1.參考論文

2023_Ge_Expressive_Text-to-Image_Generation_with_Rich_Text

2023_ITI-GEN_Inclusive_Text-to-Image_Generation

Text to Image Generation with Semantic-Spatial Aware GAN

2018_AttnGAN Fine-Grained Text to Image Generation

LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation

InstructPix2Pix Learning to Follow Image Editing Instructions

TurboEdit Instant text-based image editing 2024

2. Image Captioning - BLIP-2

BLIP-2（Bootstrapping Language-Image Pretraining）是一個結合視覺與語言理解的深度學習模型，專門用於視覺語言任務，如圖像描述生成（image captioning）、視覺問答（visual question answering，VQA）等。它由 Salesforce AI Research 提出，並在視覺與語言理解領域取得了顯著進展。

BLIP-2 的核心創新之一是引入了一種新的「bootstrapping」方法，通過將視覺特徵與語言特徵緊密結合來進行預訓練。它的工作流程大致可以分為以下幾個步驟：

1.視覺特徵提取：BLIP-2 使用輕量級的視覺編碼器（如目標檢測模型或卷積神經網絡）從圖像中提取特徵。這些特徵反映了圖像中的關鍵信息，如物體、場景等。

2.語言特徵提取：該模型使用語言理解網絡（如 Transformer）來處理文本數據。這些文本數據可以是圖像的描述、問題或與圖像相關的其他語言表達。

3.視覺-語言融合：BLIP-2 的創新之處在於，它將視覺特徵和語言特徵通過一個融合模塊（稱為「視覺-語言交互模塊」）進行結合，從而生成可以互動和理解的多模態表示。這使得模型能夠理解圖像與文本之間的語義關係。

4.增強學習：在預訓練過程中，BLIP-2 進行自監督學習，通過基於圖像的語言生成或基於文本的視覺理解任務進行學習。這一過程被稱為「bootstrapping」，即模型自我增強能力的過程。

BLIP-2 的優勢之一是其能夠以較少的計算資源進行高效訓練，並且在視覺語言理解任務上表現出色。該模型在多個視覺語言理解基準數據集（如 VQA、Flickr30k 等）上都取得了領先的結果。

3.比較模型

因為我電腦設備較差，我選擇使用：Salesforce/blip2-opt-2.7b 跟 Salesforce/blip2-flan-t5-xl 進行比較，
以圖片 pexels-photo-7019161.jpeg 為例，

Salesforce/blip2-opt-2.7b"generated_text": "a woman in a uniform standing in a warehouse"
Salesforce/blip2-flan-t5-xl"generated_text": "a woman leaning against a wall in a warehouse"
我主觀認為 Salesforce/blip2-opt-2.7b 生成的句子比較好，所以選擇用 Salesforce/blip2-opt-2.7b 作為模型

4.Template Design for Text-to-Image Generation Comparison

"generated_text": generated_text,
"prompt_w_label": f"{generated_text},focus on on {labels_str},focus on bboxes, high resolution, highly detailed",
"prompt_w_suffix": f"{generated_text}, professional quality, highly detailed"

generated_text 是從 Salesforce/blip2-opt-2.7b 生成的，
我覺得不要下太多複雜的 prompt 比較有助於生成圖片 。

5.Text-to-Image Generation - GLIGEN

GLIGEN（Grounded Language-Image Generation）是一個將語言與圖像生成結合的新型深度學習模型，專注於生成與文字指令（text prompt）相關聯的圖像內容。該模型由微軟研究院提出，旨在克服傳統圖像生成模型在語言對應性和內容控制上的不足，並大幅提升多模態生成的準確性和靈活性。

GLIGEN 的核心特點
語言指導的圖像生成： GLIGEN 可以根據自然語言描述生成圖像，並且能夠準確地在空間上「定位」語言描述的物件或場景。例如，給定「一隻紅色的鳥停在樹枝上」，GLIGEN 會在生成的圖像中準確呈現該場景。
圖像內容的可控性： 與一般的文本驅動圖像生成模型不同，GLIGEN 允許用戶指定生成圖像的結構、對象位置和空間佈局，從而實現更精細的圖像控制。例如，可以指定某個物件在特定的位置。

