Redesign Your Space with Image Inpainting Model

Group 27
王維勤、劉高甸、黃睿廷、侯貝霖、陳思如

OUTLINE

INTRODUCTION

Motivation

Pipeline Overview

MOTIVATION

"Have you ever imagined transforming your old living room furniture into your dream design in just one second?"

MOTIVATION

What’s the biggest challenge in the living room

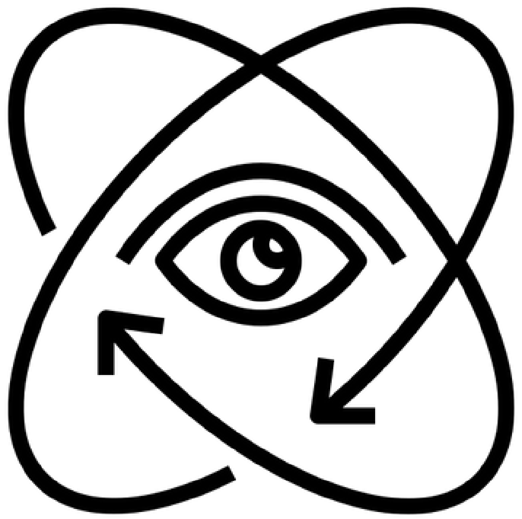
- Inspiration doesn’t translate seamlessly
- Manual edits consume hours
- Limited design flexibility

PIPELINE OVERVIEW

STEP-BY-STEP WORKFLOW

Step 1.

User specifies what to remove and replace

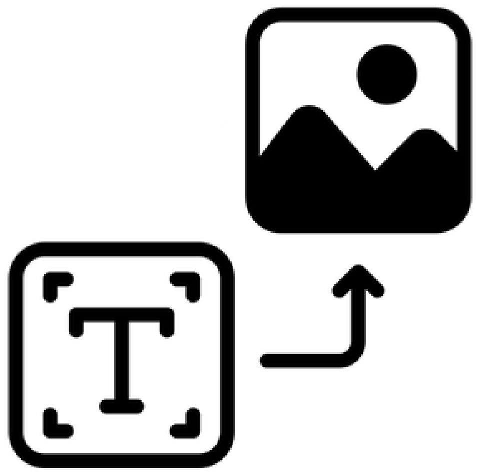


Step 2.

SegVG isolates the target area

Step 3.

PowerPaint generates the replacement



Step 4.

Outputs a redesigned image

RELATED WORK

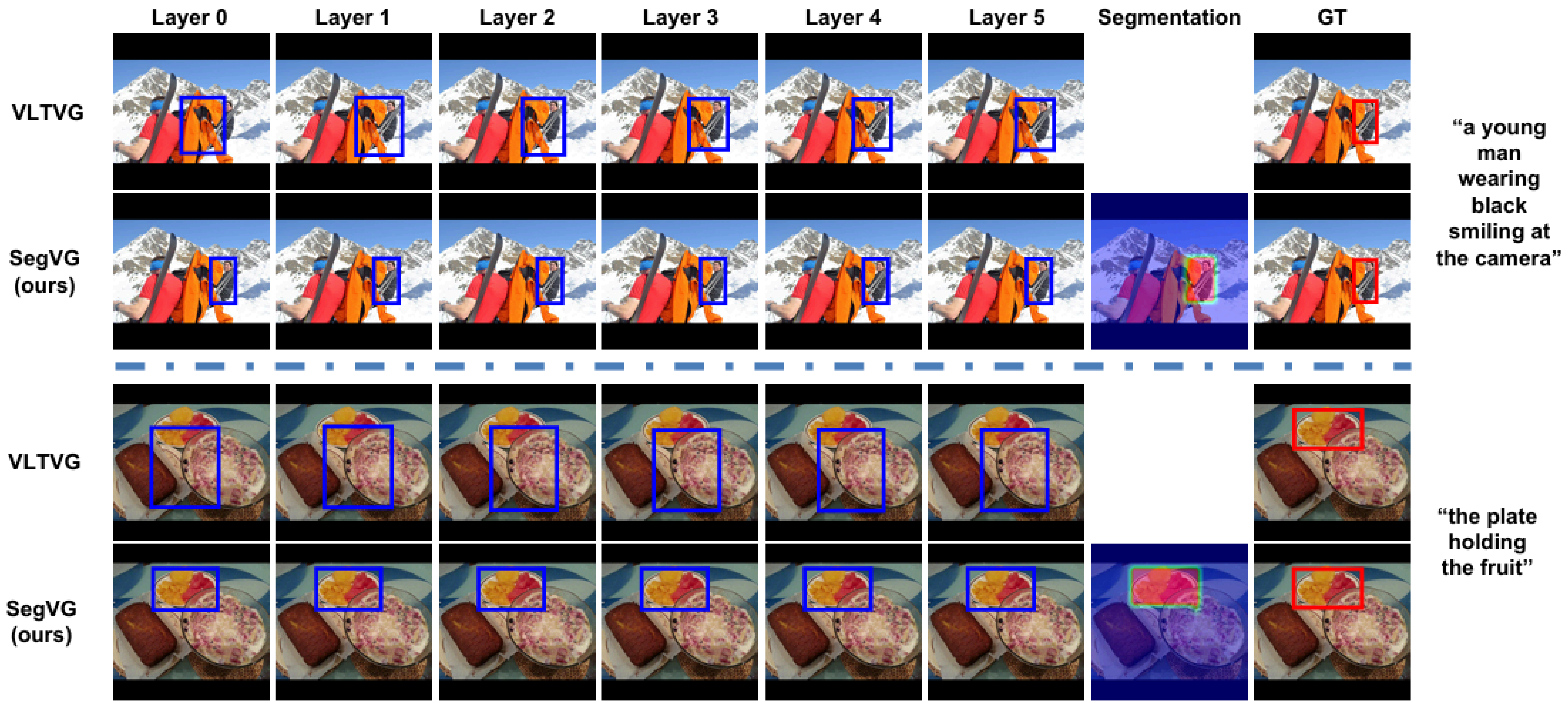
SegVG

BrushNet

PowerPaint

SegVG

Previous works' passive use of annotations, such as relying **solely on box annotations as regression ground truth**, results in suboptimal performance.

