Slot-Guided Adaptation of Pre-trained Diffusion Models for Object-Centric Learning and Compositional Generation

Adil Kaan Akan[1], Yucel Yemez[1,2]

[1]Koc University
[2]KUIS AI

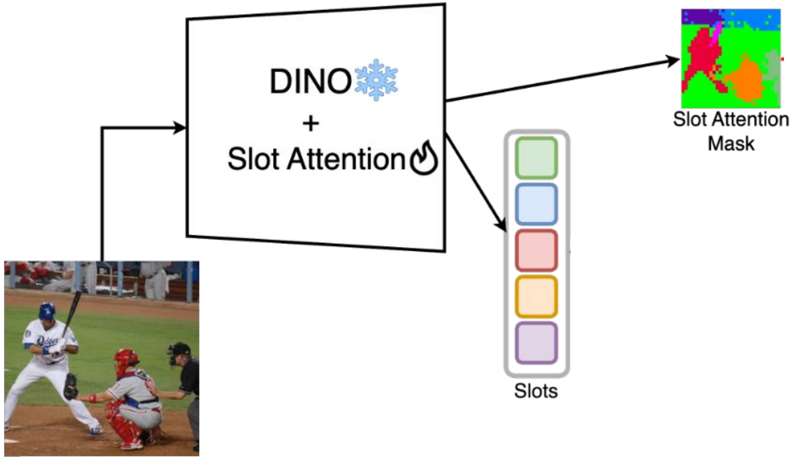
Problem & Motivation

- Real-World Complexity
  - Real-world images are challenging.
- Need for Object-Centric Learning (OCL) Methods
  - Without a structured breakdown of objects, existing methods cannot accomplish compositional generation/editing on real-world samples.
- Why It Matters
  - OCL enables better segmentation, more faithful generation, and easier editing of images

Contributions

- SlotAdapt Architecture
    - Approach that combines Slot Attention with pretrained diffusion models to boost both segmentation accuracy and generation fidelity.
- Key innovations:
    - Adapter Layers to align slot representations with the text-based diffusion model, reducing “text bias.”
    - Register Token for capturing global context/background, freeing slots to focus on distinct objects.
    - Attention Guidance to align slot masks and diffusion attention, improving object masks without external supervision.

