NYU TANDON SCHOOL OF ENGINEERING
Integrated Design & Media

# FlowGrad: Using Motion for Visual Sound Source Localization

Rajsuryan Singh, Pablo Zinemanas, Xavier Serra, Juan Pablo Bello, Magdalena Fuentes

MARL NYU

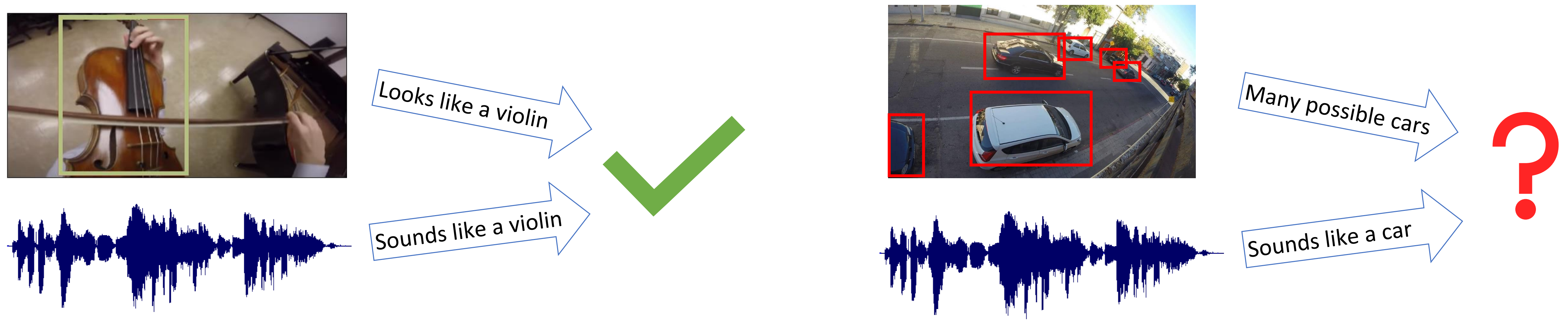