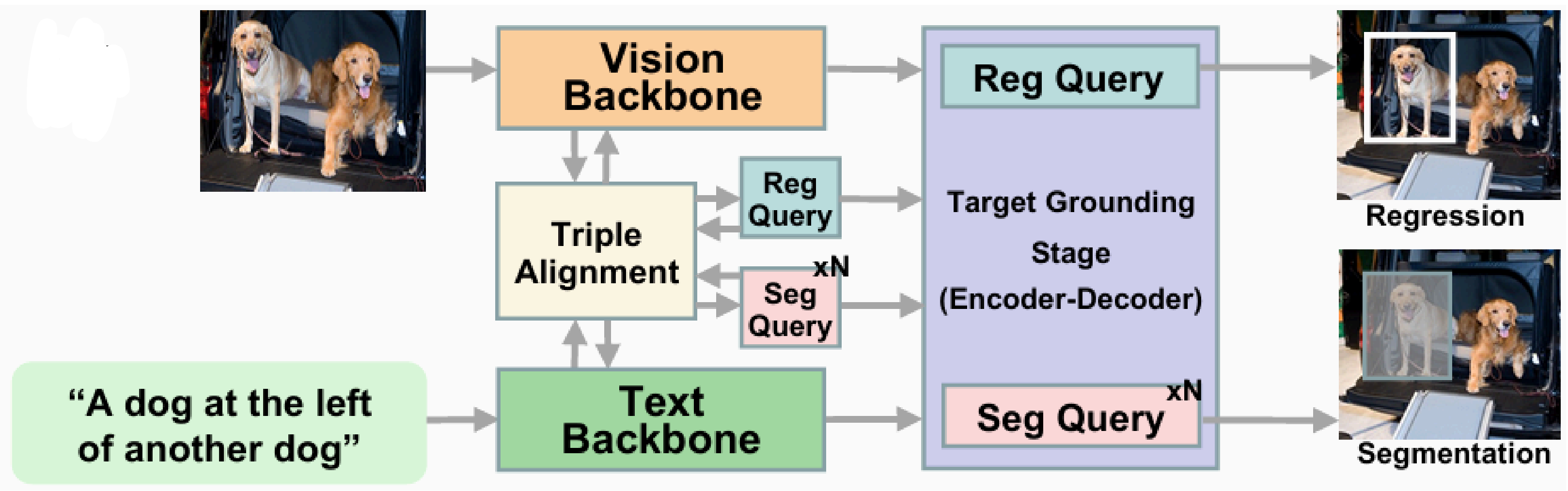


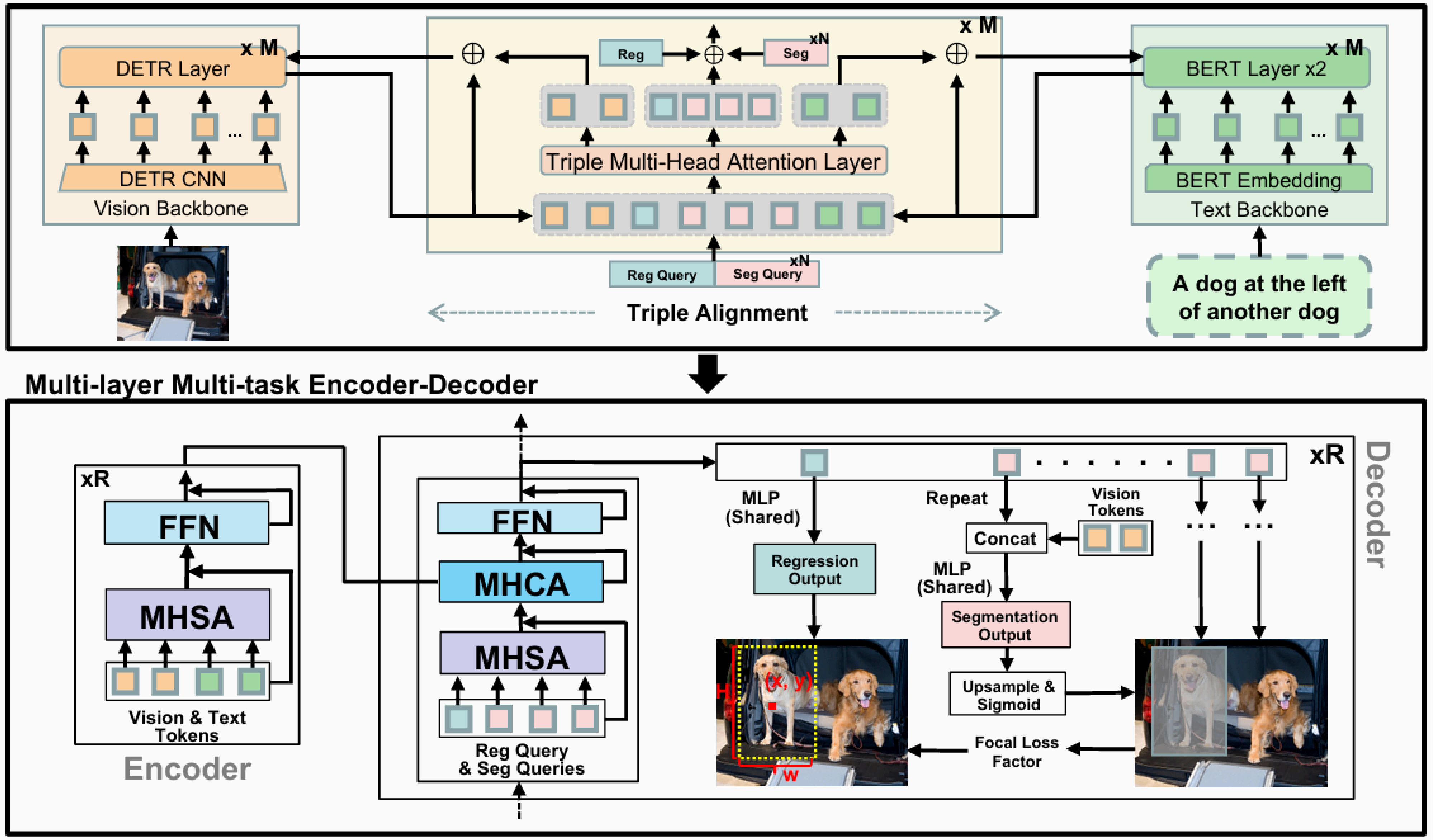
Kang, Weitai, et al. "SegVG: Transferring object bounding box to segmentation for visual grounding." (2025). In European Conference on Computer Vision (pp. 57-75).

SegVG

SegVG **transfers the box-level annotation as Segmentation signals** to provide an additional pixel-level supervision for Visual Grounding.

- Vision Backbone: DETR + Resnet
- Text Backbone: Bert
- Triple Alignment
- Multi-layer Multi-task Encoder-Decoder

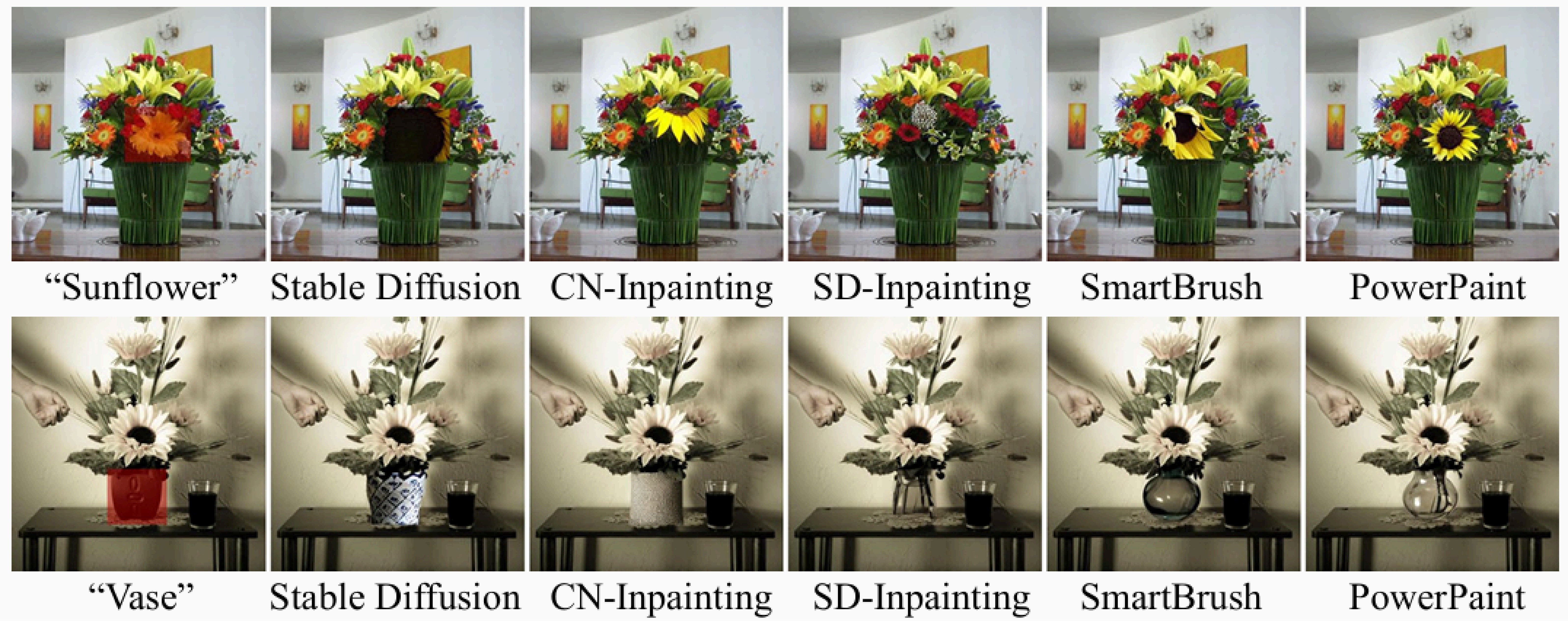




Kang, Weitai, et al. "SegVG: Transferring object bounding box to segmentation for visual grounding." (2025). In European Conference on Computer Vision (pp. 57-75).

