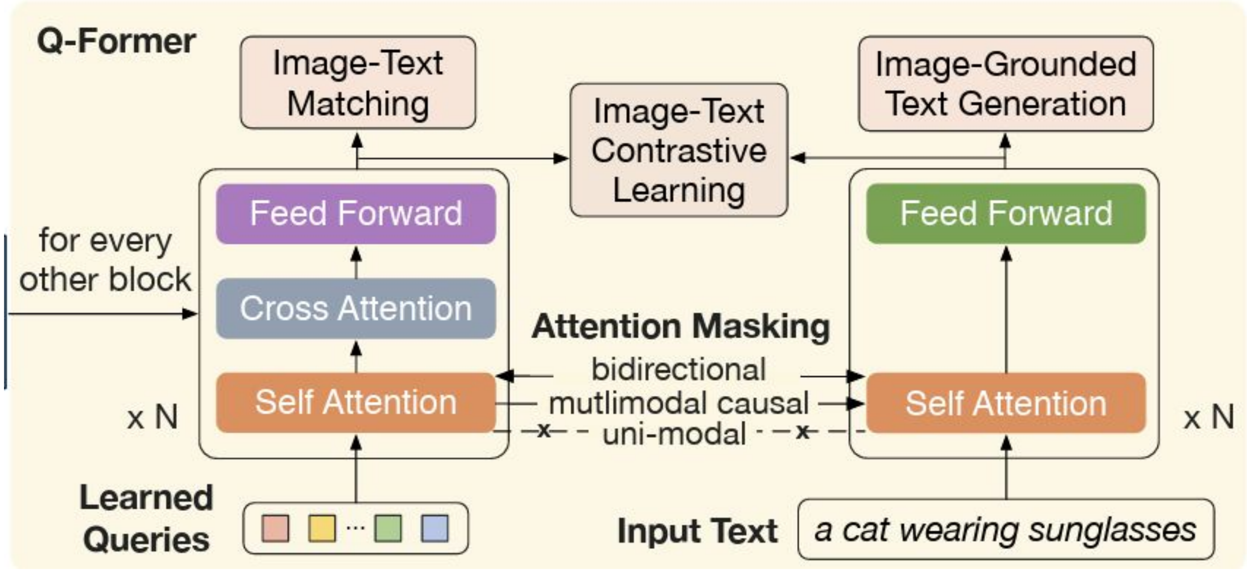
Q-Former consists of two transformer submodules that share the same self-attention layers:

→ trainable

(1) an image transformer that interacts with the frozen image encoder for visual feature extraction

(2) a text transformer that can function as both a text encoder and a text decoder.

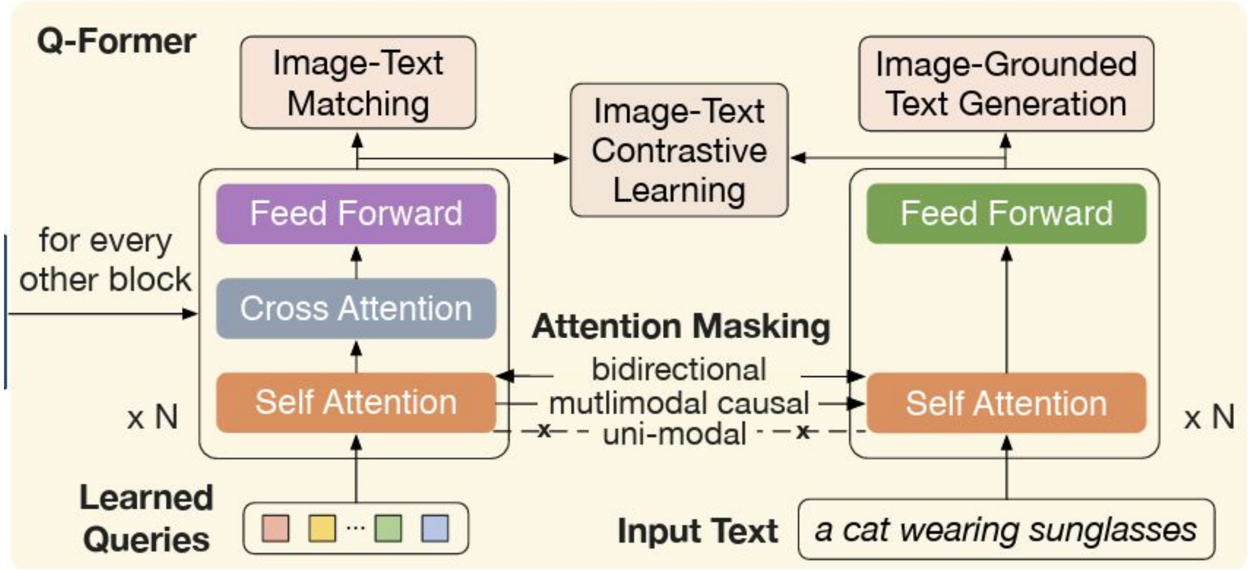
Querying Transformer: Architecture



Q-Former consists of two transformer submodules that share the same self-attention layers:

We create a set number (32) of learnable “query” embeddings as input to the image transformer. The queries interact with each other through self-attention layers, and interact with frozen image features through cross-attention layers (inserted ***every other transformer block***). The queries can additionally interact with the text through the same self-attention layers.

Querying Transformer: Architecture



We initialize QFormer with the pre-trained weights of BERT-base, whereas the cross-attention layers are randomly initialized.

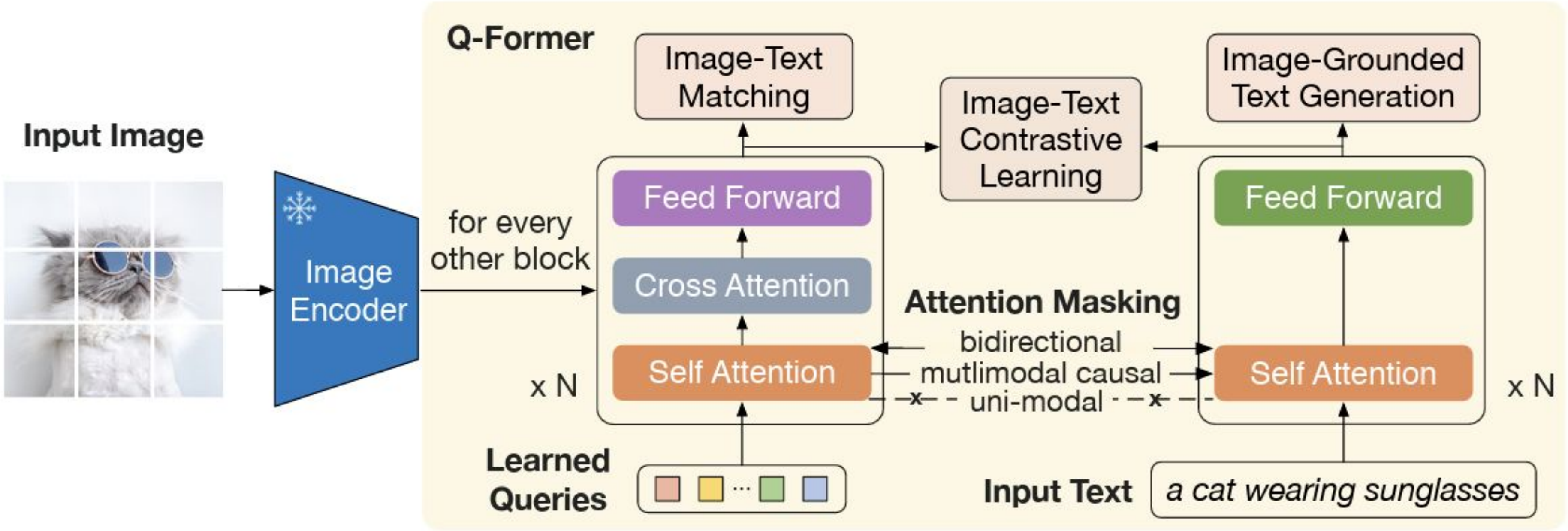
We create a set number (32) of learnable “query” embeddings as input to the image transformer. The queries interact with each other through self-attention layers, and interact with frozen image features through cross-attention layers (inserted ***every other transformer block***). The queries can additionally interact with the text through the same self-attention layers.

Querying Transformer

Q-Former is pre-trained in two stages:

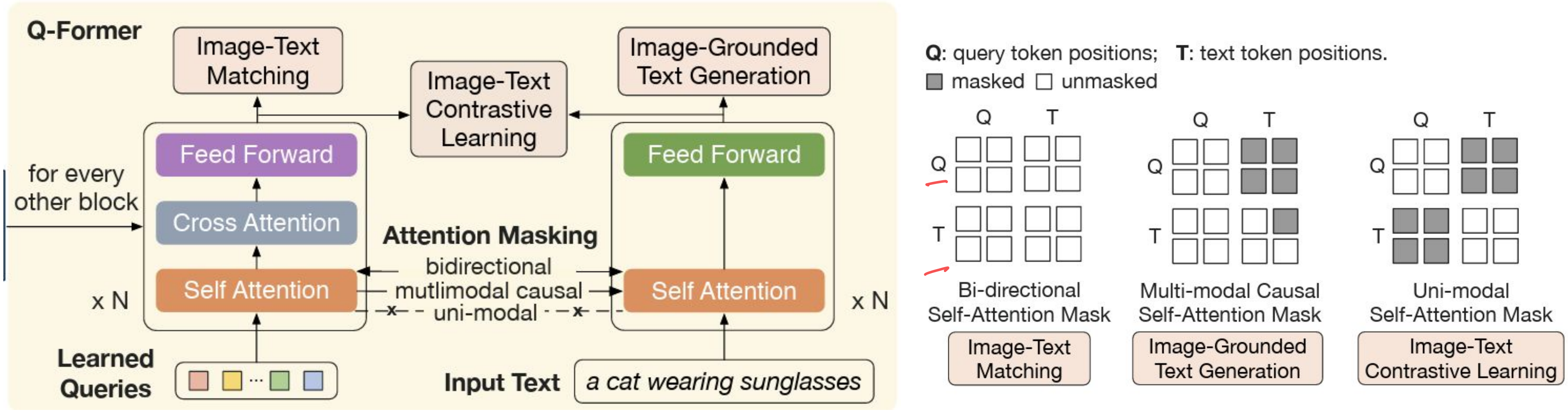
(1) vision-language representation learning stage with a frozen
image encoder and
(2) vision-to-language generative learning stage with a frozen LLM.

Bootstrap Vision-Language Representation Learning from a Frozen Image Encoder



Connect Q-Former to a frozen image encoder and perform pre-training using image-text pairs.
We jointly optimize three pre-training objectives that share the same input format and model parameters.
Each objective employs a different attention masking strategy between queries and text to control their interaction

Querying Transformer: Architecture


Q-Former
Image-Text Matching
Image-Text Contrastive Learning
Image-Grounded Text Generation
Feed Forward
Cross Attention
Self Attention
for every other block
x N
Attention Masking
bidirectional
mutlimodal causal
uni-modal
Feed Forward
Self Attention
x N
Learned Queries
Input Text
a cat wearing sunglasses
Q: query token positions; T: text token positions.
masked
unmasked
Q
T
Bi-directional Self-Attention Mask
Multi-modal Causal Self-Attention Mask
Uni-modal Self-Attention Mask
Image-Text Matching
Image-Grounded Text Generation
Image-Text Contrastive Learning

Depending on the pre-training task, we apply different self-attention masks to control query-text interaction.

Image-Text Matching (ITM) is a binary classification task. We use a bi-directional self-attention mask where all queries and texts can attend to each other. The output query embeddings Z thus capture multimodal information. We feed each output query embedding into a two-class linear classifier to obtain a logit, and average the logits across all queries as the output matching score.

