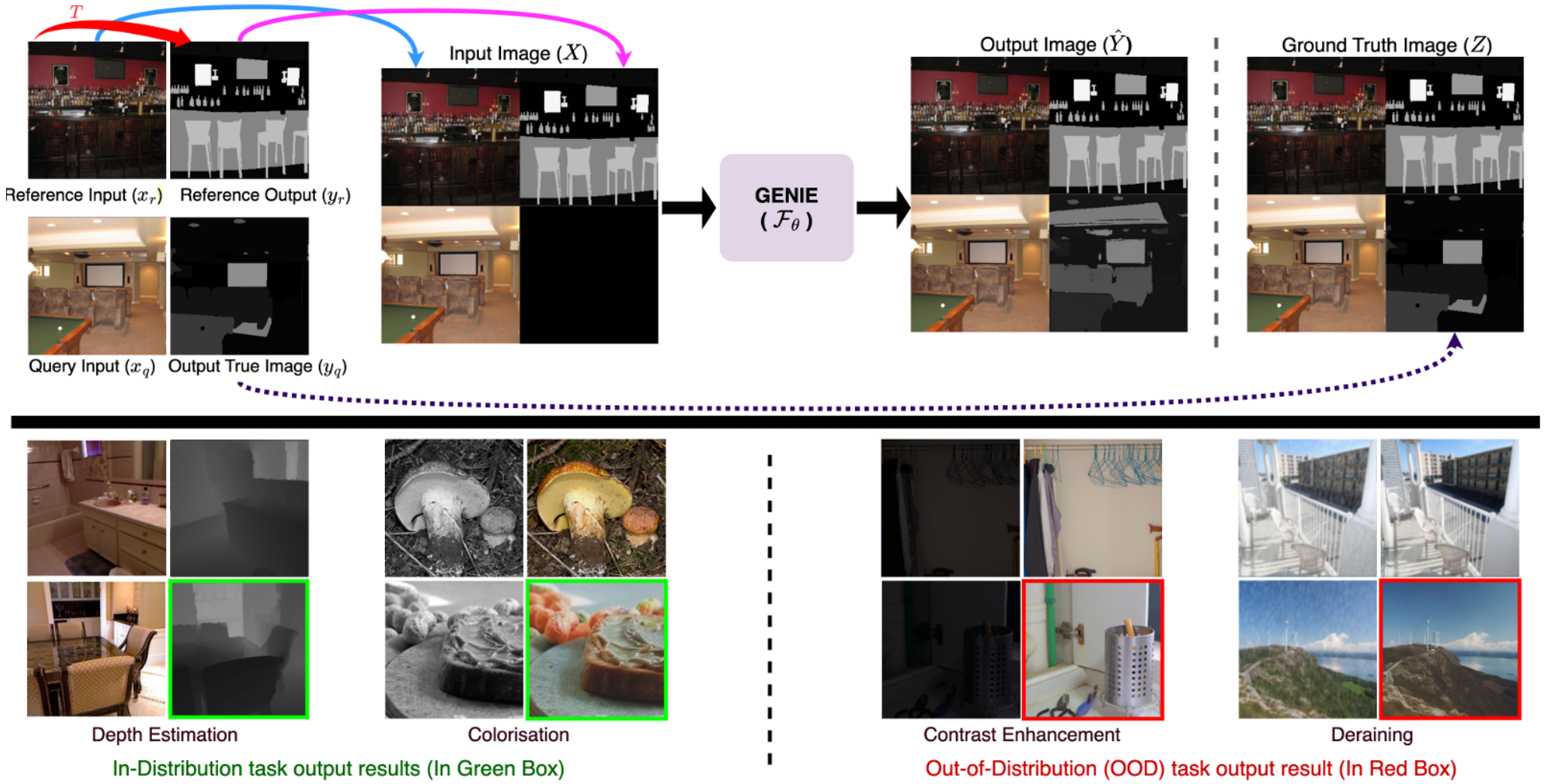
Visual Guided diffusion model as a multi-task learner

By: Aniket Thomas (22M2162)

Overview

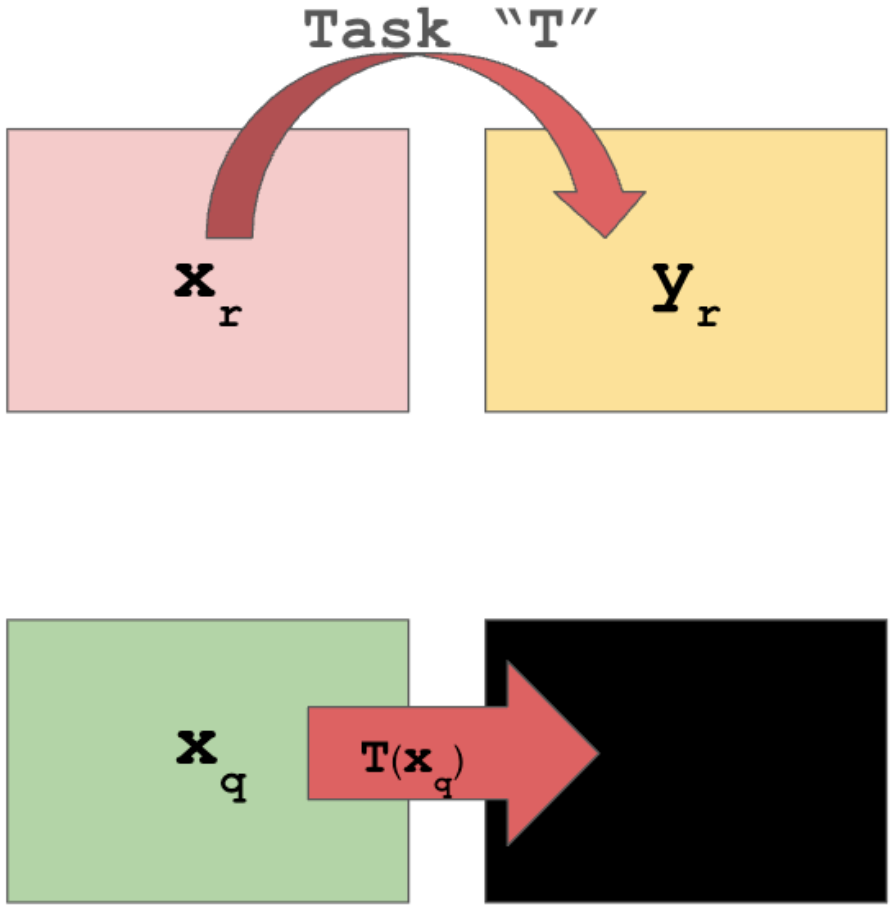

T
Reference Input (x_r)
Reference Output (y_r)
Query Input (x_q)
Output True Image (y_q)
Input Image (X)
GENIE
($\mathcal{F}_\theta$)
Output Image ($\hat{Y}$)
Ground Truth Image (Z)
Depth Estimation
Colorisation
Contrast Enhancement
Deraining
In-Distribution task output results (In Green Box)
Out-of-Distribution (OOD) task output result (In Red Box)

Outline of the presentation

- Problem formulation
- Related work
- Architectural design choices
- Replacing text conditioning
- Training details
- Quantitative results
- Qualitative results
- Future Work