PowerPaint

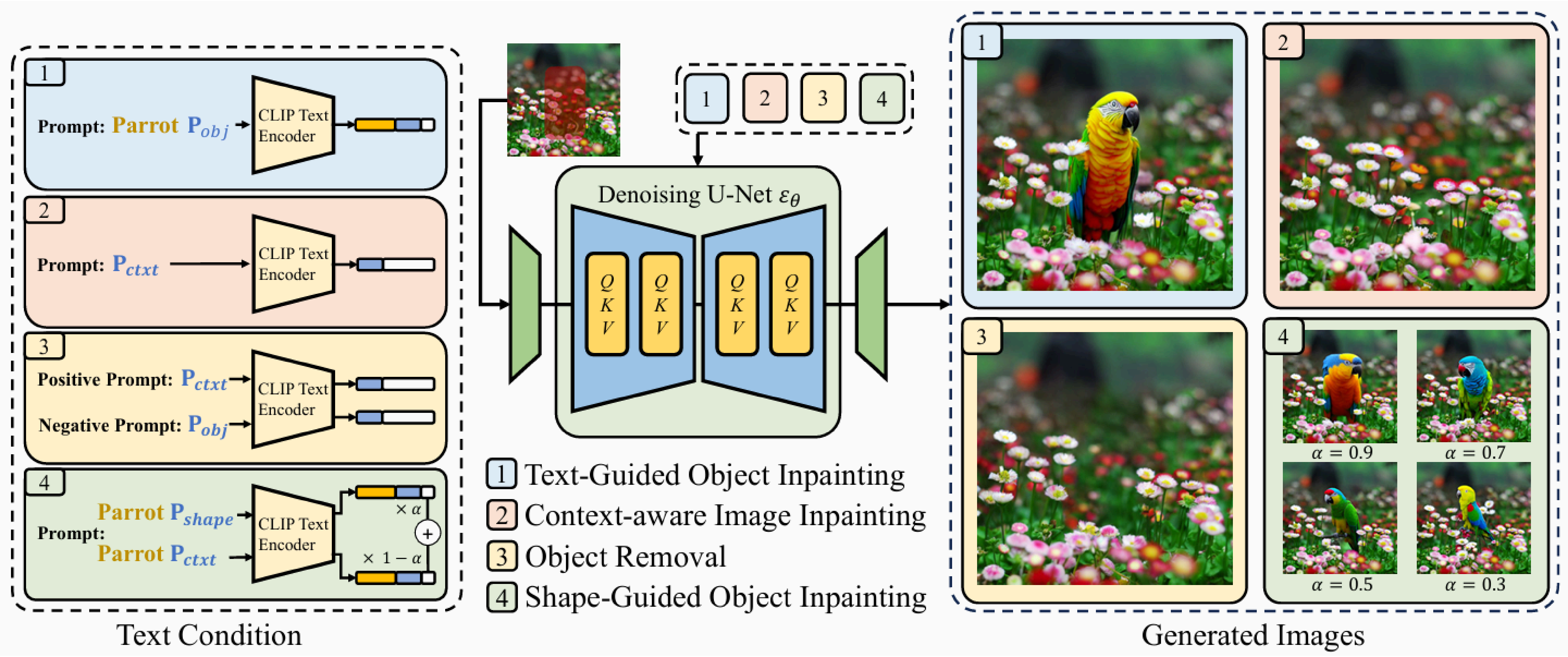
Fine-tuned text-to-image models with masks and text often generate **random artifacts incoherent with the image context**.



Zhuang, Junhao, et al. "A task is worth one word: Learning with task prompts for high-quality versatile image inpainting." (2025). In European Conference on Computer Vision (pp. 195-211).

PowerPaint

PowerPaint excels in both **text-guided object inpainting** and **context-aware image inpainting** by introducing **learnable task prompts** and **fine-tuning strategies** to focus on inpainting targets.



Zhuang, Junhao, et al. "A task is worth one word: Learning with task prompts for high-quality versatile image inpainting." (2025). In European Conference on Computer Vision (pp. 195-211).

METHODOLOGY

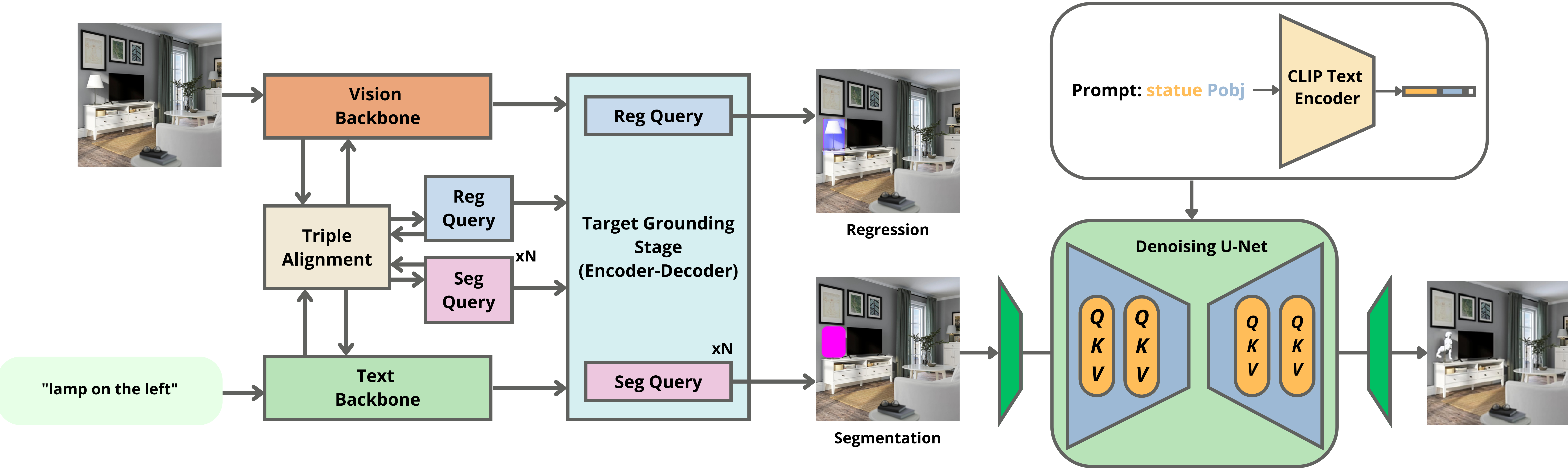
Datasets

Text Condition

Segmentation

Framework

- This is our model architecture, combining SegVG and PowerPaint.



