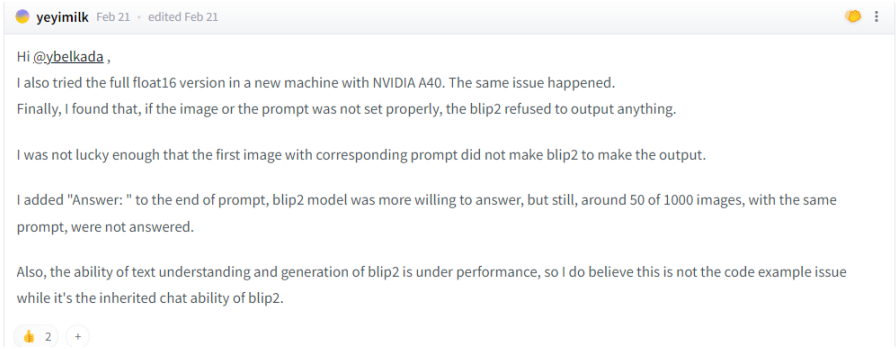
Homework #3: Data Augmentation

Image Captioning:

因為 GPU 內存的關係這邊我使用 Salesforce/blip2-opt-2.7b 以及 Salesforce/blip2-flan-t5-xl 生成每個圖像的描述。這邊我設的問題如下:

Please describe the content of this image in detail.

我最後選擇 blip2-flan-t5-xl 模型輸出的因為它生成的文字比較穩定且正確，在使用 Salesforce/blip2-opt-2.7b 時我也發現到如果不給他 Answer:的結尾或開頭設定 Question 他會不輸出文字，這個問題我也有在他的 hugging face 討論發現



後來我參考她的方式就能成功輸出了，但猶豫 blip2-flan-t5-xl 還是比較穩定且正確，所以我最後選擇他生成的描述。

Prompt Design:

這邊我參考助教 PPT 上的 prompt 以及一點點小小的更改製作 prompt_w_label 以及 prompt_w_suffix，prompt_w_label 主要就是把 generated_text 後面加上 you should have 那張圖的 label，prompt_w_suffix 就是把 prompt_w_label 加上一些東西，範例如下。

範例:

generated_text:woman in warehouse holding clipboard

Prompt Template #1(prompt_w_label**)**:

woman in warehouse holding clipboard. imagesize:512x512.Image should have Head, Person, Ear, Face, Hands

Prompt Template #2(prompt_w_suffix**)**:

woman in warehouse holding clipboard. imagesize:512x512.Image should have Head, Person, Ear, Face, Hands.HD quality, highly detailed and real.

Text-to-Image Generation

這邊生成圖片我使用 masterful/gligen-1-4-generation-text-box 生成圖片，使用剛剛製作的 prompt_w_label 以及 prompt_w_suffix 只用 text 給他生成圖片，然後用 FID 衡量生成圖像與真實圖像之間的相似度，結果如下表格

結果發現 Template #1 的 FID 是比較好的，所以我使用 Template #1 加上 layout 重新生成圖片，最後結果如下表格:

	Text grounding		Layout-to-Image
prompt	Template #1	Template #2	Template #1
FID	54.931862	55.7761929541	60.06578