Subject-Diffusion: Open Domain Personalized Text-to-Image Generation without Test-time Fine-Tuning

Review: Progressive Denoising of Monte Carlo Rendered Images

- Problem of previous denoisers
  - Loss of Detail
  - Non-converging
- Solution: Mixing parameter α
  - Generate denoised image
  - Calculate error (SURE)
  - Feed rendered image and calculate α
  - Rescale α with t-statistics

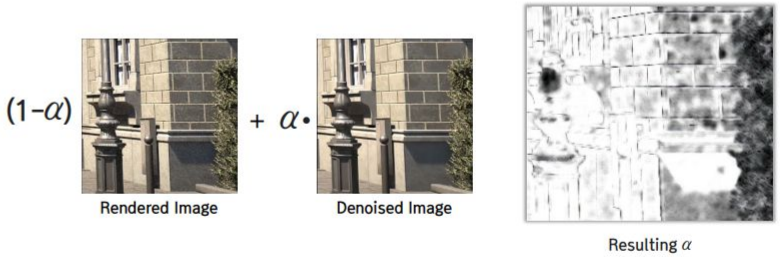


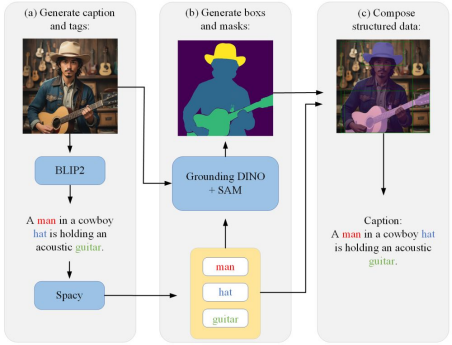
Table of Contents

Subject-Diffusion

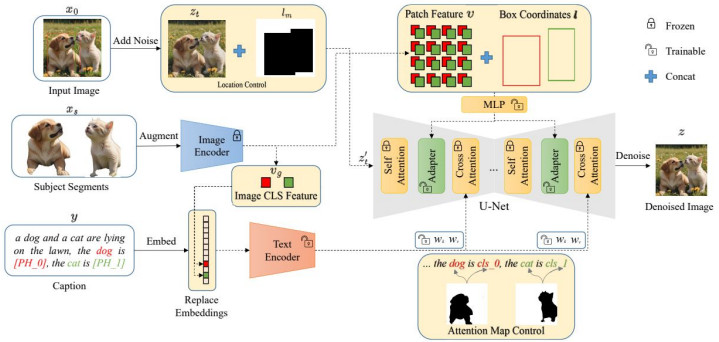
Subject-Diffusion is **personalized image generation model** that only requires a single reference image to support **personalized generation of single- or two-subjects** in any domain

