# InstaWise — Multi-Agent Instagram Captioning & Posting Assistant

Author: Deeptha Kiruba K

Capstone: Kaggle Agent Project
Date: 1 Dec 2025

## Abstract

InstaWise is a multi-agent AI system designed to assist content creators and small businesses in generating high-quality Instagram captions, optimized hashtags, and recommended posting times for images and short videos. Using a combination of computer vision (CLIP), prompt-driven LLMs, and heuristic agents, InstaWise produces a Post-Ready Kit that reduces workload and improves engagement potential.

## Introduction

Social media content creation requires creativity, time, and consistency. Many small businesses—especially home bakers and micro-brands—struggle with caption writing, hashtag research, and finding the best time to post. InstaWise automates this workflow through a coordinated set of AI agents that handle vision analysis, caption generation, hashtag optimization, and posting-time prediction.

This project demonstrates a real-world, production-ready multi-agent pipeline suitable for businesses like VV Taste Buds and scalable to any Instagram-based workflow.

## Literature Review

Automated captioning has evolved through classical image captioners to modern transformer-based models and prompt-driven multimodal LLMs. Hashtag recommendation systems range from TF-IDF keyword models to graph-based trending tag extractors. Posting-time prediction typically uses engagement forecasting, but heuristic models based on category patterns remain effective for small datasets.

CLIP (Contrastive Language–Image Pretraining) serves as an excellent backbone for zero-shot vision labeling, enabling flexible classification without custom training.

## System Requirements

Functional Requirements:

- Upload image or video

- Vision Agent must extract theme, objects, mood

- Caption Agent must generate 3 caption variants
- Hashtag Agent must output optimized tags
- Time Agent must recommend best posting times
- (Optional) Poster Agent must simulate or call Meta Graph API

Non-Functional Requirements:

- Simple UI
- Fast responses (<3–5s)
- Modular agents for reuse
- Explainable outputs

## Architecture

The system consists of:

- Streamlit UI for upload and display
- FastAPI backend for orchestration
- CLIP-based Vision Agent
- LLM-based Caption Agent
- Hashtag Agent using keyword extraction + heuristics
- Time Agent using category-based posting rules
- Optional Poster Agent for API-based publishing

Flow:

User Upload → FastAPI → Vision Agent → Caption Agent → Hashtag Agent → Time Agent → UI

## CLIP Integration

The Vision Agent uses CLIP (openai/clip-vit-base-patch32) to extract semantic labels.

Steps:

1. Load image

2. Encode using CLIP processor

3. Encode candidate labels (dessert types, bakery items, moods, colors)

4. Compute cosine similarity

5. Select top-K labels to form vision summary

The vision summary then feeds into the Caption and Hashtag agents.

See clip_integration.py for full implementation.

## Agent Design

Vision Agent:

• Extracts theme, objects, mood, food type

• Uses CLIP + candidate label set

• Returns structured JSON

Caption Agent:

• Uses prompt-engineered LLM

• Produces 3 variants: short, story, business

• Incorporates brand name (VV Taste Buds)

Hashtag Agent:

• Extracts keywords from caption + CLIP labels

• Scores using popularity dictionary

• Outputs popular + niche + local groups

Time Agent:

- Uses category heuristics
- Predicts top 3 posting slots with confidence and rationale

## Implementation Details

Backend: FastAPI

UI: Streamlit

Vision: CLIP (HuggingFace Transformers)

LLM: Prompt-driven text generation (stubbed; replace with OpenAI/Anthropic/local LLM)

Hashtags: Custom heuristic engine

Time Agent: Table-driven heuristics

Poster Agent: Mocked Meta API endpoint

## Datasets & Sample Inputs

Sample test images:

- Chocolate jar cake
- Behind-the-scenes bakery photo
- Product gift box

Each contains:

- Vision summary JSON
- Caption variants
- Hashtags
- Recommended times

## Evaluation

Quantitative Metrics:

- Hashtag relevance precision@10
- Caption creativity via human scoring (1–5)
- Time slot plausibility check

Qualitative Metrics:

- Smoothness of workflow
- Human readability and style quality of captions
- Aesthetic match between captions and image

## Sprint Plan & Scrum Artifacts

Sprint Plan (7 days):

- Days 1–3: Vision Agent, Caption Agent, UI
- Days 4–5: Hashtag Agent, Time Agent
- Days 6–7: Polish + mock Posting Agent

Scrum Artifacts:

- Product Backlog: All agents + UI + API
- Sprint Backlog: Priority subset
- Daily updates
- Sprint demo: Post-Ready Kit workflow

## Results & Demo

On 3 sample images, InstaWise correctly:

- Identified dessert objects through CLIP
- Generated engaging captions
- Suggested 10–12 relevant hashtags
- Recommended strong posting windows

Demo includes the upload-to-caption workflow.

## Future Work

- Add trending hashtag API
- Train domain-specific caption fine-tuner
- Personalized time-recommendation using user engagement logs

- Expand to multi-platform: YouTube Shorts, Facebook Reels, TikTok

## Conclusion

InstaWise successfully delivers an end-to-end multi-agent pipeline using modern AI techniques. It automates a high-value business task—Instagram content creation—while remaining explainable, modular, and easy to extend.

This system is immediately useful for real businesses and supports further academic enhancement.

## Appendix

Includes:

- Full CLIP label list
- Prompt templates for Caption Agent
- Heuristic tables for Time Agent
- Sample JSON outputs