Slot Adapt



DINO ❄

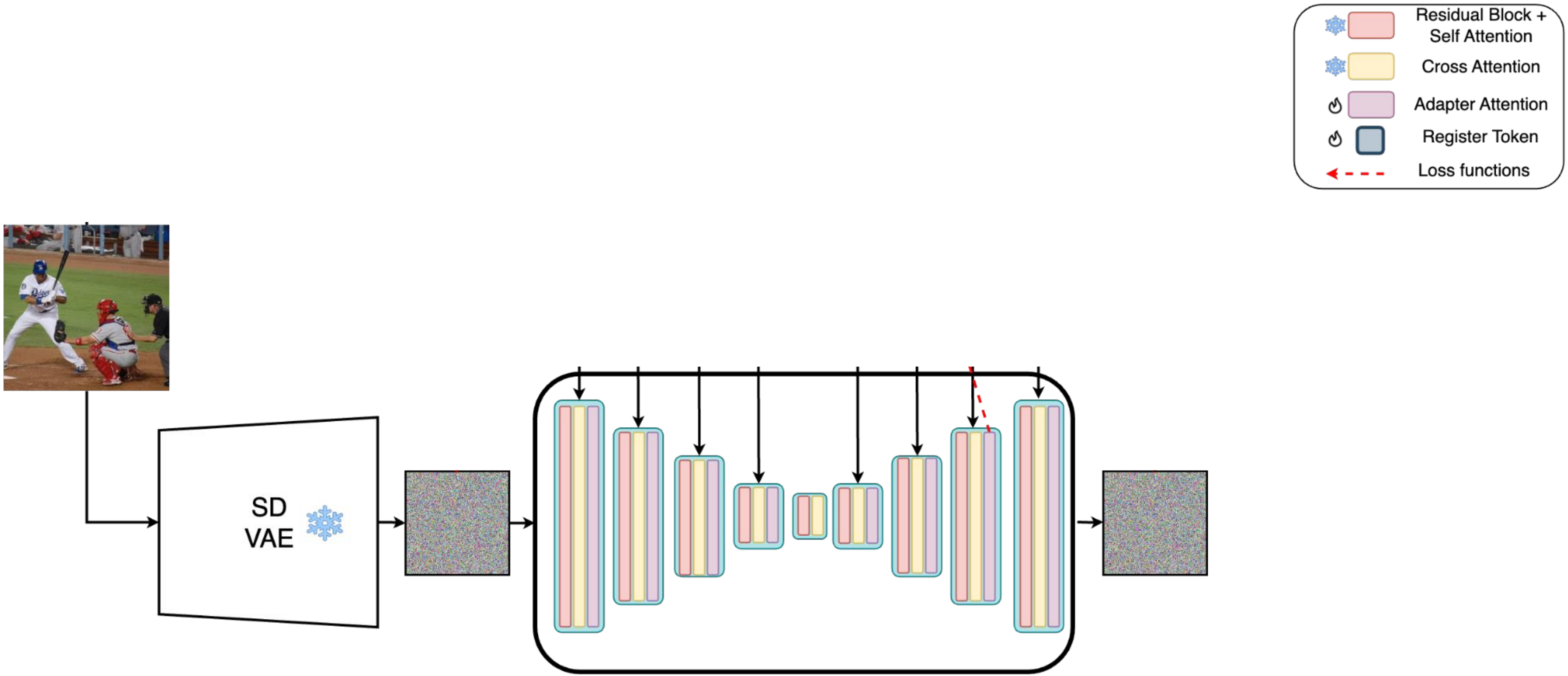
+

Slot Attention 🔥

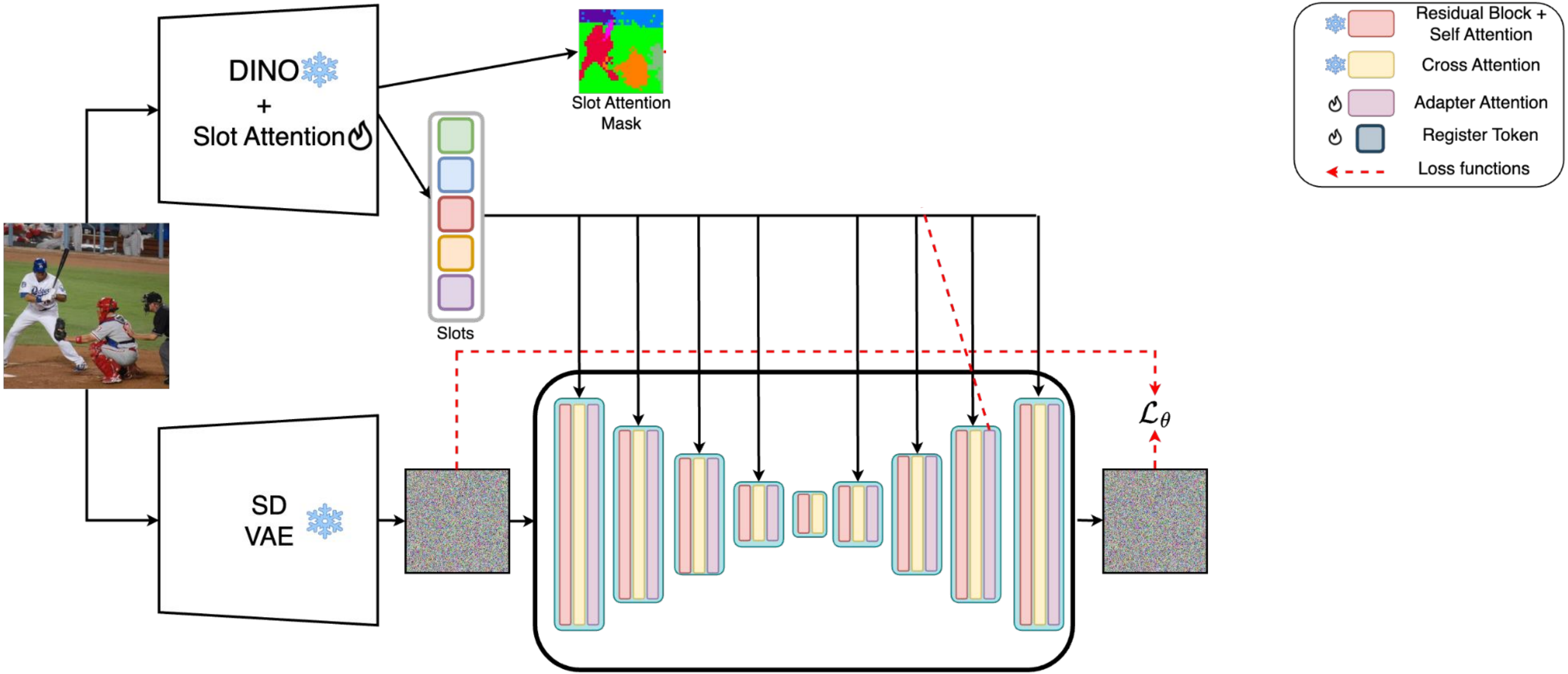
Slot Attention Mask

Slots

