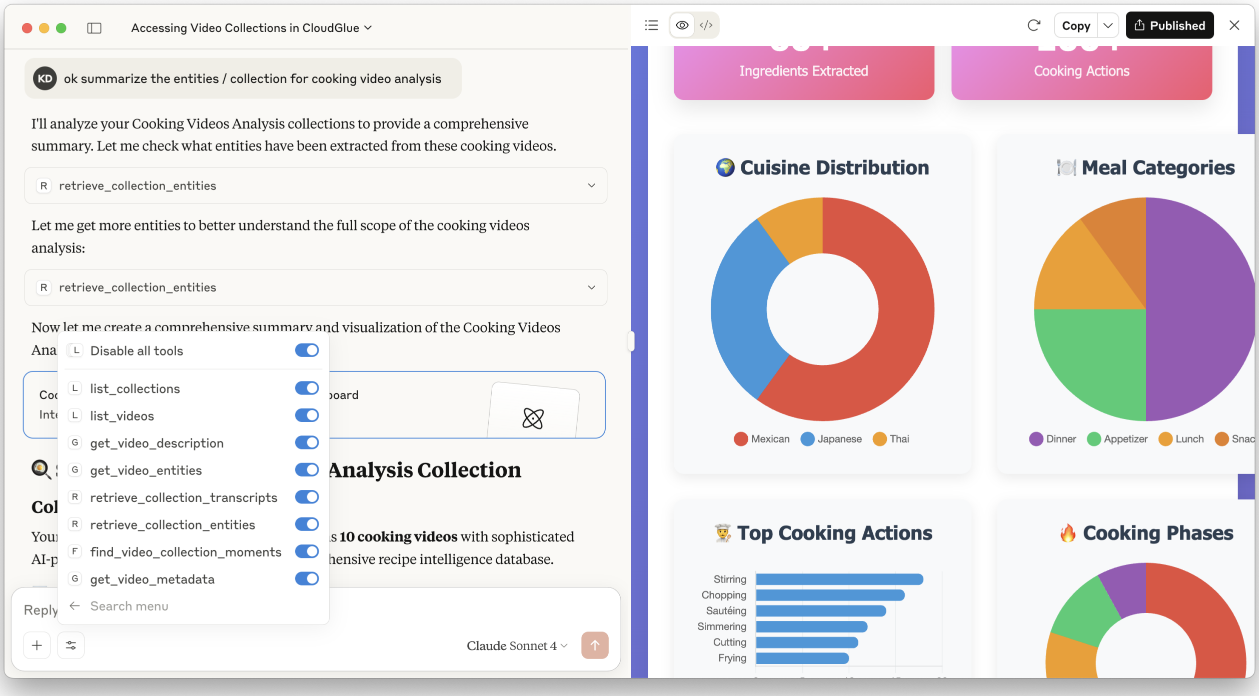
VideoMCP: MCP-Enabled Video Intelligence for Multimodal Agent Reasoning and Enterprise Video Collection Analysis

Kevin Dela Rosa
kdr@cloudglue.dev
Cloudglue
San Francisco, California, USA

Amy Xiao
amy@cloudglue.dev
Cloudglue
San Francisco, California, USA

Matt Pua
pua@cloudglue.dev
Cloudglue
San Francisco, California, USA



KD ok summarize the entities / collection for cooking video analysis

I'll analyze your Cooking Videos Analysis collections to provide a comprehensive summary. Let me check what entities have been extracted from these cooking videos.

R retrieve_collection_entities

Let me get more entities to better understand the full scope of the cooking videos analysis:

R retrieve_collection_entities

Now let me create a comprehensive summary and visualization of the Cooking Videos Ana

Coo ... oard
Inte

Analysis Collection

Col

You ...s **10 cooking videos** with sophisticated
AI-p ...hensive recipe intelligence database.

Figure 1: High-level visualization: VideoMCP-generated knowledge summary from a collection of enterprise videos.