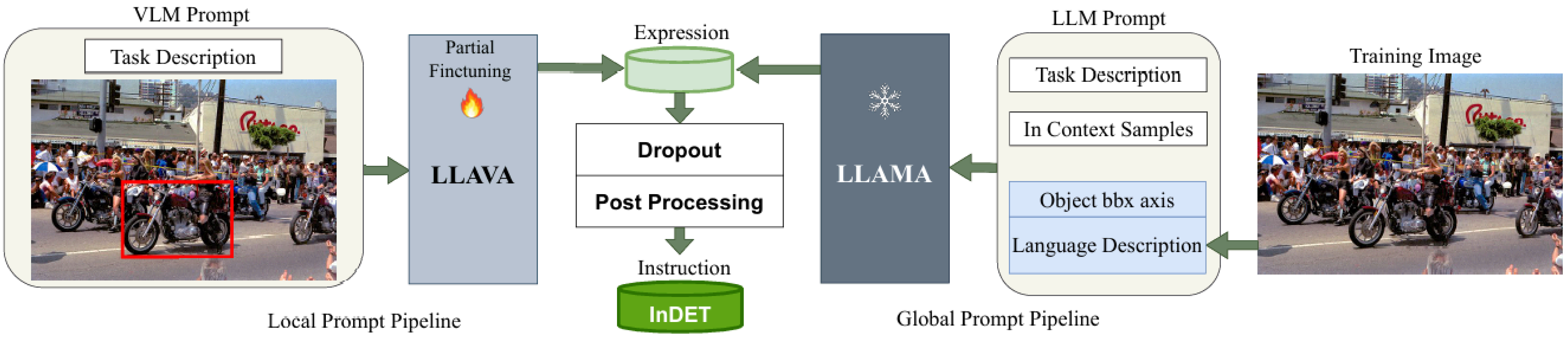
DATASETS

- Custom dataset from OpenImage, and generate expression using InstructDET
- Global Prompt

This process focuses on capturing diverse expressions with different properties of a single object.

- Local Prompt

This process generates expressions that are informative and closely related to the target object.



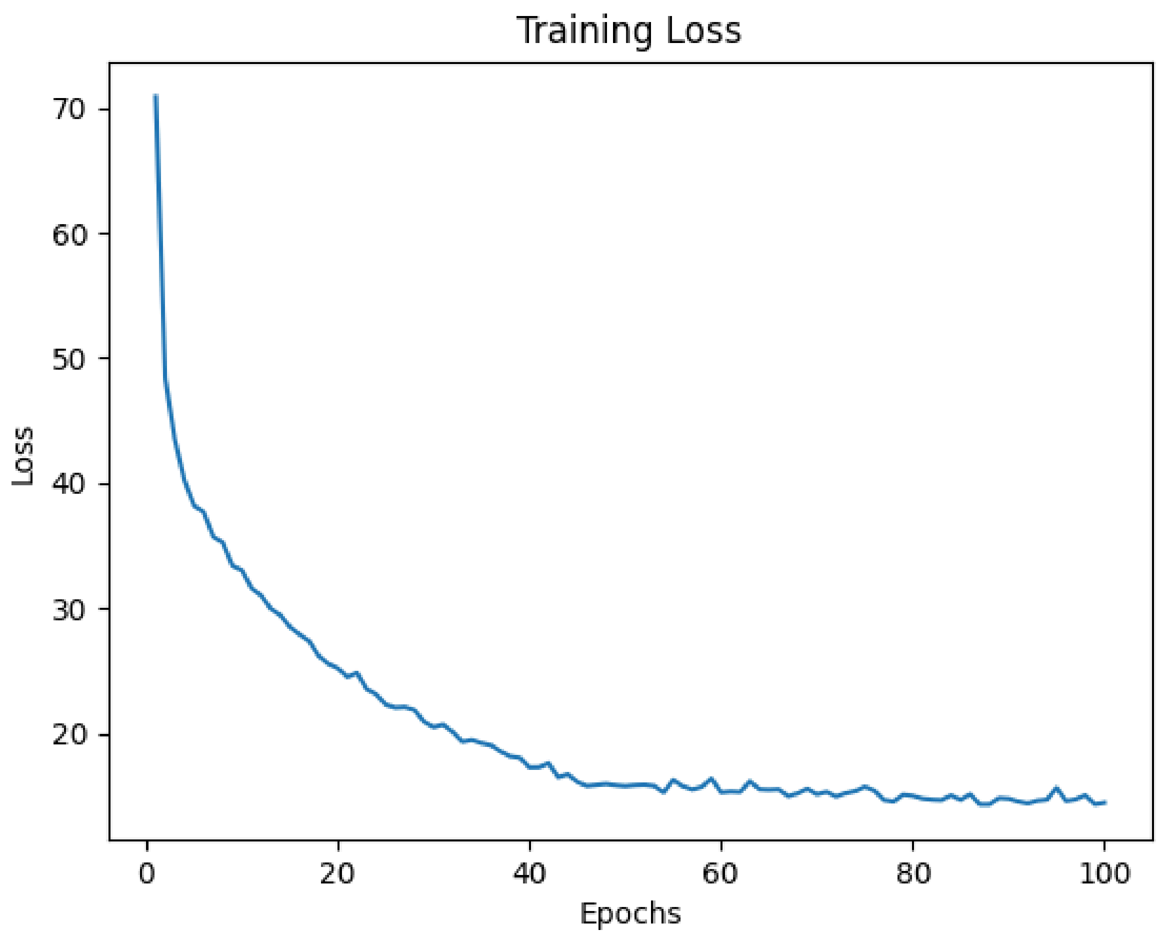
Dang, Ronghao, et al. "Instructdet: Diversifying referring object detection with generalized instructions." (2023).

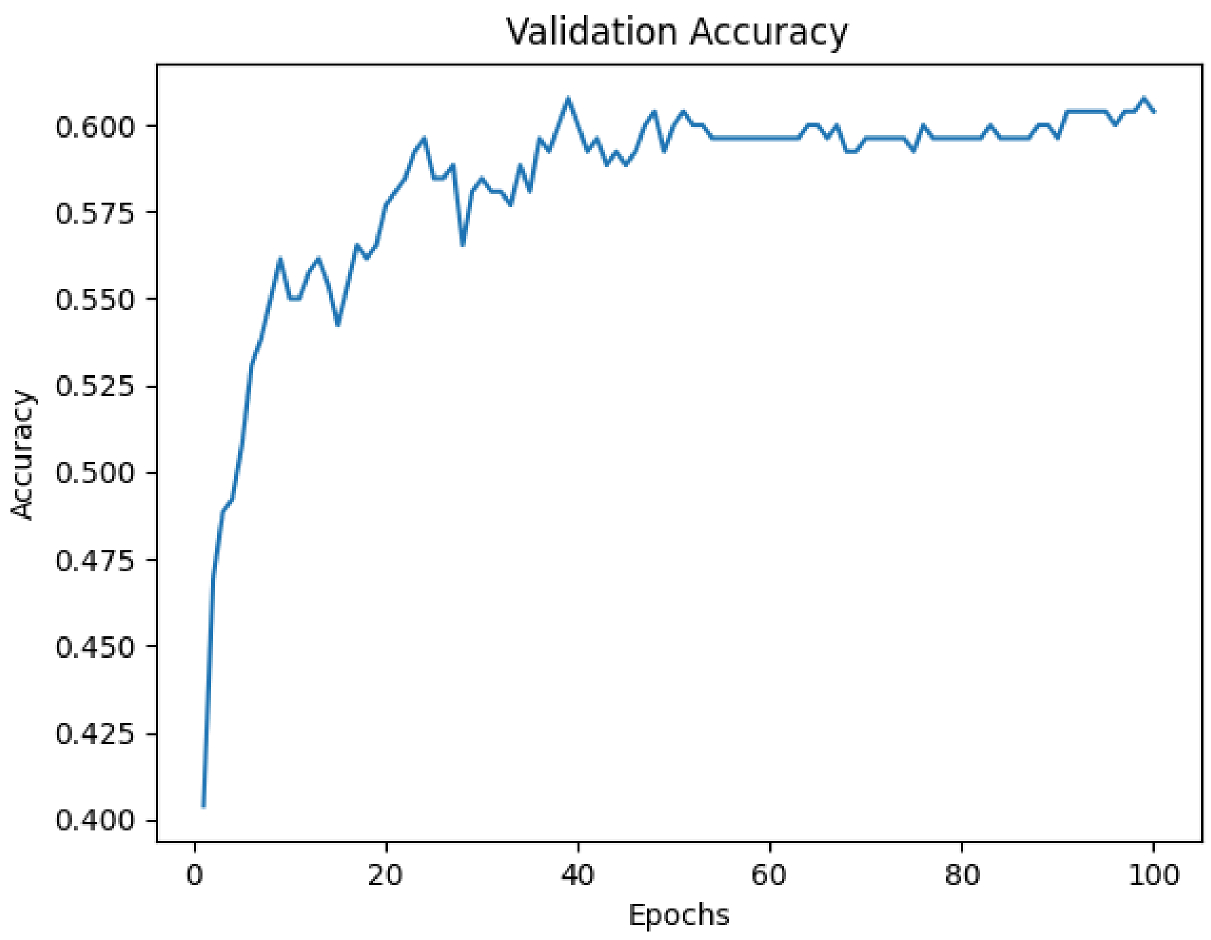
DATASETS

- Finetune indoor dataset on SegVG ReferIt Dataset Pretrain weight

Dataset:

- RefCOCO: 5,000 samples
- Our customized dataset from InstructDet OpenImage:1,300 samples





SEGMENTATION

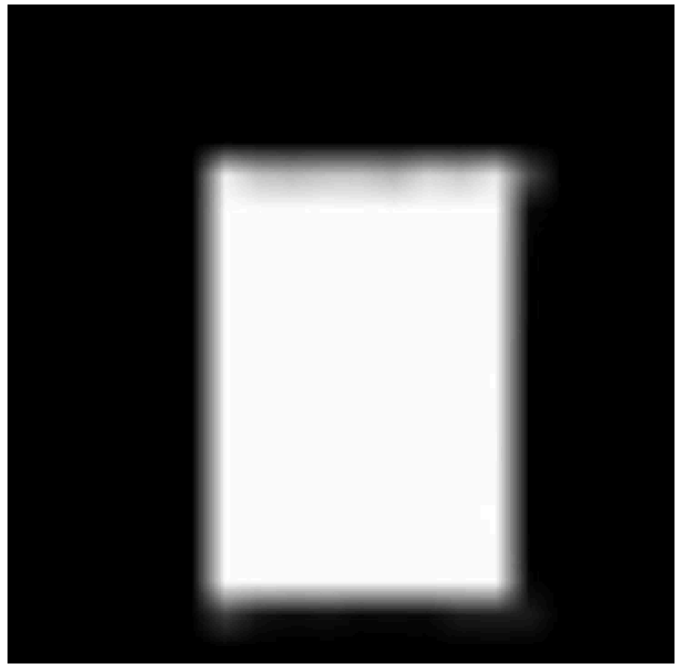
We have tried 2 kinds of inpainting masks

- Bounding box
- Precise object segmentation mask

>> Experiments have shown that **bounding boxes are more effective** for later task

Original

Bbox



mask

generation

Precise

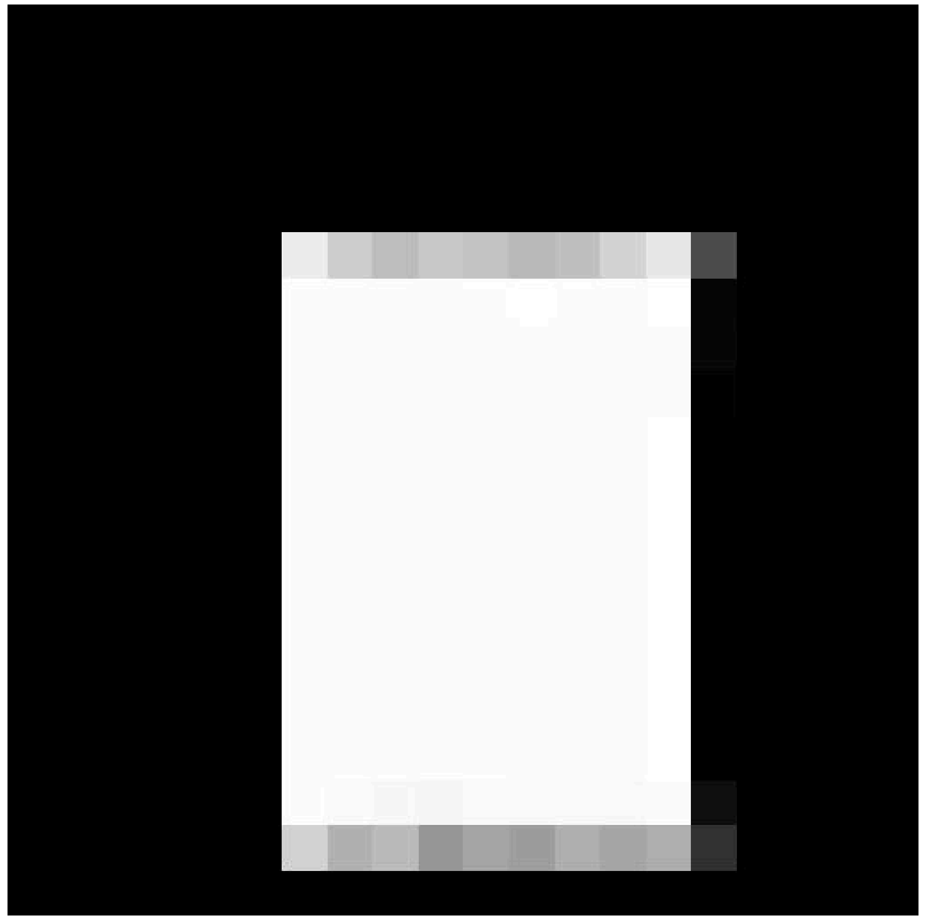
mask

generation

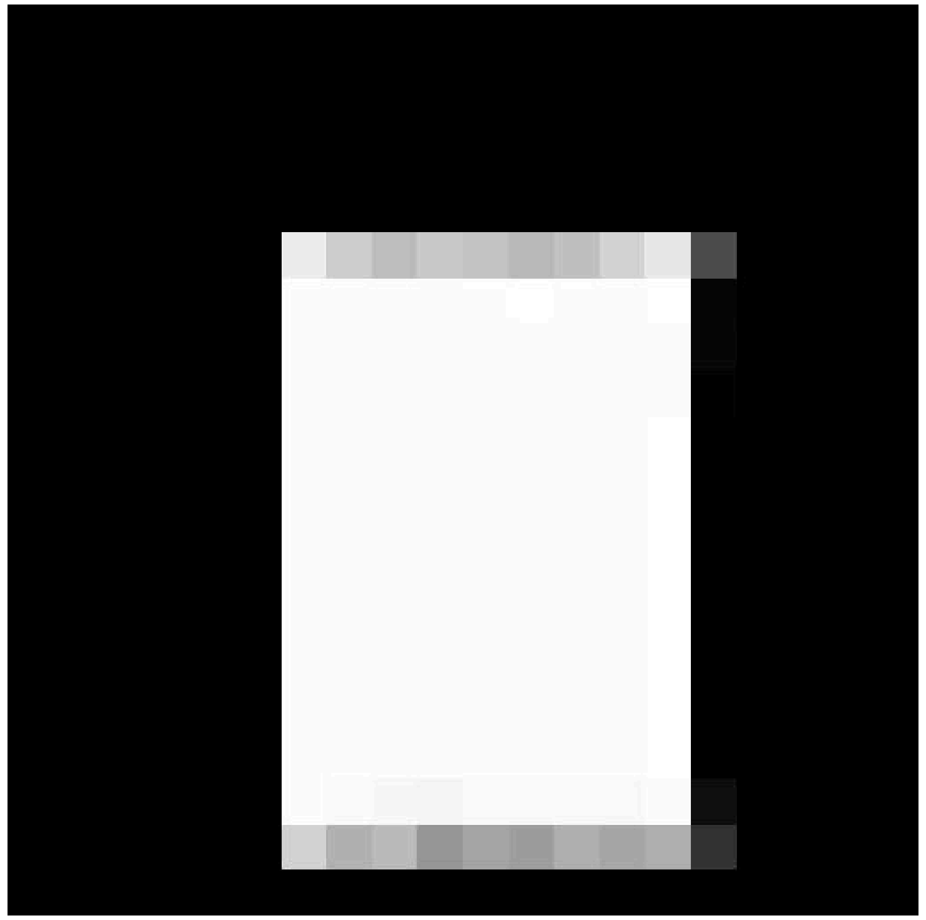
SEGMENTATION

- Different algorithms generate different mask results.

