Network Working Group
Internet-Draft
Intended status: Informational
Expires: 8 January 2026

G. Xie
Y. Li
CNIC, CAS

7 July 2025

# Multimodal Management Requirements for AI Agent Protocols
## draft-xie-ai-agent-multimodal-00

## Abstract

This document specifies the Multimodal requirements for Agent-to-Agent Protocol, which enables autonomous agents to establish multi-channel communication sessions, negotiate heterogeneous data capabilities (e.g., text, file, real-time audio/video streams, sensor streams), and exchange synchronized multimodal content with adaptive QoS policies.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 8 January 2026.

Table of Contents



# 1. Introduction

## 1.1. Purpose

This document articulates the technical imperative for multimodal interaction to be natively supported in the protocols defining and standardizing interoperability among Artificial Intelligence Agents (AI Agents). The core rationale stems from the evolution of modern Large Language Models (LLMs) into Multimodal Models capable of processing, generating, and understanding multiple media types. Consequently, establishing multimodal channels to transmit multimodal media must be supported.

## 1.2. Terminology

Multimodal:
Multimodal is a technology that integrates two or more different types of information modes (such as visual, text, and speech) to improve data processing and task execution capabilities. Its core lies in imitating the cognitive style of human multi-sensory collaboration, and enhancing the integrity and interactive experience of information by fusing data of different modes.

# 2. Use Cases

A typical use case in agent-agent communication using multimodal media:

A Housekeeping Robot Agent sends a task to a Monitoring Robot Agent to detect the incidents. The two Agents negotiate the necessary multimodal media channels and establish the session. Once the

Monitoring Robot Agent detects a glass-breaking incident in the kitchen, it transmits audio and video stream to the Housekeeping Robot Agent, simultaneously it generates and sends a text alert ("CRITICAL: Glass break at 2025-10-05T14:30:15Z") to the Housekeeping Robot Agent through the established multimodal session.

## 3. Necessity

With the rapid development of LLMs (large language models) technologies, LLMs have gradually developed from supporting single modal such as text to supporting multiple modals such as text, pictures, and video clips and their combinations. In particular, LLMs supporting real-time audio and video streams have emerged recently. The capabilities of LLMs are increasingly rich and perfect.

Agent often needs to understand multiple modal data, such as environment data, context data, audio, video, etc., to better understand and execute tasks. Therefore, various multimodal data needs to be transmitted between Agents.

Multimodal interaction capabilities supported by mainstream Agent communication protocols in the industry are shown as follows:

* A2A protocol [A2A]: Support TextPart, DataPart and FilePart carried by the part parameters in tasks/send or tasks/subscribe message. Only simple text, file, and data formats are supported.

* ANP protocol [ANP]: Only natural language (e.g., YAML and JSON-RPC format) which defined in Agent Description is supported.

* Agntcy protocol [Agntcy]: Only data with JSON format is supported.

Therefore, the general Agent communication protocol should support rich multimodal interaction including text, file, real-time audio/ video stream, etc.

## 4. Protocol Requirements

### 4.1. Multimodal Media Channel Establishment (Signaling)

* Multimodal Media Negotiation

When agents need to transmit multimodal media, especially real-time audio and video streams, a dedicated media channel is required to transmit these streams. Meanwhile, the audio and video codecs supported by different agents may be vary. Therefore, the audio and video codecs MUST be negotiated between agents before transmission.

### 4.2. Multimodal Media Transmission (Media Handling)

* Multi-stream Multiplexing

There are many types of multimodal data. The transmission requirements of different multimodal data are varied, which requires transmission using different stream.

To reduce IP port resource overhead caused by flow connections, a multi-stream multiplexing capability needs to be supported, and different transmission priorities can be set for different flows.

* Multimodal Synchronization and Collaboration

In some use cases, multimodal data transmitted between agents needs to be synchronized. As shown in the example use case, when the housekeeping robot Agent arranges the monitoring robot Agent to perform environment monitoring, the audio stream, video stream, and text events collected by the monitoring robot Agent need to be synchronized to help the housekeeping robot better understand the monitoring result.

In other use cases, multimodal data may need to coordinate transmission policies. For example, when audio and video streams are transmitted at the same time, when video frame freezing occurs due to a decrease in connection bandwidth, the video resolution and bitrate need to be automatically reduced to ensure audio stream transmission quality.

## 5. Conclusions

Multimodal interaction constitutes a critical function for multi-agent collaboration. This document discusses the necessity of introducing multimodal interaction to address Agent collaboration. Consequently, it analyzes the requirements imposed by multimodal interaction on AI Agent protocol design, specifically concerning multimodal media channel establishment and multimodal media transmission.

## 6. IANA Considerations

This memo includes no request to IANA.

## 7. Security Considerations

This document should not affect the security of the Internet.

## 8. References

### 8.1. Informative References

[A2A] Google LLC, "Agent2Agent (A2A) Protocol Specifications", 2025, <https://github.com/a2aproject/A2A/blob/main/specification>.

[ANP] ANP community, "AgentNetworkProtocol(ANP)", 2025, <https://github.com/agent-network-protocol/AgentNetworkProtocol/tree/main>.

[Agntcy] IoA community, "Agent Connect Protocol Specification", 2025, <https://github.com/agntcy/acp-spec/tree/main>.

## Authors' Addresses

Gaogang Xie
CNIC, CAS
Beijing
100083
China
Email: xie@cnic.cn

Yanbiao Li
CNIC, CAS
Beijing
100083
China
Email: lybmath@cnic.cn