Open-Vocabulary Attention Maps with Token Optimization for Semantic Segmentation in Diffusion Models

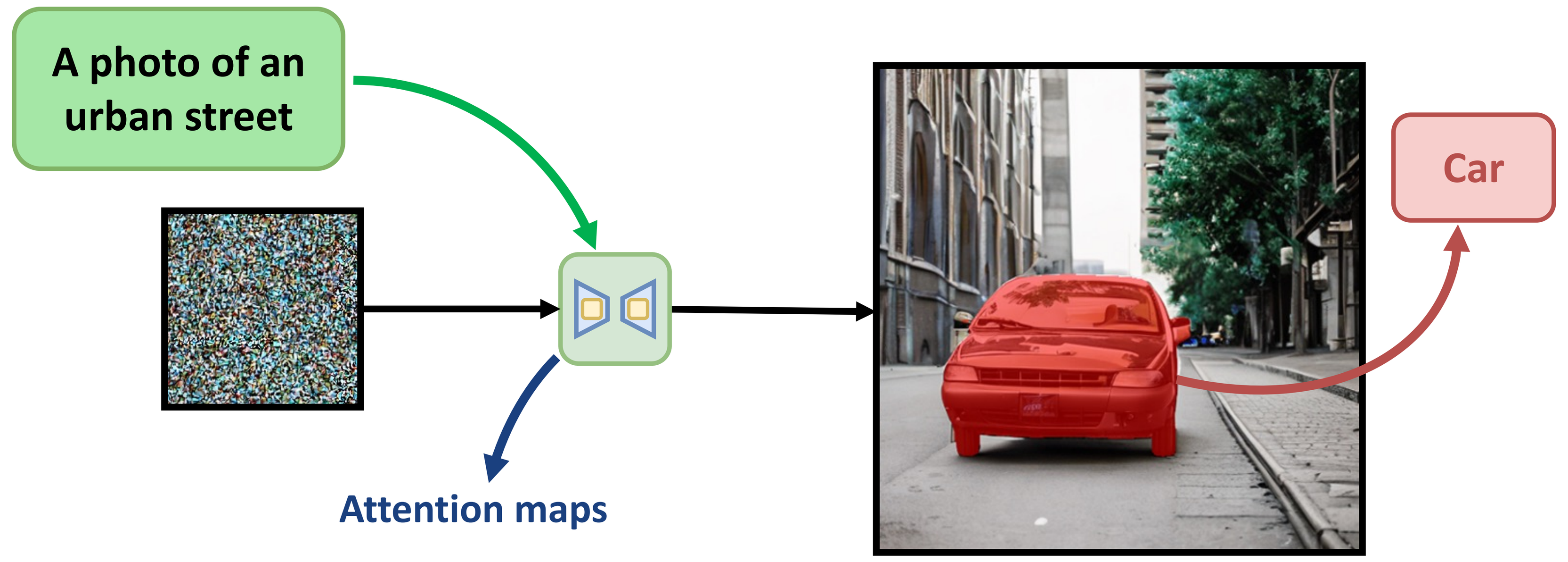
CVPR JUNE 17-21, 2024 SEATTLE, WA

Pablo Marcos-Manchón[1,2], Roberto Alcover-Couso[1], Juan C. SanMiguel[1], Jose M. Martínez[1]

[1]VPULab, Autonomous University of Madrid [2]Dynamics of Memory Formation, University of Barcelona