Nearest

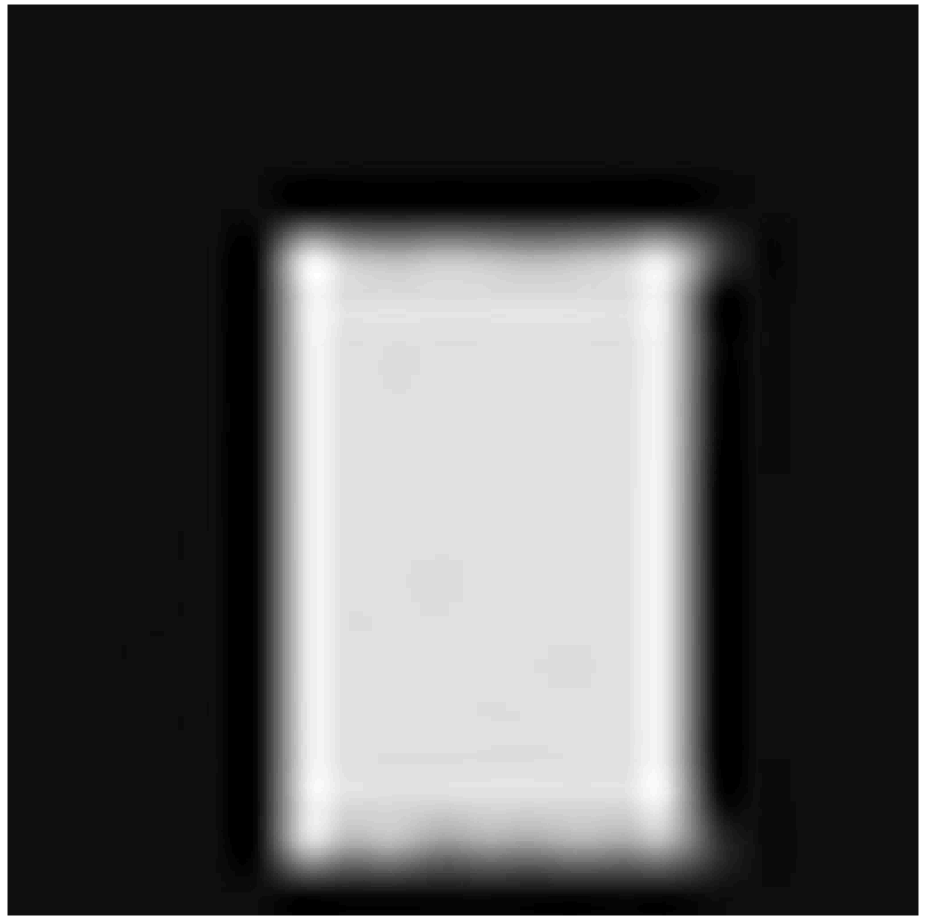


Nearest Exact



Bilinear

Bicubic



SEGMENTATION

If there are multiple identical objects in the image

- Users can specify objects and their **positions** for precise segmentation

lamp

lamp **on the left**

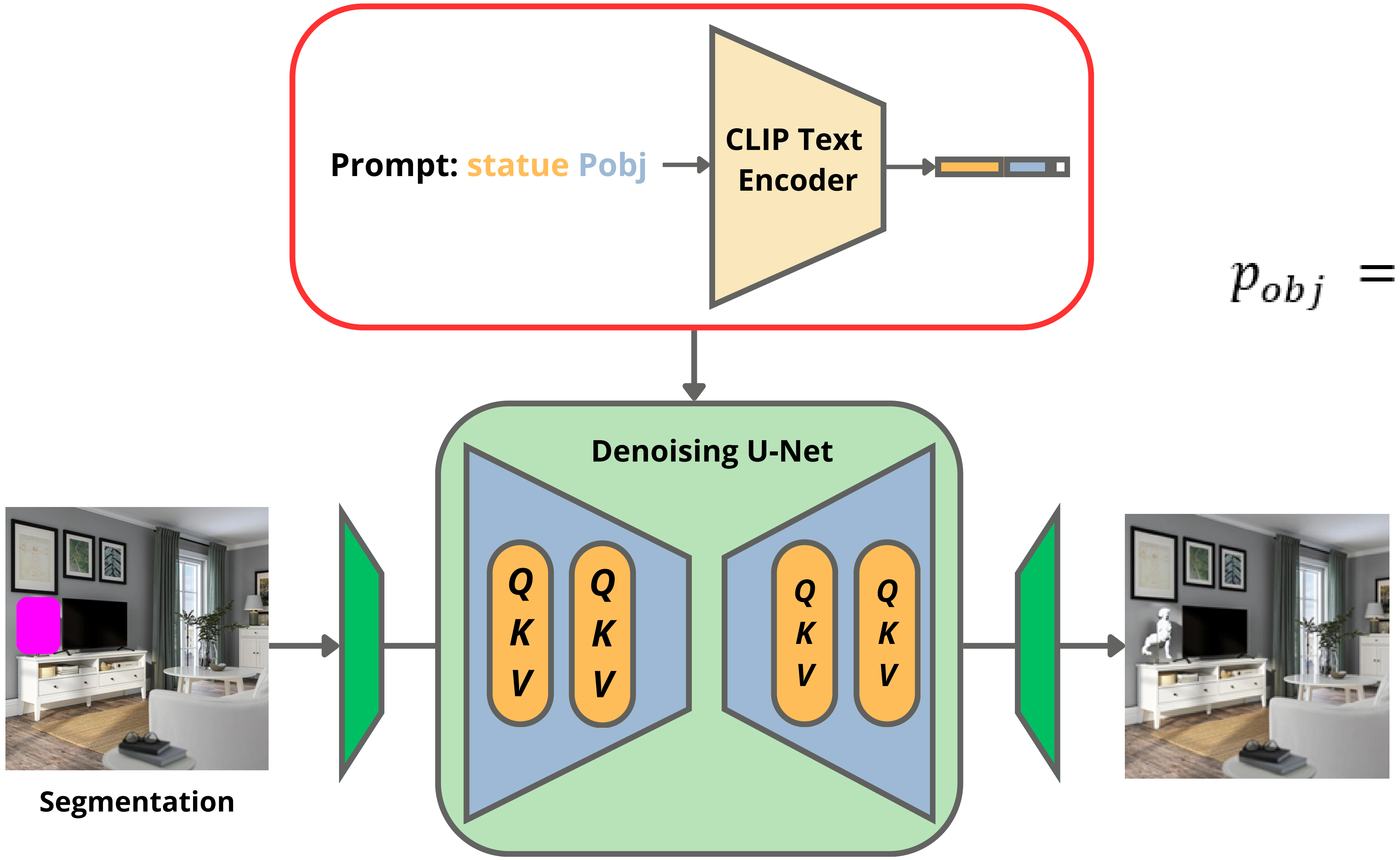
juice machine

juice machine **on the top right**

TEXT CONDITION

We refer to the text-guided object inpainting task in PowerPaint. They Introduce a learnable task prompt, denoted as **Pobj**.

- **Pobj** is appended as a suffix to the masked region's text description, guiding the model to inpaint based on text.



