CS M213A / ECE M202A (Fall 2025)

From Words to Shields: Agentic Privacy for Smart Sensors

Sam E. Seban & Daniel Luzzatto

Motivation and Objectives

Project Outline:

- Smart devices constantly collect audio/video, often sending raw data to the cloud.
- Users can't control what is filtered out or how privacy is enforced.
- We want a system where users can simply *say what they want protected* ("blur children's faces", "mute medical terms"), and the system automatically enforces it.

Audience:

- Protects people in homes, hospitals, workplaces…
- Helps meet changing regulations (e.g., GDPR).
- Reduces engineering effort for companies.

Goals:

- A prototype "agentic privacy hub" where an LLM builds and executes privacy pipelines.
- Multimodal support: audio, video, text, tabular.
- Verification system ensuring sanitized data actually respects the request.
- Quantitative evaluation of correctness, robustness, latency, and adaptability.

Technical Approach and Novelty

How it’s done today:

Systems like Peekaboo use fixed, developer-built pipelines. They provide strong guarantees, but low flexibility, as new privacy rules require new code.

Our high-level approach:

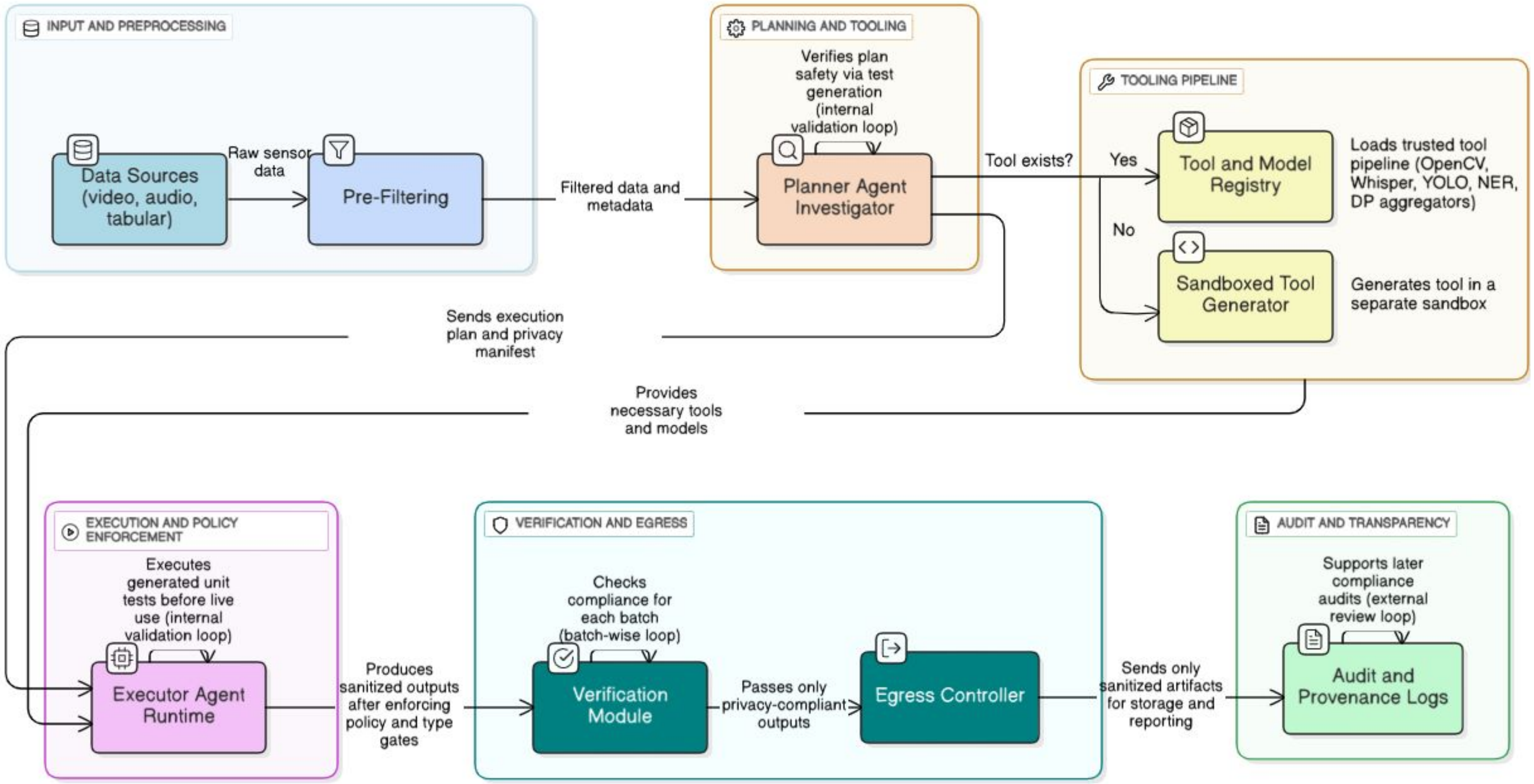
- LLM agent interprets natural-language privacy rules.
- It builds a pipeline: selecting tools, generating missing ones, executing, and verifying.
- Closed-loop: replans if verification fails.