Universidad Autónoma de Madrid
Video Processing and Understanding Lab

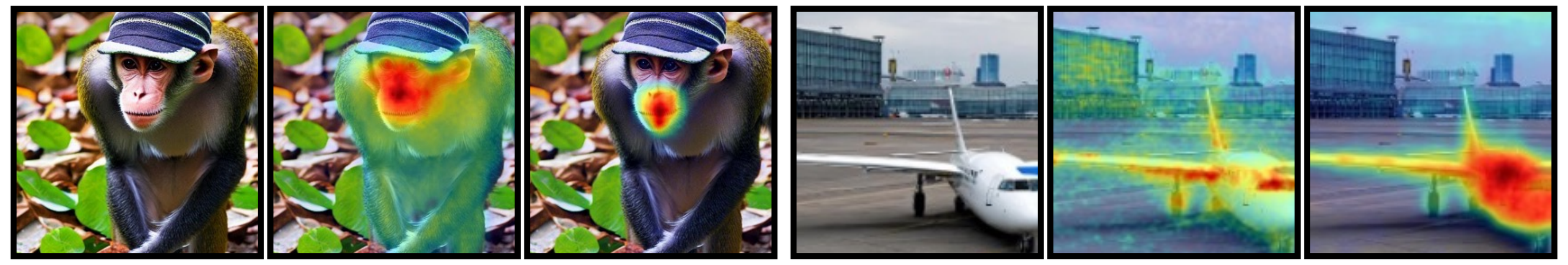
Problem & Contributions



- How can we extract pixel-level semantic information?
  → Using the **cross-attention maps** yielded during the diffusion process.
- How can we obtain semantic information from **non-prompt** words?
  → By introducing an independent **attribution prompt** to compute open-vocabulary attention maps (OVAM) in a training-free manner.
- How to choose the attribution prompt to segment a given class?
  → Learning a specific embedding for the attribution prompt employing our **token optimization**.

Visual Summary

Open-Vocabulary Attention Maps | Attention Maps with Optimized Token



Prompt: *monkey with hat walking*. **Attentions**: *monkey, mouth*

Prompt: *aeroplane parked at an airport in the city*. **Attentions**: *Aeroplane*.

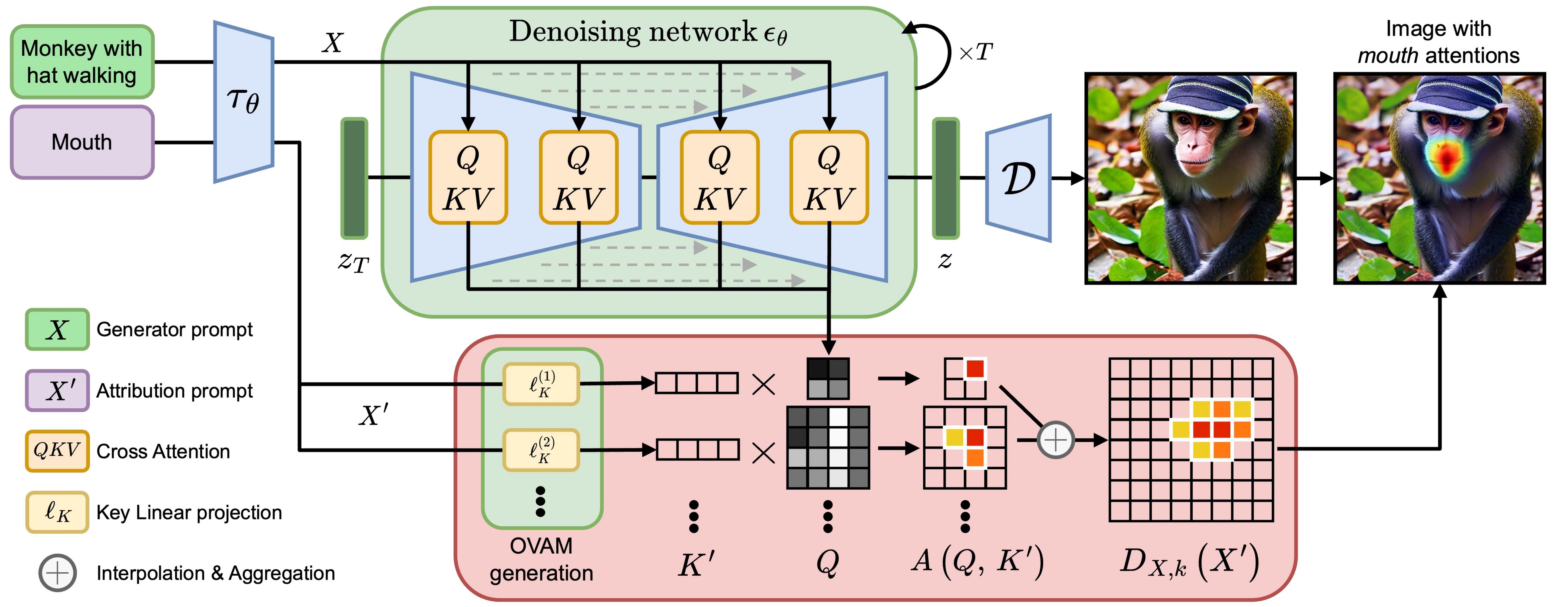
Open-vocabulary pixel-level annotations generated using OVAM

