Comparing Vision Token Interpretability Methods (Layer 0)

Input Image

Static NN (Input Embed.)	LogitLens (LM Head)	LN-Lens (Ours) (Contextual Embed.)
"Michele"	"aber"	**"ouver"** → "steeple white louvered windows." L8
"blinds"	"wat"	**"ouver"** → "siding of building with louvered ty..." L8
"rien"	"Hag"	**"ouver"** → "a dark green louver" L8
"White"	"marvin"	**"ouver"** → "kitchen window with black curtain a..." L8
"handwriting"	"wav"	**"-window"** → "gabled roof on multi-windowed brow..." L8