Novelty:

- Autonomous tool generation (sandboxed).
- Unified policy for video, audio, text, tabular.
- Dynamic verification + recovery (retry → replan).
- Strong auditability (manifests, logs).

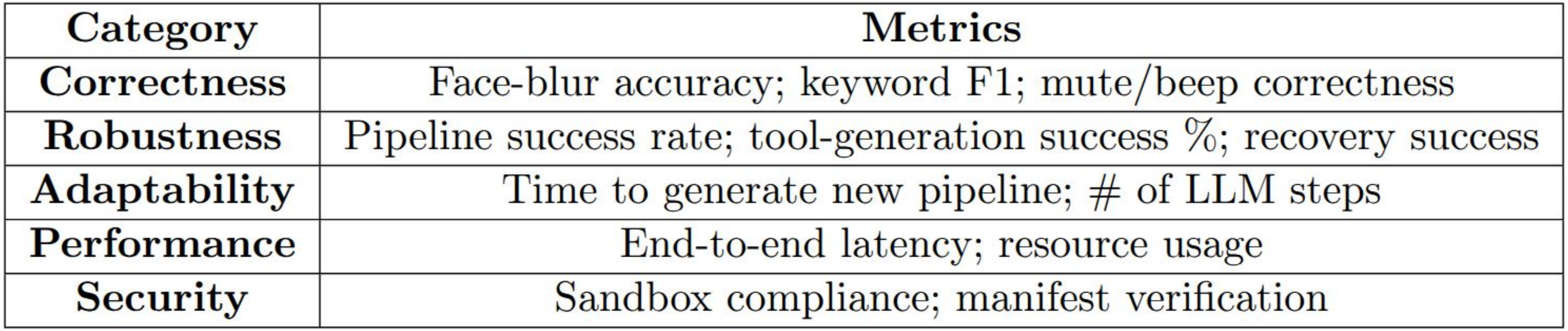
Methods

- Vision: face detection and blur using OpenCV, YuNet, Kalman Filter.
- Speech: keywords detection and beep/silence using Whisper.
- LLM for pipeline and tool generation: llama-3.3-70b-versatile.



INPUT AND PREPROCESSING
Data Sources (video, audio, tabular)
Raw sensor data
Pre-Filtering
Filtered data and metadata
PLANNING AND TOOLING
Verifies plan safety via test generation (internal validation loop)
Planner Agent Investigator
Tool exists?
TOOLING PIPELINE
Yes
Tool and Model Registry
Loads trusted tool pipeline (OpenCV, Whisper, YOLO, NER, DP aggregators)
No
Sandboxed Tool Generator
Generates tool in a separate sandbox
Sends execution plan and privacy manifest
Provides necessary tools and models
EXECUTION AND POLICY ENFORCEMENT
Executes generated unit tests before live use (internal validation loop)
Executor Agent Runtime
Produces sanitized outputs after enforcing policy and type gates
VERIFICATION AND EGRESS
Checks compliance for each batch (batch-wise loop)
Verification Module
Passes only privacy-compliant outputs
Egress Controller
Sends only sanitized artifacts for storage and reporting
AUDIT AND TRANSPARENCY
Supports later compliance audits (external review loop)
Audit and Provenance Logs

Evaluation and Metrics



Category	**Metrics**
Correctness	Face-blur accuracy; keyword F1; mute/beep correctness
Robustness	Pipeline success rate; tool-generation success %; recovery success
Adaptability	Time to generate new pipeline; # of LLM steps
Performance	End-to-end latency; resource usage
Security	Sandbox compliance; manifest verification