Detailed Information of Subject Diffusion

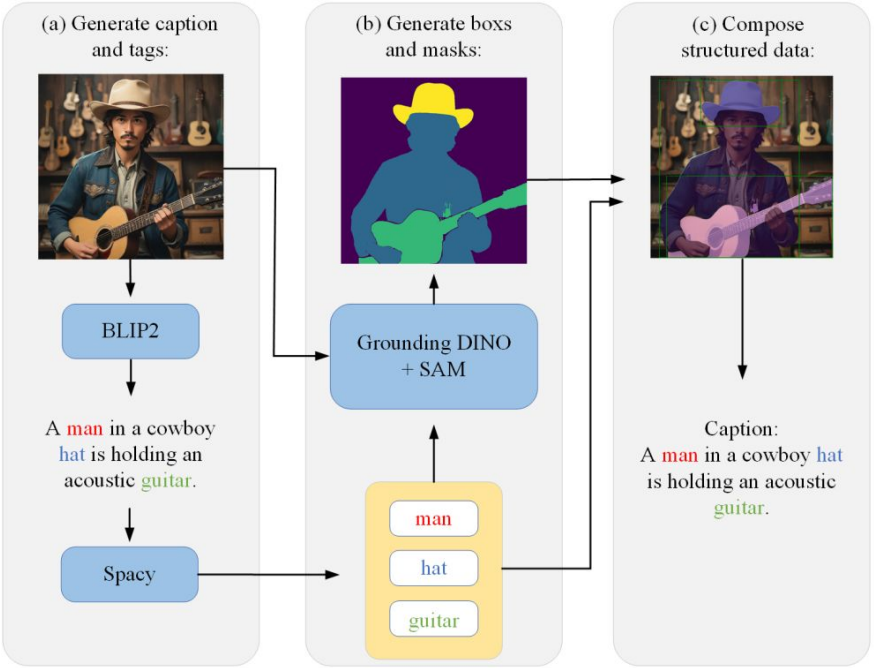
1. Dataset Construction



2. Model of Subject Diffusion



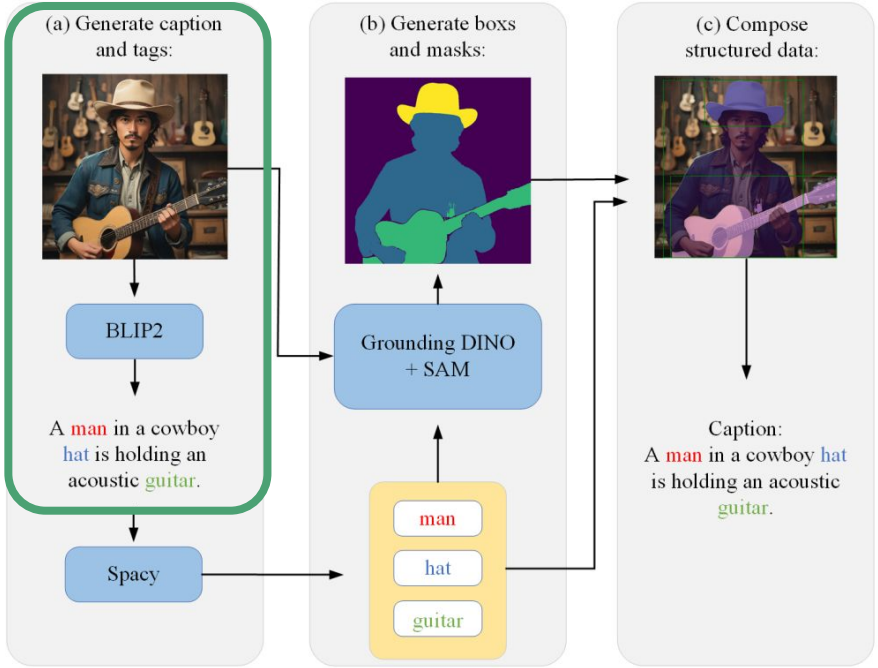
Dataset Construction



Input : picture
Output : structured data which contains (**S**ubject **D**iffusion **D**ataset)

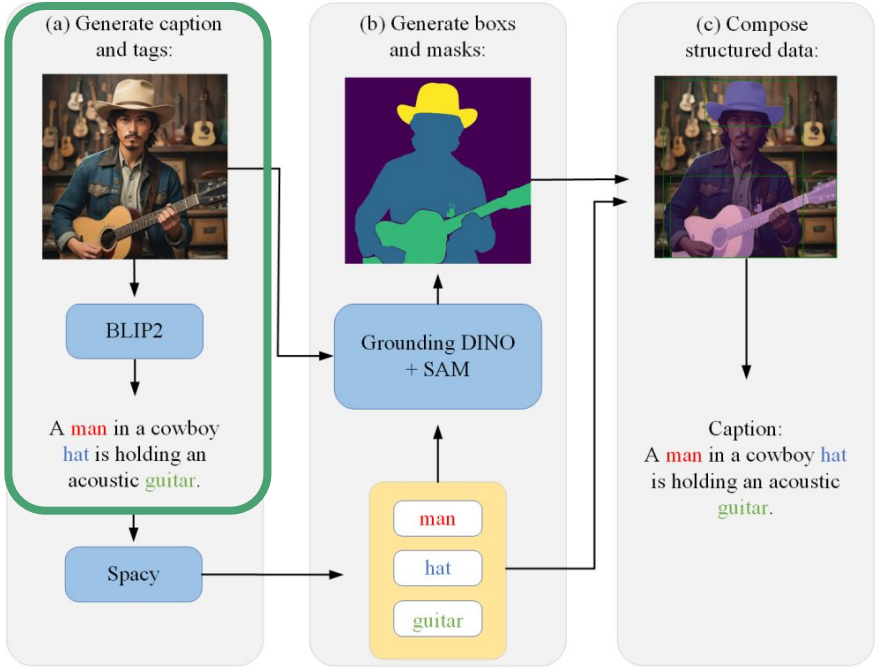
- **image-text pairs**
- **detection boxes**
- **segmentation masks**

