# MontageAgent – Architecture & Rationale

**Project:** MontageAgent — AI Agent for Event Photo Curation & Instagram Publishing
**Author:** Prem Kondru (BTech Engineering Physics, IIT Guwahati; Photography Club: Montage)

---

## Purpose & Problem

At IIT Guwahati's Photography Club (Montage), every event generates hundreds of photos. The manual workflow involves:

1. Removing duplicates
2. Grouping by theme/moment
3. Writing captions in a consistent style
4. Assembling Instagram carousel posts

This process is repetitive, error-prone, and consumes hours of human effort.

**MontageAgent automates the workflow end-to-end:**
**Ingest → Dedupe → Categorize → Cluster → Caption → Export/Publish**

---

## Interaction Flow

1. **User Input:** Upload images, set labels, choose captioner mode, define event name, adjust `max_images_per_post`.
2. **Pipeline Execution:** Ingest → Embed → Deduplicate → Categorize → Cluster → Caption → Export.
3. **Captioning:**
   - **BLIP-2 mode:** Generate per-image captions, extract common words, produce abstract caption.
   - **Template mode:** Deterministic, rule-based caption.
   - Hashtags pulled from base + labels + historical RAG hints.
4. **Preview & Export:** IG-style preview, per-image inclusion/exclusion, JSON export for carousel.

---

## How the Agent Works

### Reasoning

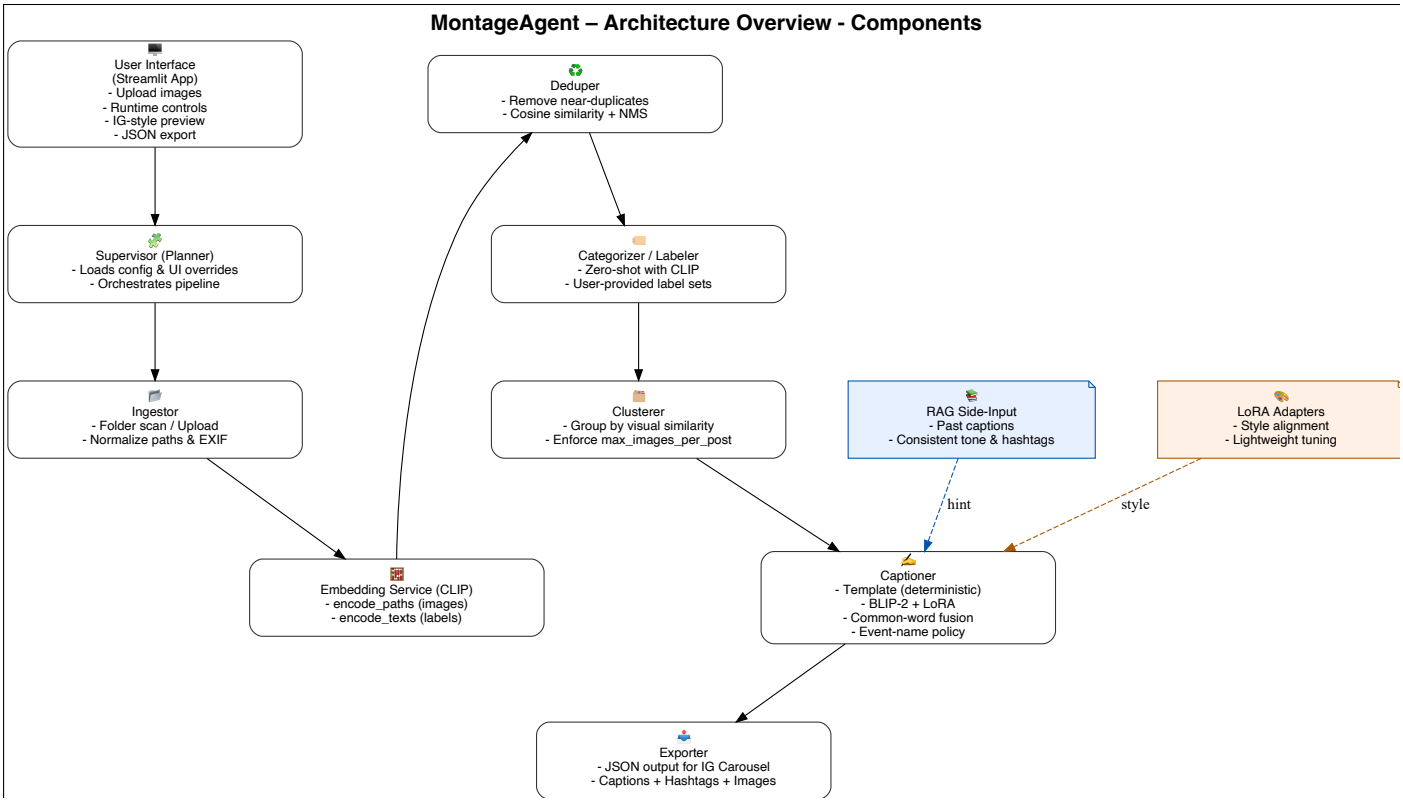
- **Visual Understanding:** CLIP embeddings assess similarity, remove near-duplicates, and infer labels (zero-shot).
- **Captioning:** Two captioner modes:
  - **Template Mode** → Deterministic, rule-based captions.
  - **BLIP-2 (with LoRA)** → Generates captions based on a set of cluster images.
- **Stylistic Consistency:** Retrieval-augmented input (RAG) over past captions maintains tone and ensures continuity across posts.