ABSTRACT

Cloudglue VideoMCP enables enterprise agents to extract actionable insights from organizational video archives by integrating automatic speech recognition (ASR), optical character recognition (OCR), and visual scene analysis tools via the Model Context Protocol (MCP). Leveraging an LLM-based agent, VideoMCP demonstrates how protocol-driven multimodal integration delivers accurate, evidence-grounded answers, summarization, and search over sales calls, product demos, and training sessions. We describe the system's architecture, core functionalities, user interface, and results from a comparative evaluation with an ASR-only baseline.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Visual content-based indexing and retrieval**.

KEYWORDS

Retrieval Augmented Generation, Cross-modal Retrieval, Multimodal Retrieval, Agentic AI Frameworks, Enterprise Search Systems

KDD AAE '25, August 04, 2025, Toronto, ON
2025. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of 2nd Workshop on Agentic AI for Enterprise @ KDD (KDD AAE '25)*.

ACM Reference Format:
Kevin Dela Rosa, Amy Xiao, and Matt Pua. 2025. VideoMCP: MCP-Enabled Video Intelligence for Multimodal Agent Reasoning and Enterprise Video Collection Analysis. In *Proceedings of 2nd Workshop on Agentic AI for Enterprise @ KDD (KDD AAE '25)*. ACM, New York, NY, USA, 4 pages.

1 MOTIVATION, CONTEXT, AND USE CASES

1.1 The Knowledge Trap in Enterprise Video

Enterprises generate hundreds of hours of video content per month, but most valuable information remains buried in recordings. Traditional approaches—automatic transcription or manual review—fail to capture visual context (e.g., slide text, UI walkthroughs) and the interplay between spoken and visual signals.

1.2 Enabling Agentic Multimodal Intelligence

VideoMCP addresses this by exposing robust video intelligence tools through the standardized Model Context Protocol (MCP). LLM agents can invoke specialized modules—transcription, OCR, visual captioning, structured extraction, and retrieval—on demand, enabling queries about what was said, shown, and when.

Use Cases:

- *Sales Enablement:* Extract competitor mentions, objections, and product demo highlights across all sales calls, grounded in both speech and visuals.
- *Compliance:* Verify that regulated topics are covered in speech and on screen, with timestamped citations.
- *Training:* Summarize when and how key features or workflows are explained, linking speech to interface actions.

1.3 Design and Deployment Considerations

VideoMCP uses MCP for secure, extensible agent integration, ensuring plug-and-play deployment and privacy for sensitive data. New tools or modules can be added without disrupting agent workflows.

1.4 Limitations

VideoMCP's performance depends on upstream transcription, OCR, and captioning accuracy. Complex visual reasoning (e.g., interpreting charts, diagrams, or handwriting) remains an open challenge. Our current agent operates in a single-session, single-user setup; collaborative, multi-agent, and live streaming support are left to future work.

2 SYSTEM ARCHITECTURE

VideoMCP connects an LLM agent to video intelligence tools via MCP, inspired by modular agent frameworks [3, 4]:

- **Transcribe Module:** Captures speech, slide text, and visual captions.
- **Extraction Module:** Structures video content via vision-language models.
- **Retrieve Module:** Fetches transcripts, extractions, or metadata on demand.
- **Find Module:** Orchestrates results to surface moments relevant to user queries.
- **VideoMCP Server:** Exposes all tools as MCP services.
- **LLM Agent:** Orchestrates tool use for complex queries.
- **MCP Client:** Enterprise chat interface interacting via MCP.

2.1 Design Rationale

Each module addresses a unique challenge—capturing rich signals, structuring knowledge, enabling fast retrieval, and agentic reasoning. The MCP-driven separation supports modular deployment and easy extensibility.

3 FUNCTIONALITIES & USER INTERFACE

VideoMCP enables:

- *Multimodal QA*: Evidence-grounded answers using speech and visual signals [2].
- *Summarization*: Merges spoken and visual highlights.
- *Contextual Search*: Finds keywords in transcripts or scene descriptions.
- *Interactive Drill-down*: Supports follow-up queries and multi-turn search.

When integrated with a user's existing AI chat client, for example in Figure 3, we show Claude Desktop integrating with VideoMCP as an MCP server. The agent leverages VideoMCP tools to provide grounded answers using both speech and visual text.