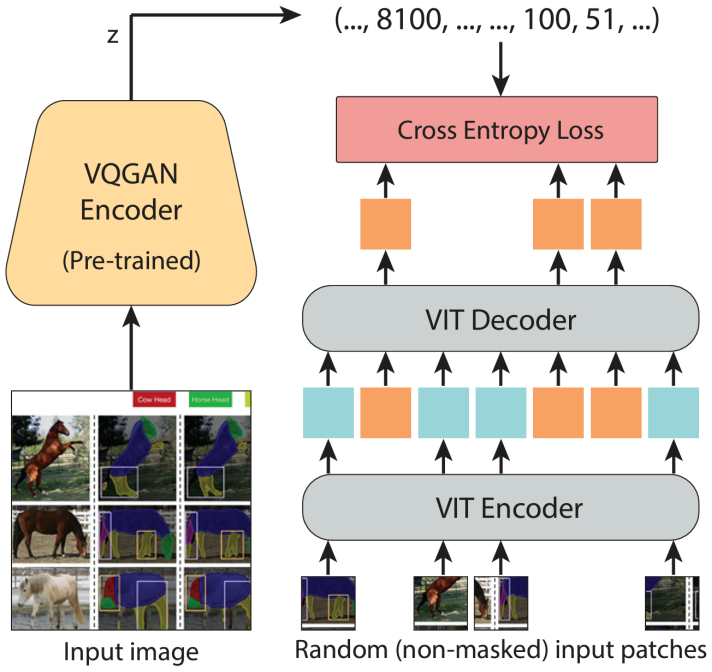
Problem Formulation

- The goal is to transform one image into another, referred to as tasks, in a multi-task framework relying on visual guidance only (absence of textual description)
- Formally:
  - $\mathbf{x_r}$: Reference input image
  - **T**: An unknown image transformation (task) from a set of predefined "K" tasks.
  - $\mathbf{y_r = T(x_r)}$: Transformed reference image, obtain by applying transformation T to query image.
  - $\mathbf{x_q}$: Query input image
  - $\mathbf{y_q = T(x_q)}$: Ground truth transformation of the query image using task T
  - $\mathbf{y_{pred}}$: Model generated transformation of $\mathbf{x_q}$
- The model learns to generate $\mathbf{y_{pred}}$ from query image $\mathbf{x_q}$,based on the reference pair ($\mathbf{x_r}$, $\mathbf{y_r}$) , to match the ground truth $\mathbf{y_q}$.

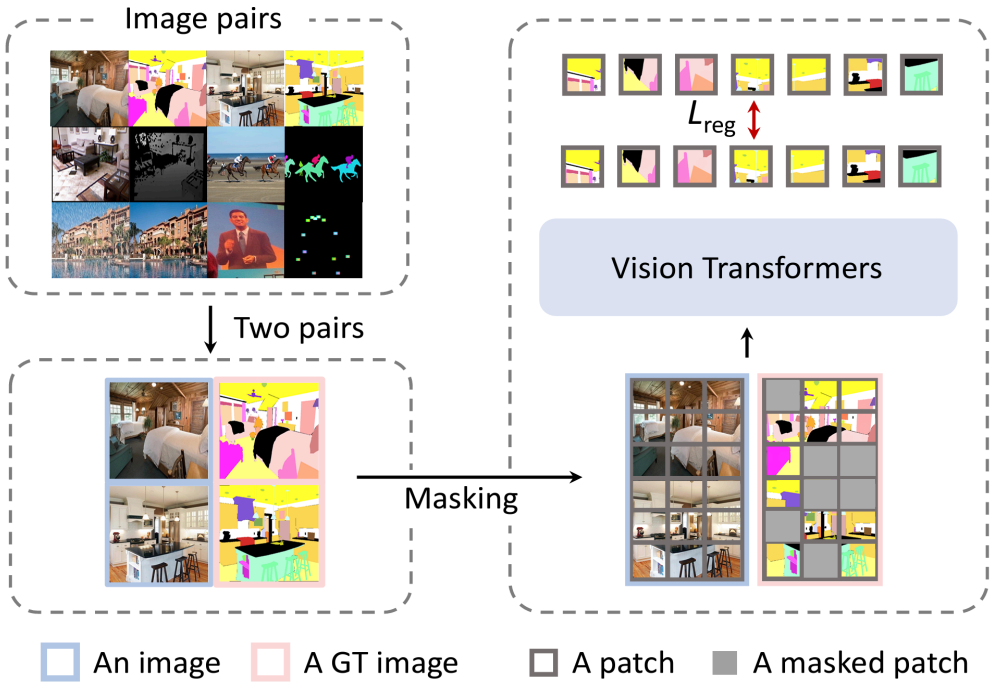


Related Work

- Two related work that closely resembles visual prompting are:
  - Visual Prompting via Image Impainting (Left)
  - Painter (Right)
- Both architecture utilizes impainting approach using transformers to generate the desired output.



