CVPDL HW3

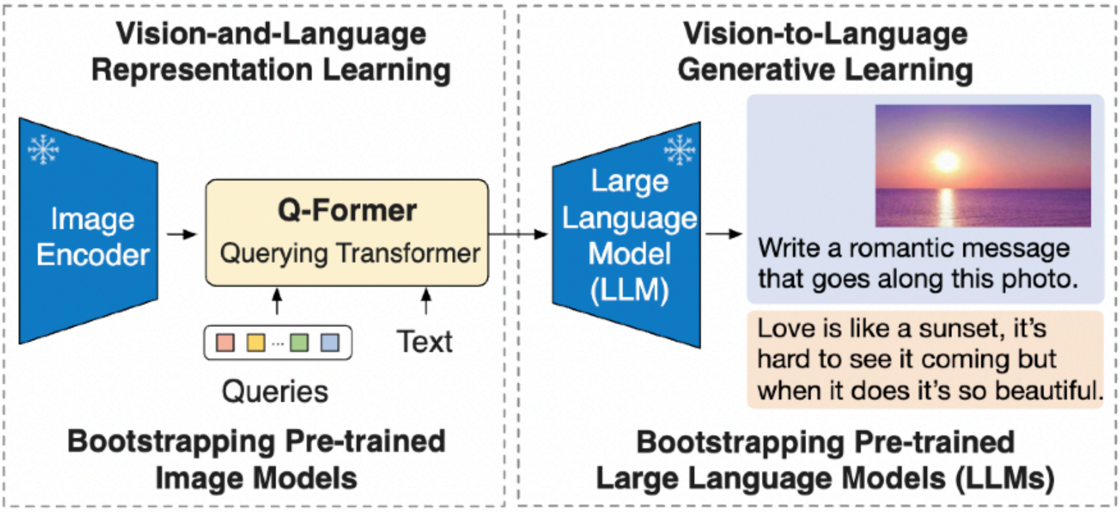
1. Related Works:

In this homework, the pipeline is to apply ‘image2text’ model and then apply ‘text2image’ model to generate more diverse images. For ‘image2text’, I mainly used `blip2-opt-2.7b` and `blip2-flan-t5-xl` while for ‘text2image’, I mainly used `pure stable-diffusion` and `gligen` model. In the following, I will introduce the method of my used model:

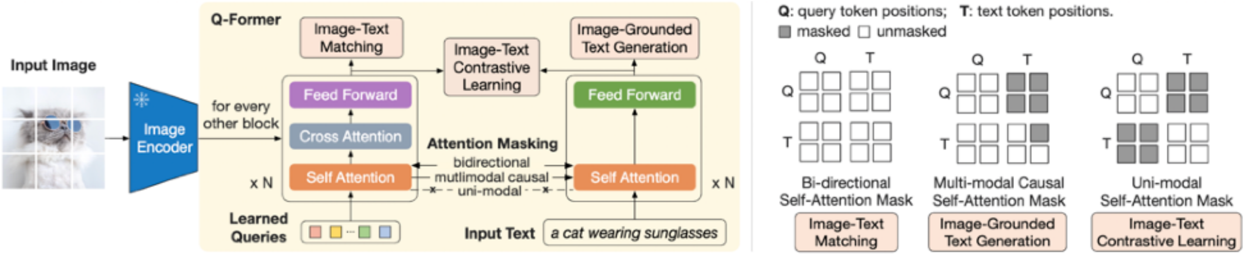
A. BLIP-2[1]:

As shown in the following two figures, the BLIP contains three main module, one is the Image Encoder, which encodes the image to the latent features; another is the Q-Former, which applies the attention mechanism to match the image spaces to the text spaces; the other is the LLM part, which is fed with the text output from the Q-Former and generated the final output. Given an input image and a text description, the model will give texts output to match both the image and text.

