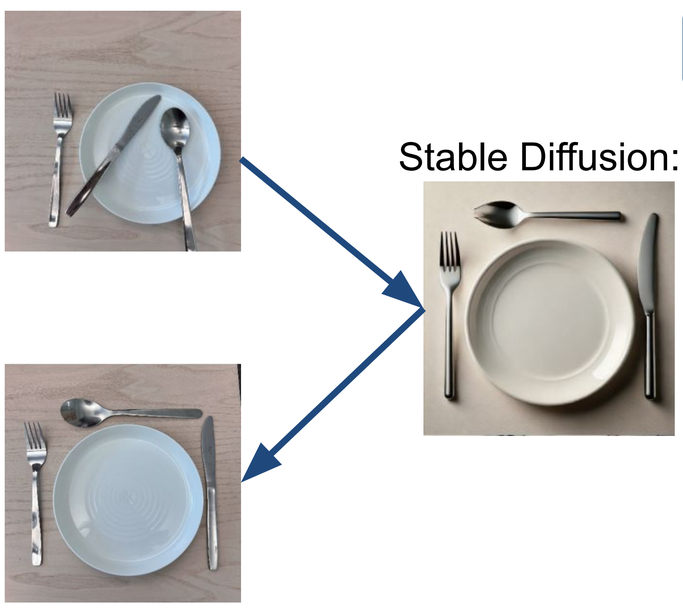
PACA: Perspective-Aware Cross-Attention Representation for Zero-shot Scene Rearrangement

Shutong Jin[*1], Ruiyu Wang[*1] , Kuangyi Chen[2], Florian T. Pokorny[1]
[1]KTH Royal Institute of Technology, [2]Graz University of Technology, [*]Equal Contribution

Motivation

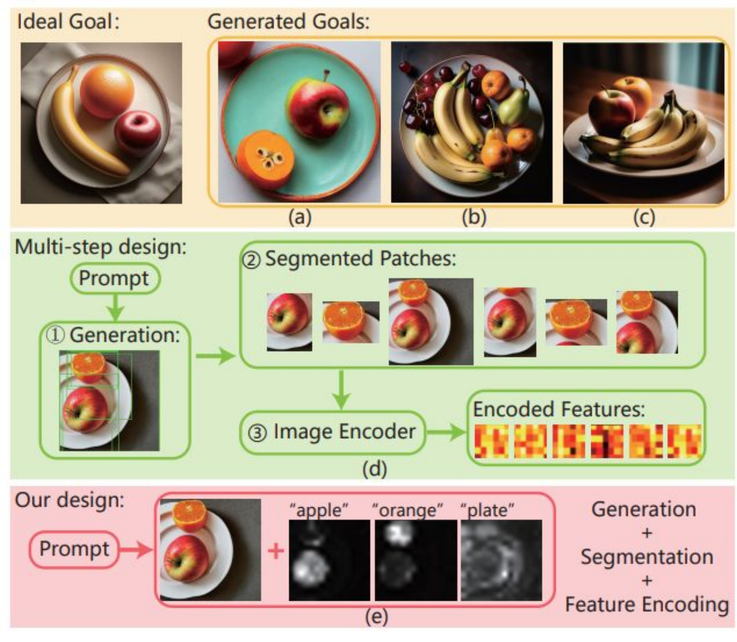
Generative models can **generate scene goals for robotic manipulation**, but current approaches face two key **challenges**:

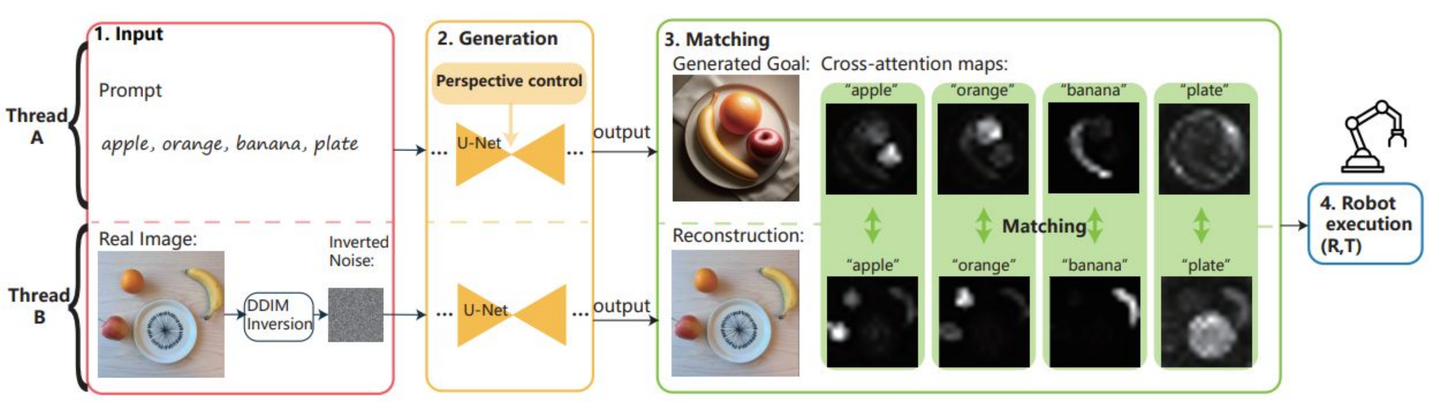
1. **Multi-step error accumulation** due to separate generation, segmentation, and encoding steps.
2. Confined to **3-DoF top-down** operation.