https://arxiv.org/abs/2209.00647



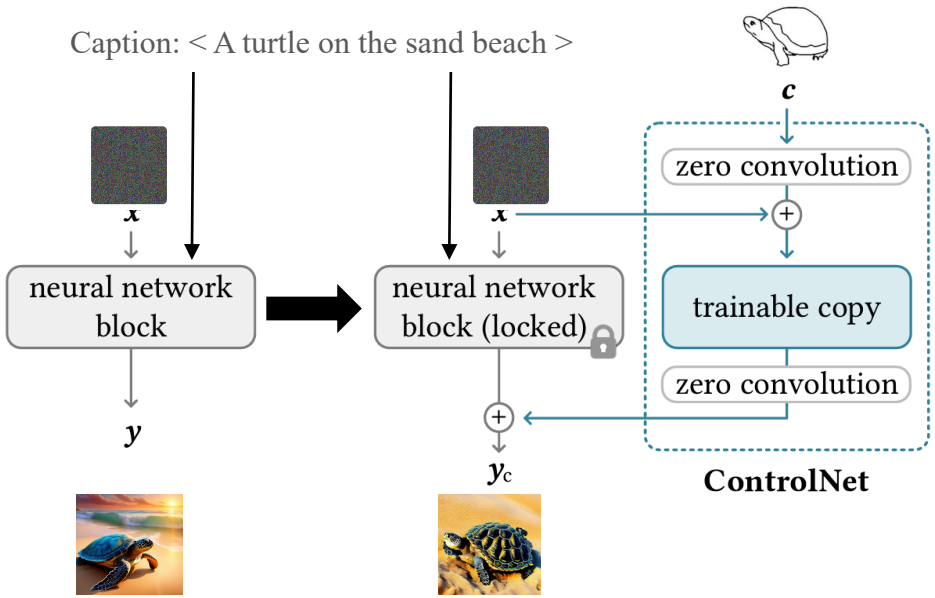
https://arxiv.org/abs/2212.02499

Related Work

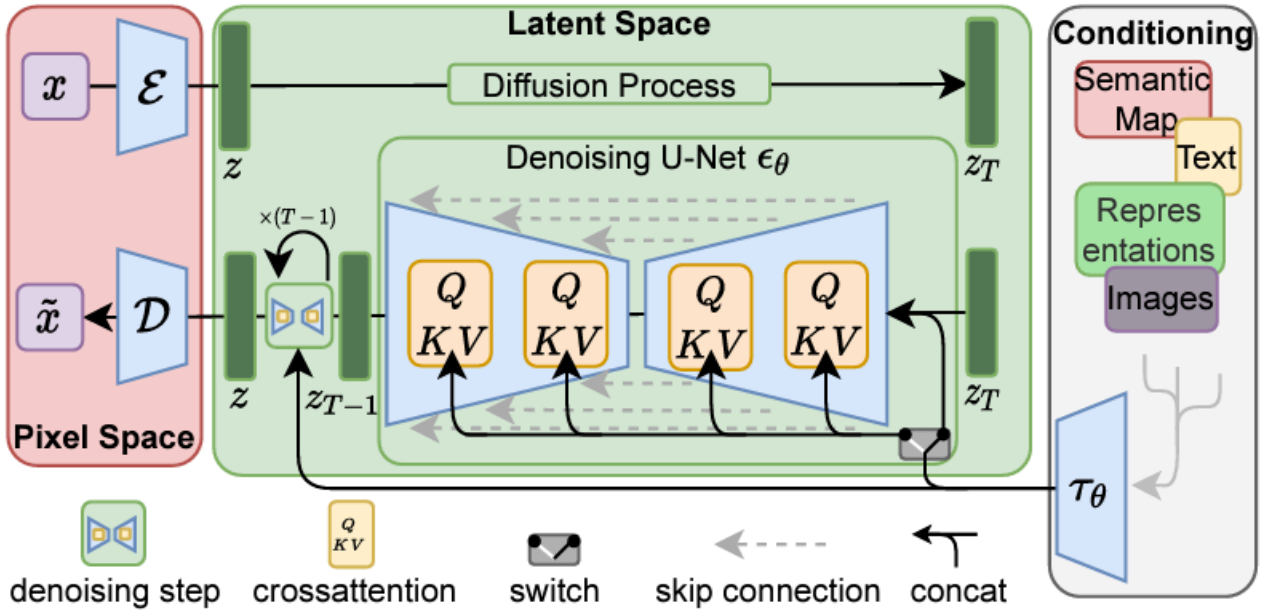
- Visual Prompting performs well for both in-distribution & Out-of-distribution (OOD) tasks but the quality of images produced is of low-quality.
- Painter performs well as a multi-task learner for in-distribution data & tasks but struggles with OOD tasks.
- Both the work lack a conditioning mechanism which tries to enforce alignment between the intended task in reference and output

Architectural Design Choices

- Instead of traditional vision transformer-based architecture from the related work, we select SOTA latent diffusion model for generating the output of query input image.
- For fine tuning we utilized ControlNet framework which in addition to textual input also uses a conditional image



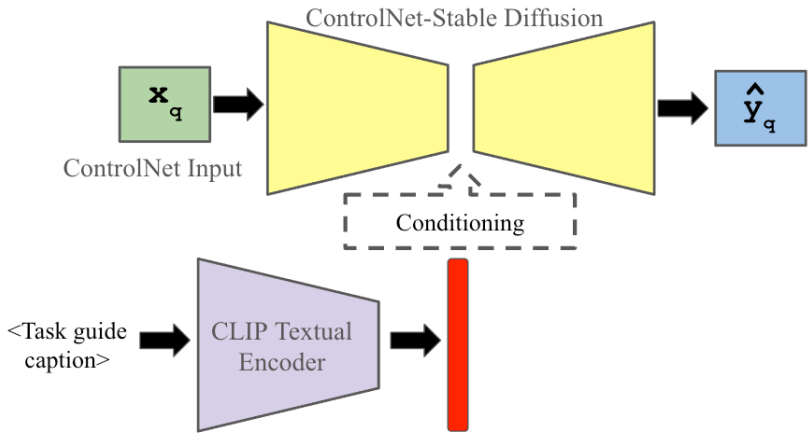
https://arxiv.org/abs/2302.05543

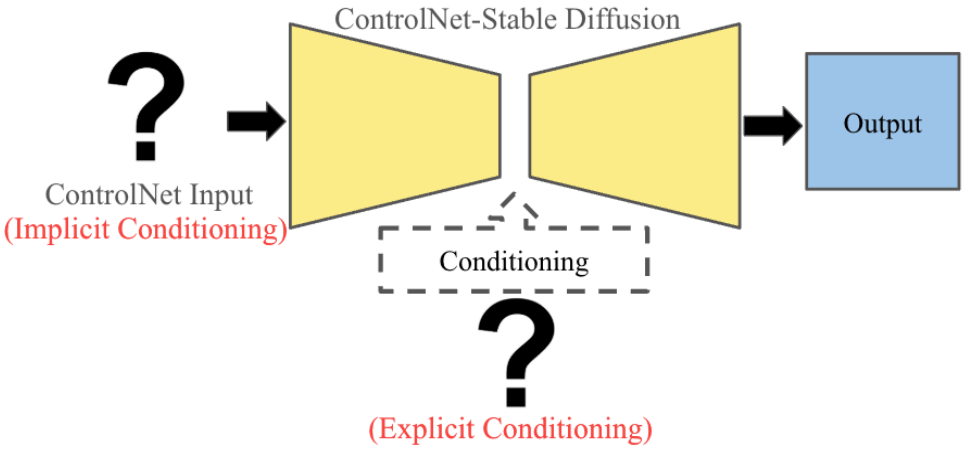


https://arxiv.org/abs/2112.10752

Design Choices Questions

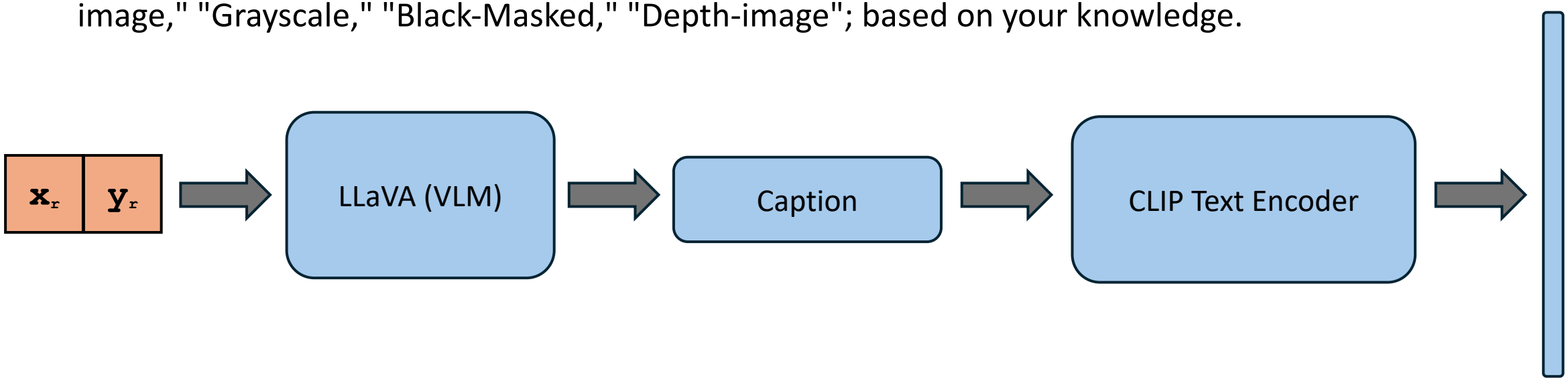
- What should be fed as an input to ControlNet?
- How to replace textual conditioning effectively?