Querying Transformer: Architecture

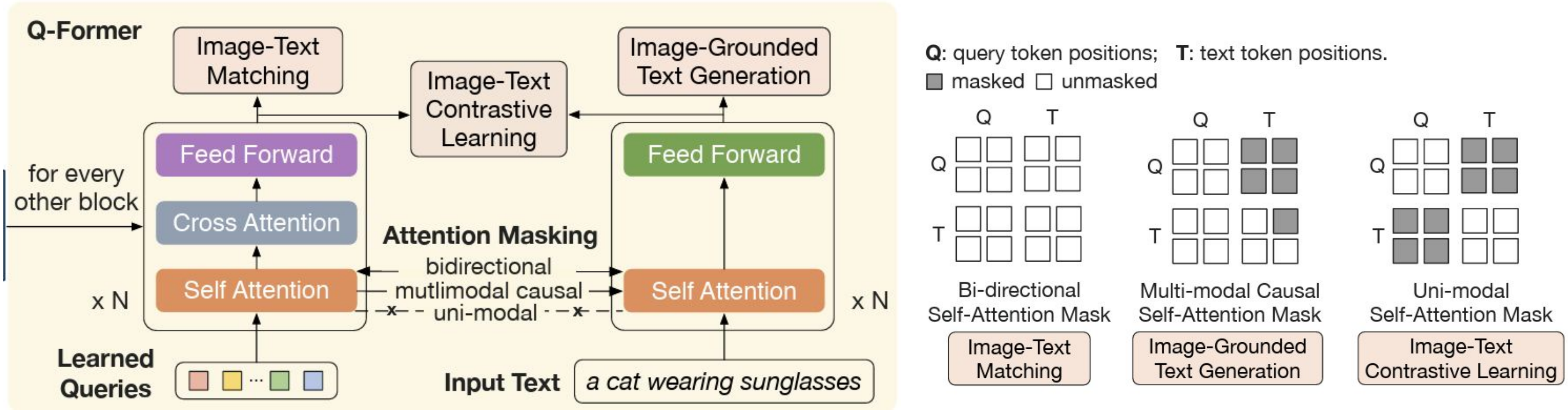
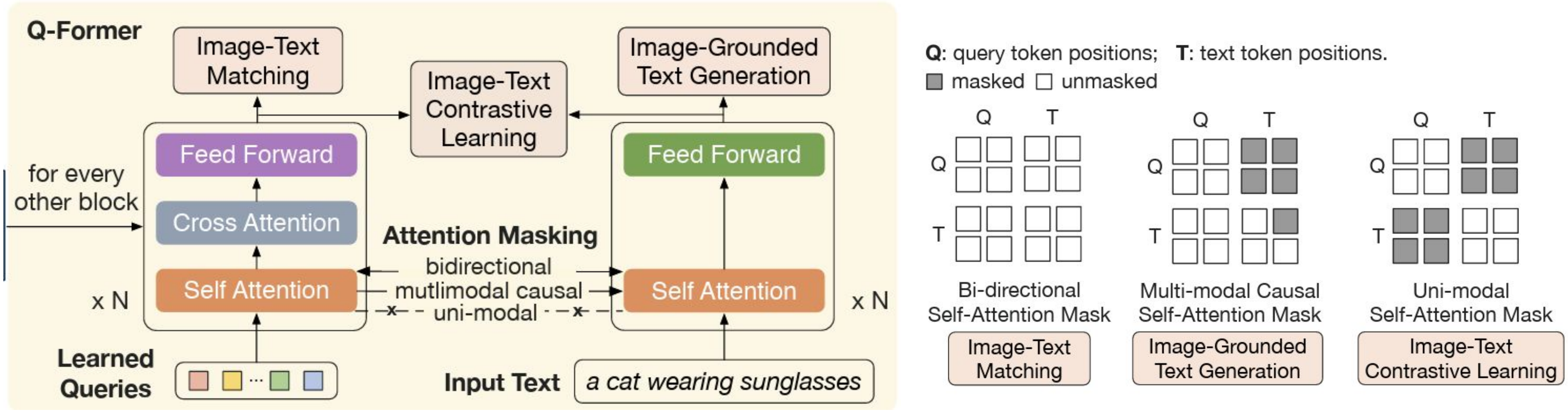

Q-Former
Image-Text Matching
Image-Text Contrastive Learning
Image-Grounded Text Generation
for every other block
Feed Forward
Cross Attention
Self Attention
x N
Feed Forward
Self Attention
x N
Attention Masking
bidirectional
mutlimodal causal
uni-modal
Learned Queries
Input Text
a cat wearing sunglasses
Q: query token positions; T: text token positions.
masked
unmasked
Q
T
Bi-directional Self-Attention Mask
Image-Text Matching
Multi-modal Causal Self-Attention Mask
Image-Grounded Text Generation
Uni-modal Self-Attention Mask
Image-Text Contrastive Learning

Image-grounded Text Generation (ITG). Since the architecture of Q-Former does not allow direct interactions between the frozen image encoder and the text tokens, the information required for generating the text must be first extracted by the queries, and then passed to the text tokens via self-attention layers. Therefore, the queries are forced to extract visual features that capture all the information about the image. We employ a multimodal causal self-attention mask to control query-text interaction. The queries can attend to each other but not the text tokens. Each text token can attend to all queries and its previous text tokens. We also replace the [CLS] token with a new [DEC] token as the first text token to signal the decoding task.

Querying Transformer: Architecture



	Q	Q	T	T
Q	□	□	□	□
Q	□	□	□	□
T	□	□	□	□
T	□	□	□	□

Bi-directional Self-Attention Mask

Image-Text Matching

	Q	Q	T	T
Q	□	□	▨	▨
Q	□	□	▨	▨
T	□	□	□	▨
T	□	□	□	□

Multi-modal Causal Self-Attention Mask

Image-Grounded Text Generation

	Q	Q	T	T
Q	□	□	▨	▨
Q	□	□	▨	▨
T	▨	▨	□	□
T	▨	▨	□	□

Uni-modal Self-Attention Mask

Image-Text Contrastive Learning

Image-Text Contrastive Learning (ITC). We align the output query representation *t* from the text transformer, where *t* is the output embedding of the [CLS] token. Since ***Z*** contains multiple output embeddings (one from each query), we first compute the pairwise similarity between each query output and *t*, and then select the highest one as the image-text similarity. To avoid information leak, we employ a unimodal self-attention mask, where the queries and text are not allowed to see each other.