Master in Deep Learning for Audio and Video Signal Processing

Synthetic Data Generation using Latent Diffusion Models for Semantic Segmentation of Urban Scenes

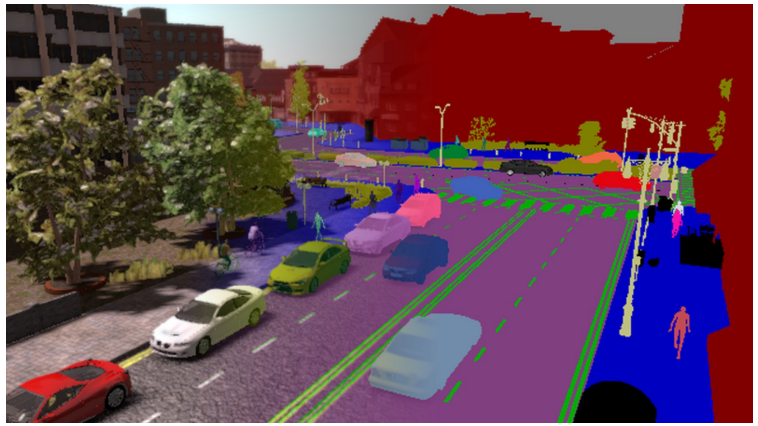
Pablo Marcos Manchón

pablo.marcosm@estudiante.uam.es

INTRODUCTION: SEMANTIC SEGMENTATION

Urban Scene Image (Cityscapes)

Semantic Segmentation ground truth



Synthetic data from a simulator (Synthia)

- **Task**: Pixel-wise classification of semantic labels.
- **High annotation cost**. One fined-grained annotation can take 90' of expert work.

- **Solution:** Synthetic data from simulators, data augmentation.
- **Issues**: Lack of variability and realism. Domain adaptation problems.

INTRODUCTION: GENERATIVE MODELS

- Rising on generative models for image, video, audio, and text.
- Powerful tool for synthetic data: diverse and controllable scenes.
- **Text-to-image** architectures: DALL·E 2, Midjourney, or Stable Diffusion.
- Responsible by rising of **Latent Diffusion Models**

"A photograph of an astronaut riding a horse"

"Darth Vader is surfing on waves"

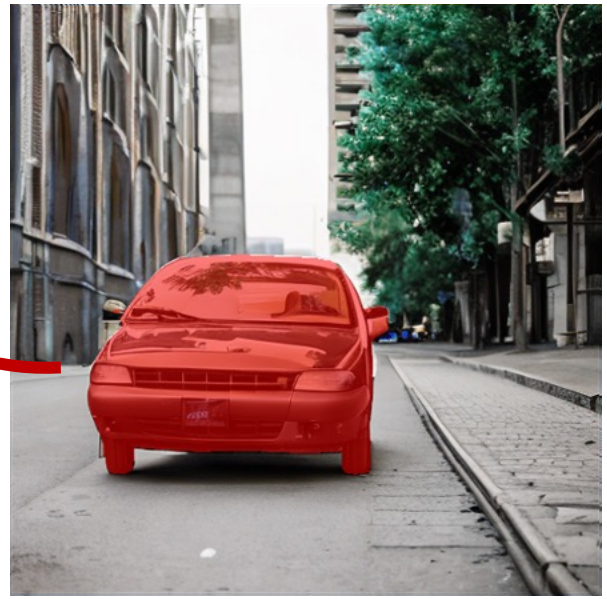
Astronaut image generated with Stable Diffusion [source]. Darth Vader surfing video generated with ZeroscopeV2 text-to-vide [huggingface]

INTRODUCTION: RESEARCH QUESTION

Is it possible to use
Diffusion Models to generate
Synthetic Data for semantic segmentation?

"A car in an urban environment"

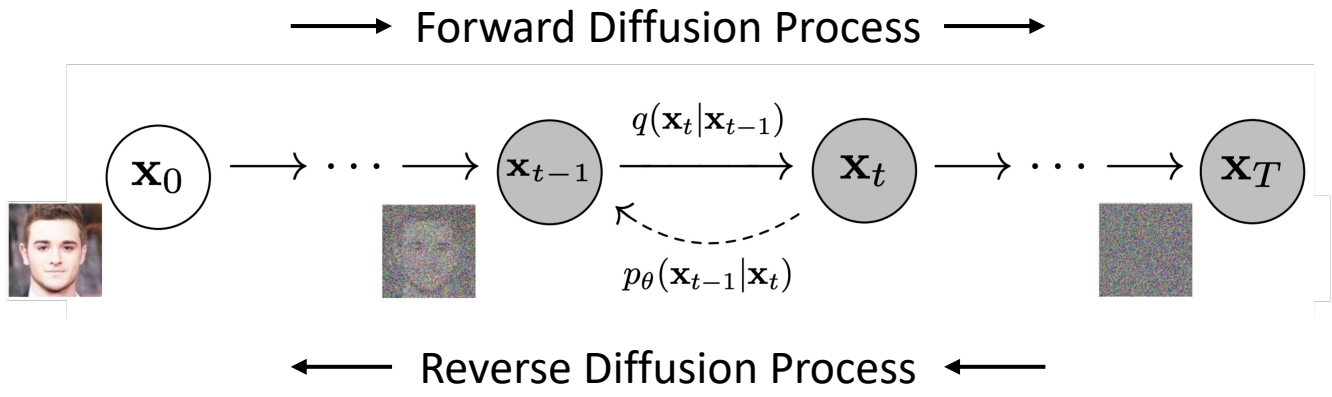
Semantic Information
Car



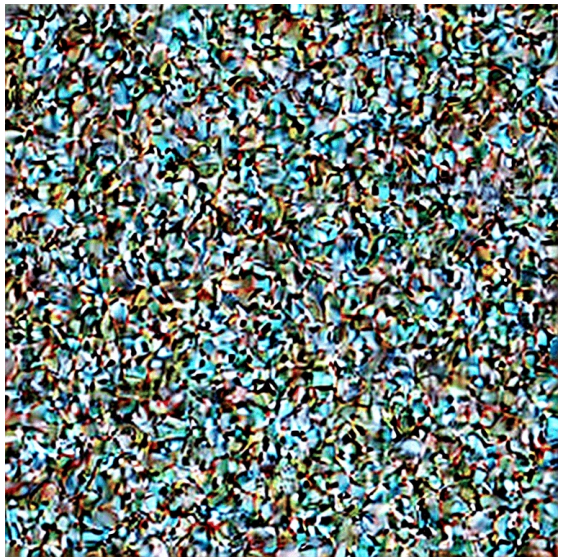
- Exploratory work
  1. Extensive related work study
  2. In depth exploration of diffusion models
  3. Design and proposal of approach
  4. Experimental validation
  5. Potential use and future research directions

RELATED WORK: LATENT DIFFUSION MODEL (LDM)

- Latent diffusion model for **text-to-image** generation.
- Inspired in thermodynamics of a **forward diffusion process**.
- Model learns the unknown **reverse diffusion process**.
- Process is guided by text **semantic information**.
- Implementation: **Stable Diffusion**.