Also refer to Figure 1, which shows a summary of information extracted across a collection of videos using VideoMCP, visualized for rapid knowledge transfer.

4 ENTERPRISE SCENARIOS AND DEPLOYMENT

Sales: Managers used to review calls by hand; now, agents instantly extract objections, competitor mentions, and demo reactions with citations.
Compliance: Teams can confirm all regulated topics were covered, with traceable evidence.
Training: Sessions are summarized with links between spoken instructions and UI actions.

VideoMCP can be deployed on-premises or as a cloud microservice; it is agent-agnostic and easy to extend.

5 BROADER IMPACT AND EXPLAINABILITY

VideoMCP generalizes to support customer support, operations, or legal teams. The MCP-based design allows new tools and workflows to be added quickly. Grounded, evidence-cited answers improve trust and make compliance review and knowledge transfer more efficient. Pilot feedback shows confidence gains with explainable agentic outputs.

Table 1: Example Query Types Supported by VideoMCP

Query	Signals Used
Who raised pricing?	Speech
Metric on slide 4?	Visual
Customer reaction?	Cross-modal
Competitor mentions?	Speech
Compliance slides?	Visual
Feedback summary?	Speech + Visual

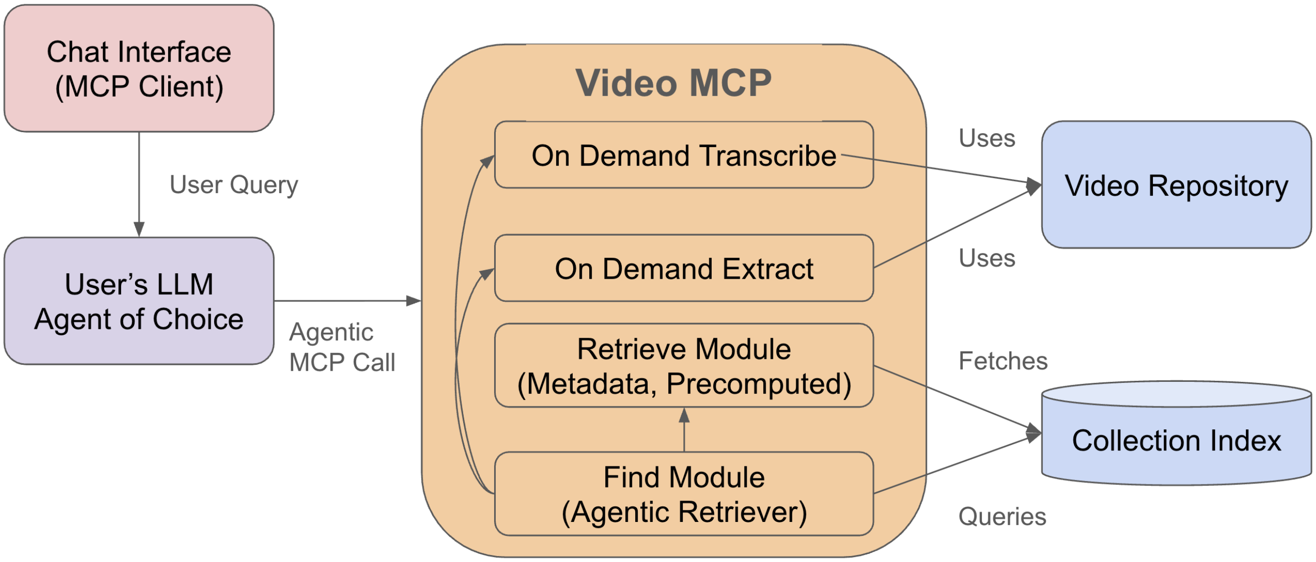


Figure 2: VideoMCP architecture: ASR, OCR, and vision tools are MCP-exposed and orchestrated by an LLM agent, interfacing with users' preferred chat AI.

6 EVALUATION

VideoMCP and an ASR-only baseline were evaluated on 15 queries (5 per modality) over a 2 hour video collection. Queries were semi-automatically generated from sales and product demo recordings, and then manually categorized as speech, visual, or cross-modal. Human annotators scored correctness (1/0.5/0) and grounding (evidence cited=1/0). Evidence was defined as a timestamped transcript excerpt or screenshot from the relevant video segment.