Replacing textual condition (VLM Based)

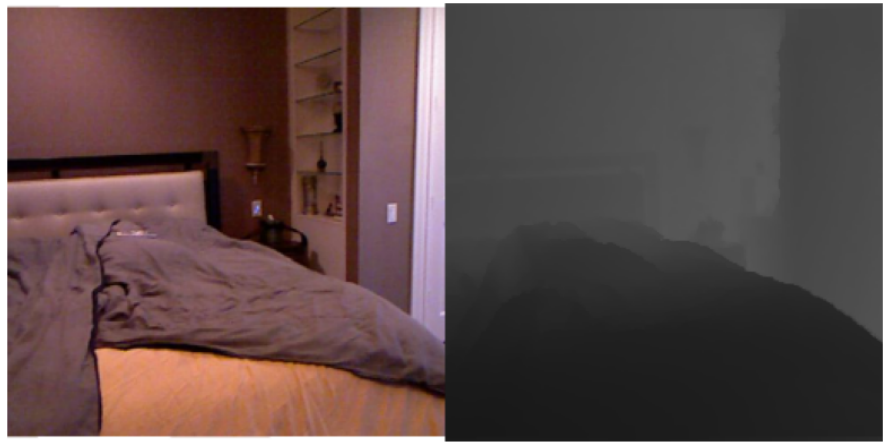
- LLaVA Vision-Language Model (VLM) was used to generate caption corresponding to the transformation of reference pair of images in zero-shot setting
- We can utilize the caption generated to generate textual embeddings in diffusion model, effectively replacing textual dependency.
- **Caption**: The image is a concatenation of two images side by side. Tell me the relationship between the images. You can instruct like "The right image is <task> of the left image." Choose the closest <task> out of "Segmentation," "Denoised," "Colorization," "Hed map," "Boundary image," "Grayscale," "Black-Masked," "Depth-image"; based on your knowledge.



VLM Success Cases

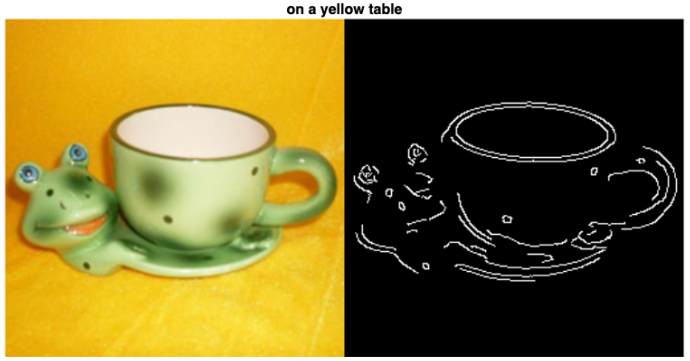
The right image is a colorized version of the left image.

The right image is a black-masked, depth-image of the left image.



VLM Failure Cases

The right image is a black and white image of a coffee cup. The left image is a green coffee cup sitting on a yellow table

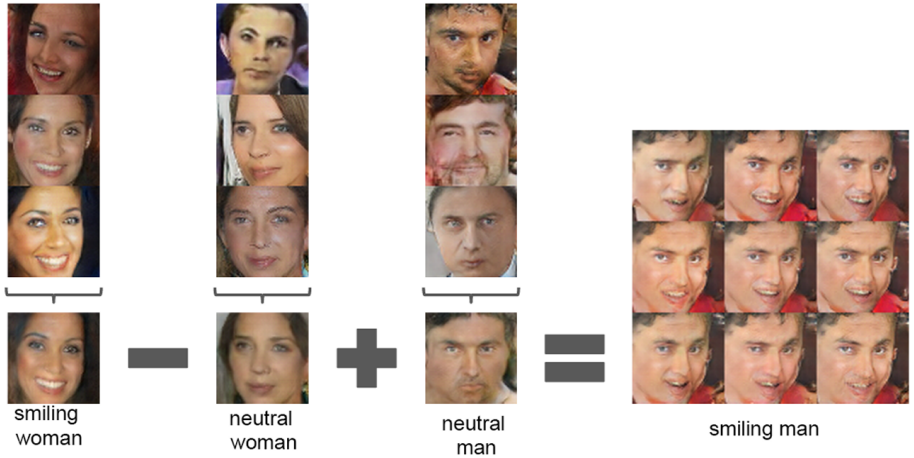


The right image is a black-masked, depth-image of the left image.

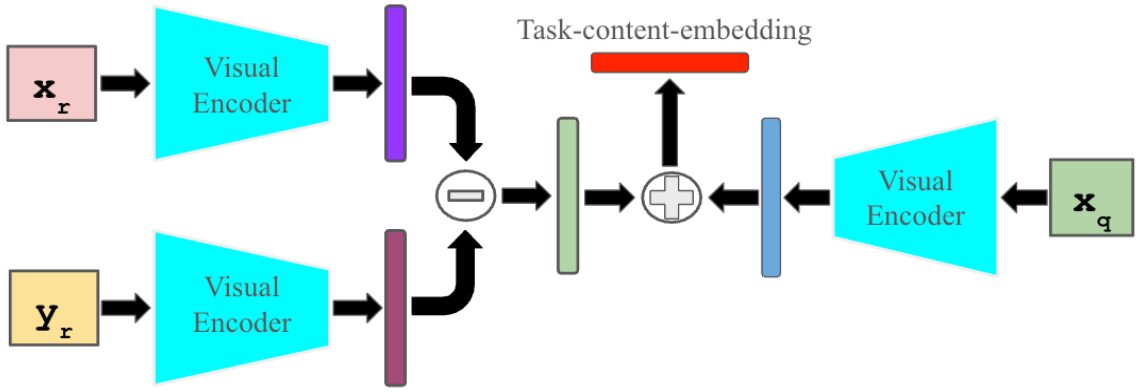
The right image is a denoised version of the left image.

Task Arithmetic Embedding Based

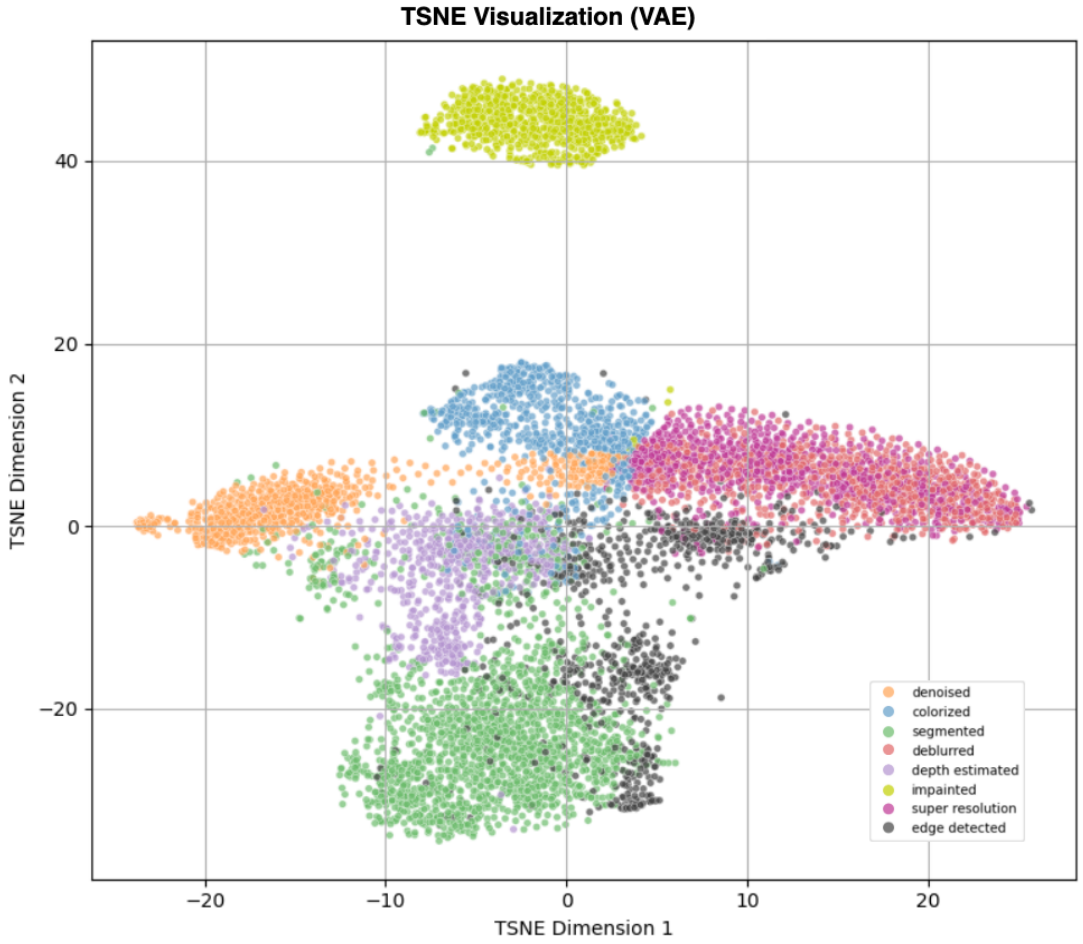
- Inspired from latent task arithmetic of GAN, the aim is to utilize embeddings to fetch task and content from reference and query image respectively.
- Encoder model needs to be decided of which VAE and CLIP Image Encoder are the primary candidates
- Also, to fetch task information an external penalization in loss might be required.