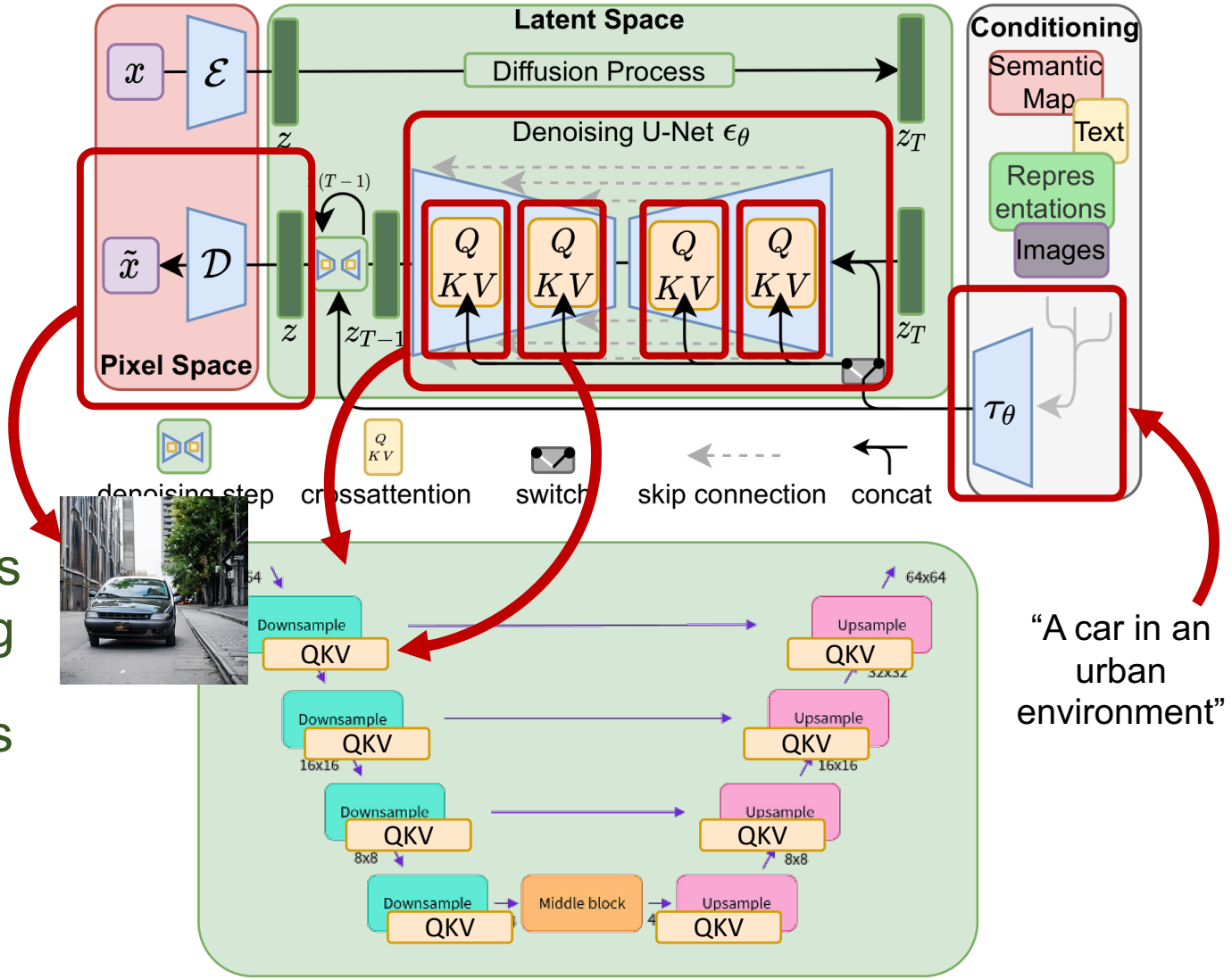
"A car in an urban environment"



Diffusion diagram modified from Ho J. et al. : "Denoising Diffusion Probabilistic Models"

RELATED WORK: STABLE DIFFUSION

- **Stable Diffusion**: text-to-image latent diffusion model.
- Compose of three subnetworks:
  - **Denoiser network**. A U-NET applies the reverse diffusion
  - **Text encoder**. A **CLIP Model** maps the text prompt to a text embedding
  - **Latent decoder**. A **VAE** transforms the latent variable to a final image



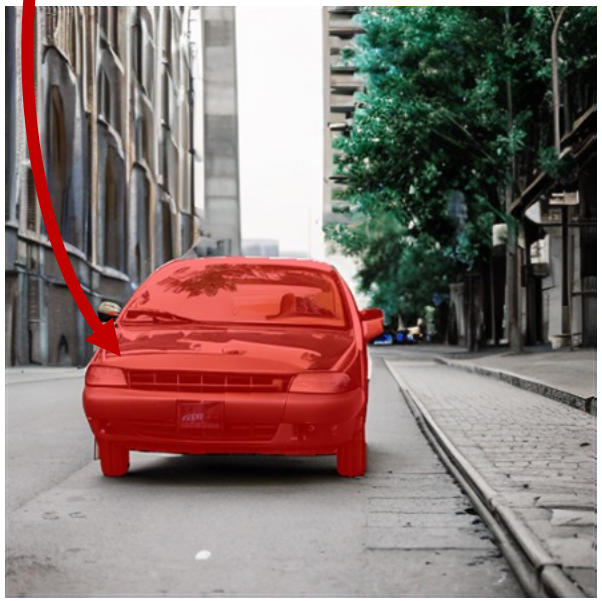
Stable Diffusion diagram from Rombach R. et al. : "High-Resolution Image Synthesis with Latent Diffusion Models"

RELATED WORK: LDM EXPLAINABILITY

How to extract pixel semantic information?

Explainability approach: **Text Attribution**

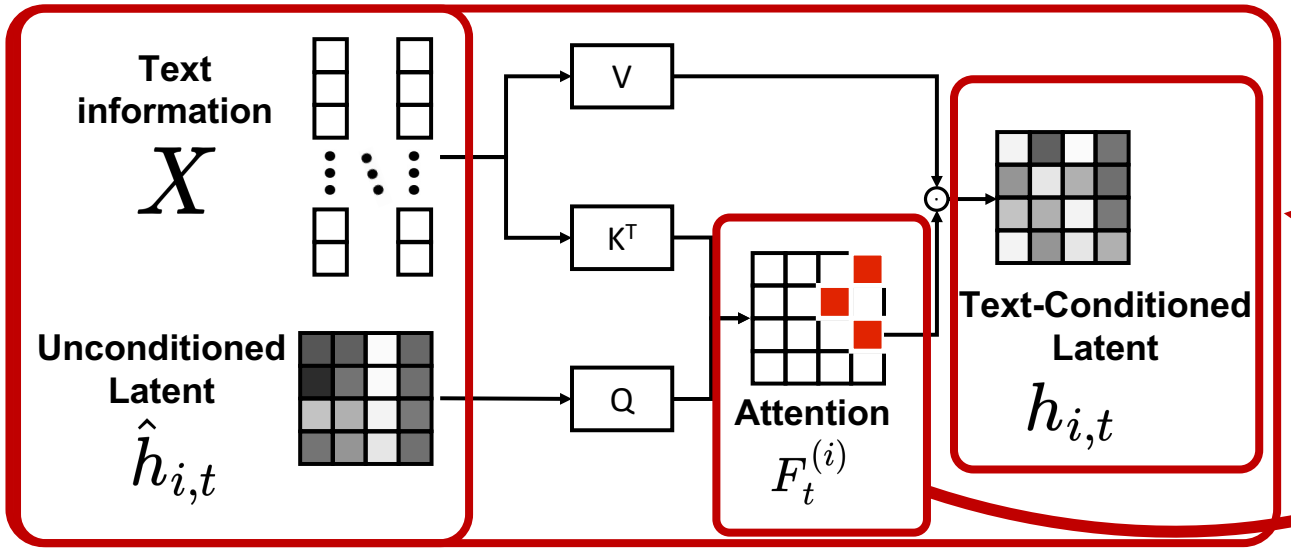
"A car in an urban environment"