Current Status and Next Steps

Current status:

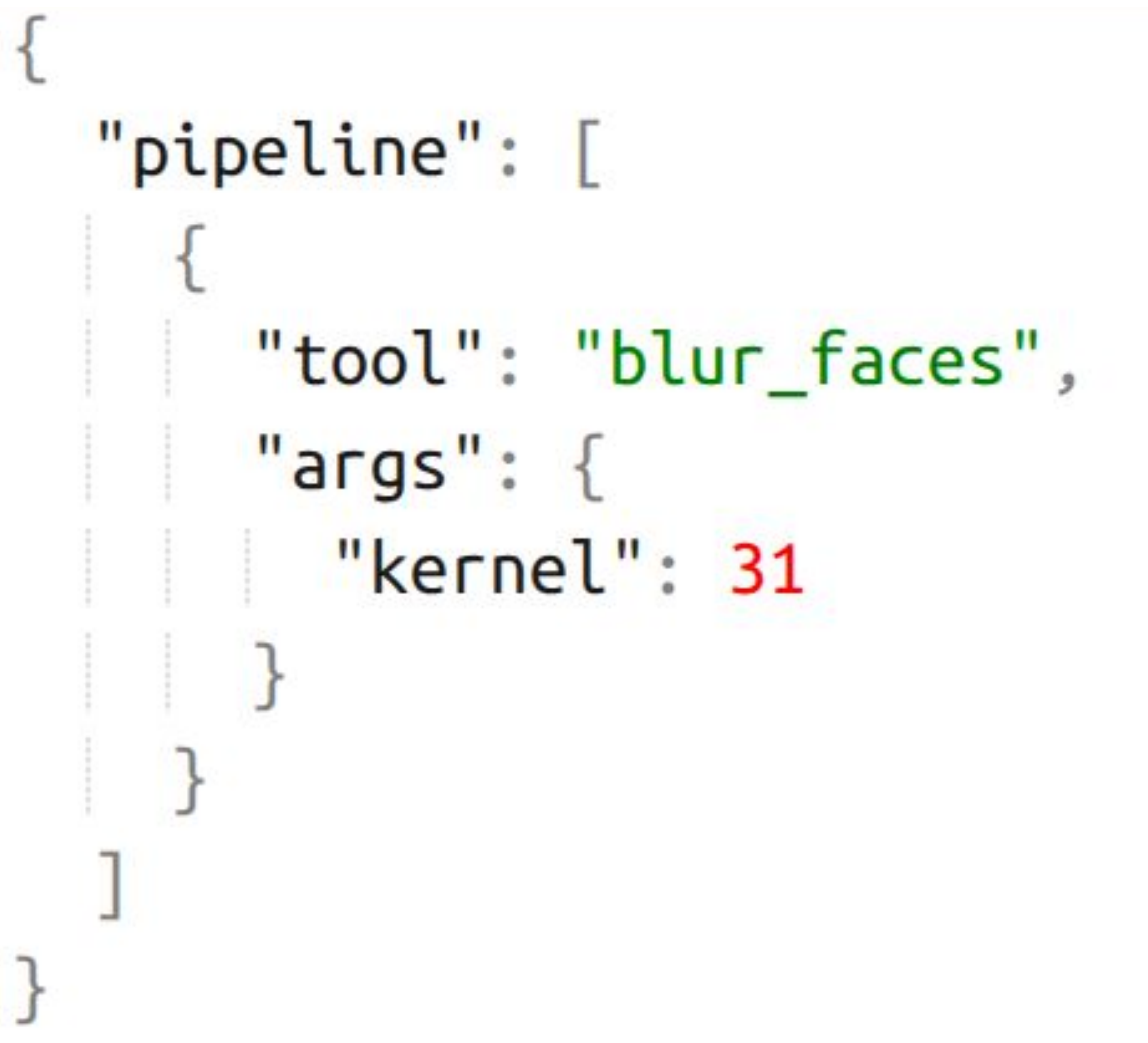
- Implemented face recognition and blur tool and verification for non-live videos.
- Implemented keyword detection and beep/silence and verification for non-live audio files.
- Implemented pipeline planner.
- Implemented tool generator.