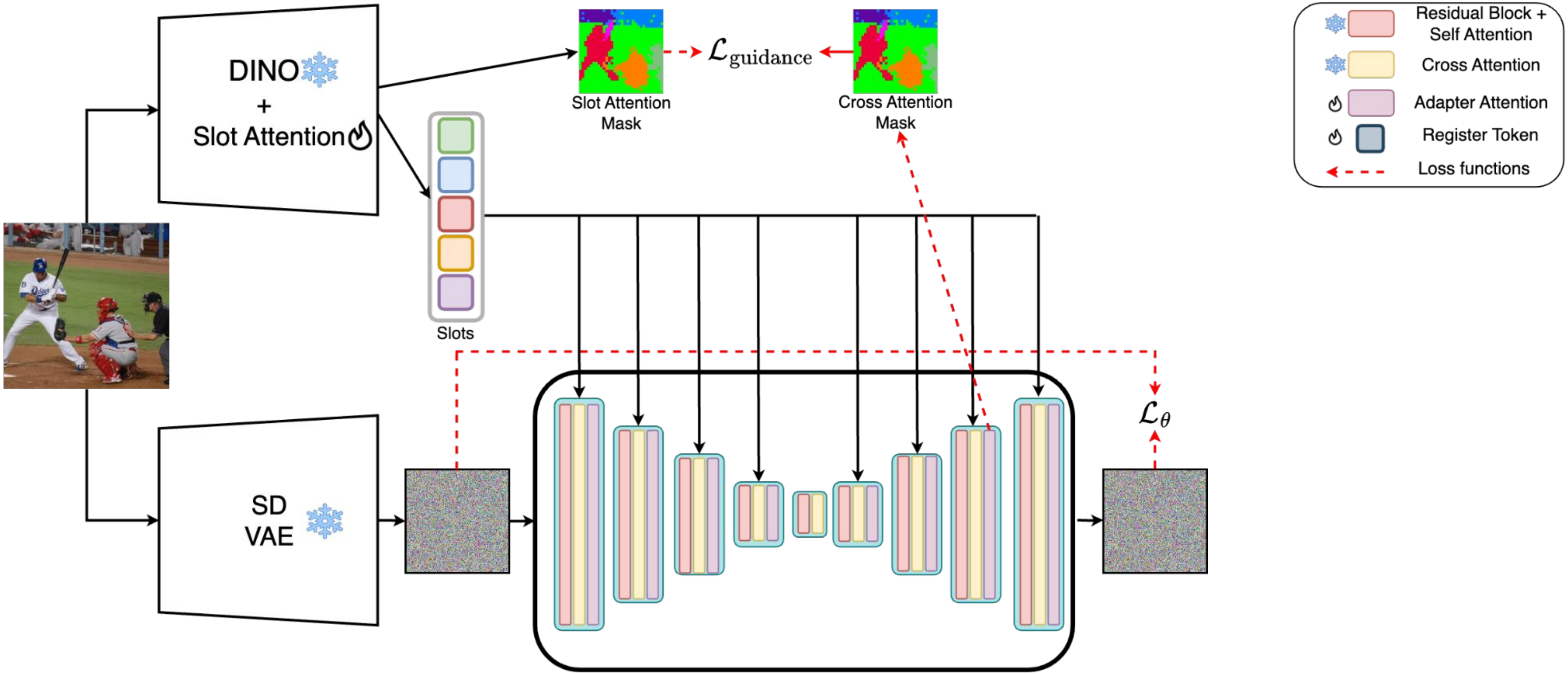
Slot Adapt


Residual Block + Self Attention
Cross Attention
Adapter Attention
Register Token
Loss functions
SD VAE