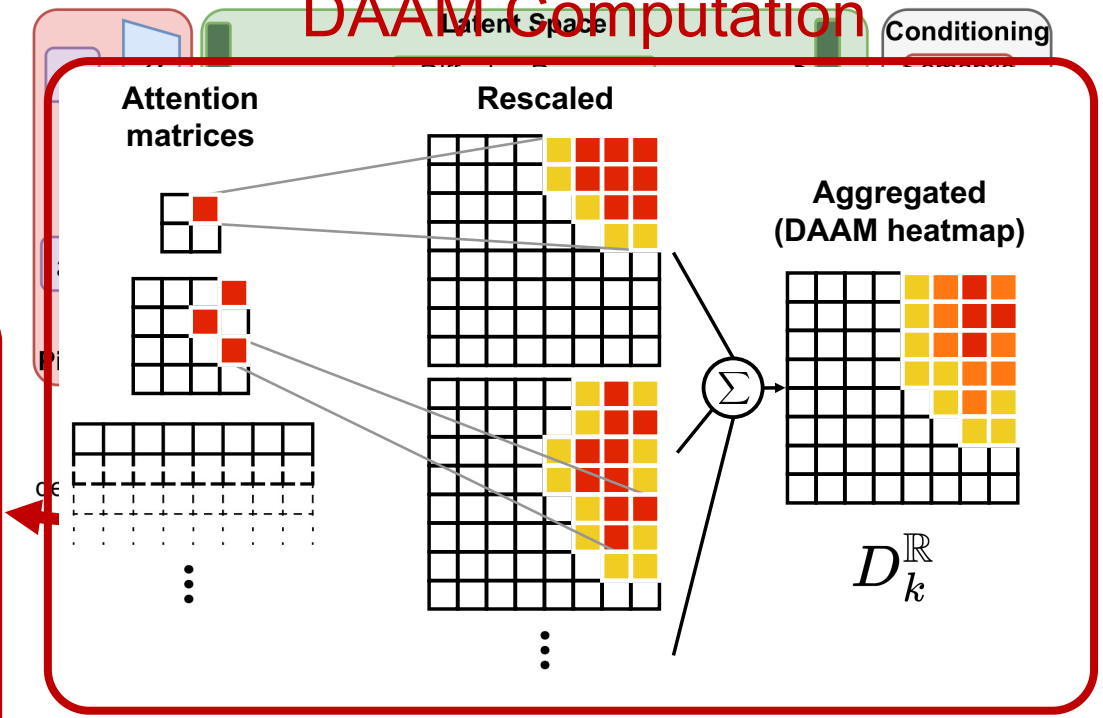
- Computer Vision Explainability methods:
  - **Gradient-based**: Not computationally viable (Grad-CAM, Saliency Maps).
  - **Perturbation-based**: Not feasible. E.g. (Extremal perturbations).
  - **Attention-based**. Takes ideas from Natural Language Processing. **Diffusion Attentive Attribution Maps (DAAM).**

METHOD BASELINE: DAAM

- DAAM: Text attribution explanations
- Aggregates cross-attentions during the diffusion process



DAAM Computation



$$h_{i,t}^{\downarrow} = F_t^{(i)}\left(\hat{h}_{i,t}^{\downarrow}, X\right) \cdot \left(W_v^{(i)\downarrow} X\right)$$

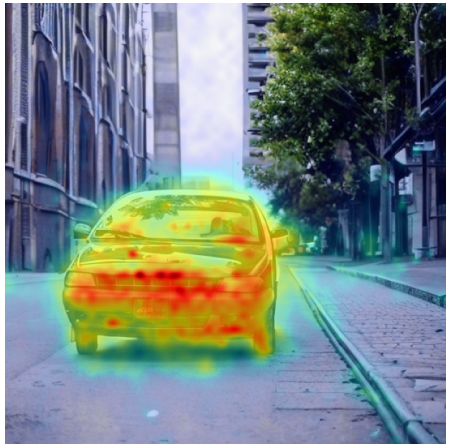
$$F_t^{(i)}\left(\hat{h}_{i,t}^{\downarrow}, X\right) = \text{softmax}\left((W_q^{(i)\downarrow}\hat{h}_{i,t}^{\downarrow})(W_k^{(i)\downarrow}X)^T/\sqrt{d}\right).$$

$$D_k^{\mathbb{R}}[x,y] := \sum_{t,i,l} \tilde{F}_{t,k,l}^{(i)\downarrow}[x,y] + \tilde{F}_{t,k,l}^{(i)\uparrow}[x,y]$$

Tang R. et al. : "What the DAAM: Interpreting Stable Diffusion Using Cross Attention"

METHOD BASELINE: DAAM EXAMPLES

Car



In

Urban

Environment

X = A person in an urban environment

- Attribution maps can be used as segmentation masks
- Limitations to extend beyond explainability:
  - The words should appear in the text prompt
  - Attention depends on the other words of the text: cannot evaluate words alone
  - Difficult to control semantic information

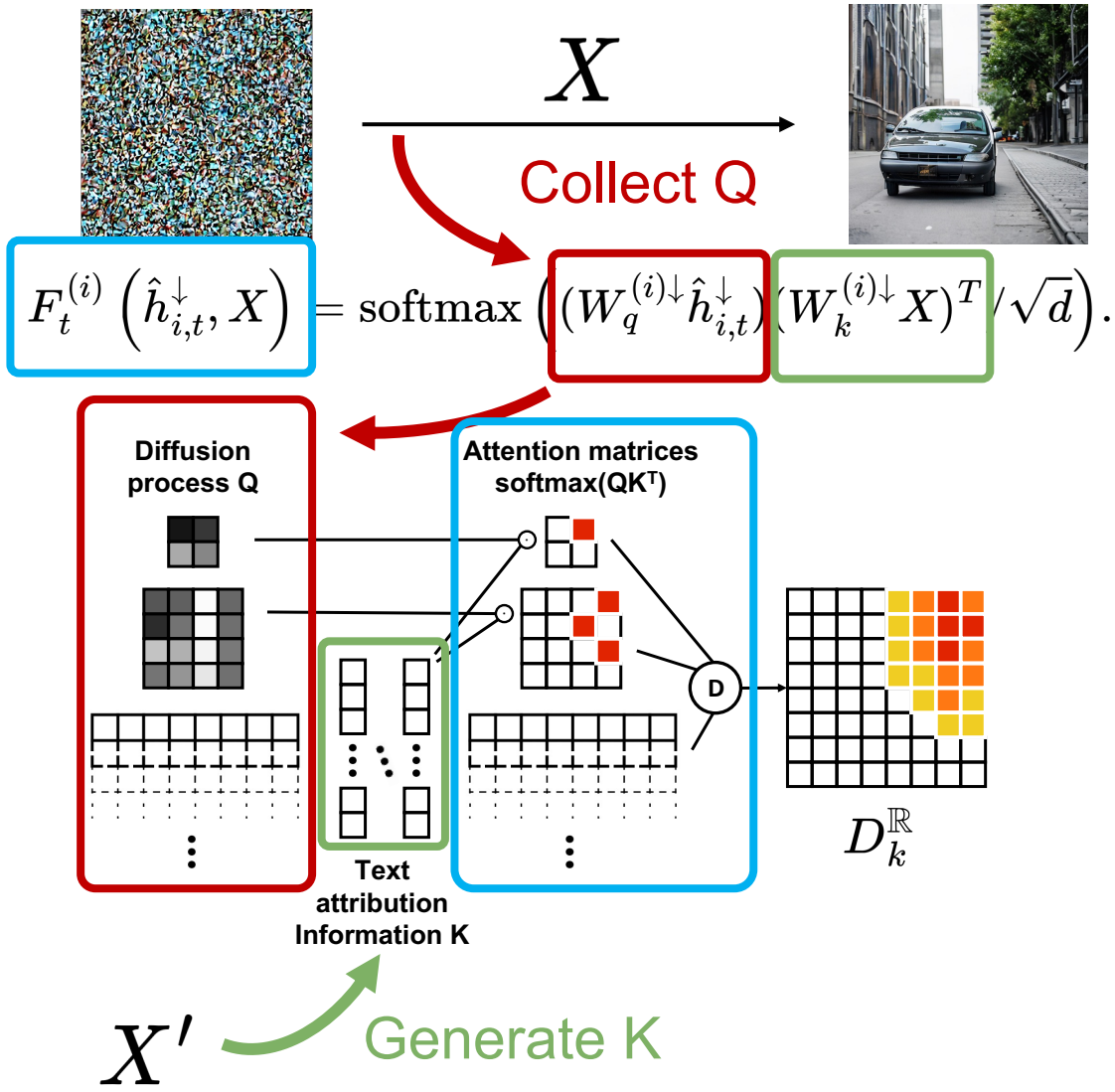
Design and implement extensions

PROPOSAL: OPEN-VOCABULARY DAAM

Only can evaluate text prompt tokens

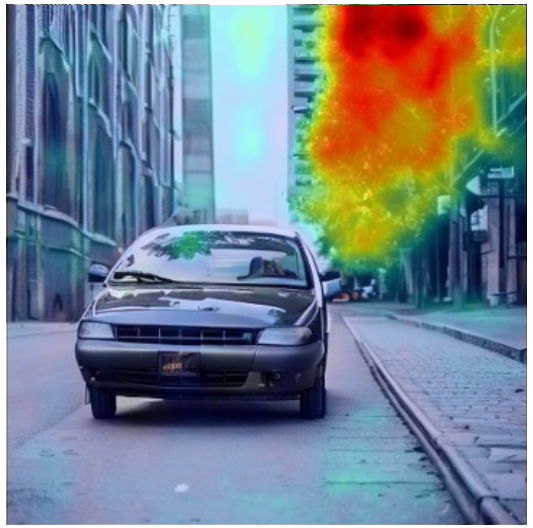
Open-vocabulary DAAM

- Extension to create attribution maps of arbitrary text prompts
- Dissociate text that generate image with text that generates heatmaps
- Redesign DAAM as a 2-step process:
  1. Query collection process from X
  2. Generate Keys from X` and compute DAAM