The main objective is to train a good Q-Former. Firstly, it is connected with a frozen Image Encoder to train the parameters and align the Queries to texts. The main optimization techniques are



Vision-and-Language
Representation Learning
Image Encoder
Q-Former
Querying Transformer
Text
Queries
Bootstrapping Pre-trained
Image Models
Vision-to-Language
Generative Learning
Large Language Model (LLM)
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)

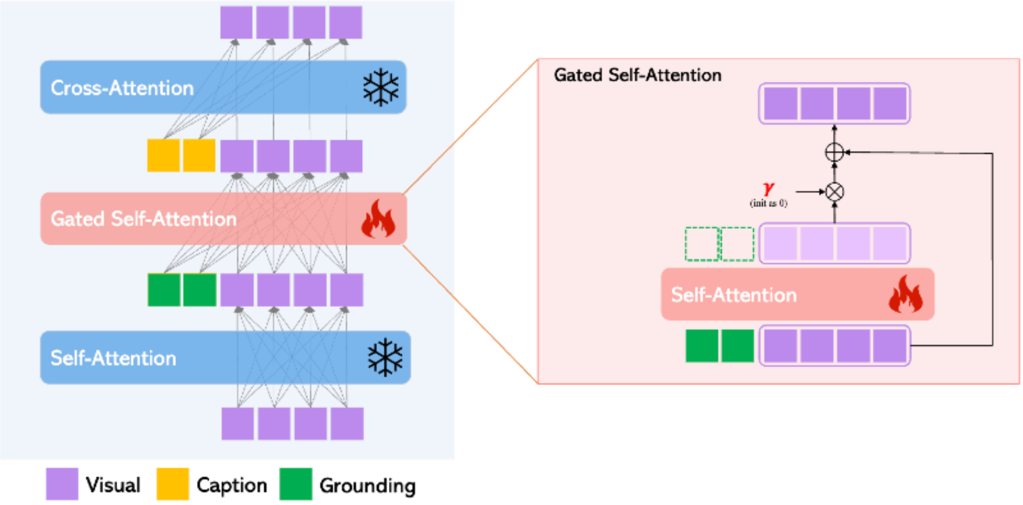


1. **Image-Text Contrastive Learning:** This aims to align the text representation (right in the figure of Q-Former) and the image representation via maximizing their mutual information.
2. **Image-Ground Text Generation:** This aims to train the model to predict the text via input images. Since Image Encoder is frozen, the queries are forced to extract the features that are related to the text generations.
3. **Image-Text Matching:** This aims to judge if the image and text match to each other. It is a binary judgement.

Afterwards, the whole pipeline is connected to a frozen LLM, leveraging the ability of LLM to train the Q-Former parameters.

B. Gligen Model[2]:

Gligen Model tries to include layout into the Diffusion Model, which contains the new idea of grounding token and the Gated Self Attention Layer, as shown in the figure. The grounding token is constructed via the fusing process of layouts (eg: bounding box), through Fourier Embedding method. And the Gated-Self-Attention is the self-attention layer with the concatenation of visual tokens and grounding tokens. Then, a token selection operation will act on the vision tokens and provide the input of the next steps (a.k.a cross attention with the caption token). During the training stage of the Gated Self-Attention layer, the Self-Attention and Cross-Attention are frozen. Through out the grounding token and Gated Self-Attention



layer, the final representation of the visual tokens can be embedded with the information from both the caption and the information of layout.

2. Method, Prompt Design, Result and Visualization

For image captioning, I used `blip2-opt-2.7b` and `blip2-flan-t5-xl` with the inputs:

`Question: Write an overall description of the picture. Answer:`

to generate the caption of the image (a.k.a the 'generated_text'). Afterwards, I followed the default setting from TA to generate 'prompt_w_label' (the caption together with the labels) and 'prompt_w_suffix' (the 'prompt_w_label' together with ', height: 512, width: 512, HD quality, highly detailed.') Since 'prompt_w_label' is similar to 'prompt_w_suffix', I finally chose 'generated_text' and 'prompt_w_suffix' as my two prompt templates for text-only generation.