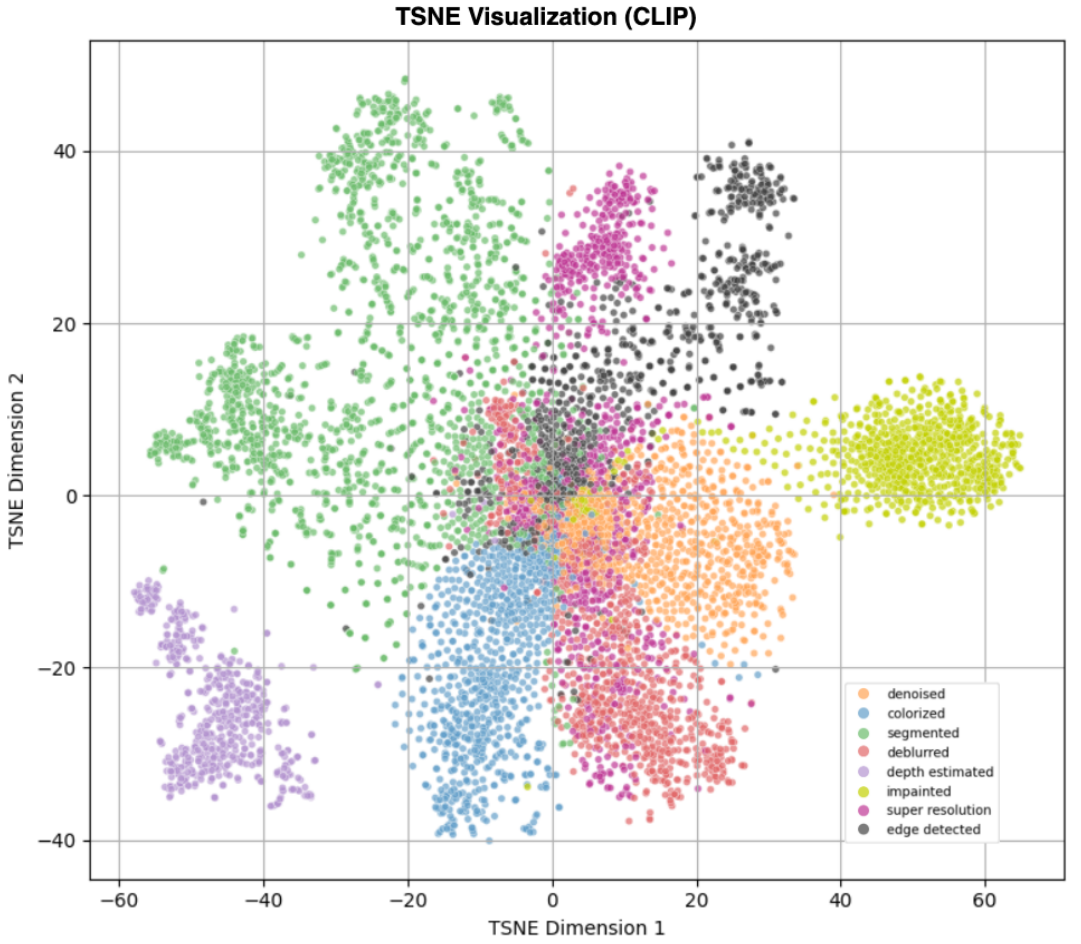


https://arxiv.org/abs/1511.06434v2

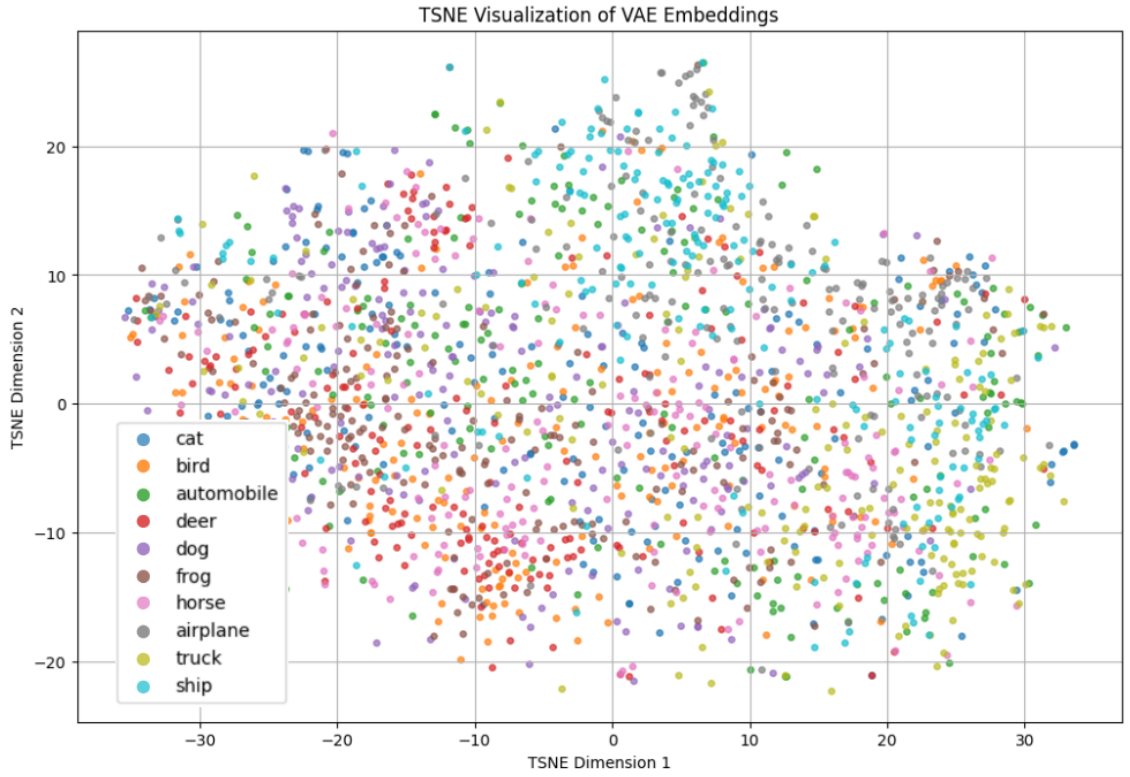


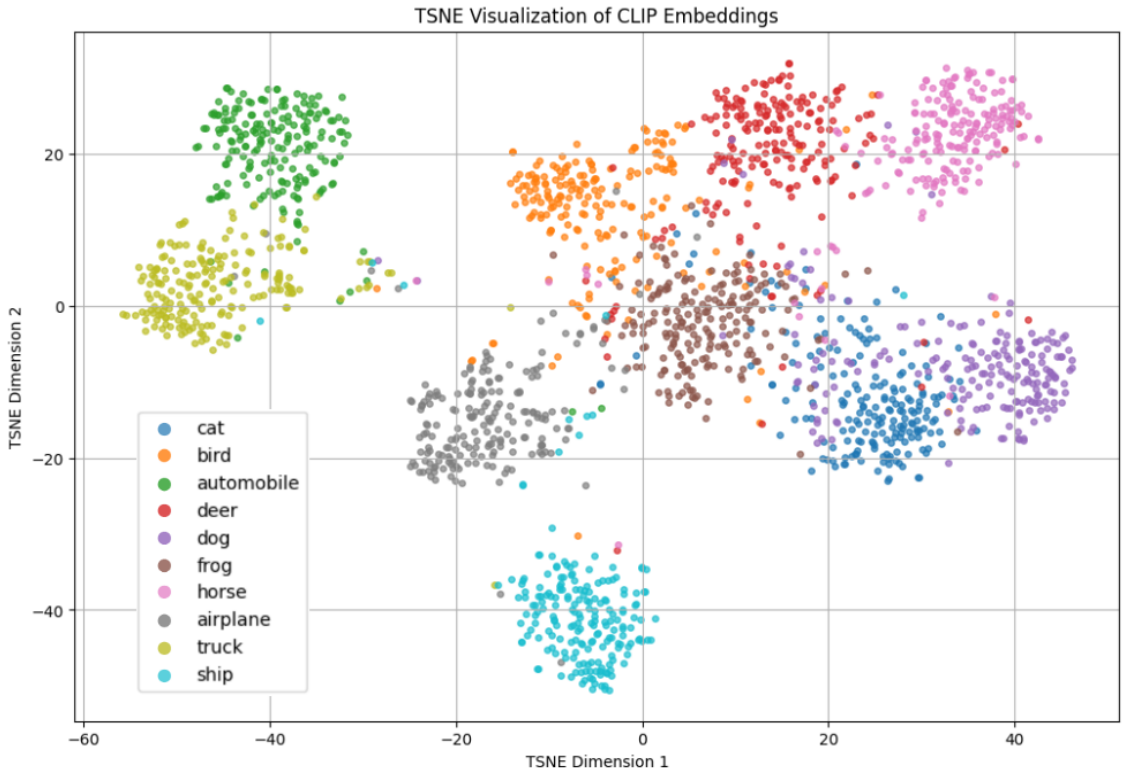
Task Specific Embeddings





Content Specific Embeddings





Training Details (Dataset)

- We categorize tasks and datasets into in-distribution and out-of-distribution (OOD).
- For a given model, in-distribution tasks refer to those included in training, while OOD tasks are not encountered during training.

Model	In-Distribution Tasks	In-Distribution Datasets	OOD Tasks
Our model	Impainting, Deblurring, Colorization, Edge Det., Super-Res., Denoising, Segmentation, Depth Est.	ImageNet-22K ADE20K NYU Depth V2	Deraining, Contrast Enhance
Visual Prompt	Impainting	-	All tasks except Impainting
Painter	Depth Est., Denoising, Segmentation, Contrast Enhance	NYU Depth V2, SIDD, ADE20K, LOL	Super-Res., Deblurring, Colorization, Impainting

Dataset	Number of Training Samples	Number of Testing Samples	Tasks
ImageNet	15000	15000	Impainting, Deblurring, Colorization, Edge Detection, Super-Resolution, Denoising
ADE 20K	15000	3000	Segmentation
NYU V2	15000	654	Depth Estimation
LOL	-	15	Contrast Enhancement
Deraining	-	500	Deraining

Training Details (Choice of Task)

- Mixture of high-level tasks and self-supervision tasks were chosen by comparing with the tasks in literature
- "K" tasks were chosen for training purposes:
  - Impainting
  - Colorization
  - Edge Detection
  - Super Resolution
  - Deblurring
  - Denoising
  - Semantic Segmentation
  - Depth Map