Dataset Construction



1. use **BLIP2** to generate the caption of the given image

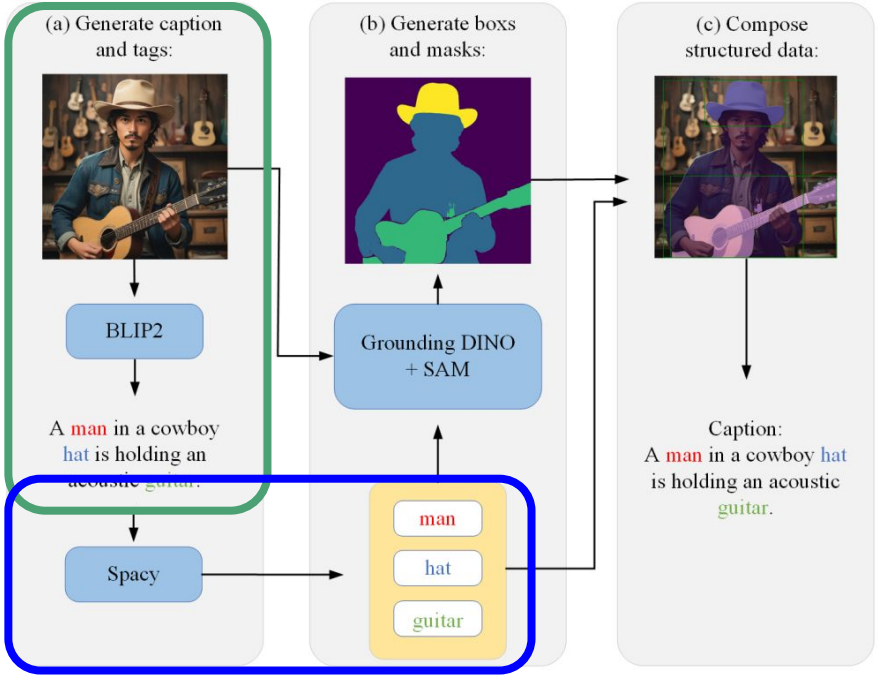
Dataset Construction



1. use **BLIP2** to generate the caption of the given image

BLIP2 : The **LLM**(Language Learning Model Which Receives **Image as input** and return the **comments of the image as output**

