HW4 - “Analyst Track” How Multimodality Changes Agents: Analysis Report

Course: MAS.665 AI Studio: AI Agents and Agentic Web **Author:** Rong Wu
Date: October 08, 2025

---

1. Planner: How Multimodality Affects Action Decisions and Reasoning

The Planning Challenge

Text-only planning is straightforward. You receive text input, process it, and generate text output. Multiple modalities make everything complex.

When I added ElevenLabs speech synthesis, I had to rethink how the system decides what to do next. The planner now makes several new decisions:

Modality Selection: The system must choose which modality to use for each interaction. Speech works better for explanations. Text works better for detailed feedback. This wasn’t a decision I had to make before.

Cross-Modal Reasoning: Students perform differently across modalities. Good at reading but poor at listening. The planner must coordinate responses across all modalities while maintaining consistency. This is harder than handling each modality independently. I added an option to give English scores (“A1-C2”) based on both OCR + Text or Speech + Text. A combination yields a score with extra weight given to a predefined modality.

Resource Management: Different modalities have different computational costs. OCR processing for images is expensive. Speech synthesis requires real-time processing. Text generation needs careful prompt engineering. The planner must allocate resources efficiently.

Temporal Coordination: With multiple modalities, timing becomes critical. The system must coordinate when to switch between speech and text. It must decide when to process images. It must handle delays in different processing pipelines.

What I Learned

Multimodality transforms planning from a simple pipeline into a complex orchestration system. Instead of linear decision-making, the planner now needs to manage multiple parallel processes. It must handle modality-specific failures. It must maintain consistency across different interaction modes.

2. Retriever: How Multimodality Affects Knowledge and Data Retrieval

The Retrieval Challenge

Traditional text retrieval uses semantic similarity. You embed the query and find similar documents. Multiple modalities make retrieval much more complex.

Cross-Modal Indexing: When I added OCR to my writing module, I realized images containing text need multiple indexing ways. The system must maintain both the original image and the extracted text. It must handle cases where OCR fails completely. Failing OCR slightly lowers the score for 'readability'.

Modality-Specific Strategies: Different modalities require different retrieval approaches. Text queries use semantic search. Audio content needs feature extraction and matching. Images require visual similarity combined with textual metadata. There's no one-size-fits-all approach. I used more powerful models (Gemini 2.5 vs Gemini 2.0) for tasks considered more difficult: image > speech > text.

Quality Degradation: Multimodal retrieval introduces new failure modes. OCR errors can corrupt text retrieval. Audio transcription mistakes affect speech retrieval. Cross-modal mismatches occur when image content doesn't match its textual description.

Efficiency Concerns: Maintaining multiple embedding spaces is computationally expensive. Computing cross-modal similarities is expensive. The system must balance retrieval quality with processing speed.

What I Learned

Multimodal retrieval isn't just about adding new search capabilities. It's about fundamentally rethinking how information is organized and accessed. The system needs to maintain multiple types of indexes. It must handle quality issues gracefully. It must optimize for the specific requirements of each modality.

3. Guardrails: How Multimodality Affects Safety, Moderation, and Constraints

The Safety Challenge

Cyber Space: There are new AI attacks that use hidden prompts in images to steal user data. I am not going into that here, but there are interesting reads.

Cross-Modal Violations: Safety violations can span multiple modalities. An image might be inappropriate while its text description is benign. Audio content could be harmful while its transcription appears neutral. The system must detect violations that emerge when modalities are considered together, not only alone.

Consistency Challenges: Ensuring consistent safety policies across all modalities is difficult. A policy that works for text might not apply to images. Audio content has different cultural and contextual considerations.

What I Learned

Multimodal safety requires a completely different approach than single-modality systems. Instead of simple content filtering, the system needs multi-layered safety mechanisms. It must detect violations across modalities. It must handle privacy concerns appropriately. It must maintain consistent policies while accounting for modality-specific requirements.

Conclusion

Each new modality doesn't just add capabilities. It fundamentally changes how the entire system operates. Successful multimodal agents require holistic architectural redesigns rather than simple modality additions.

Multimodality introduces both opportunities and challenges. It requires sophisticated orchestration across planning, retrieval, and safety dimensions. While it enables richer, more natural interactions, it also introduces significant complexity that must be carefully managed. The future of AI agents will depend on our ability to design systems that can effectively coordinate multiple modalities while maintaining consistency, safety, and performance.