Next steps:

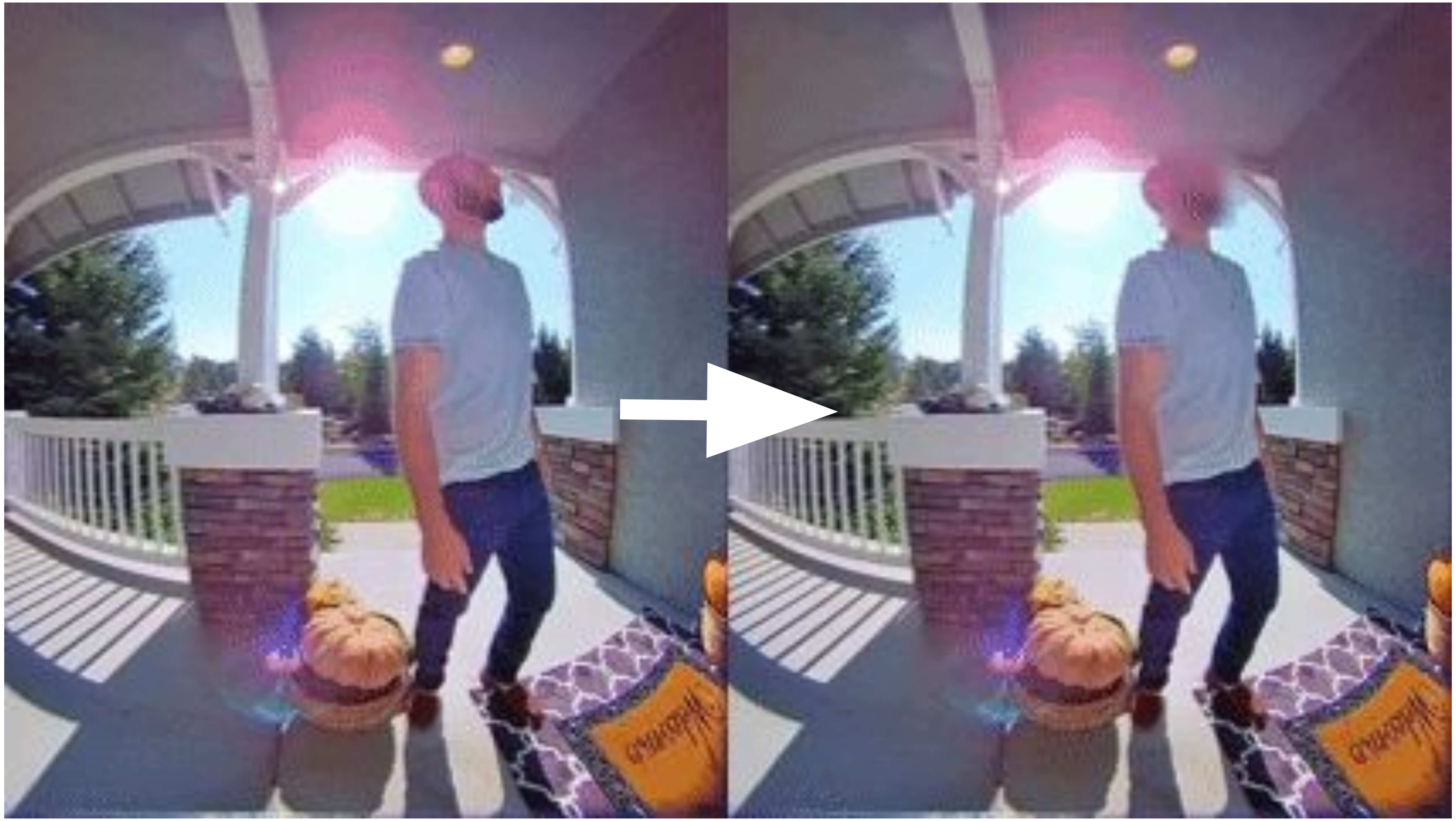
- Add multimodal, and tabular input tools.
- Adapt the existing tools for live applications.
- Implement verification of the generated tools.
- Implement the closed-loop architecture.
- Run the generated code in a sandbox environment for security.