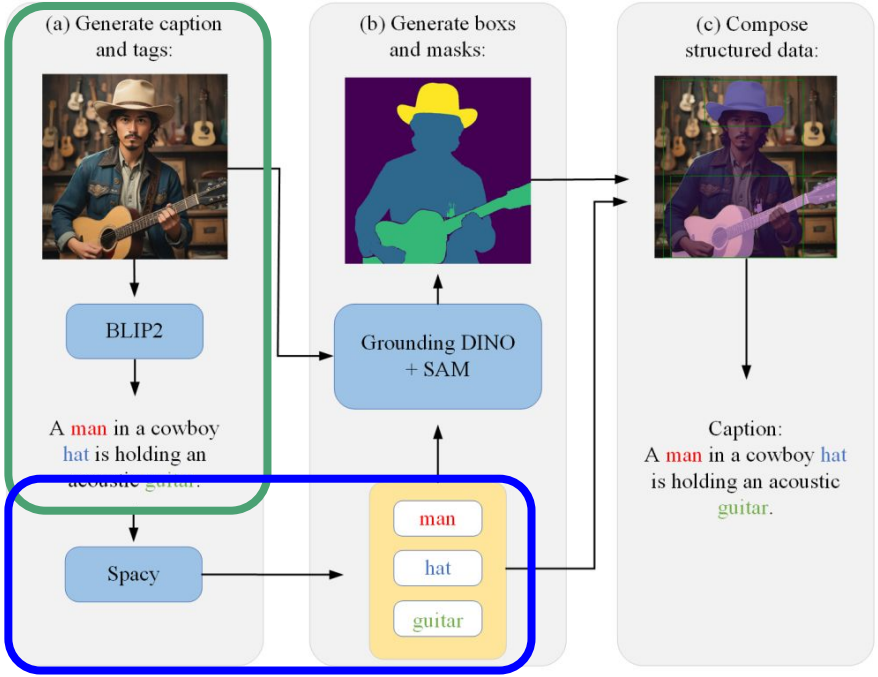
Dataset Construction



1. use **BLIP2** to generate the caption of the given image

2. use **spacy** to extract tags based on the part of each word in the caption sentence

Dataset Construction

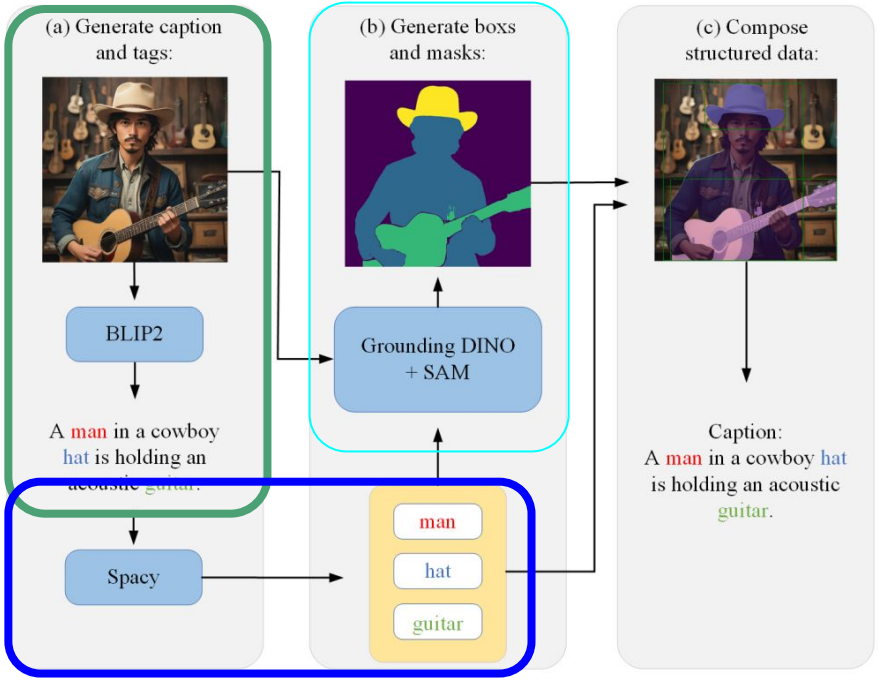


1. use **BLIP2** to generate the caption of the given image

2. use **spacy** to extract tags based on the part of each word in the caption sentence

Spacy : The Python library which is used to do the **Natural Language Processing**
While doing the dataset construction
It gets the **caption sentence as input** and returns **set of tags as output**