Table 2: Evaluation Results

Modality	ASR (%)	VideoMCP (%)
Speech	80	90
Visual	20	100
Cross-modal	60	100
Overall	53.3	96.7

VideoMCP matches ASR on speech queries and dramatically outperforms on visual and cross-modal queries, also providing evidence in nearly all cases.

7 DISCUSSION AND FUTURE DIRECTIONS

While VideoMCP provides strong gains in multimodal QA, there remain open challenges. Our evaluation relies on a fixed set of human-annotated queries and a modest dataset size; future studies will target larger, more diverse enterprise collections and real-time applications.

Limitations: VideoMCP's performance depends on upstream transcription, OCR, and captioning accuracy. Complex visual reasoning (e.g., interpreting charts, diagrams, or handwriting) remains an open challenge. Our current agent operates in a single-session, single-user setup; collaborative, multi-agent, and live streaming support are left to future work.

Broader Impact: VideoMCP has the potential to empower non-technical users across sales, compliance, and training. Early adopters report reduced manual review time and higher confidence in findings due to evidence citation. However, further human factors studies are required to assess trust, transparency, and adoption barriers.

Future Work:

- **Scalability:** Deploy VideoMCP on large-scale, multi-tenant enterprise datasets.
- **Real-Time Analysis:** Extend to live meeting/video call settings with low latency.
- **Explainable Reasoning:** Integrate richer explanations (e.g., visual attention heatmaps).
- **Multi-Agent Collaboration:** Support multiple agents for complex retrieval and synthesis.
- **Workflow Integration:** Connect outputs to downstream reporting, knowledge graph, or RAG pipelines.

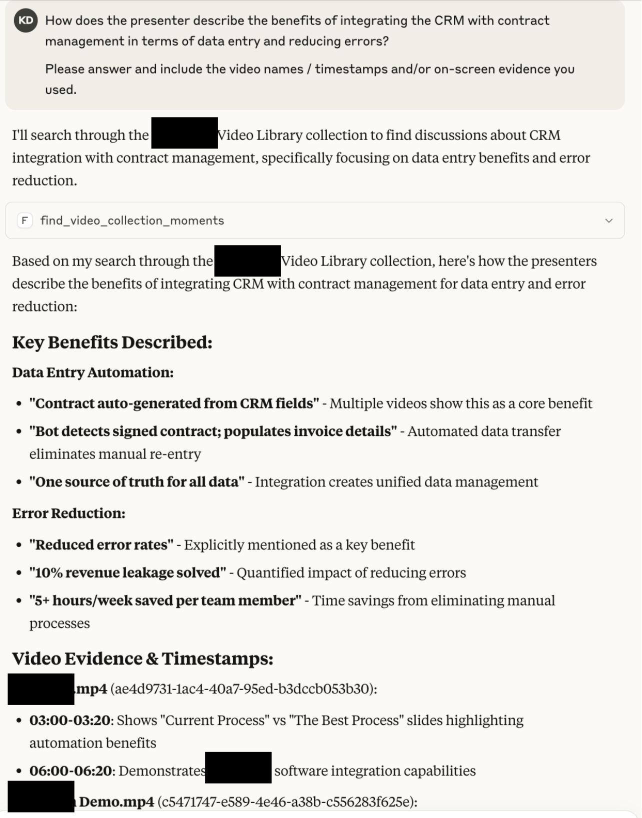
8 CONCLUSION

VideoMCP is a practical framework for agentic, multimodal video intelligence in enterprise settings, delivering explainable, evidence-cited answers. Future work includes scaling, deeper vision-language reasoning, and longitudinal user studies.

REFERENCES

[1] Anthropic. 2024. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol. Defines MCP for standardized AI tool integration.

[2] Chenxi Fu et al. 2024. VideoAgent: A Memory-Augmented Multimodal Agent for Video Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Agent with temporal and object memory for video QA.