Our Contributions:

- We introduce PACA, a **training-free** pipeline for **scene rearrangement** that utilizes web-scale trained Stable Diffusion.
- Leveraging the lossy denoising process of diffusion models, we develop an object-level representation that integrates **generation, segmentation, and feature encoding into a single step.**
- We expand the image-goal-based method **from 3-DoF to 6-DoF**.





Training-free Pipeline:

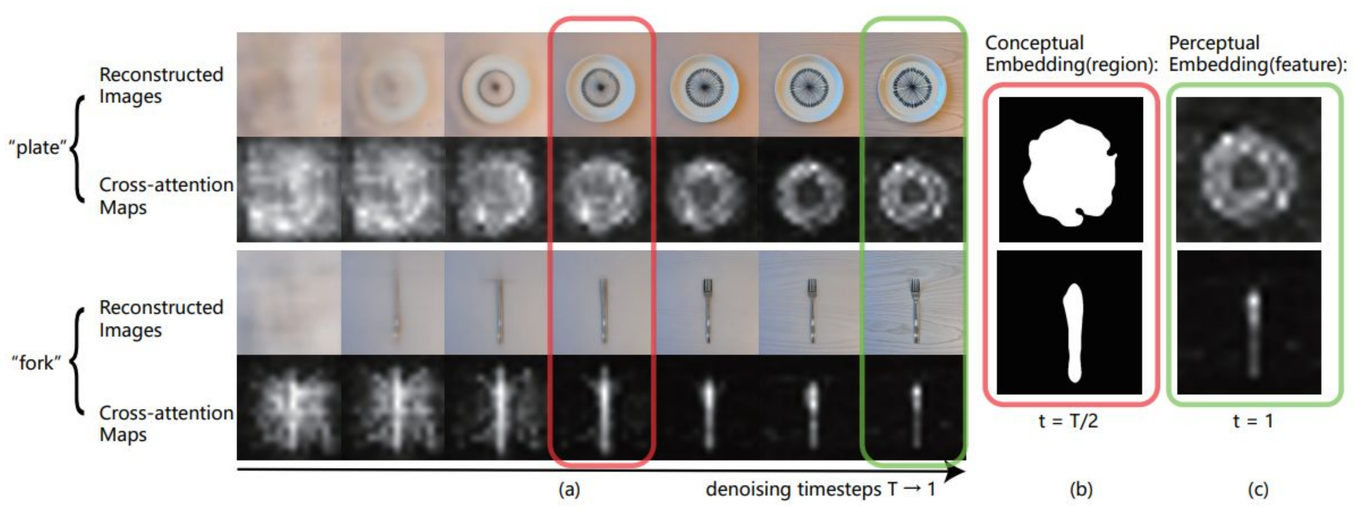
- **Input** acquisition.
- Goal **generation** with perspective control.
- Representation **matching**.
- Robot **execution**.

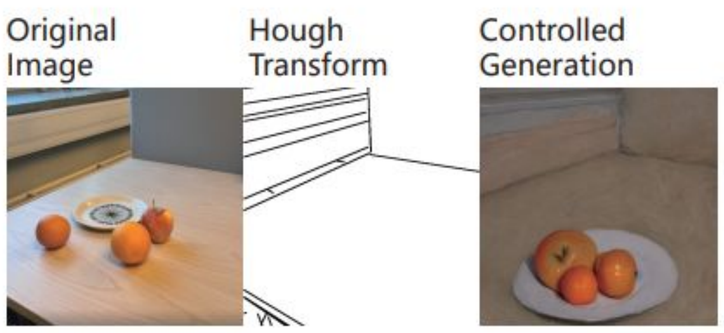
Observations:

$$\hat{z}_0 \approx z_0 = \left(z_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon_\theta(z_t)\right)/\sqrt{\bar{\alpha}_t},$$

Viewing the denoising process from an **information theory perspective**:

At any time t, partial information z_t becomes available and can be used to estimate the final reconstruction or generation z_0. **Perceptual details gradually added to the generated images.**





3-DoF to 6-DoF:

$$C_h = \text{HOUGHTRANSFORM}(x^{real}),$$

$$x^{goal} = \text{CONTROLNET}(C_h, P, s, \beta_{cfg}),$$

Demo:

Homepage:

Accepted by Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025