upf. Music Technology Group

Code Available

## State-of-the-art methods are inadequate for urban scenes

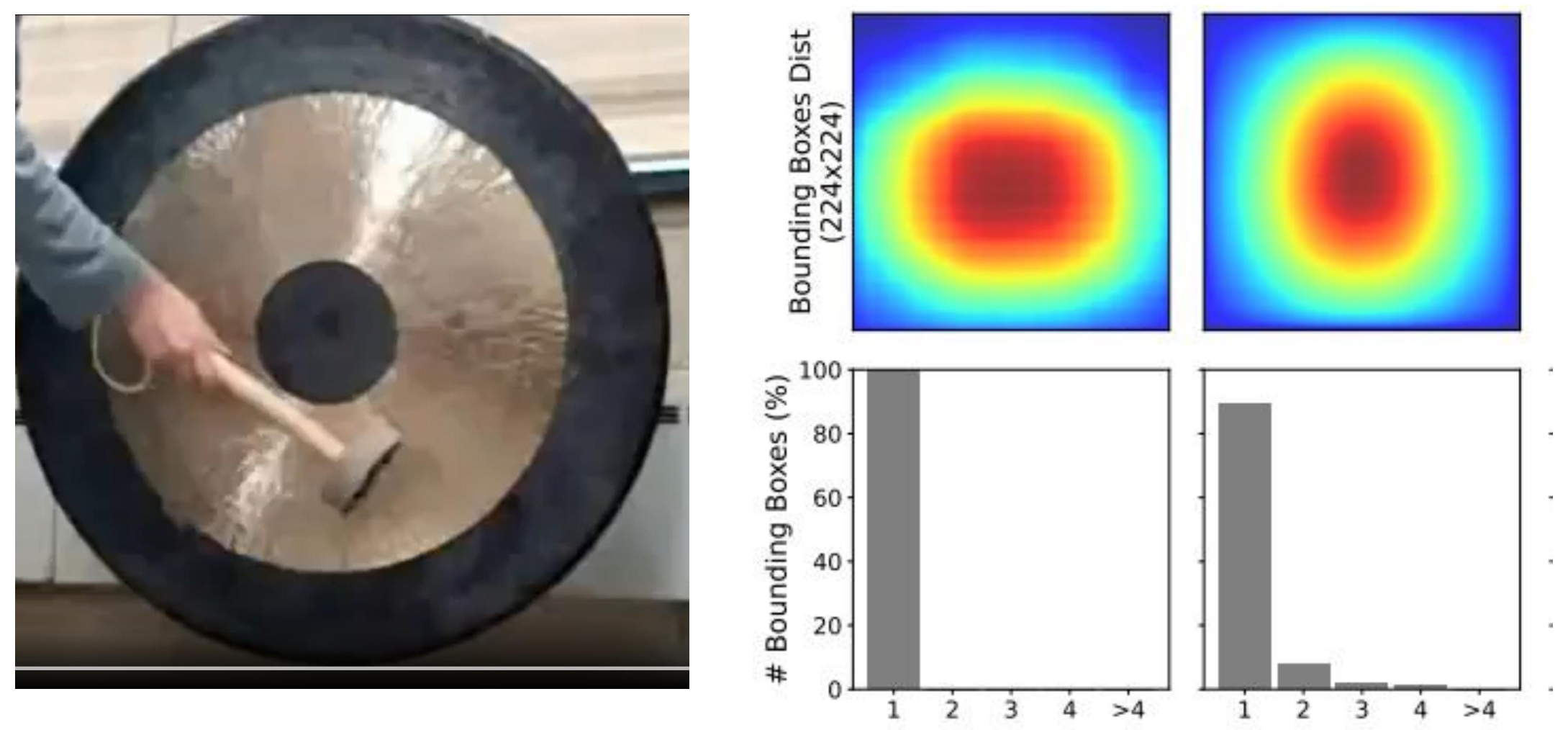
Purely semantic representations

Impossible to disambiguate between multiple potential sound sources