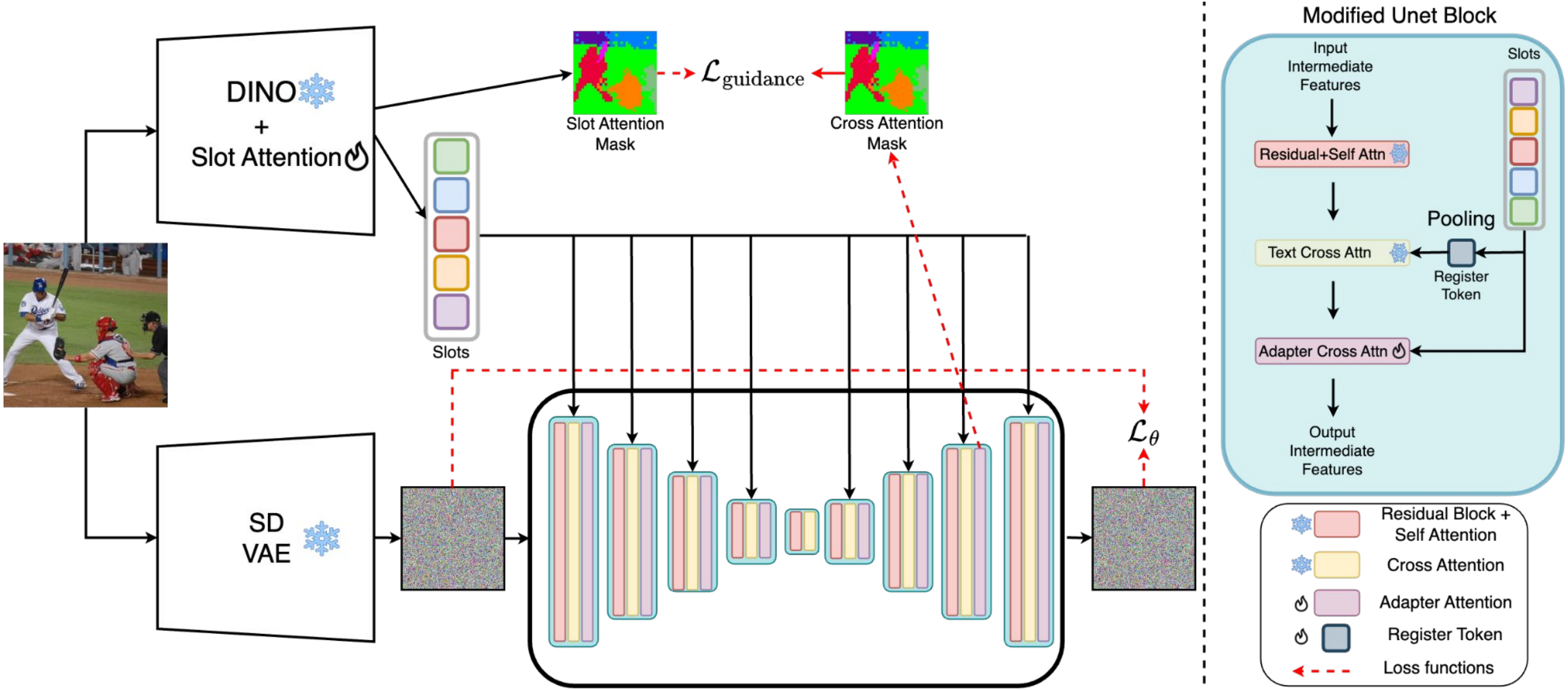
Slot Adapt


DINO
+
Slot Attention
Slot Attention Mask
Slots
SD VAE
$\mathcal{L}_\theta$
Residual Block + Self Attention
Cross Attention
Adapter Attention
Register Token
Loss functions