Dataset Construction

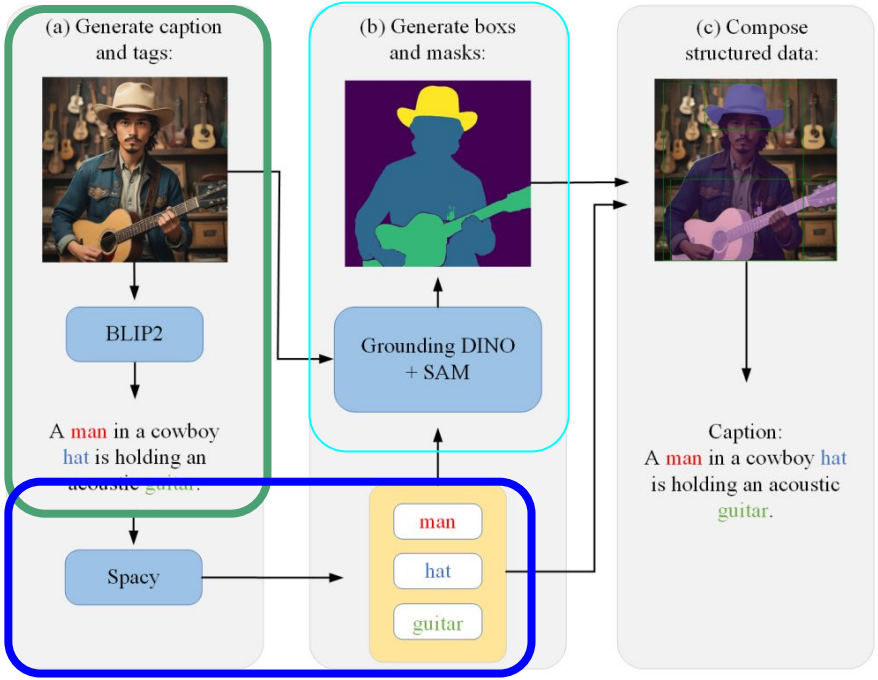


1. use **BLIP2** to generate the caption of the given image

2. use **spacy** to extract tags based on the part of each word in the caption sentence

3. use **Grounding DINO** to obtain detection boxes for each object and use the detection boxes are used as input for **SAM**

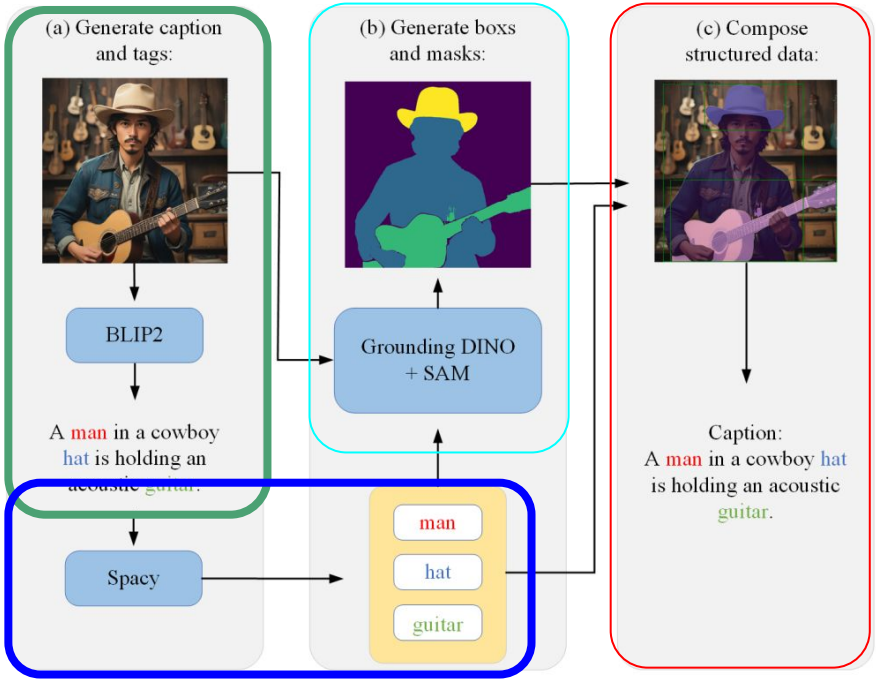
Dataset Construction



3. use **Grounding DINO** to obtain detection boxes for each object and use the detection boxes are used as input for **SAM**

Grounding DINO : Is the **pre-trained self-supervised model** which is used to do the **Zero-Shot Object Detection**
Input : image + set of tags
Output : detection box which each box contains the object related to tag