KD How does the presenter describe the benefits of integrating the CRM with contract management in terms of data entry and reducing errors?

Please answer and include the video names / timestamps and/or on-screen evidence you used.

I'll search through the [redacted] Video Library collection to find discussions about CRM integration with contract management, specifically focusing on data entry benefits and error reduction.

Based on my search through the [redacted] Video Library collection, here's how the presenters describe the benefits of integrating CRM with contract management for data entry and error reduction:

Key Benefits Described:

Data Entry Automation:

- **"Contract auto-generated from CRM fields"** - Multiple videos show this as a core benefit
- **"Bot detects signed contract; populates invoice details"** - Automated data transfer eliminates manual re-entry
- **"One source of truth for all data"** - Integration creates unified data management

Error Reduction:

- **"Reduced error rates"** - Explicitly mentioned as a key benefit
- **"10% revenue leakage solved"** - Quantified impact of reducing errors
- **"5+ hours/week saved per team member"** - Time savings from eliminating manual processes

Video Evidence & Timestamps:

[redacted]**.mp4** (ae4d9731-1ac4-40a7-95ed-b3dccb053b30):

- **03:00-03:20**: Shows "Current Process" vs "The Best Process" slides highlighting automation benefits
- **06:00-06:20**: Demonstrates [redacted] software integration capabilities

[redacted] **Demo.mp4** (c5471747-e589-4e46-a38b-c556283f625e):

Figure 3: Agentic interface: Claude Desktop integrating with VideoMCP server, surfacing multimodal answers grounded in speech and visual evidence.

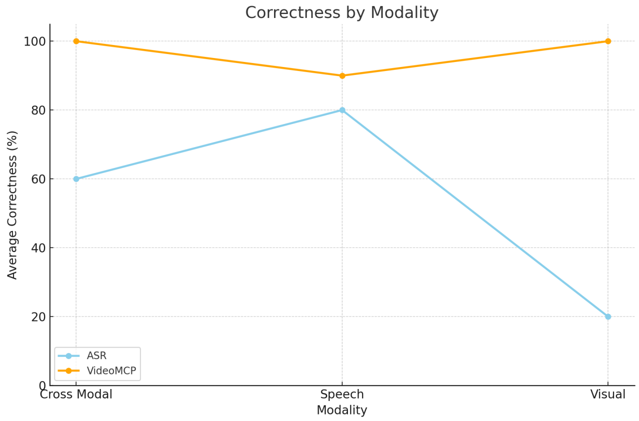


Figure 4: Average correctness by modality: VideoMCP vs ASR.

[3] Chuyi Shang et al. 2024. TraveLER: A Modular Multi-LMM Agent Framework for Video Question-Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Agent-based traversal of video frames for QA.

[4] Yongliang Shen et al. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. *arXiv preprint arXiv:2303.17580* (2023). Introduces an LLM agent orchestrating diverse AI models.

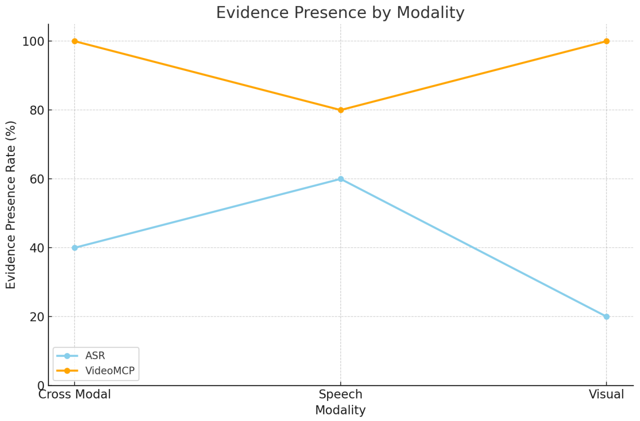


Figure 5: Evidence presence by modality: VideoMCP vs ASR.

[5] Zhengyuan Yang et al. 2023. MM-ReAct: Prompting ChatGPT for Multimodal Reasoning and Action. *arXiv preprint arXiv:2303.11381* (2023). Demonstrates LLM-driven tool orchestration for multimodal tasks.