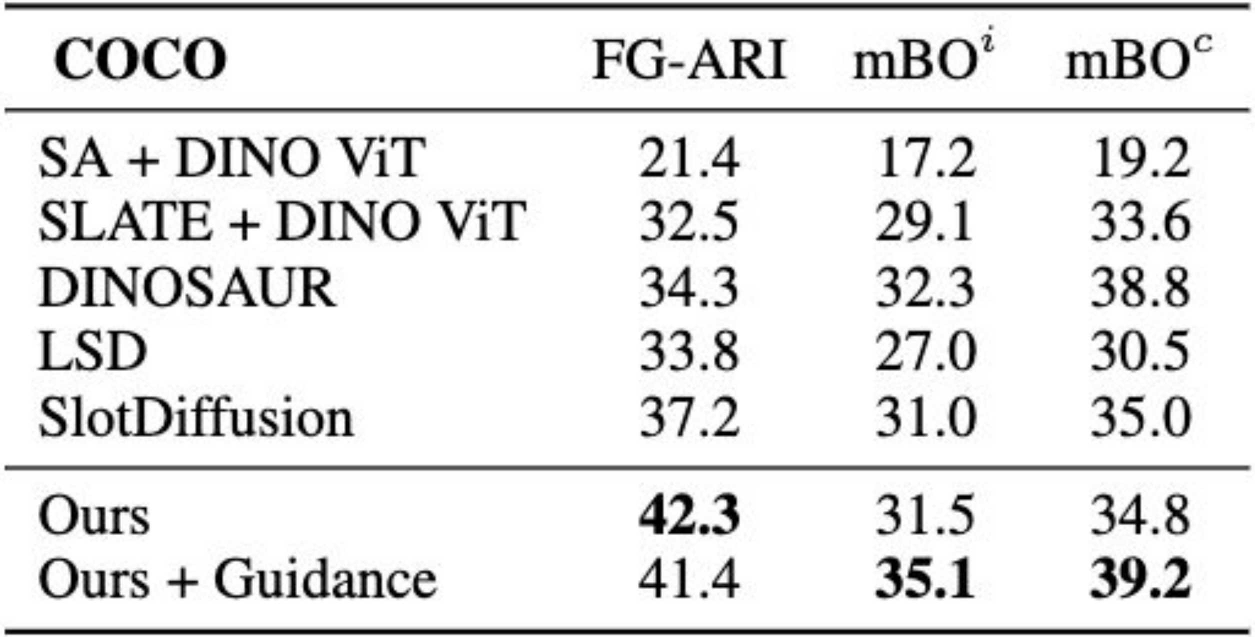
Slot Adapt


DINO
+
Slot Attention
Slot Attention Mask
$\mathcal{L}_{\text{guidance}}$
Cross Attention Mask
Slots
SD VAE
$\mathcal{L}_\theta$
Residual Block + Self Attention
Cross Attention
Adapter Attention
Register Token
Loss functions

Slot Adapt

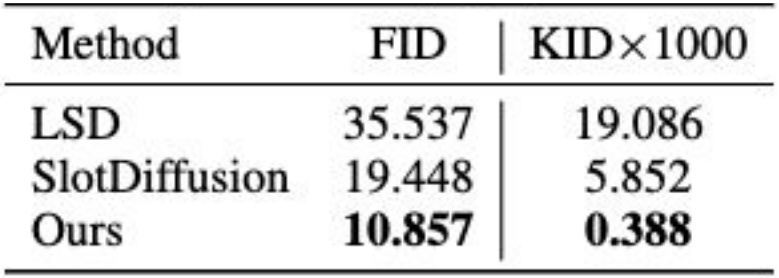

DINO
+
Slot Attention
SD VAE
Slots
Slot Attention Mask
$\mathcal{L}_{\text{guidance}}$
Cross Attention Mask
$\mathcal{L}_\theta$
Modified Unet Block
Input Intermediate Features
Slots
Residual+Self Attn
Pooling
Text Cross Attn
Register Token
Adapter Cross Attn
Output Intermediate Features
Residual Block + Self Attention
Cross Attention
Adapter Attention
Register Token
Loss functions

Results on COCO



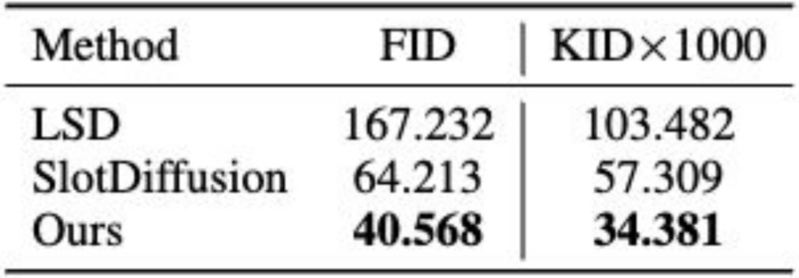
COCO	FG-ARI	mBO^i	mBO^c
SA + DINO ViT	21.4	17.2	19.2
SLATE + DINO ViT	32.5	29.1	33.6
DINOSAUR	34.3	32.3	38.8
LSD	33.8	27.0	30.5
SlotDiffusion	37.2	31.0	35.0
Ours	**42.3**	31.5	34.8
Ours + Guidance	41.4	**35.1**	**39.2**

Reconstruction & Compositional Generation



Method	FID	KID×1000
LSD	35.537	19.086
SlotDiffusion	19.448	5.852
Ours	**10.857**	**0.388**