OPEN-VOCABULARY DAAM EXAMPLES

X = A car in an urban environment



Tree

X' = A car and a **tree** in an urban environment

Building

X' = A car and a **building** in an urban environment

Sidewalk

X' = A car on a **sidewalk** in an urban environment

PROPOSAL: LINEAR OPEN-VOCABULARY DAAM

Attention depends on other words

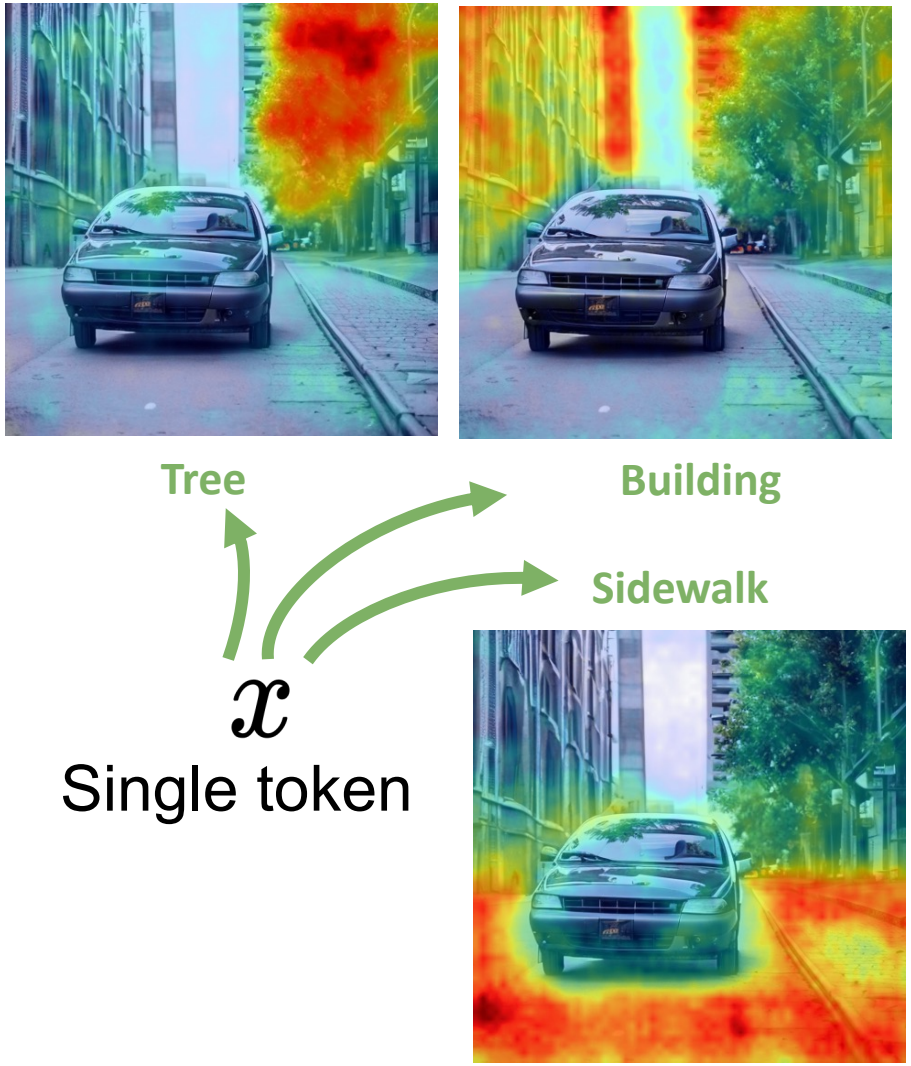
Linear Open-vocabulary DAAM

- Given the proposed OV-DAAM we remove softmax between tokens

$$F_t^{(i)}\left(\hat{h}_{i,t}^{\downarrow}, X\right) = \text{softmax}\left((W_q^{(i)\downarrow}\hat{h}_{i,t}^{\downarrow})(W_k^{(i)\downarrow}X)^T/\sqrt{d}\right).$$

$$L_{X,t}^{(i)}(x) = (W_q^{(i)\downarrow}\hat{h}_{i,t}^{\downarrow})(W_k^{(i)\downarrow}x)^T$$

$$LD_X^{\mathbb{R}}[x,y] := \sum_{t,i,l} \tilde{L}_{X,t,l}^{(i)\downarrow}[x,y] + \tilde{L}_{X,t,l}^{(i)\uparrow}[x,y].$$



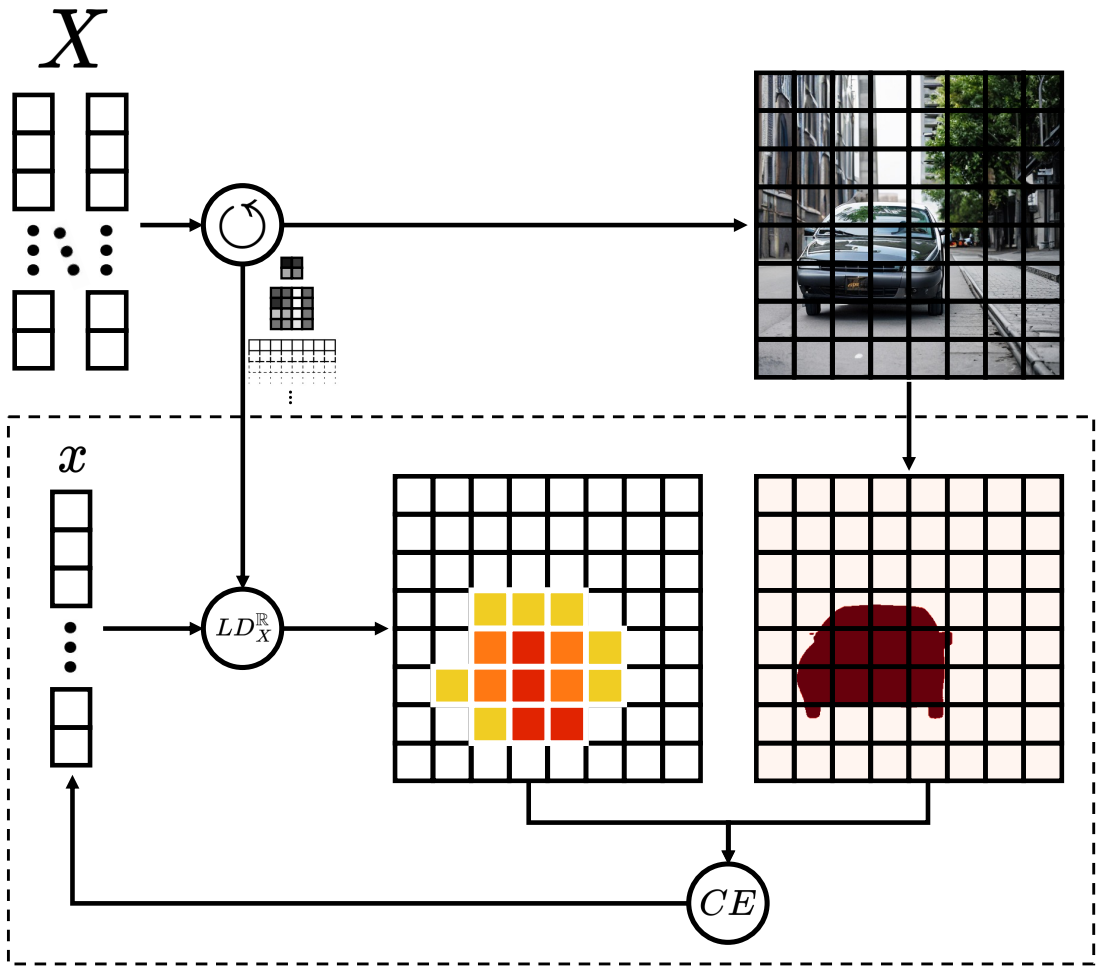
PROPOSAL: TEXT PROMPT OPTIMIZATION

How to find word with semantic information of region?