Prompt: *someone jumping in the air on their snowboard*

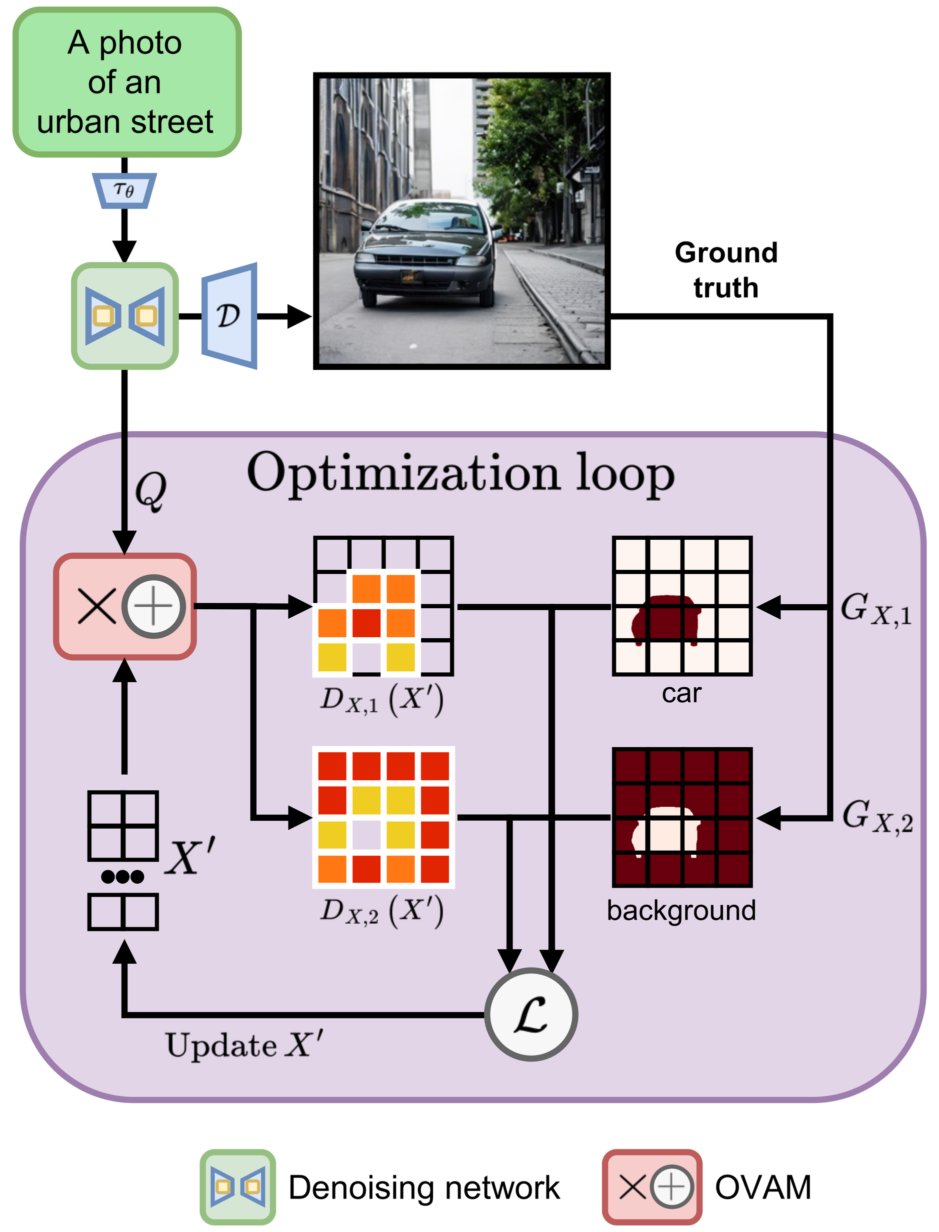
Prompt: *a photo of a lion on a mountain top at sunset*