Dataset Construction



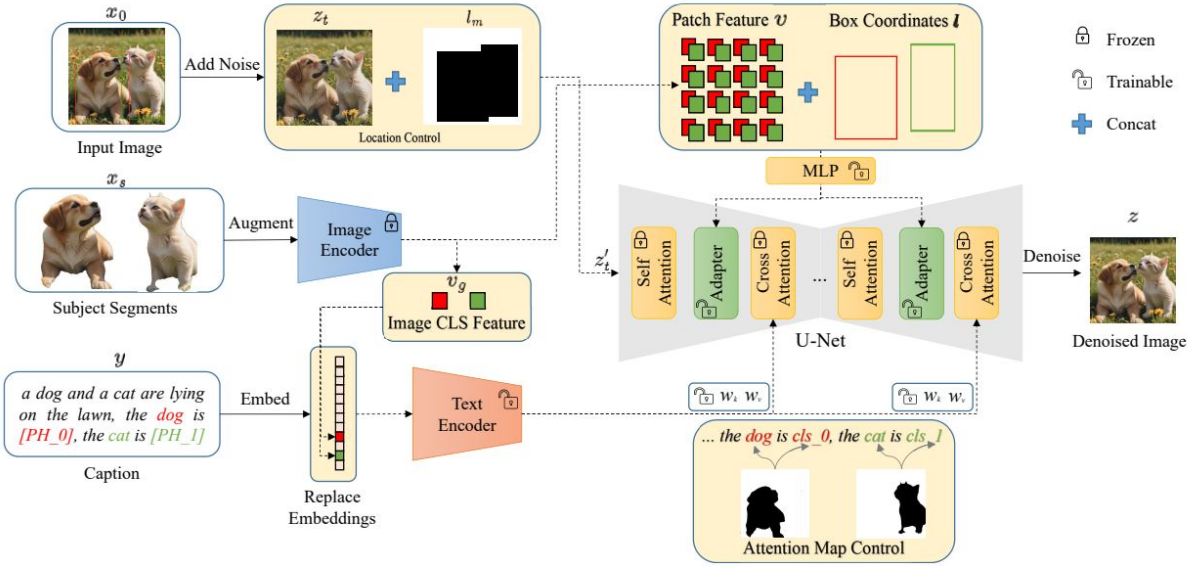
1. use **BLIP2** to generate the caption of the given image

2. use **spacy** to extract tags based on the part of each word in the caption sentence

3. use **Grounding DINO** to obtain detection boxes for each object and use the detection boxes are used as input for **SAM**

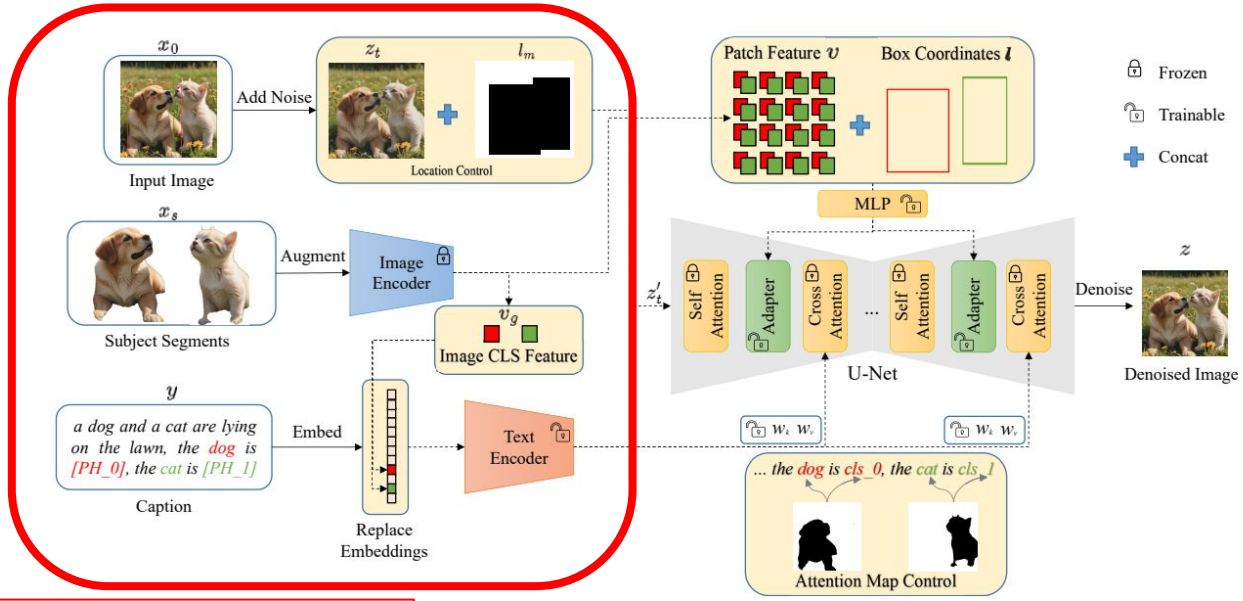
4. all of the different modalities are combined into **structured data**.