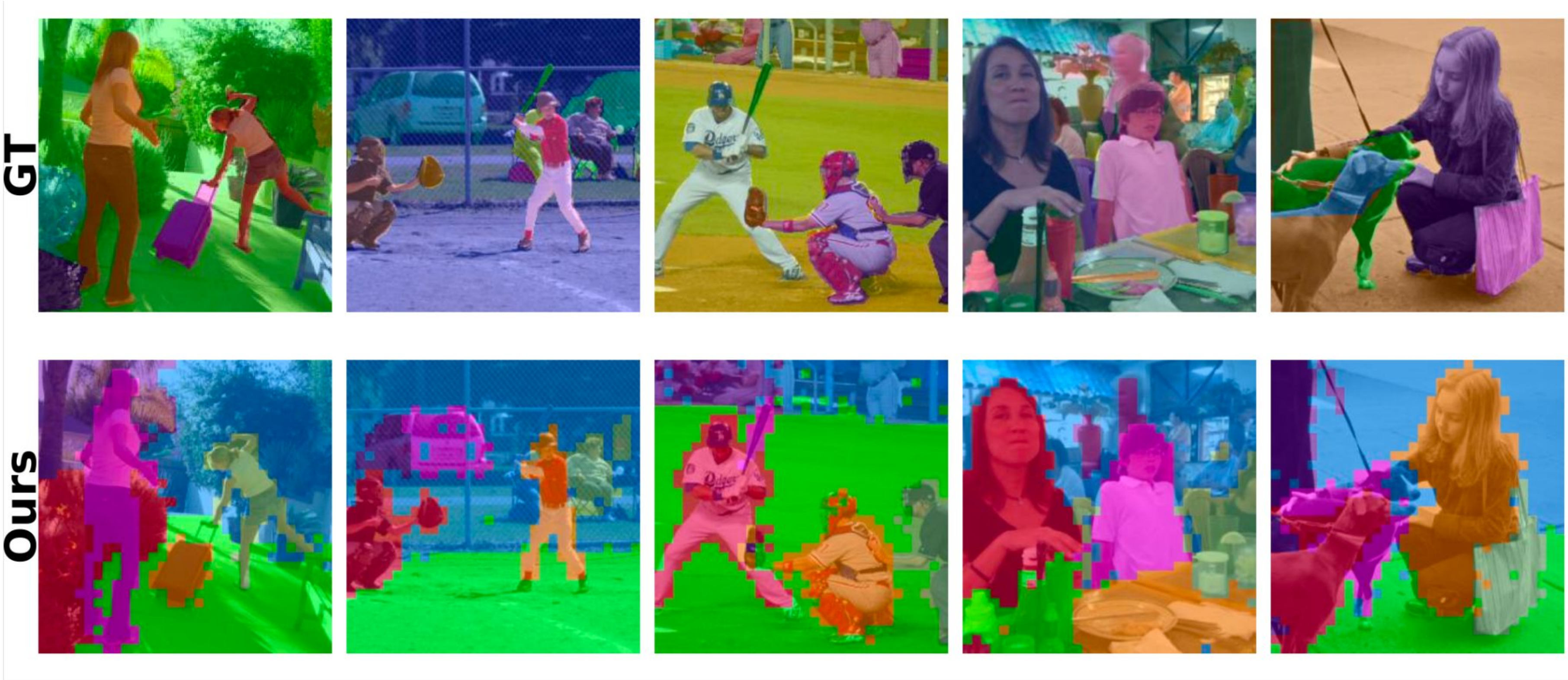
Reconstruction



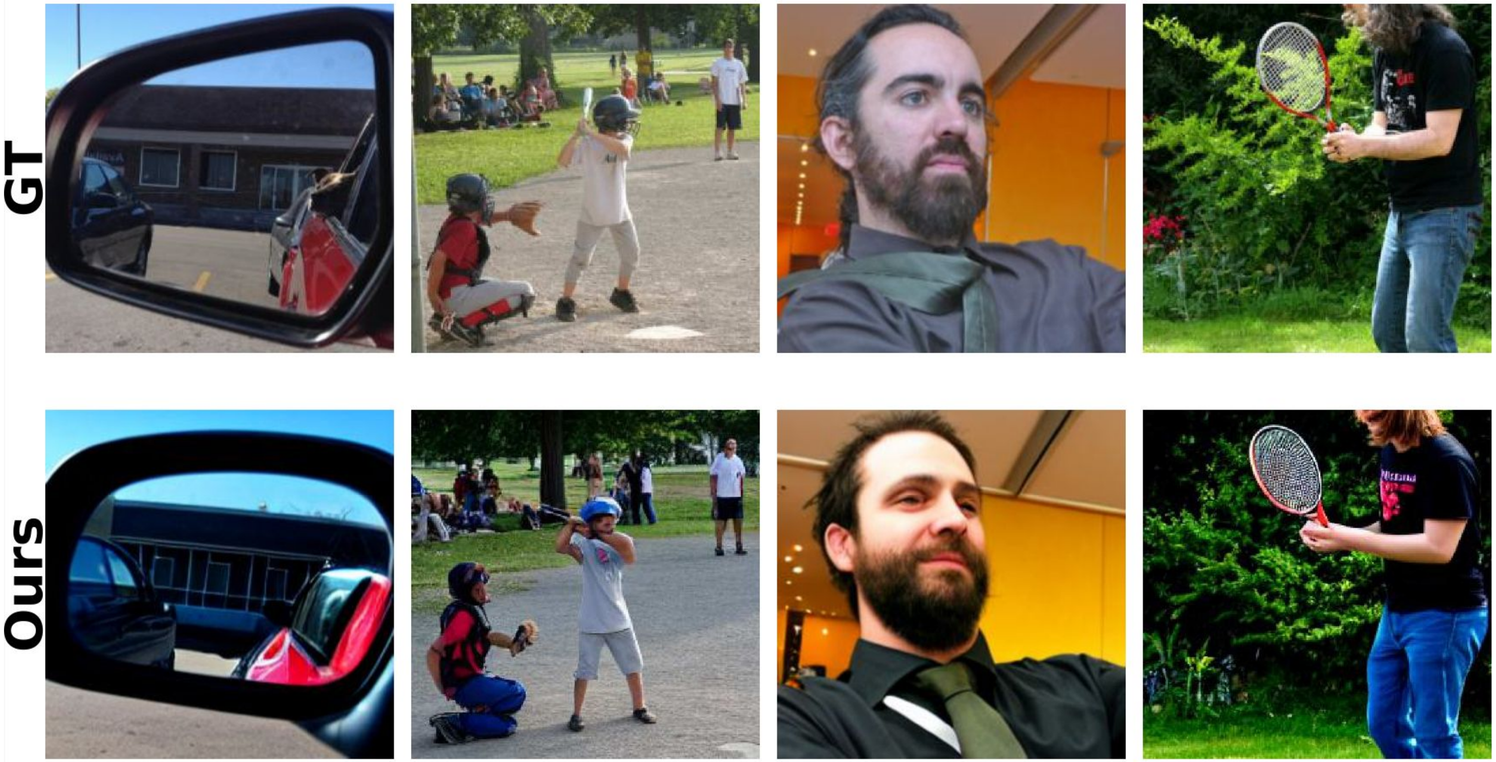
Method	FID	KID×1000
LSD	167.232	103.482
SlotDiffusion	64.213	57.309
Ours	**40.568**	**34.381**

