Model name	Input feature	Prediction type	Training target
Sound model	Log mel-spectrogram	Sound activation	Video-level instrument lab
Object model	RGB image	Object activation	Video-level instrument lab
Video tag as target (VT)	Dense optical flow	Action activation	Video-level instrument lab
Sound as target (ST)	Dense optical flow	Action activation	Sound activation
Object as target (OT)	Dense optical flow	Action activation	Object activation
Sound×Object as target (SOT)	Dense optical flow	Action activation	Sound activation×Object