Model of Subject Diffusion



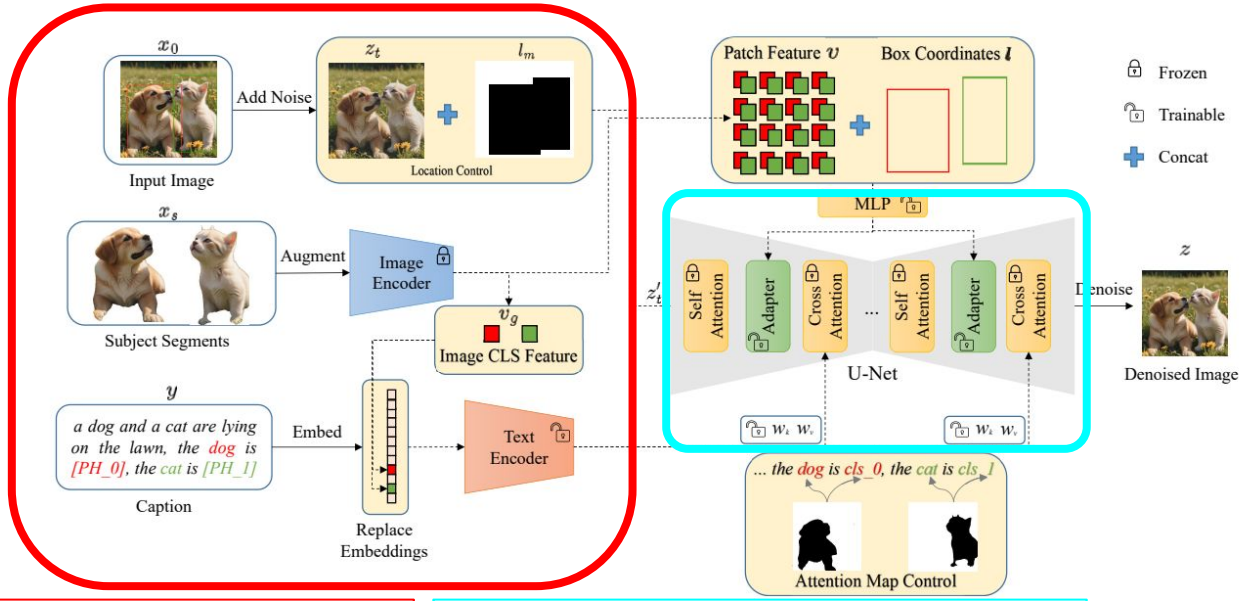
Input : Input image, subject segments, caption
Output : Denoised images