Optimization

- Approach as a optimization of the token of Open Vocabulary DAAM
- Is differentiable w.r.t. X thus we can use gradient descent to optimize the token

$$\hat{x} = \arg\min_{x} C\left(LD_X^{\mathbb{R}}(x), G_X\right)$$



TEXT PROMPT OPTIMIZATION RESULTS

X = A person in an urban environment

LDAAM

x = “Person”

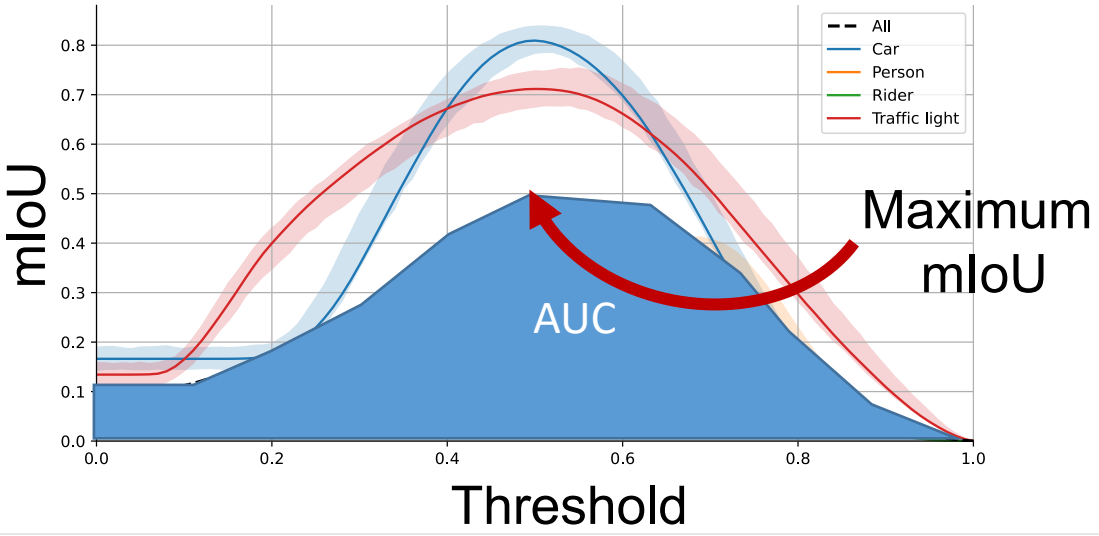
$\hat{x}$

Optimized token with person semantic information

Information transferred

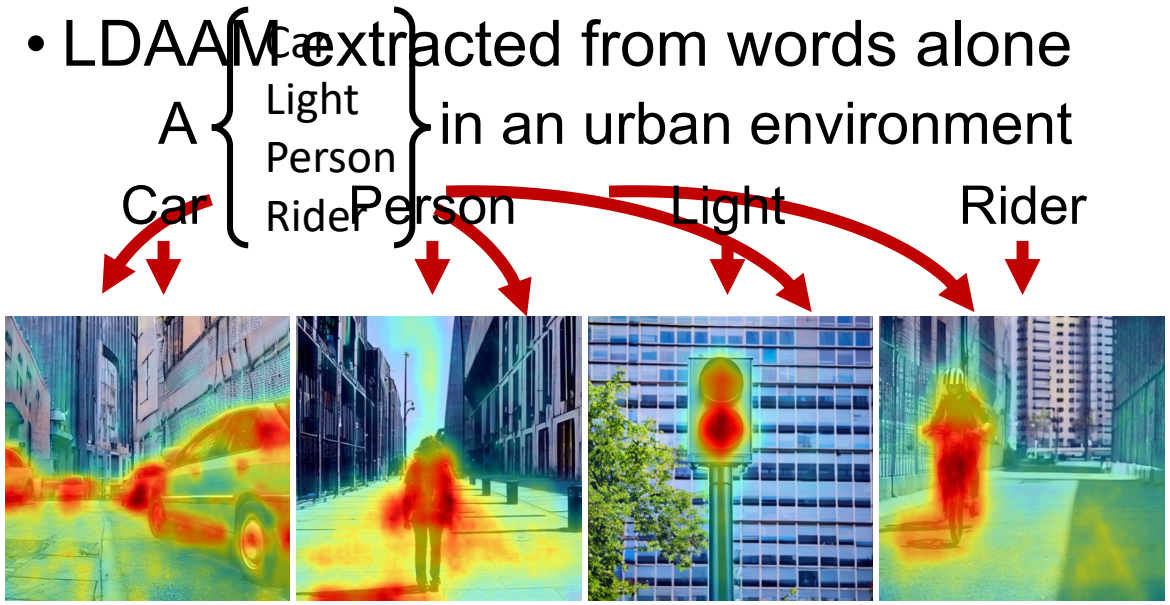
EXPERIMENT: METHODOLOGY

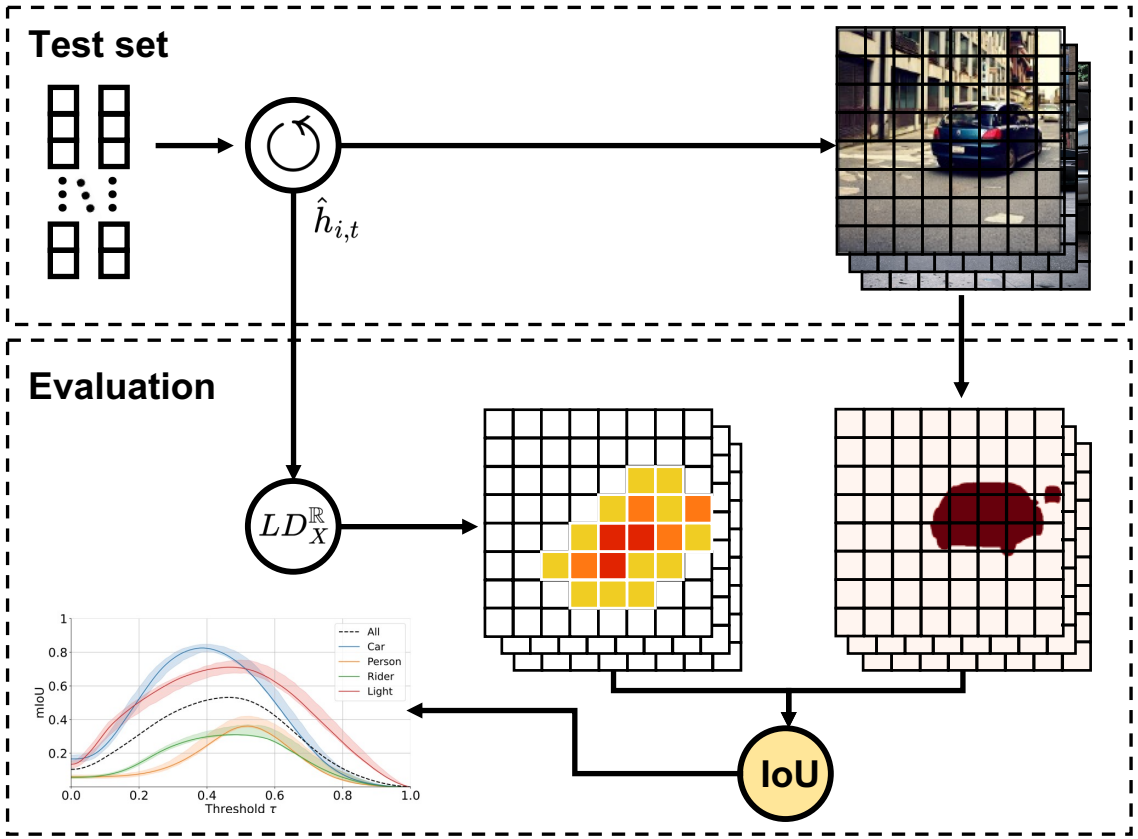
- Experiments designed to compare DAAM and LDAAM with optimized versions
- Generated a synthetic dataset with 200 images with 4 object classes
- Use mIoU to compare segmentation mask
- Curves with mIoU vs threshold



