Separating the “Chirp” from the “Chat”: Self-supervised Visual Grounding of Sound and Language

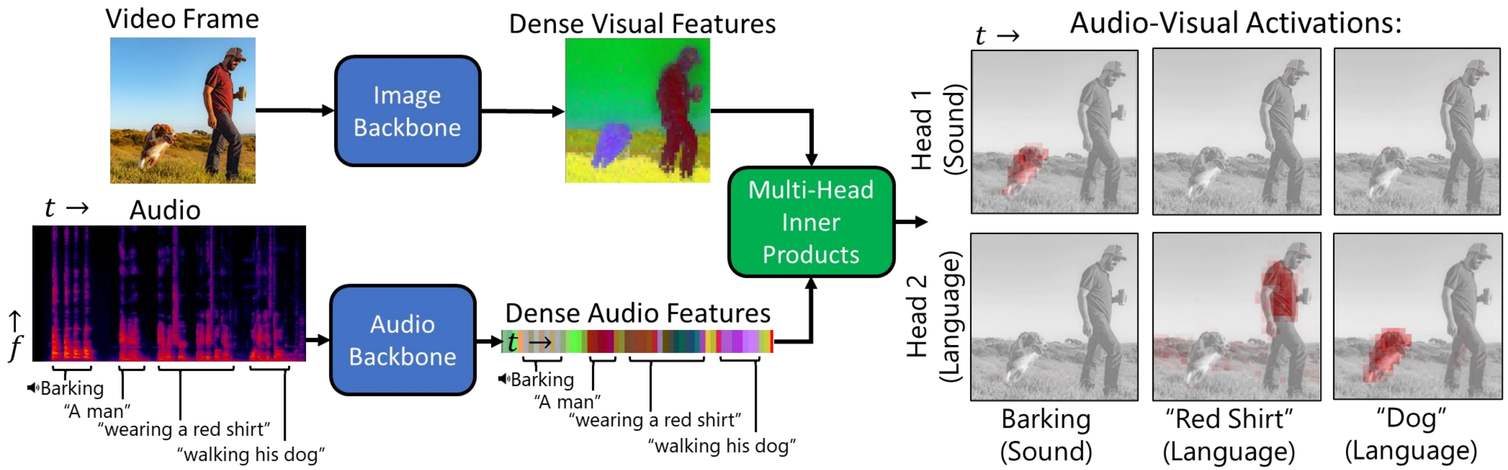
Mark Hamilton[1,2], Andrew Zisserman[3,4], John R. Hershey[3], William T. Freeman[1,3]

[1] MIT [2] Microsoft [3] Google [4] University of Oxford

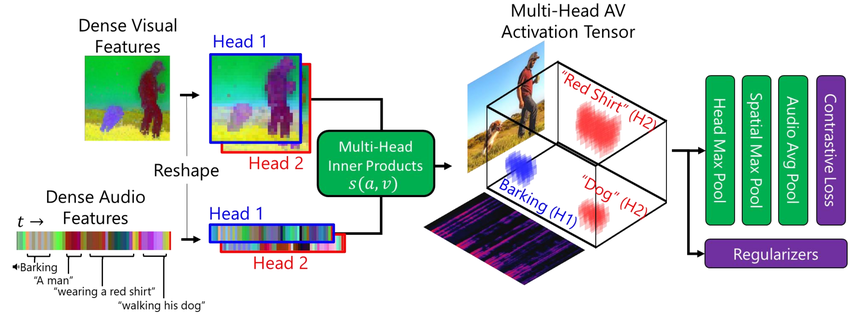
Demo Video iPad Here

aka.ms/denseav

Our model, DenseAV, learns the meaning of words and the location of sounds (visual grounding) **without supervision or text.**



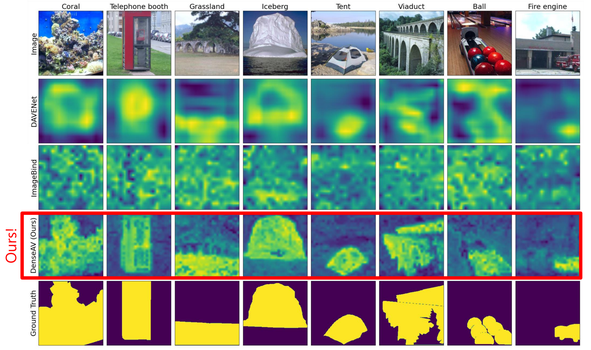
Multi-Head Dense Contrastive Learning

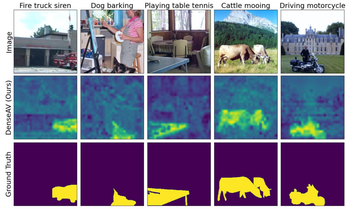


Speech and Sound Prompted Semantic Segmentation

Method	Speech Semseg. mAP	Speech Semseg. mIoU
DAVENet [23]	32.2%	26.3%
CAVMAE [18]	27.2%	19.9%
ImageBind [17]	20.2%	19.7%
Ours	**48.7%**	**36.8%**

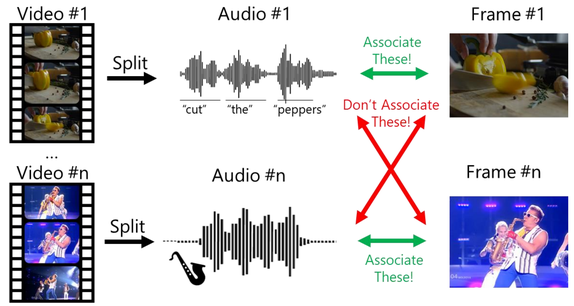
Method	Sound Semseg. mAP	Sound Semseg. mIoU
DAVENet [23]	16.8%	17.0%
CAVMAE [18]	26.0%	20.5%
ImageBind [17]	18.3%	18.1%
Ours	**32.4%**	**25.5%**





Two New High-Quality Datasets for Speech + Sound Prompted Segmentation

Audio-Visual Contrastive Learning



CLS Tokens are not Enough for Localization!