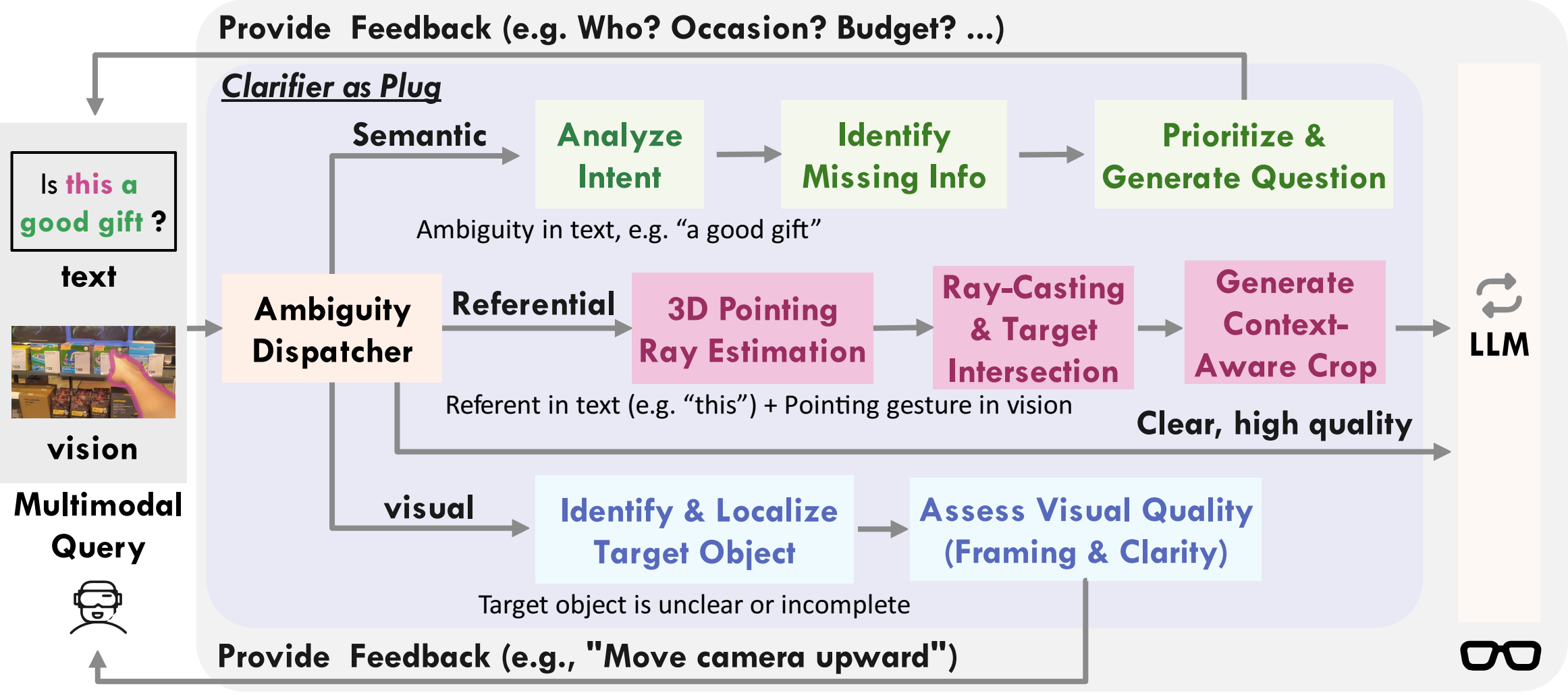
Provide Feedback (e.g. Who? Occasion? Budget? ...)
Clarifier as Plug
Is this a good gift ?
text
vision
Multimodal Query
Ambiguity Dispatcher
Semantic
Analyze Intent
Identify Missing Info
Prioritize & Generate Question
Ambiguity in text, e.g. “a good gift”
Referential
3D Pointing Ray Estimation
Ray-Casting & Target Intersection
Generate Context-Aware Crop
Referent in text (e.g. “this”) + Pointing gesture in vision
Clear, high quality
visual
Identify & Localize Target Object
Assess Visual Quality (Framing & Clarity)
Target object is unclear or incomplete
LLM
Provide Feedback (e.g., "Move camera upward")