EXPERIMENT: DAAM & LDAAM

- Compute mIoU curves from DAAM and LDAAM
- Use the words “car”, “light”, “person” and “rider” for generating the heatmaps
- DAAM extracted from the text prompt
- LDAAM extracted from words alone





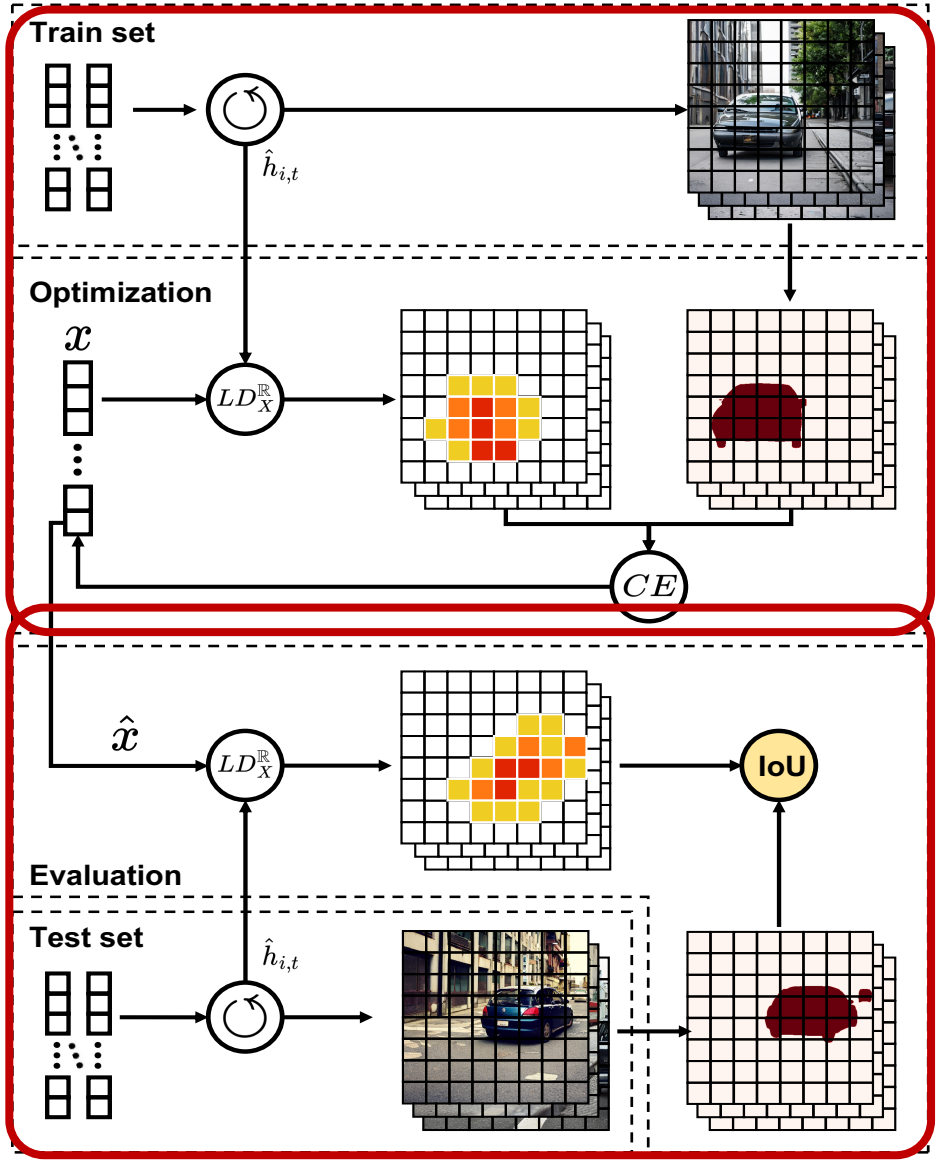
EXPERIMENT: DAAM & LDAAM OPTIMIZATION

- Compute mIoU curves from DAAM and LDAAM using an optimized token
- Optimize token using a train set

	1 train sample		2 train samples		5 train samples	
	Train	Test	Train	Test	Train	Test
Car	88.4	85.9	86.1	86.0	86.3	**86.6**
Person	65.1	70.0	65.5	**71.8**	65.7	70.9
Traffic Light	88.3	69.7	70.3	**72.2**	70.3	70.3
Rider	41.1	38.6	50.7	**47.8**	47.4	38.7
All	69.8	63.1	69.8	**67.3**	63.6	65.4

- Evaluate the heatmap generated by the token in the test set
- DAAM: optimize full text prompt. LDAAM: optimize single token.

Table results: LDAAM optimization, maximum mIoU of curves



EXPERIMENT: QUANTITATIVE RESULTS

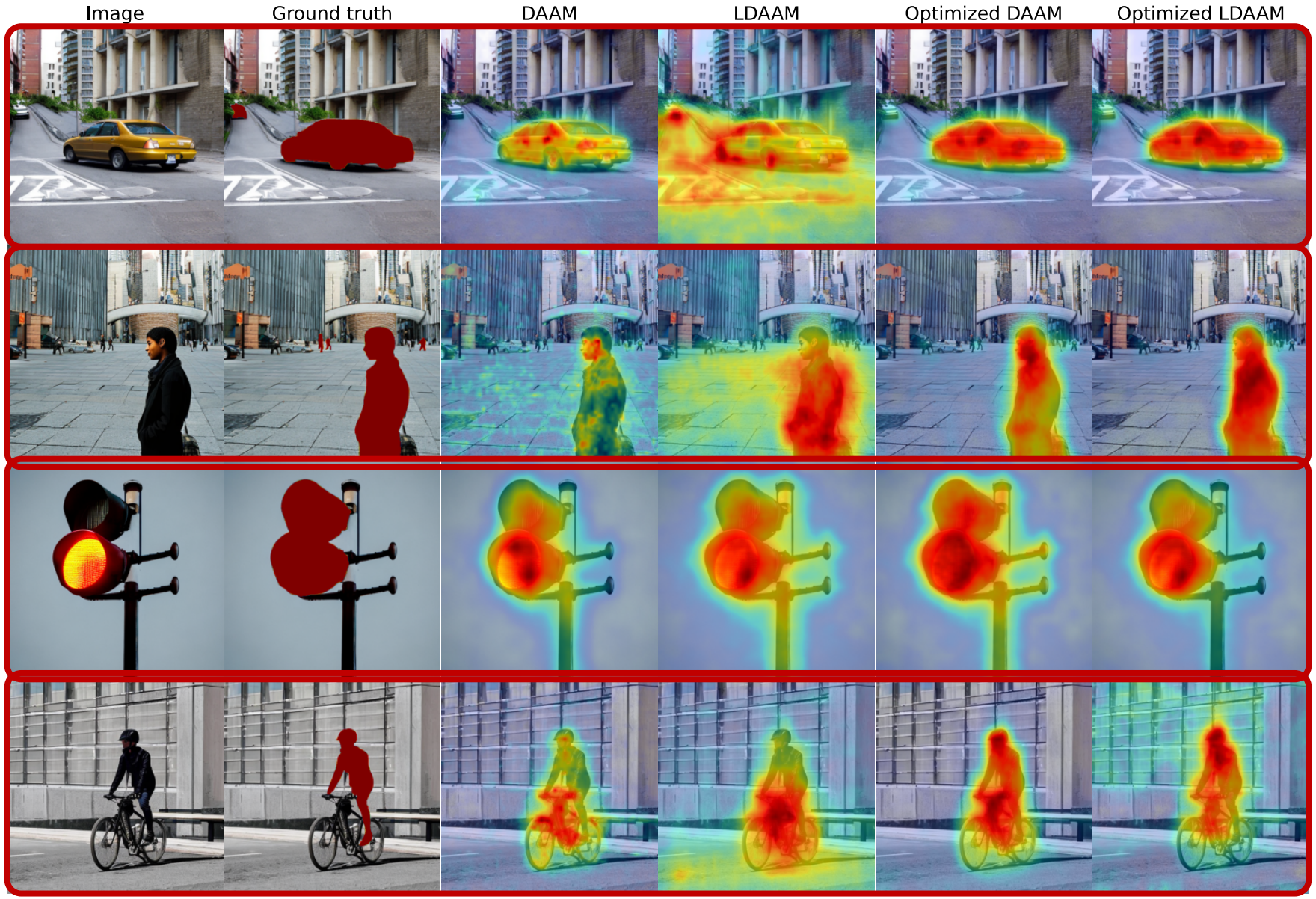
	Non-Optimized		Optimized	
	DAAM	LDAAM	DAAM	LDAAM
Car	82.5	41.7	**86.0**	85.8
Person	36.2	37.4	**71.6**	71.3
Traffic Light	71.1	68.2	**72.5**	72.3
Rider	31.0	21.3	**47.6**	43.1
All	53.2	38.3	**67.3**	66.6