For the generation model, I used `runwayml/stable-diffusion-v1-5`[3] for text-only generation and `anhnct/Gligen_Text_Image` for generation together with layout. The FID score is shown as the following table. As you can see, the FID score of the 'generated_text' is far better than 'prompt_w_suffix'. Hence, I used 'generated_text' as the generation for "layout", which I choose to include boxes position as well as the reference images.

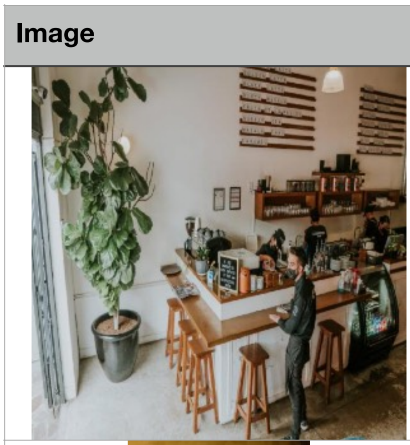
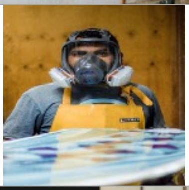
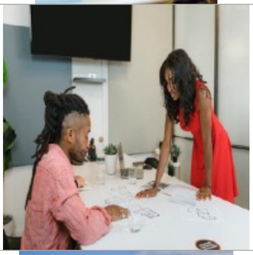
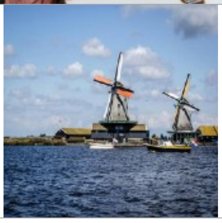
FID Score Table	Text Only 'generated_text'	Text Only 'prompt_w_suffix'	Layout (Text ['generated_text'] + Boxes + Reference Images)
blip2-opt-2.7b	46.97	57.40	43.72
blip2-flan-t5-xl	45.33	55.87	43.98

In this approach, we can find that with a complex but confused details provided to the image generation model will probably bias the model generation. The model may probably generate a sequence of repeated objects, based on the 'generated_text' and the extra provided labels (in the text) again. Also, it does not really match the request form in the real-world, which may be a problem since these models are pertained using real-world samples.

3. Discussion on the generated prompts

I will start the discussion about the generated prompts first by showing the examples generated from the two models.

The following table shows some examples of the prompts about 'generated_text' and 'prompt_w_suffix' using blip2-flan-t5-xl.

Image	generated_text	Prompt_w_suffix
	a man is standing at the counter of a coffee shop	a man is standing at the counter of a coffee shop Head Head Head Head Head Hands Hands Hands Hands Face-mask-medical Face-mask-medical Face-mask-medical Face-mask-medical Ear Ear Ear Ear Ear Ear Shoes Shoes Person Person Person Person Person, height: 512, width: 512, HD quality, highly detailed.
	A man is holding a surfboard in a workshop	A man is holding a surfboard in a workshop Head Head Earmuffs Face-guard Person, height: 512, width: 512, HD quality, highly detailed.
	A woman is sitting at a table with a man, who is looking at a piece of paper	A woman is sitting at a table with a man, who is looking at a piece of paper Head Head Face Face Ear Hands Hands Hands Hands Person Person, height: 512, width: 512, HD quality, highly detailed.
	The picture shows a windmill in the middle of a lake with a boat in the background	The picture shows a windmill in the middle of a lake with a boat in the background Head Face Person Person, height: 512, width: 512, HD quality, highly detailed.

The following table shows some examples of the prompts about 'generated_text' and 'prompt_w_suffix' using blip2-opt-2.7b.