Prompt: “Blur faces”

Generated manifest:

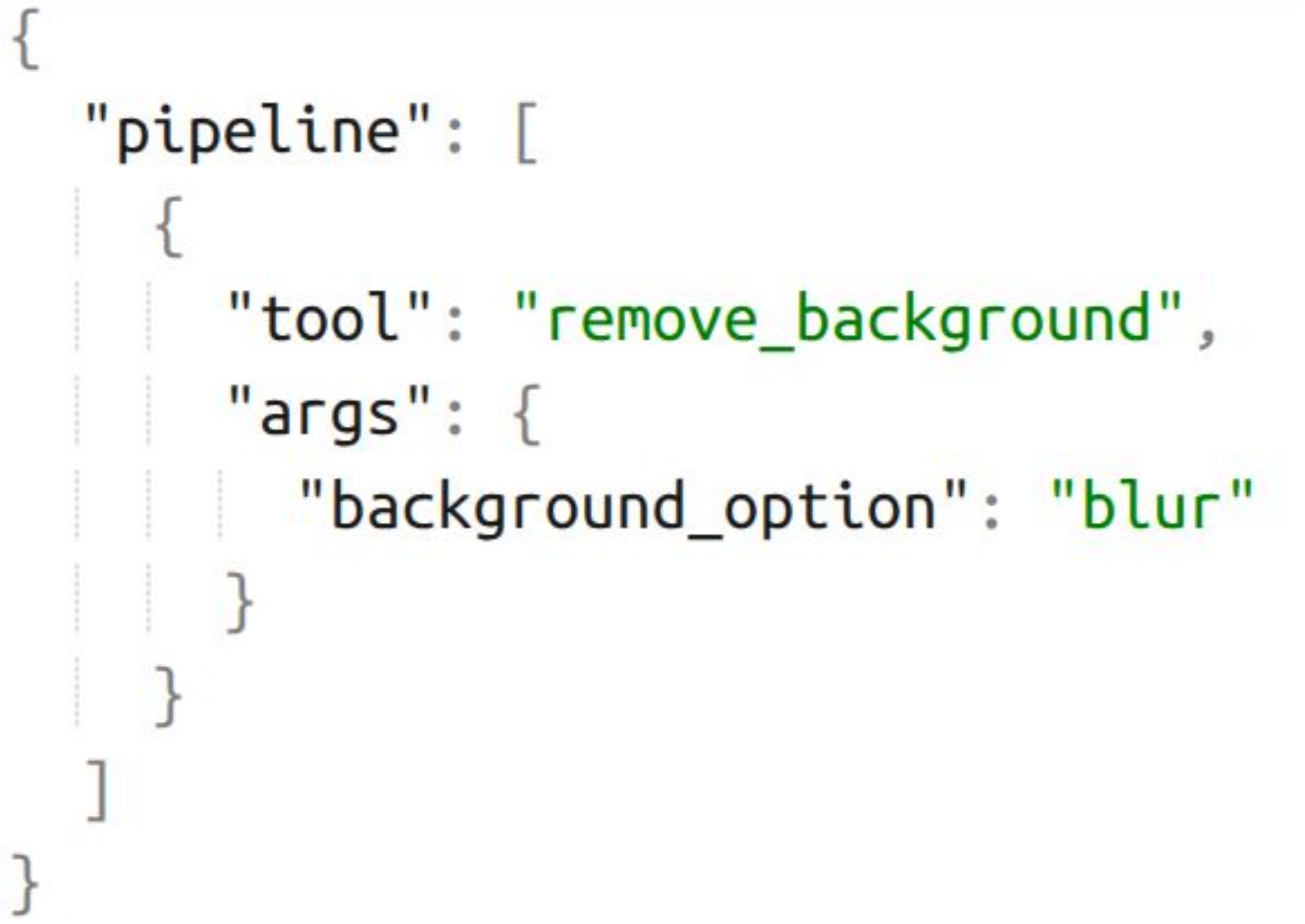


```
{
  "pipeline": [
    {
      "tool": "blur_faces",
      "args": {
        "kernel": 31
      }
    }
  ]
}
```



Prompt: “Hide the background”

Generated manifest:



```
{
  "pipeline": [
    {
      "tool": "remove_background",
      "args": {
        "background_option": "blur"
      }
    }
  ]
}
```

Tool doesn’t exist

↓

Trigger tool generation