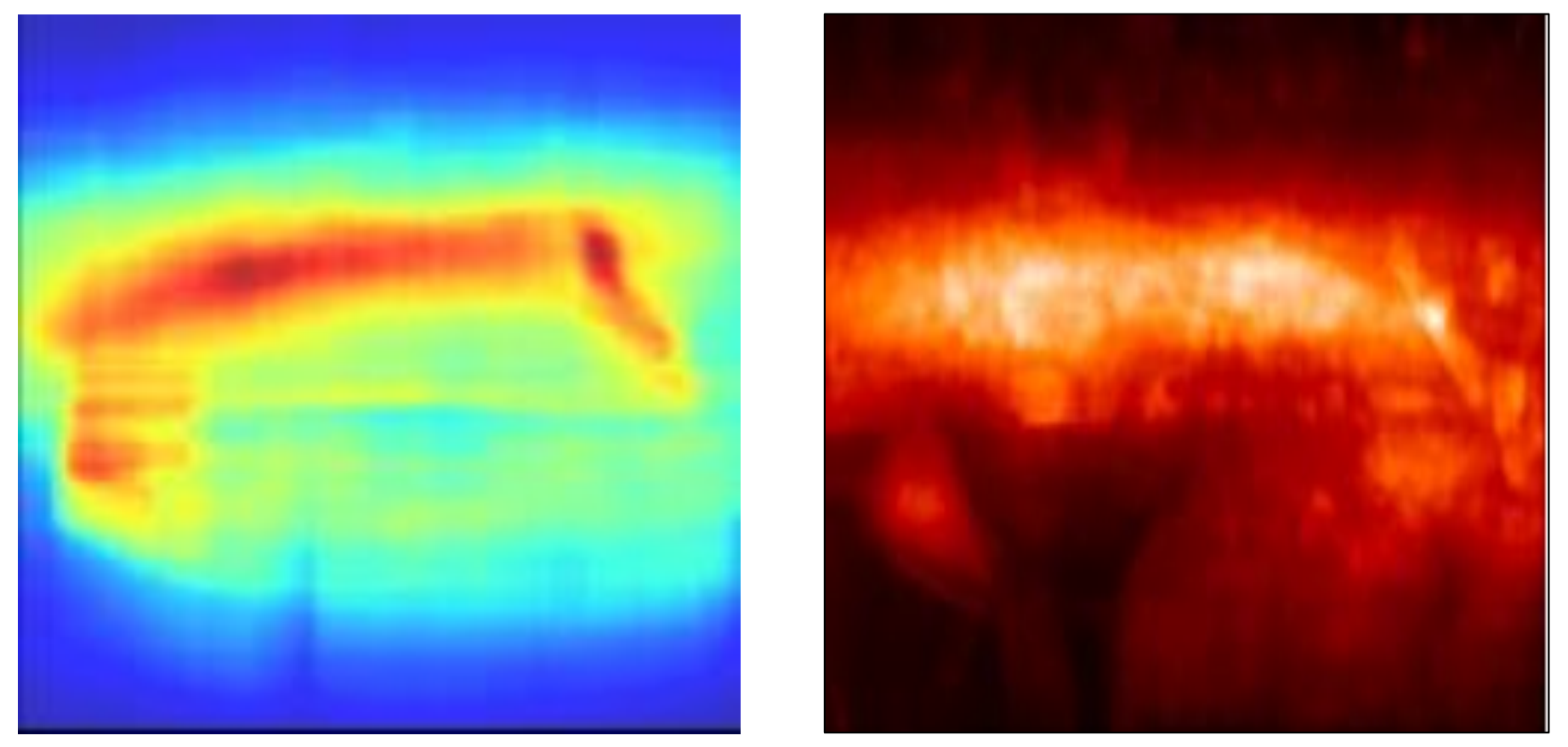


## All datasets are biased, urban scenes uniquely so

Single-centered-object bias in benchmarks



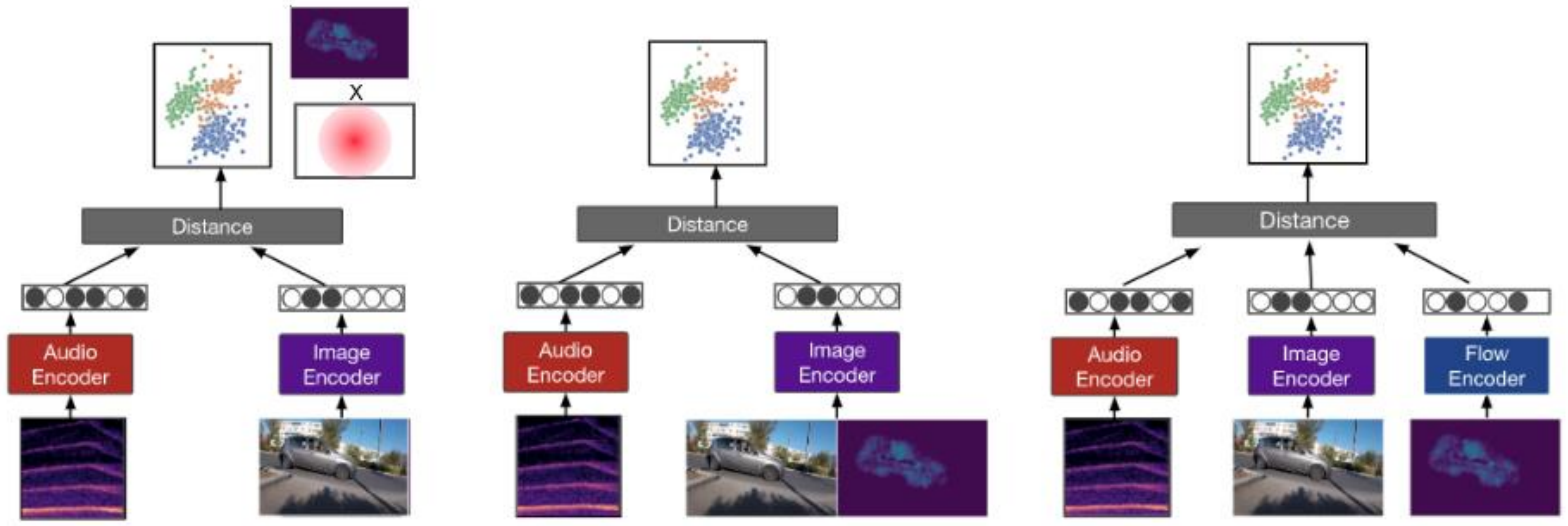
Motion bias in Urbansas



Bounding Boxes

Optical Flow

## Training



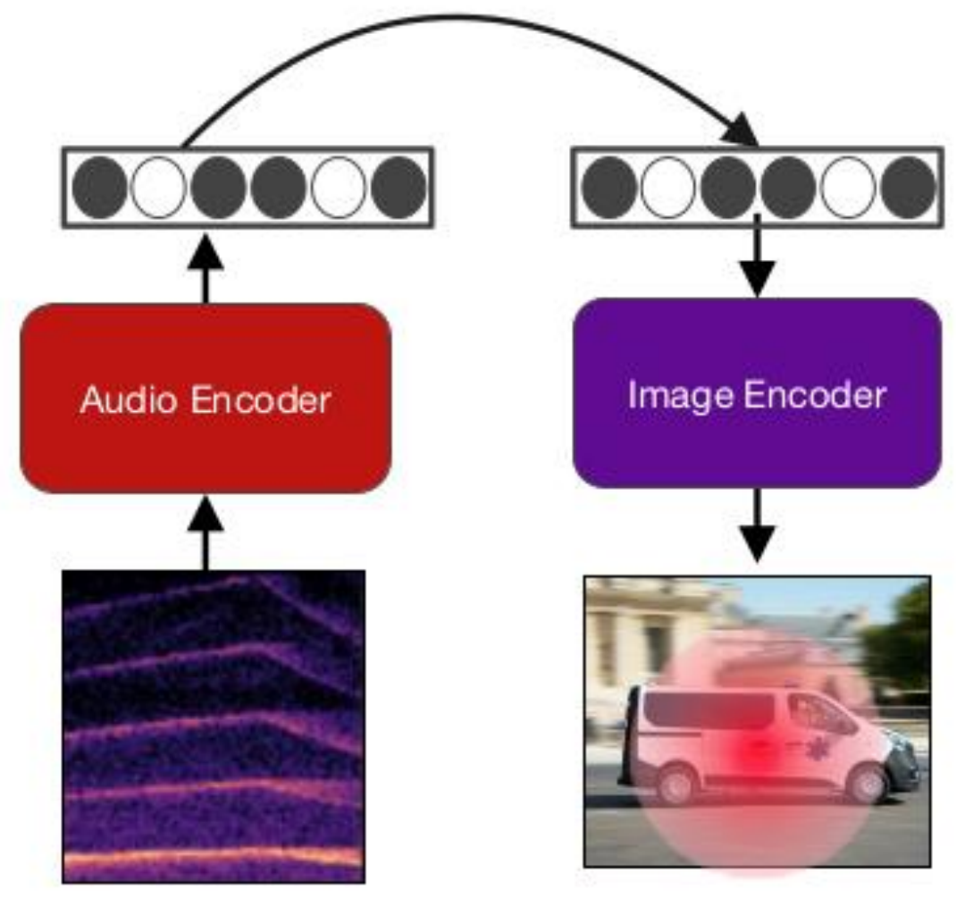