可擴展性與模塊化設計： GLIGEN 基於預訓練的大型生成模型（如 Stable Diffusion 或 DALL-E）進行微調，並引入了可插拔的「語言-圖像對應模塊」。這些模塊可以動態加載，用於處理特定任務，讓模型更加靈活。
支持多種輸入模式： 除了文字提示外，GLIGEN 還可以接受結構化信息作為約束條件，例如物件的邊界框（bounding box）或草圖等，這使得它能夠適應更多應用場景。

GLIGEN 的技術架構
GLIGEN 的架構主要包括以下幾部分：
預訓練模型： 使用大規模預訓練的生成模型作為基礎，這些模型在大規模文本-圖像配對數據上訓練，具備強大的生成能力。
Grounding Module（語言-圖像對應模塊）： 這是 GLIGEN 的核心創新，該模塊負責將語言提示與生成圖像的

局部特徵關聯起來，從而確保生成內容與語言描述的對應性。
跨模態學習： 通過自監督學習，GLIGEN 在大規模的語言-圖像對應數據集上進行調整，學習語言和圖像特徵的對應關係。

6.Text Grounding and Image Grounding

Text Grounding: masterful/gligen-1-4-generation-text-box
Image Grounding: anhnct/Gligen_Text_Image

7.FID

	Text grounding		**Layout-to-Image**
prompt	Template #1	Template #2	Template #2
FID	114.03	113.45	113.62

text grounding 使用 prompt_w_label、prompt_w_suffix，因為 prompt_w_suffix 的 FID 分數比較低，所以採用 prompt_w_suffix 用在 Layout to image。

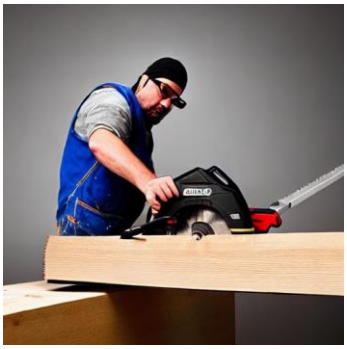
使用 layout to image 後 FID 分數反而提高了，我覺得可能是因為 Layout to image 多了 gounding bboxes 的條件所以才讓 FID 分數提高。

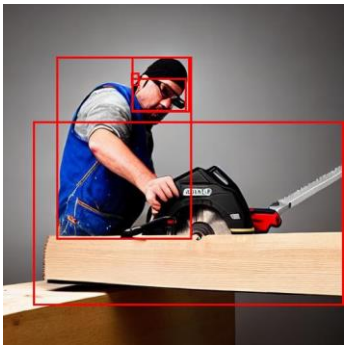
我認為 FID 分數雖然很重要，但這次作業的要求是要符合 bboxes，兼顧照片品質的同時，也要盡可能使照片符合 bboxes，我認為最關鍵變數是 guidance_scale=11.0，這個變數越高模型越會根據 prompt 的內容生成圖片，此變數最大值是 12，我設定 11，如此生成的圖片才能完美符合 bboxes。

8.Visual comparison

原圖 pexels-photo-8817849

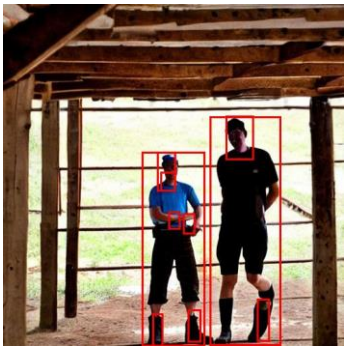
生成 pexels-photo-8817849



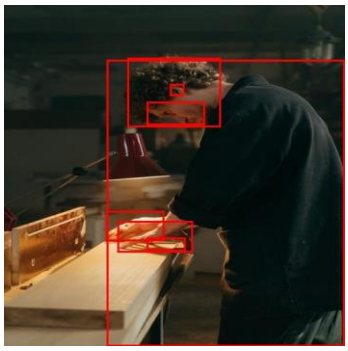


原圖 pexels-photo-16053596

生成 pexels-photo-16053596



原圖 pexels-photo-5089160



生成 pexels-photo-5089160