Computer Vision applications often require a textual grounding module with precision, interpretability, and resilience to counterfactual inputs/queries. To achieve