Table 4.2: Summary of experiments: maximun mIoU

	Non-Optimized		Optimized	
	DAAM	LDAAM	DAAM	LDAAM
Car	39.7	23.7	**51.6**	51.5
Person	13.8	15.4	36.6	**41.3**
Traffic Light	43,3	38.3	**42.8**	42.7
Rider	15.6	11.0	**23.5**	22.5
All	28.1	22.1	38.6	**39.5**

Table 4.3: Summary of experiments: AUC

EXPERIMENT: QUALITATIVE RESULTS



CONCLUSIONS

- Work recap
  - Extensive Literature review of semantic segmentation, generative and diffusion models and explainability
  - In-depth exploration of Stable Diffusion
  - Formulation and implementation of Open-Vocabulary DAAM and Linear-DAAM
  - Design of framework to optimize semantic information in diffusion models
  - Experimental evaluation
  - Study of possible research directions
- Contributions
  - Preliminary study on the extraction of semantic ground truth in Stable Diffusion
  - Extension of method for explainability of Diffusion Models
  - Formulation of optimization approach to analyze semantic information learnt by a diffusion model

FUTURE WORK

- Enhancing Semantic Segmentation with DAAM
  - Synthetic Data Generation
  - Improve Open-Vocabulary existing models
- Advancing the Explainability of Text-to-Image LDMs
  - Study biases on LDMs
  - Study internal mechanisms of LDMs
- Exploring Multi-Modal Extensions of DAAM
  - DAAM extension for text-to-audio LDMs
  - Explainability of fMRI-to-Image models