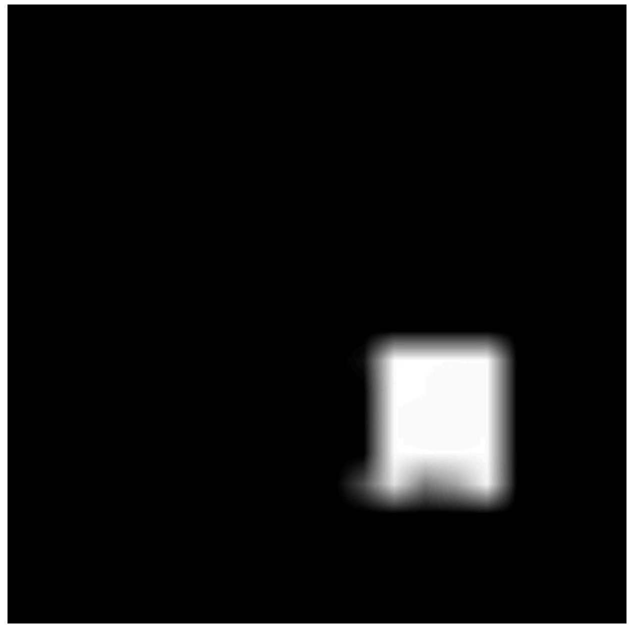
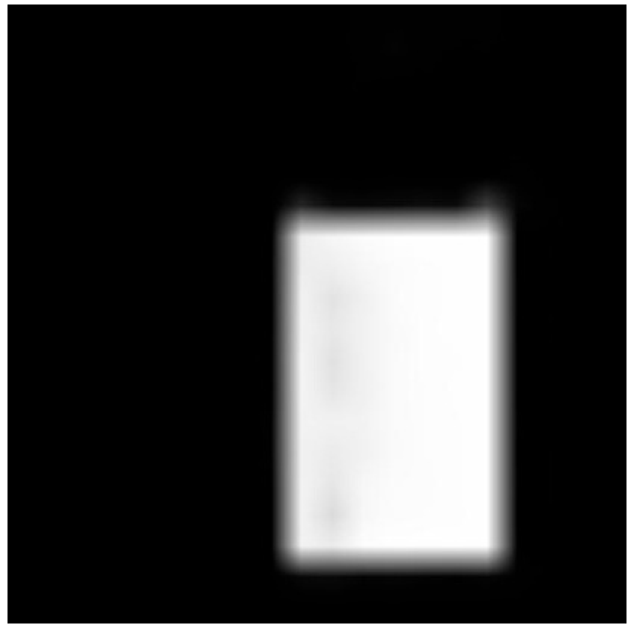
$$p_{obj} = \arg\min_{p} \mathbb{E}_{x_0,m,t,p,\epsilon_t} \|\epsilon_t - \epsilon_\theta(x'_t, \tau_\theta(p), t)\|_2^2,$$

RESULTS

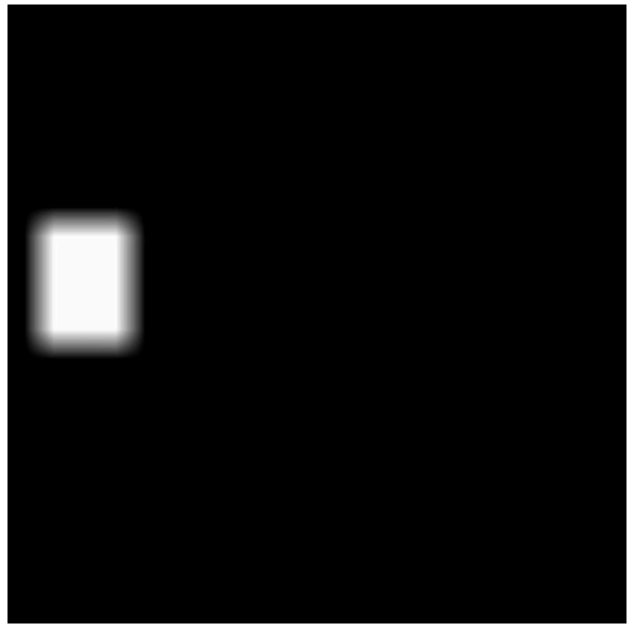
Visualization

Issues

VISUALIZATION

Input Image	*Remove Prompt*	*Replace Object Prompt*
	clock	telephone
	 	
	cabinet	shoe rack
	 	

VISUALIZATION

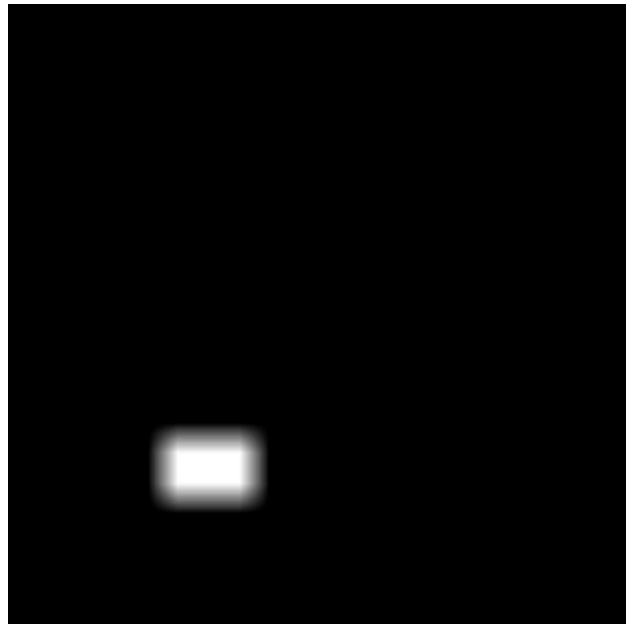
Input Image	*Remove Prompt*		*Replace Object Prompt*
	right painting		mirror
			