Prompt: *a car driving under the sea on the ocean floor*

Open-Vocabulary Attention Maps

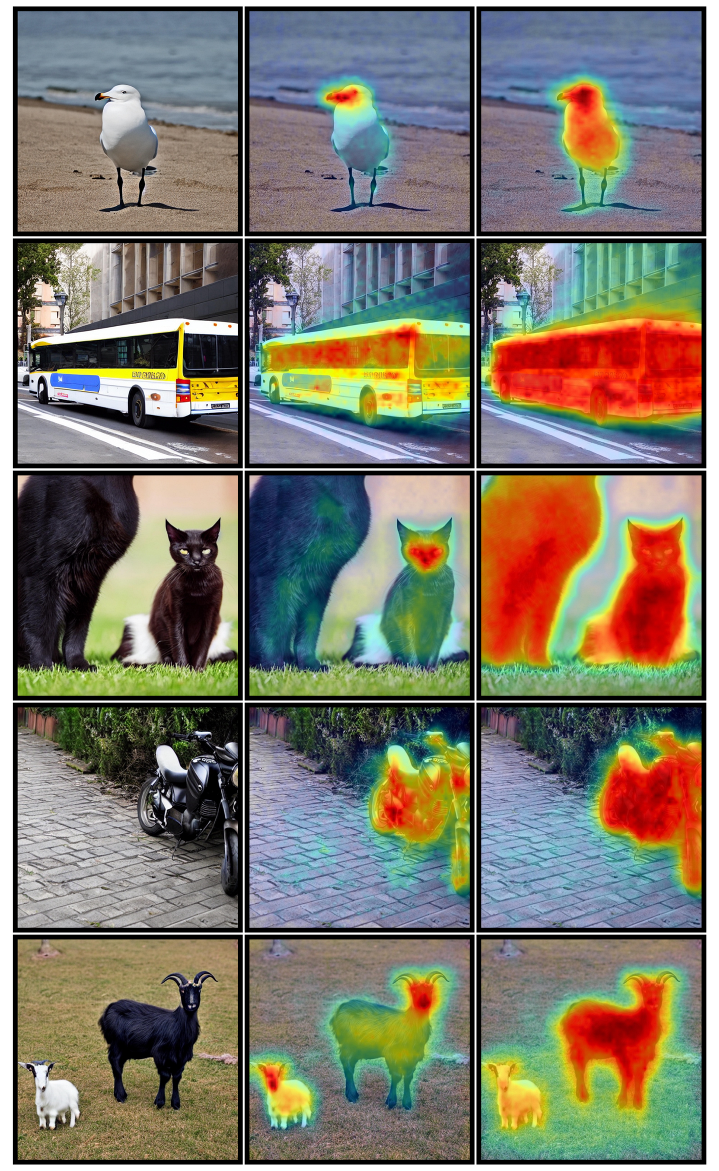


Token Optimization



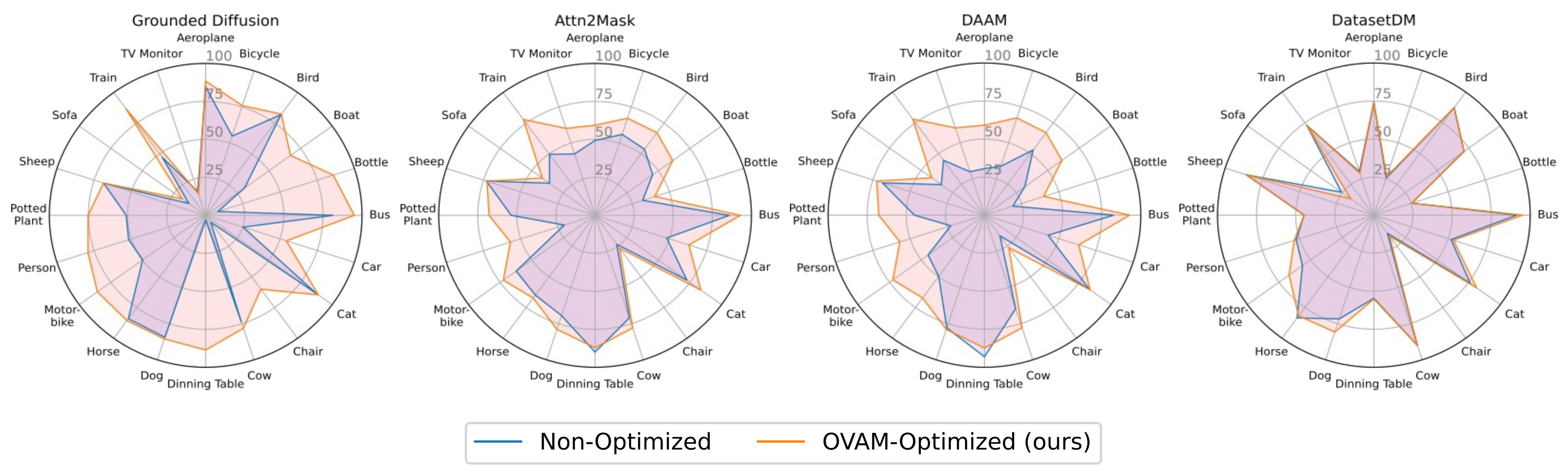
Attention Maps

Attention with (right) and without (left) optimized tokens



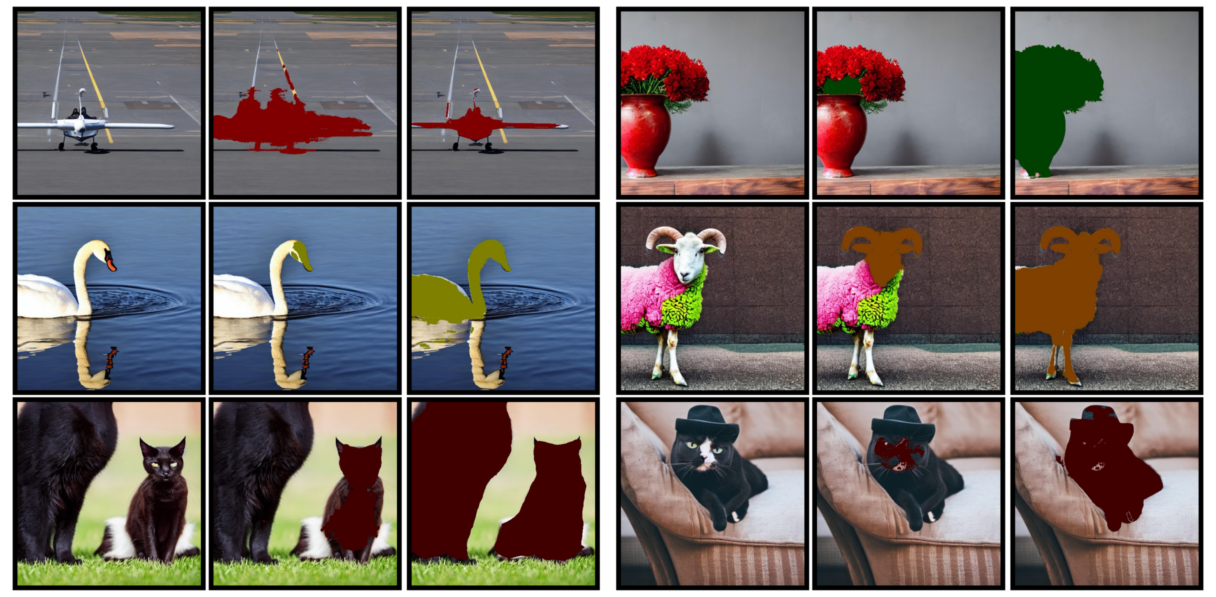
Evaluation of Optimized Tokens

Method	Selected classes (COCO-cap IoU %)										Dataset (mIoU %)	
	aeroplane	bicycle	boat	bus	car	cat	cow	dog	motorbike	person	VOC-sim	COCO-cap
Training-free methods												
DAAM	30.6	33.8	31.9	82.6	42.8	83.0	64.6	77.9	44.6	22.7	66.2	48.4
DAAM + *token optimization*	**59.1**	**67.2**	**61.2**	**92.8**	**63.4**	**83.6**	**77.9**	**79.0**	**72.4**	**56.9**	**79.7**	**66.1**
Attn2Mask	49.3	56.0	45.5	85.8	48.2	72.6	71.0	69.4	62.4	20.7	68.7	55.0
Attn2Mask + *token opt.*	**59.3**	**67.2**	**61.4**	**92.9**	**63.1**	**83.6**	**77.9**	**78.9**	**72.5**	**56.9**	**81.9**	**66.1**
OVAM (ours)	65.1	64.3	51.9	84.9	47.5	67.9	76.5	65.8	69.4	19.7	70.4	58.2
OVAM + *token optimization*	**67.8**	**68.4**	**64.6**	**94.5**	**63.2**	**87.6**	**82.4**	**81.9**	**74.2**	**60.9**	**82.5**	**69.2**
Methods with additional training												
DatasetDM	**74.1**	25.7	**71.3**	91.4	51.9	76.2	**90.1**	71.7	56.4	52.2	80.3	59.3
DatasetDM + *token opt.*	73.7	**26.7**	71.2	**95.1**	**53.5**	**81.2**	89.9	**80.8**	**67.0**	**53.5**	**80.6**	**60.5**
Grounded Diffusion	84.6	54.9	30.8	81.4	25.1	87.3	73.7	84.4	49.8	51.8	62.1	50.2
Grounded Diffusion + *token opt.*	**88.3**	**75.9**	**67.1**	**95.0**	**54.3**	**89.0**	**78.1**	**85.5**	**85.6**	**79.1**	**86.6**	**73.3**



Segmentation Masks

Segmentation Masks with (right) and without (left) optimized tokens



Website

Code and examples

github.com/vpulab/ovam