Compositional Generation

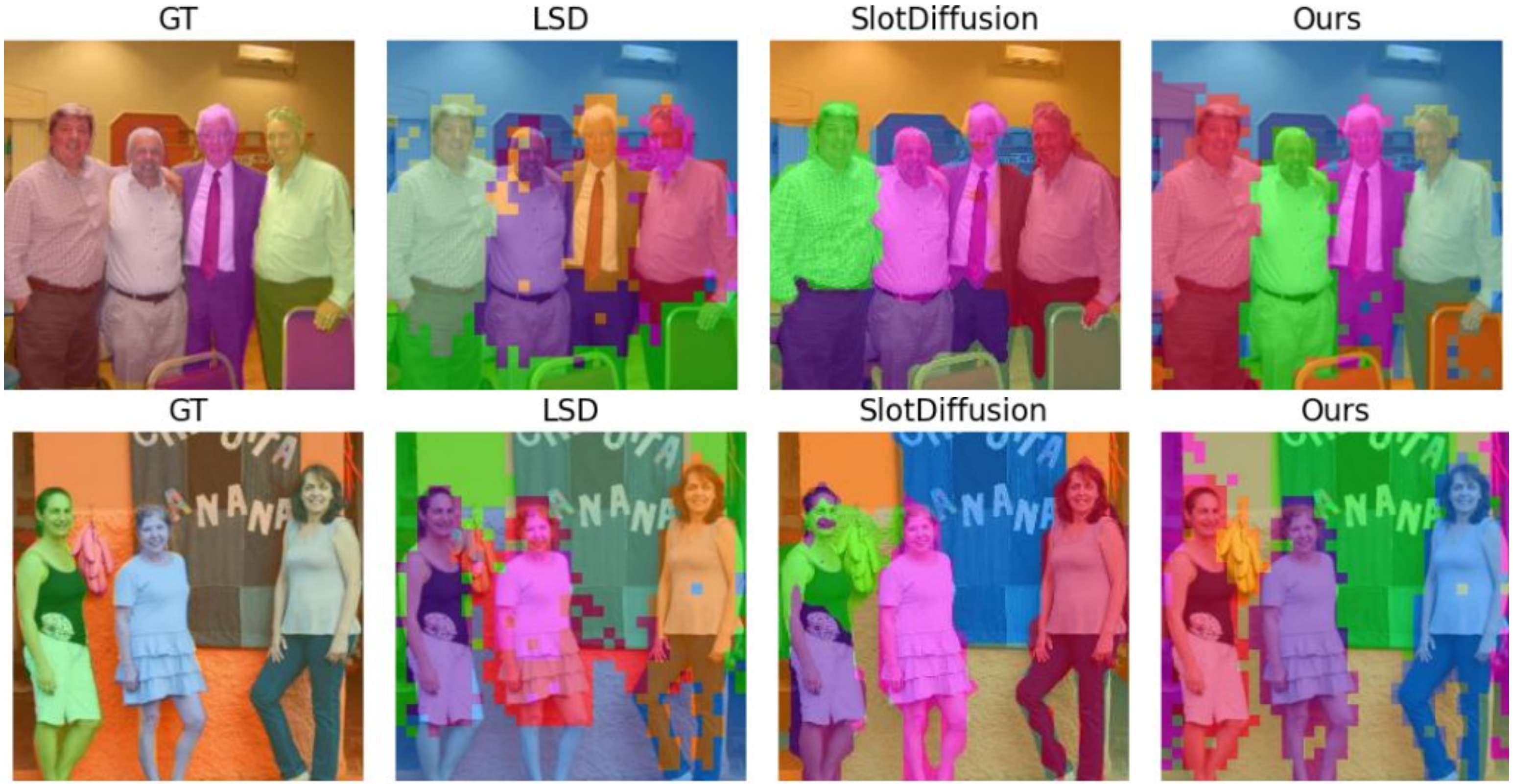
Visual Results - COCO



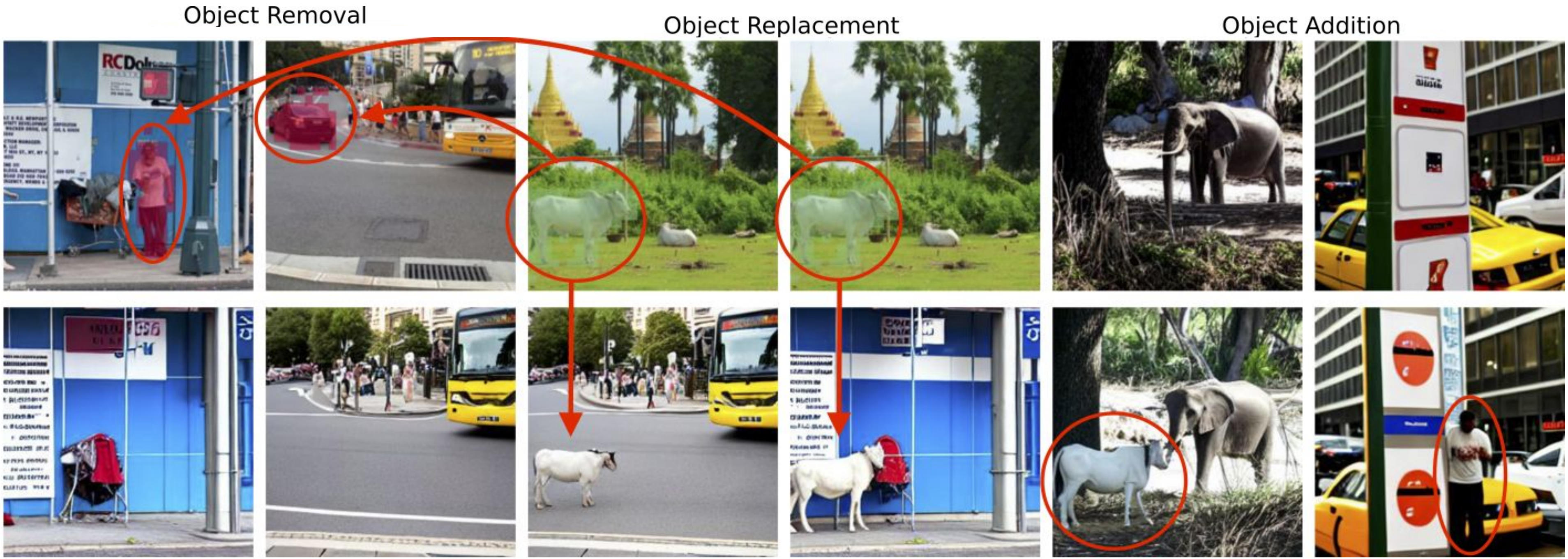
Visual Results - COCO



Visual Results - COCO



Compositional Generation - COCO



Conclusion

- We introduce SlotAdapt
    - Adapters for slot-based conditioning
    - A register token for capturing background context
    - Attention guidance to align slot attention with diffusion cross-attention
- Experiments show that:
    - State-of-the-art results on real-world datasets for object discovery, segmentation, and compositional editing
    - Fully unsupervised approach—first to demonstrate compositional editing on COCO.

Project Page