	lamp on the left		statue
			

VISUALIZATION

Input Image	*Remove Prompt*	*Replace Object Prompt*

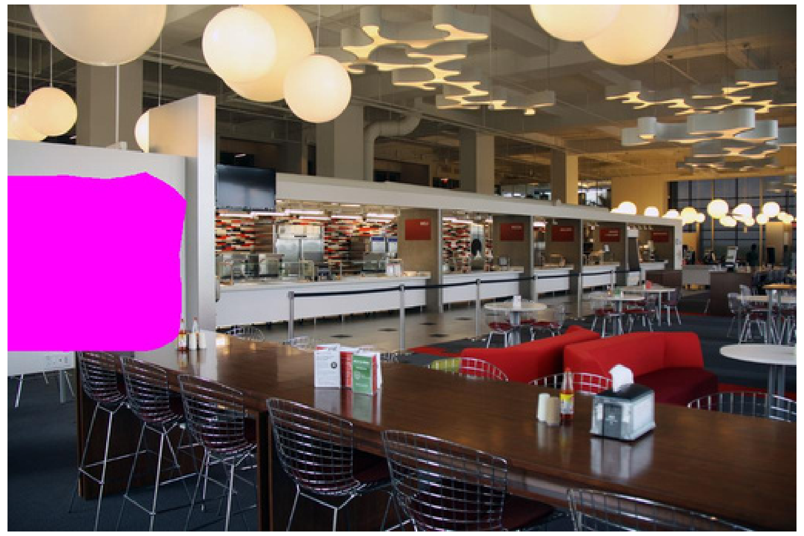
the yellow bowl

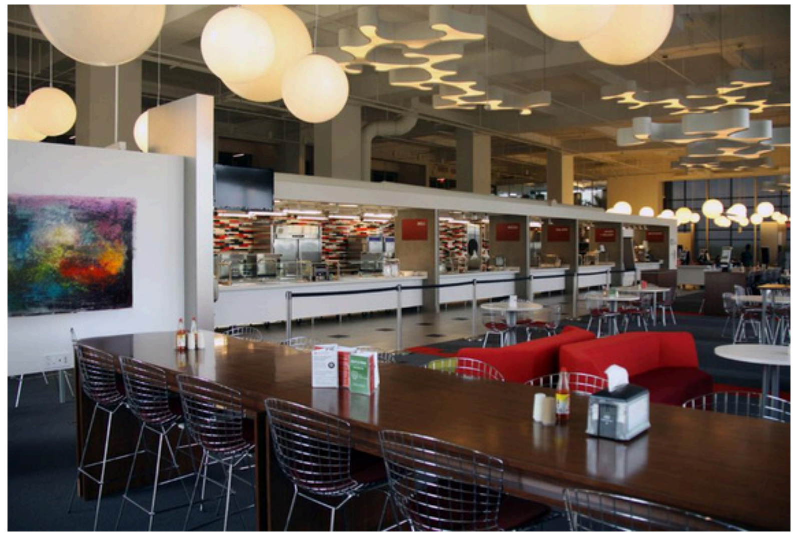
mug



TV

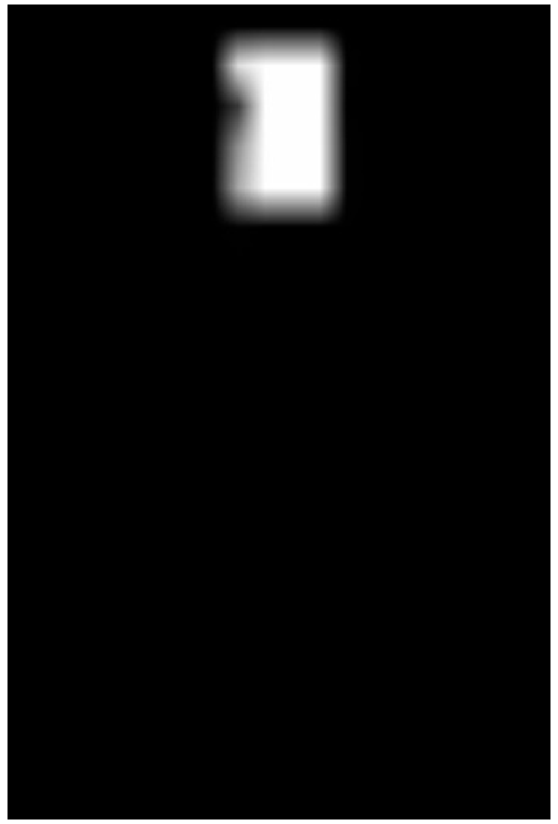
painting





ISSUE

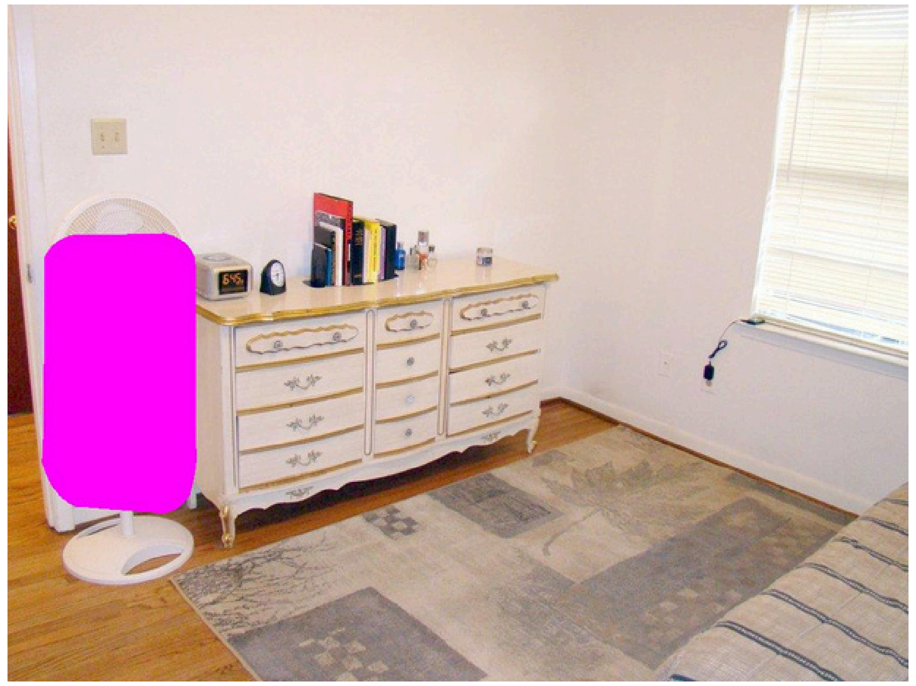
Mask did not cover the whole item

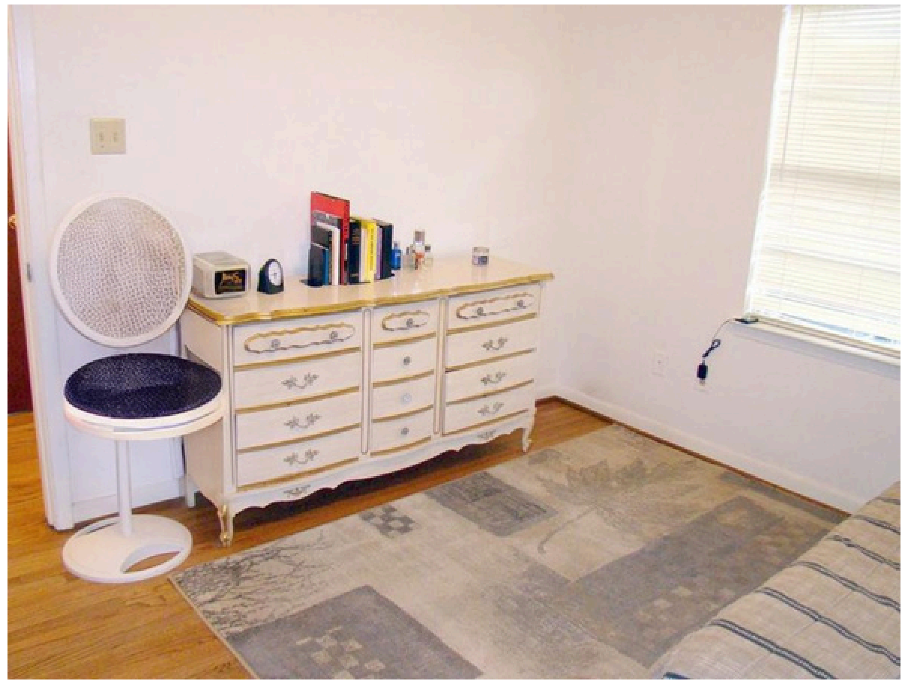
Input Image	*Remove Prompt*	*Replace Object Prompt*
	lamp	clock
	 	

ISSUE

Mask did not cover the whole item

Input Image	*Remove Prompt*	*Replace Object Prompt*
	fan on the left	chair





CONCLUSION

Contribution

Future Work

CONTRIBUTION

- *Integration of SOTA Models*

Combine SOTA models for seamless object replacement in interior design.

- *Fine-Tuning SegVG*

Adapt SegVG on indoor datasets for better interior-specific segmentation.

- *Semantic User Input*

Develop a pipeline to interpret semantic cues for precise object manipulation.

- *Text-to-Image Object Replacement*

Enable text-based object replacement for effortless modifications.

FUTURE WORK

Short Term

1. Utilize more metrics for qualitative results. **(ex. CLIP similarity)**
2. Compare with other proposed models.
3. Pack our pipeline into UI interface.

Long Term

1. Develop comprehensive Indoor training dataset.
2. Finetune diffusion models with LoRA.
3. Perform fine-tuning on the integrated pipeline of two models.

REFERENCES

Dang, R., Feng, J., Zhang, H., Ge, C., Song, L., Gong, L., Liu, C., Chen, Q., Zhu, F., Zhao, R., & Song, Y. "InstructDET: Diversifying Referring Object Detection with Generalized Instructions." arXiv preprint arXiv:2310.05136.

Kang, W., Liu, G., Shah, M., & Yan, Y. "SegVG: Transferring Object Bounding Box to Segmentation for Visual Grounding." European Conference on Computer Vision. Springer, Cham, 2025.

Zhuang, J., Zeng, Y., Liu, W., Yuan, C., & Chen, K. "A task is worth one word: Learning with task prompts for high-quality versatile image inpainting." European Conference on Computer Vision. Springer, Cham, 2025.

Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., & Xu, Q. "BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion." arXiv preprint arXiv:2403.06976 (2024).

THANK YOU