Model of Subject Diffusion



1.craft a specific prompt format and utilize a text encoder

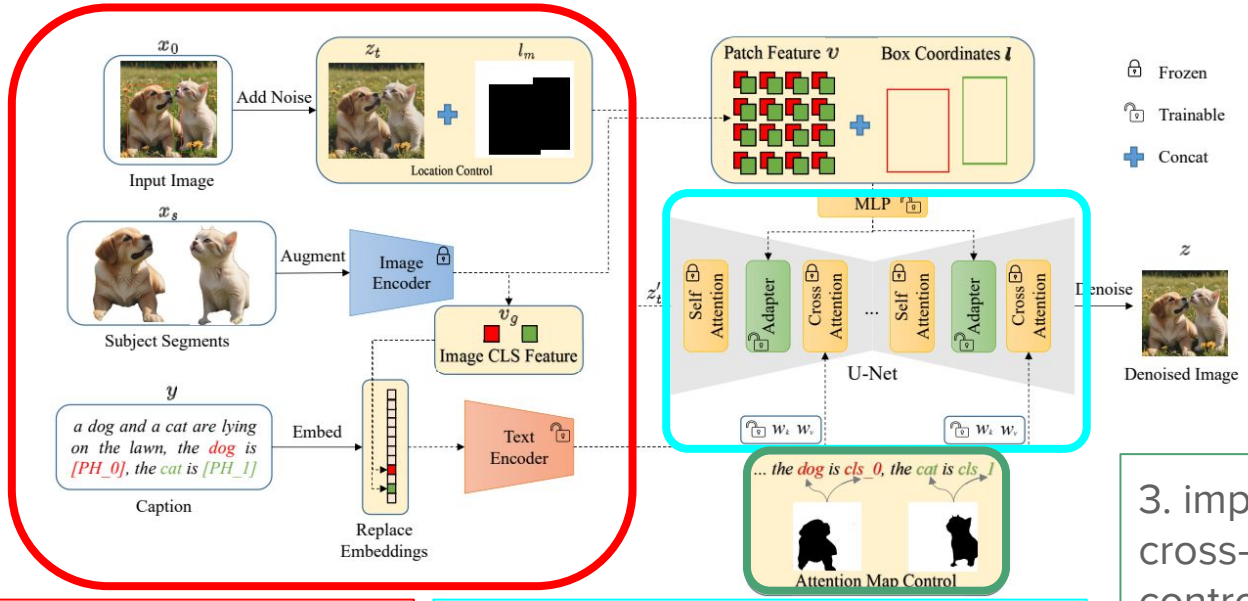
Model of Subject Diffusion



1.craft a specific prompt format and utilize a text encoder

2. integrating an adapter among each cross-attention block

Model of Subject Diffusion



1.craft a specific prompt format and utilize a text encoder

2. integrating an adapter among each cross-attention block

3. implement a cross-attention map control strategy

Experiments Using the Subject Diffusion

1. Single-subject Diffusion

Experiments Using the Subject Diffusion

1. Single-subject Diffusion

similarity index:
higher number is better

Methods	Type	Testset	DINO	CLIP-I	CLIP-T
Real Images †	-	-	0.774	0.885	-
Textual Inversion †	FT	DB	0.569	0.780	0.255
DreamBooth †	FT	DB	0.668	0.803	0.305
Custom Diffusion	FT	DB	0.643	0.790	0.305
ELITE	ZS	DB	0.621	0.771	0.293
BLIP-Diffusion †	ZS	DB	0.594	0.779	**0.300**
IP-Adapter †	ZS	DB	0.667	**0.813**	0.289
Subject-Diffusion	ZS	DB	**0.711**	0.787	0.293
		OIT	0.668	0.782	0.303

Experiments Using the Subject Diffusion

2. Two - subjects Diffusion

Experiments Using the Subject Diffusion

2. Two - subjects Diffusion

similarity index:
higher number is better

	Index	Methods	DINO	CLIP-I	CLIP-T
	(a)	**Subject-Diffusion**	**0.506**	**0.696**	**0.310**
	(b)	trained on OpenImage	0.491↓	0.693↓	0.302↓
Two	(c)	w/o location control	0.477↓	0.666↓	0.281↓
	(d)	w/o box coordinates	0.464↓	0.687↓	0.305↓
Subjects	(e)	w/o adapter layer	0.411↓	0.649↓	0.307↓
	(f)	w/o attention map control	0.500↓	0.688↓	0.302↓
	(g)	w/o image cls feature	0.457↓	0.627↓	0.309↓

Limitation of the Subject-Diffusion

1. subject-diffusion **face challenges in editing attributes and accessories** within user-input images ▶ limitations in the scope of the model's applicability
2. **fail** to make the harmonious images **which has more than two subjects**

Summary

1. What is Subject-Diffusion?
    a. Subject-Diffusion is **personalized image generation model** that only requires a single reference image to support **personalized generation of single- or two-subjects** in any domain
2. Detailed Information of Subject Diffusion
    a. Dataset construction
    b. Model Overview
3. Limitations
    a. scope of the model’s applicability is small
    b. fail to make the harmonious images which has more than two subjects

Q&A

Quiz