Master in Deep Learning for Audio and Video Signal Processing

Thank you

Pablo Marcos Manchón

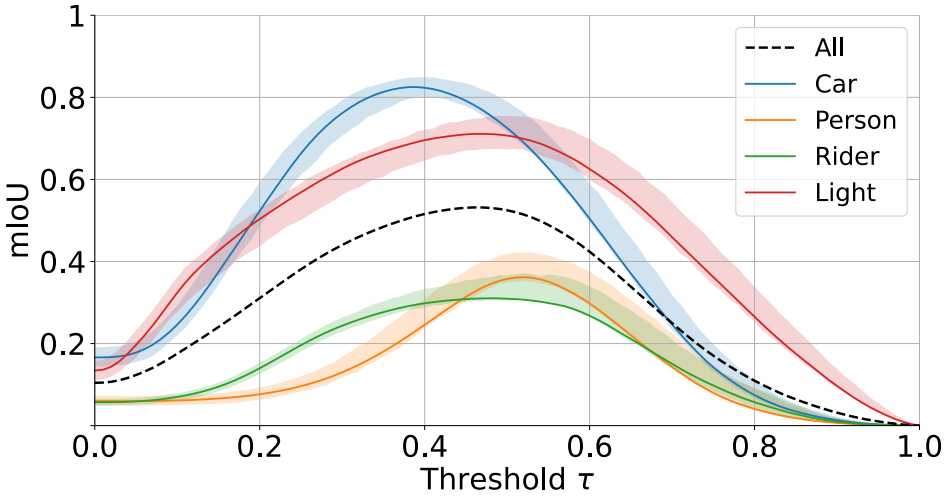
pablo.marcosm@estudiante.uam.es

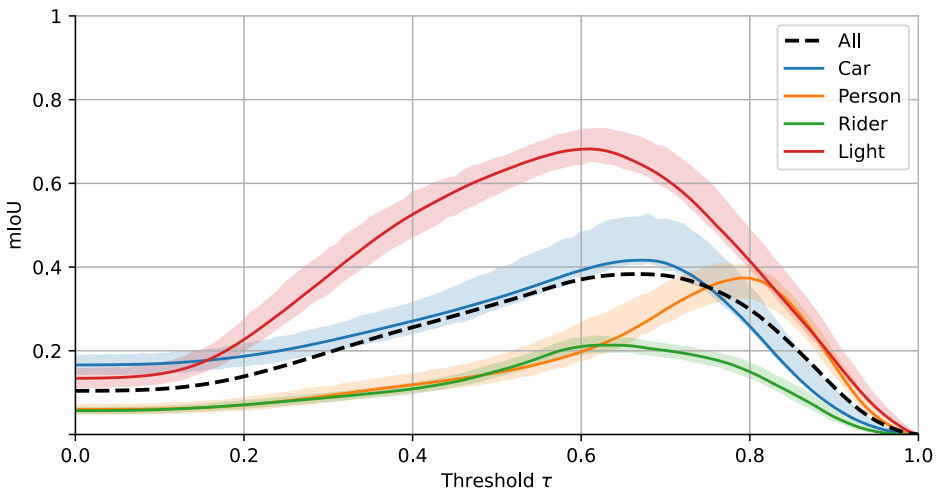
EXPERIMENT CURVES

DAAM

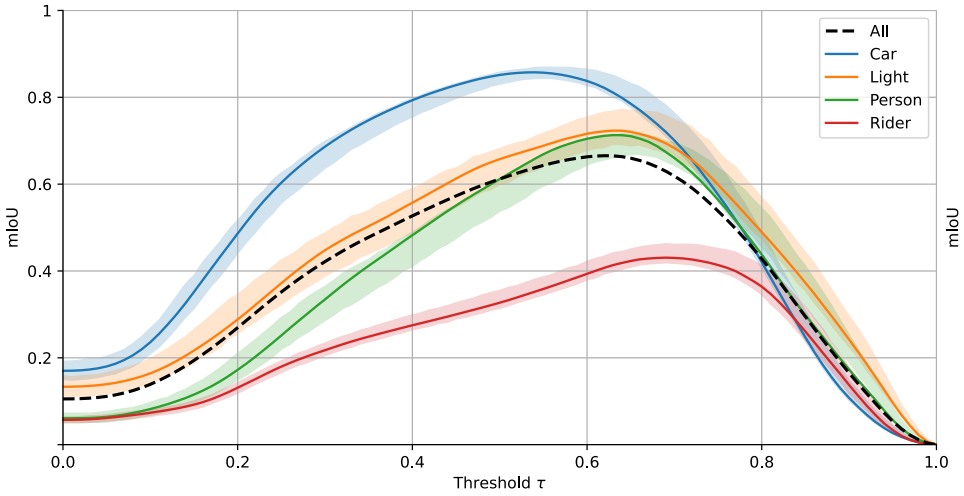
LDAAM

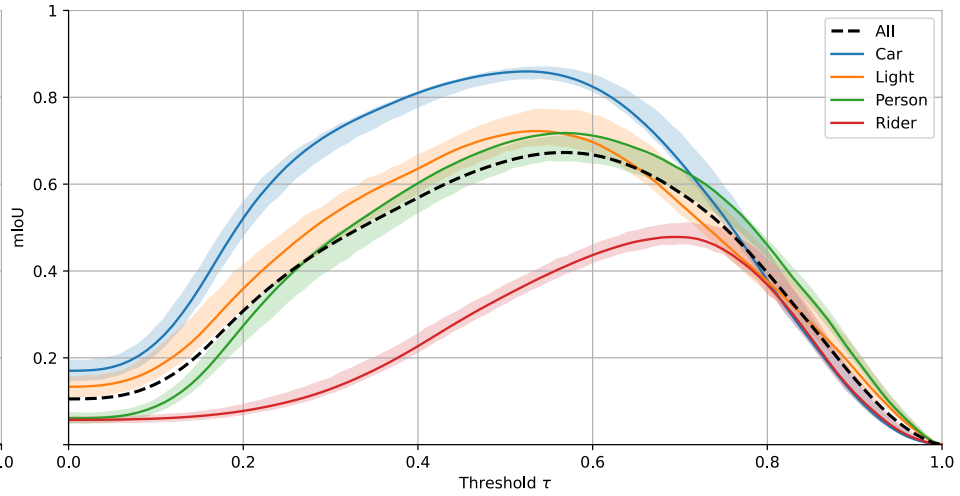
Non-Optimized





Optimized

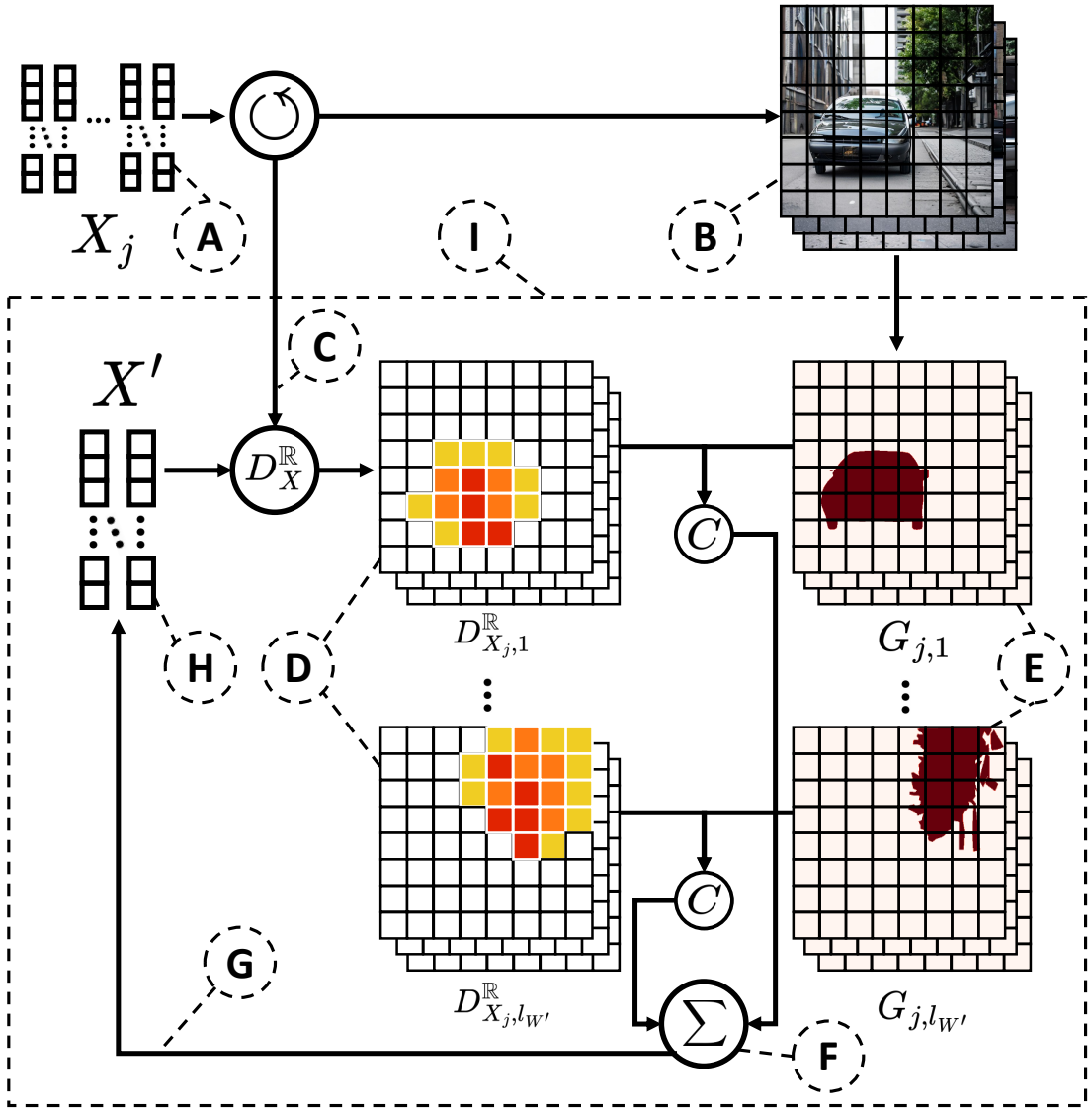




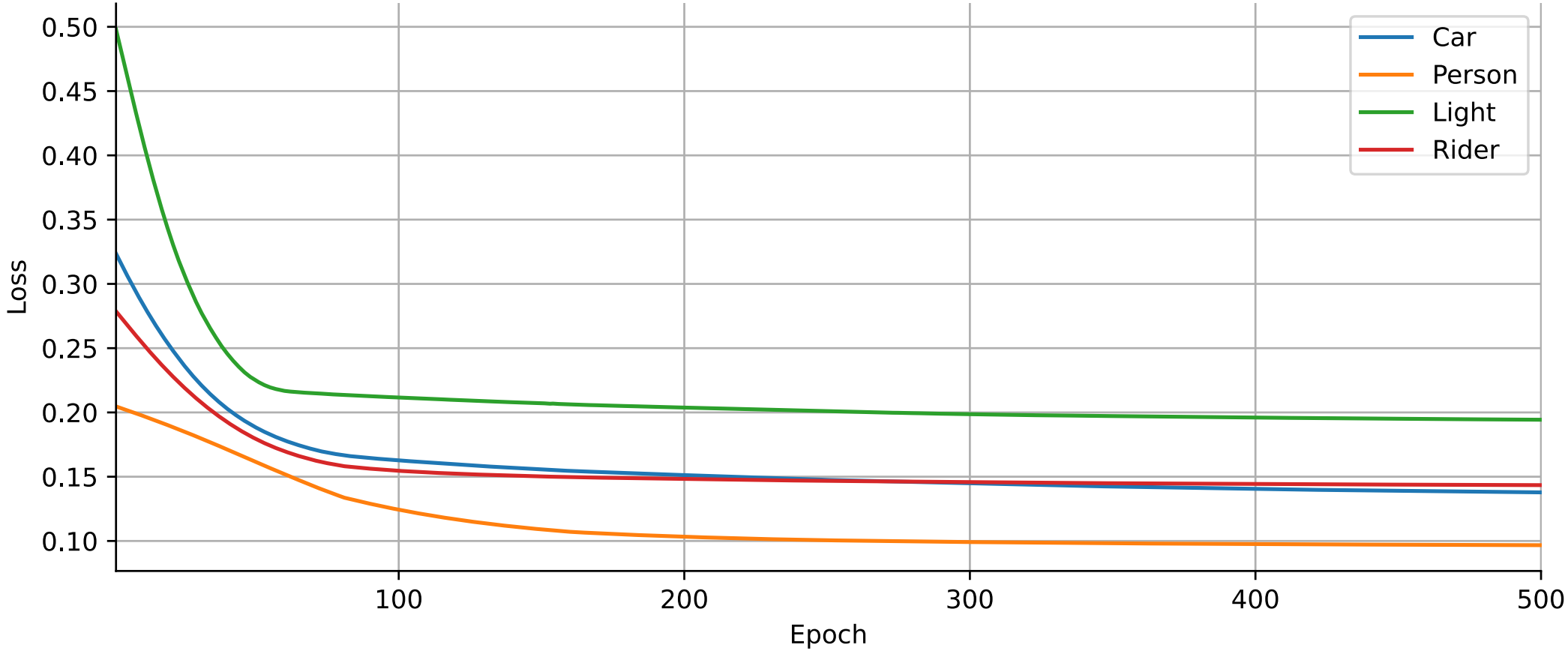
GENERAL OPEN-VOCABULARY DAAM OPTIMIZATION

- A. Text that generates image
- B. Images generated
- C. Query attentions used in OV-DAAM
- D. Heatmaps generated
- E. Ground truth annotated per token
- F. Compute loss
- G. Backpropagate loss gradients
- H. Optimize Key text prompt
- I. Optimization Loop

$$\hat{X}' = \arg\min_{X'} \sum_{j=1}^{N} \sum_{k=1}^{l_{W'}} C\left(D^{\mathbb{R}}_{X_j,k}\left(X'\right), G_{X_j,k}\right).$$



OPTIMIZATION PROCESS CURVES



Optimization of DAAM token for 2 images. Loss is computed as cross-entropy between segmentation masks and heatmaps. Approximate computation time: 0.3s/epoch (Mac Pro M1).