Colorisation

Edge Detection

Impainting

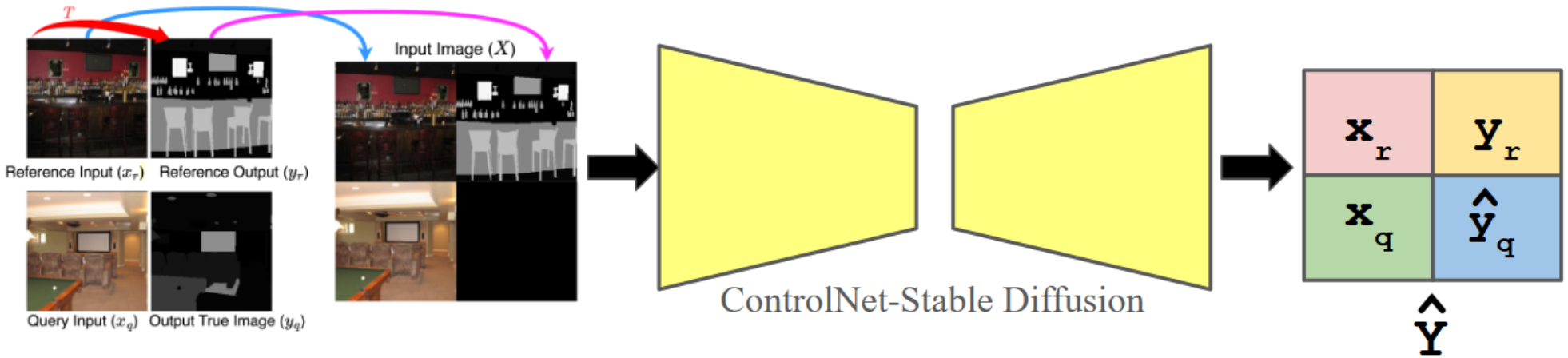
Depth Estimation

Denoising

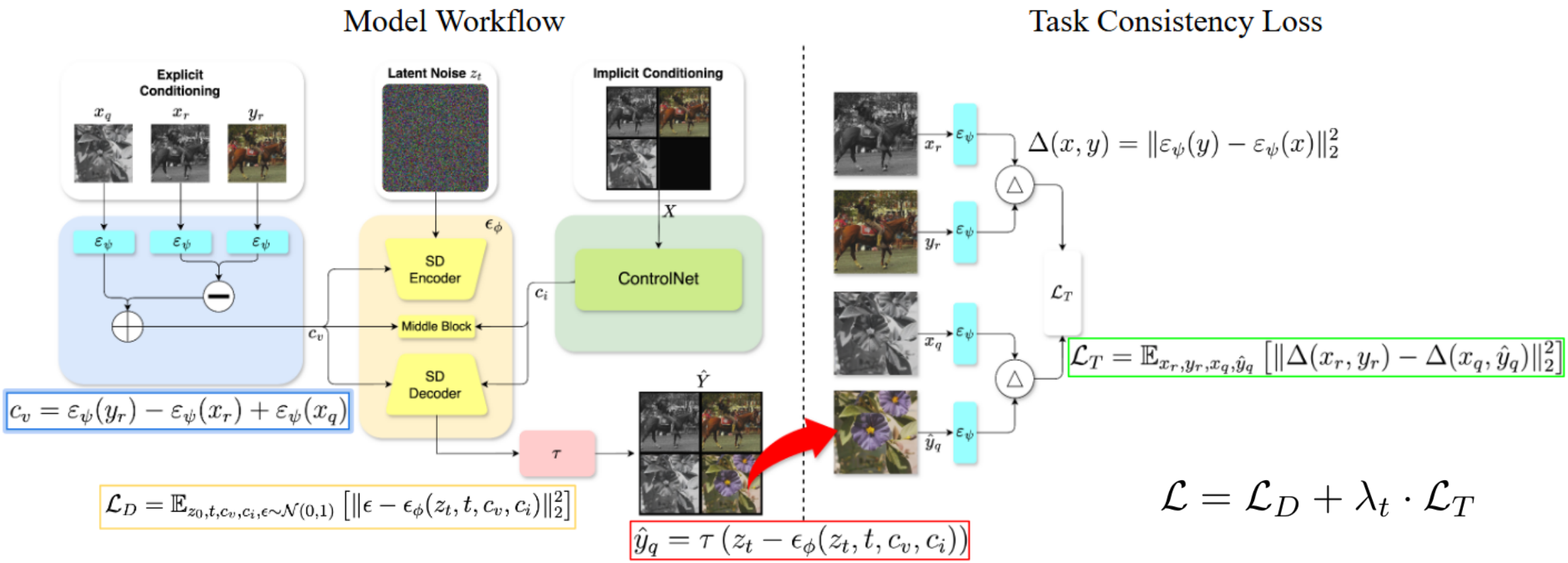
Semantic Segmentation

Training Details (Implicit Condition)

- Input to Control Net is a 2x2 grid with reference images and masked output of query image.
- Due to self-attention mechanism in control block, it is denoted as implicit conditioning.
- Diffusion Model aims to "impaint" the masked output of query image



Training Details (Architecture and Loss)



Training Details (Hyperparameters)

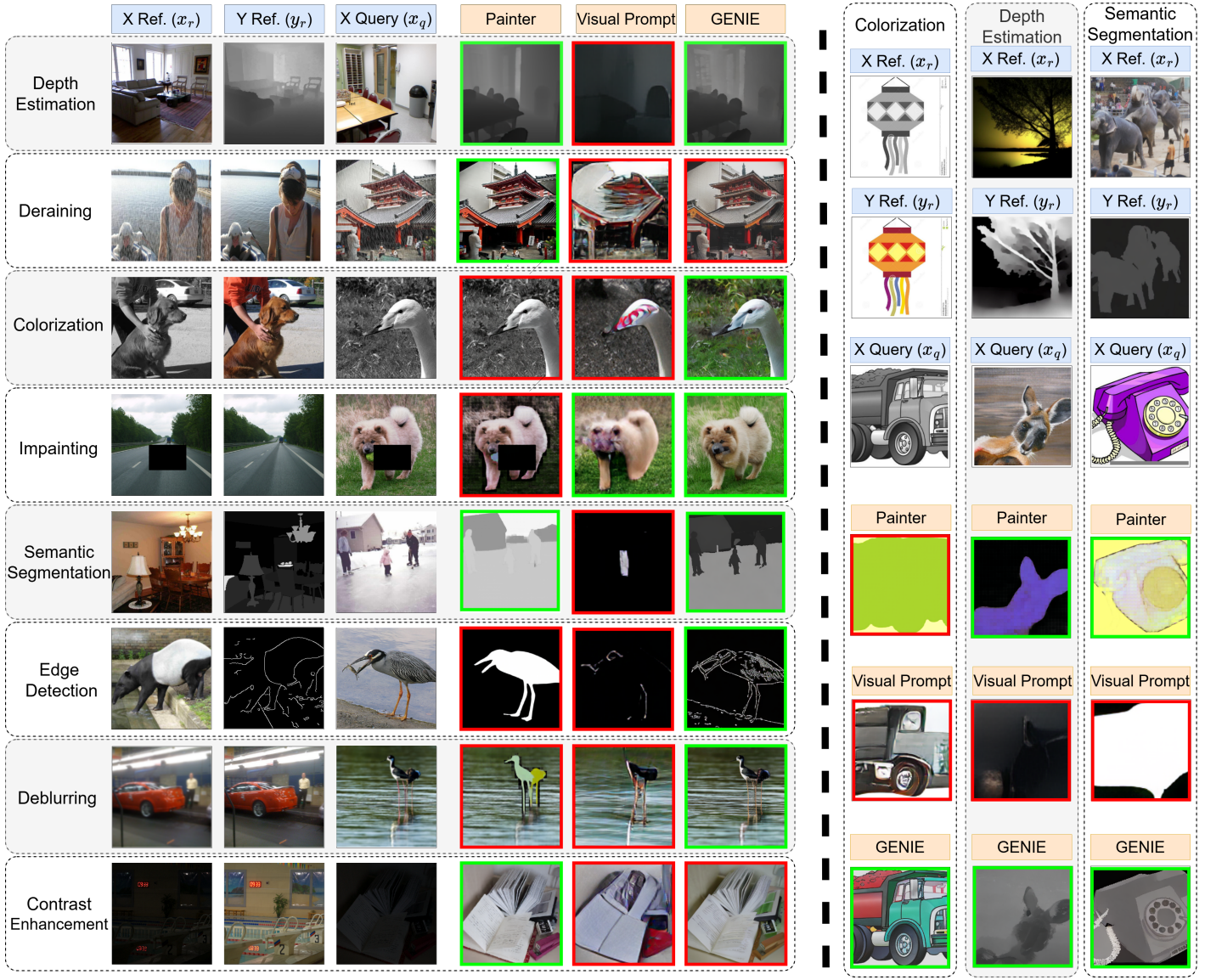
- Optimizer: Adam
- Learning Rate: 10^{-5}
- Loss Weighing Parameter λ_t : 0.05
- Batch Size: 16
- Training Time: 15 days