a) FlowGrad-H

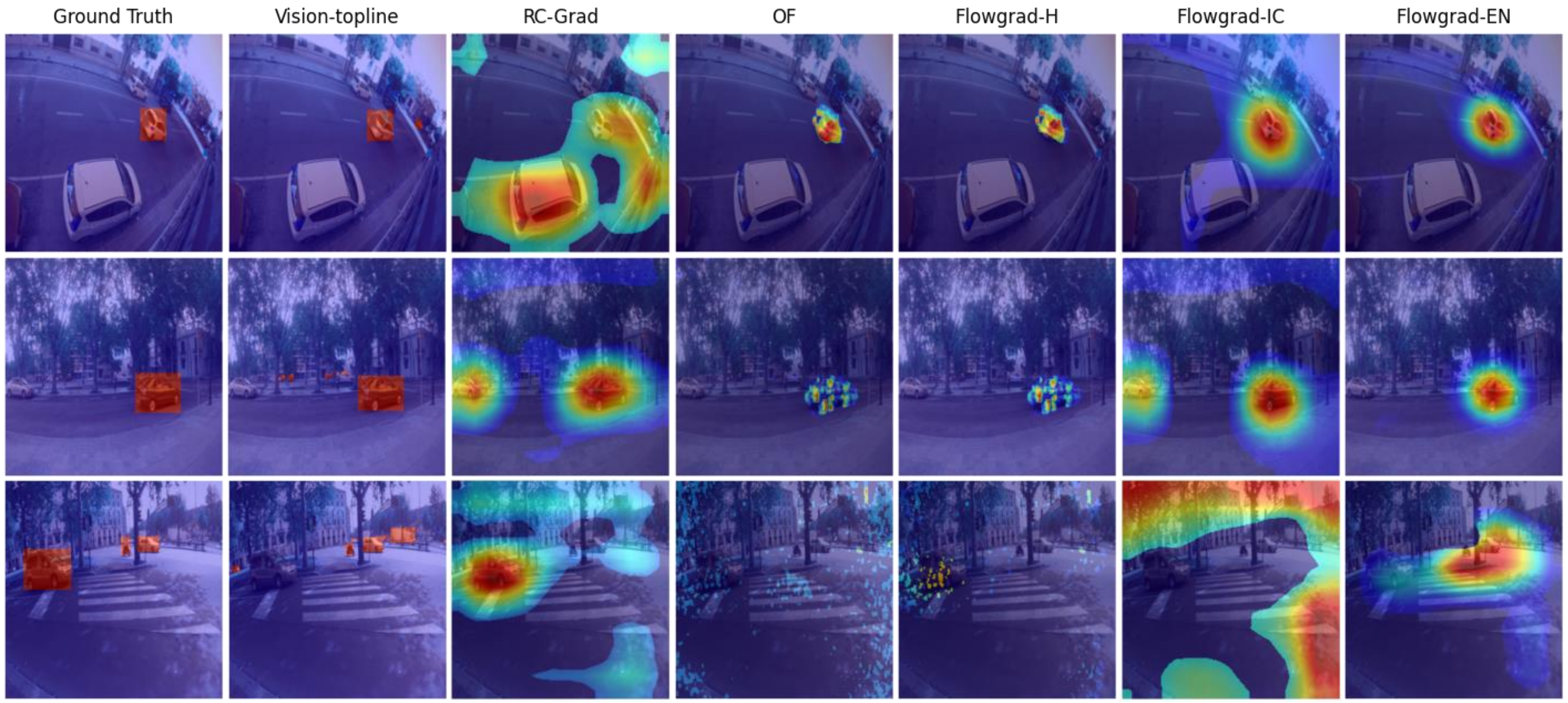
b) FlowGrad-IC

FlowGrad-EN

## Localization

**Grad-CAM** - Backpropagate through the vision subnetwork wrt the audio embedding





## Results

Heuristics can outperform learning based methods

| *model* | IoU | AUC |
|---|---|---|
| **Vision-topline** | **0.68** | **0.51** |
| Optical Flow | 0.33 | 0.23 |
| RCGrad | 0.16 | 0.13 |
| **FlowGrad-H** | **0.50** | **0.30** |
| FlowGrad-IC | 0.26 | 0.18 |
| FlowGrad-EN | 0.37 | 0.23 |