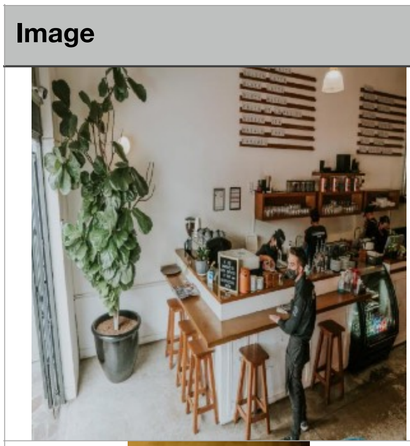
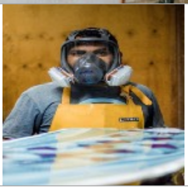
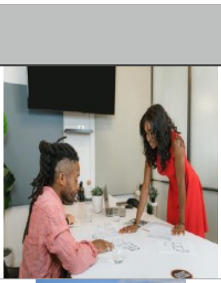
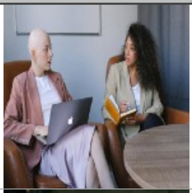
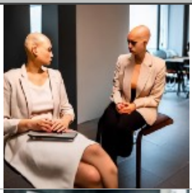
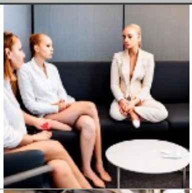
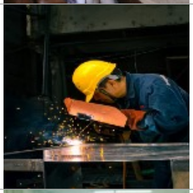
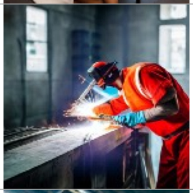
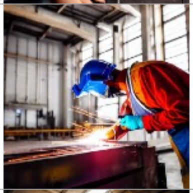
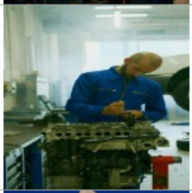
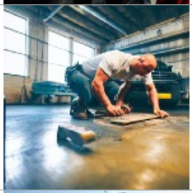
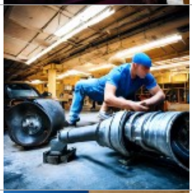
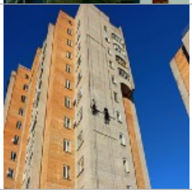
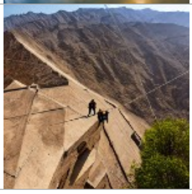
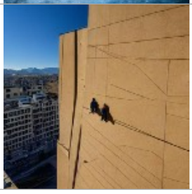
Image	generated_text	Prompt_w_suffix
	a coffee shop with a bar counter, a table, and a potted plant	a coffee shop with a bar counter, a table, and a potted plant Head Head Head Head Head Hands Hands Hands Hands Face-mask-medical Face-mask-medical Face-mask-medical Face-mask-medical Ear Ear Ear Ear Ear Ear Shoes Shoes Person Person Person Person Person, height: 512, width: 512, HD quality, highly detailed.
	A man is wearing a gas mask and holding a surfboard	A man is wearing a gas mask and holding a surfboard Head Head Earmuffs Face-guard Person, height: 512, width: 512, HD quality, highly detailed.

Image	generated_text	Prompt_w_suffix
	a woman and man are working on a project together	a woman and man are working on a project together Head Head Face Face Ear Hands Hands Hands Hands Person Person, height: 512, width: 512, HD quality, highly detailed."
	two windmills on the water	two windmills on the water Head Face Person Person, height: 512, width: 512, HD quality, highly detailed.

As you can see, the performance of these two models are almost the same, via the concentration on the second and the third generations. blip2-flan-t5-xl returned a more detailed caption on the third pictures while blip2-opt-2.7b did well on the second. This phenomenon can also be seen via the FID score, which they share almost the same results w/ or w/o layouts. Hence, in the end, I just chose blip2-opt-2.7b for my vis_200 output.

4. Generated Images Visualization

This section will show several generated images as some visualization examples.

Original Images	Generated Images (blip2-opt-2.7b)	Generated Images (blip2-flan-t5-xl)
		
		
		
		

Reference

[1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (ICML'23), Vol. 202. JMLR.org, Article 814, 19730–19742.

[2] Li, Yuheng & Liu, Haotian & Wu, Qingyang & Mu, Fangzhou & Yang, Jianwei & Gao, Jianfeng & Li, Chunyuan & Lee, Yong Jae. (2023). GLIGEN: Open-Set Grounded Text-to-Image Generation. 22511-22521. 10.1109/CVPR52729.2023.02156.

[3] Rombach, Robin, A. Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models.” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021): 10674-10685.