Quantitative Result

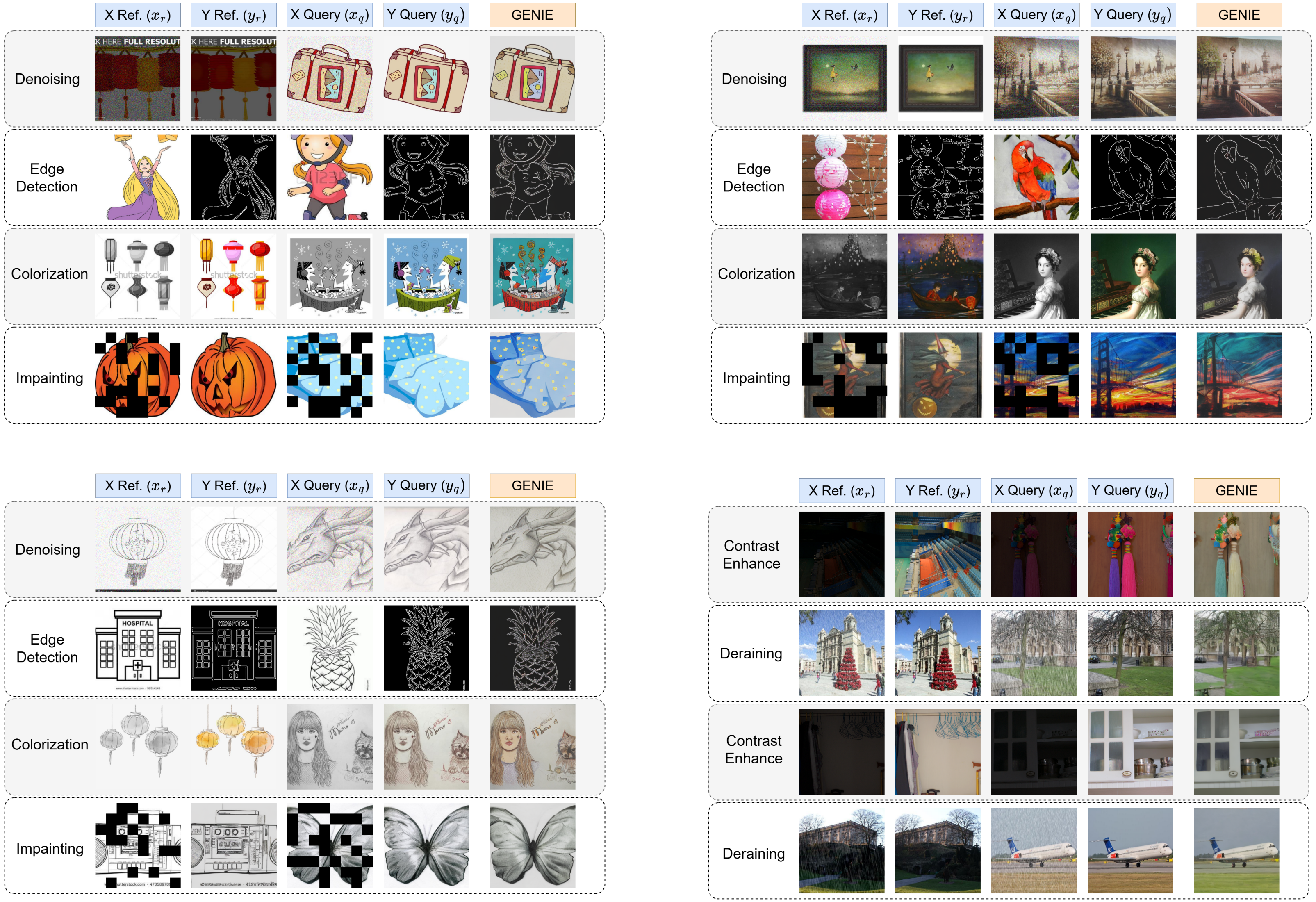
Models	Depth Estimation	Denoising		Super Resolution	Deblurring	Colorization	Impainting		Semantic Segmentation	Contrast Enhancement		Deraining	
	RMSE ↓	PSNR ↑	SSIM ↑	PSNR ↑	PSNR ↑	MSE ↓	SSIM ↑	MSE ↓	mIoU ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Visual Prompt [4]	0.348†	16.20†	0.3490†	16.45†	16.53†	0.037†	0.245*	0.105*	- *	13.915†	0.410†	14.91†	0.303†
Painter [30]	0.288*	27.21*	0.839*	- †	- †	- †	- †	- †	49.9*	22.40*	0.872*	29.49*	0.868*
GENIE	0.136*	26.61*	0.831*	19.94*	21.33*	0.017*	0.598*	0.018*	- *	20.623†	0.771†	25.41†	0.826†

- "-" indicates that the model outputs unrelated or ambiguous results, making quantitative evaluation unsuitable.
- The model can effectively segment multiple objects within a scene its reliance on latent space reconstruction without explicit class-specific constraints results in inaccurate class label assignments

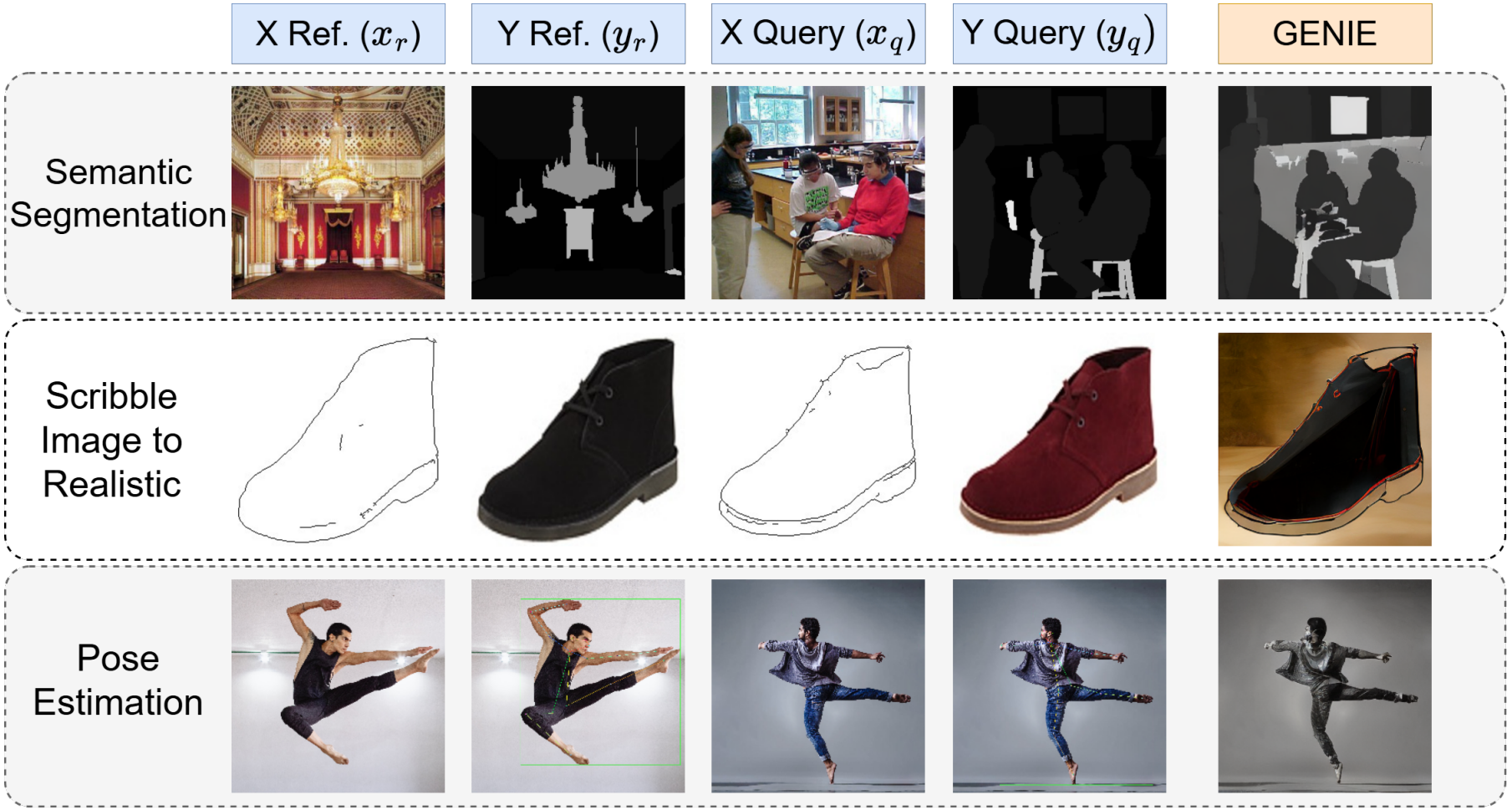
Qualitative Result



Qualitative Result (OOD Task)



Qualitative Result (Failure Cases)



Future Work

- Adapt the model to map effectively to specific classes in semantic segmentation
- While GENIE demonstrates strong multi-task capabilities and adaptability to OOD tasks, its performance on extremely deviated OOD task is limited, often yielding in suboptimal result.

Thank You