Grounded Robot Perception and Auditability in Multimodal Foundation Models

Name: Atuhurra Jesse
Laboratory' s name: Natural Language Processing
Supervisor' s name: Professor Taro Watanabe

Abstract (should be within 1st page)

Vision-language models and large language models increasingly power interactive systems and multimodal agents, yet they often fail in the same underlying way: they produce fluent outputs without reliable grounding, and their intermediate reasoning is difficult to audit under noisy, multilingual, and deployment-realistic conditions. This dissertation establishes a unified, evidence-centered framework for grounding and auditability across two complementary domains: embodied robot perception in Japanese human-robot interaction, and multimodal long-text grounding across typologically diverse languages. First, we address attribute-grounded robot scene understanding with vision-language models using J-ORA, an evaluation suite of Japanese dialogue-driven scenes requiring object identification, reference resolution, affordance understanding, and next-action prediction. Drawing on our extended task-decomposed analysis for robot perception, we show that structuring perceptual commitments into explicit object attributes improves multimodal grounding under realistic clutter, supports affordance-aware reasoning, and exposes systematic robustness gaps between open and proprietary models. Second, we introduce VLURes, a multilingual benchmark of 4,000 web-curated image and long-text pairs across English, Japanese, Swahili, and Urdu. By separating core visual perception from bimodal reasoning and introducing the Unrelatedness task, we show that article-length distractors induce severe context over-trust and demonstrate that textual multilinguality in base models does not ensure vision-language grounding readiness. Taken together, these projects support a single thesis claim: robust grounding requires explicit intermediate structure, such as visual attributes and negative alignment boundaries, paired with evaluation protocols that measure not only task accuracy but also support validity and auditability. We synthesize shared failure modes across physical and web-scale multimodal domains into a unified failure taxonomy, providing practical methods and empirical evidence that grounded, auditable foundation-model systems are achievable even in low-resource, long-context, and safety-relevant settings.