## Planning

- A **Supervisor (Planner)** module orchestrates a fixed pipeline:
  *Ingest → Embed → Dedupe → Categorize → Cluster → Caption → Export*.
- The plan is deterministic, simplifying debugging and evaluation.

## Execution

- **Executors (Workers):** Each step runs as an independent tool with clear inputs/outputs.
- **Human-in-the-loop UI:** Built with Streamlit, allowing users to include/exclude photos, adjust labels, and preview posts before export.




# Components

- **User Interface:** Streamlit app with IG-style previews, zoom, toggles, runtime configs, and JSON export.
- **Supervisor (Planner):** Loads config, applies UI overrides, sequences executors.
- **Ingestor:** Scans folders/uploads, builds image list.
- **Embedding Service (CLIP):** Provides `encode_paths` and `encode_texts` for similarity, clustering, and label assignment.
- **Deduper:** Removes near-duplicates via cosine similarity.
- **Categorizer/Labeler:** Zero-shot label assignment using CLIP and user-provided label sets.
- **Clusterer:** Groups images by visual similarity; enforces `max_images_per_post` with balanced sampling.
- **Captioner:**
  - **Template mode:** Deterministic phrasing.

    - **BLIP-2 mode:** Batch captioning + LoRA fine-tuning; fuses outputs into one abstract caption. Supports event-name injection policies (off | hint | only_proper_noun).
- **Exporter:** Builds JSON for Instagram Posts upload: `{caption, hashtags, images[]}`.

---

# Key Design Choices & Agent Patterns

https://www.anthropic.com/engineering/building-effective-agents
https://arxiv.org/pdf/2405.1046

- **Role-based cooperation (multi-agent): Supervisor + tool workers (ingest, embed, dedupe, cluster, caption, export).**
- **Fixed pipeline:** Single path improves reliability, reproducibility, and transparency.
- **Prompt/response optimisation:** strict caption prompts; output guards (no hashtags, abstract, proper-noun policy).
- **CLIP** for structural tasks (dedupe, clustering, labels): Efficient, lightweight, and non-generative.
- **RAG:** style/hashtag hints from nearest past captions.
- **BLIP-2 + LoRA for style:** Strong caption quality with minimal compute/storage overhead.
- **Config-driven + UI overrides:** Enables repeatability, A/B testing, and flexible runtime control.
- **Guardrails:** Caption length limits, “event name only” injection, and exclusion of proper nouns ensure alignment with club style.
- **Evaluator:** CLIPScore & silhouette surfaced to the user; iterative feedback loop.

---

# Models & Rationale

- **OpenCLIP (model ViT-B/32 pretrained laion2b_s34b_b79k):**
    - Lightweight, efficient visual similarity.
    - Ideal for deduplication, clustering, zero-shot labels, CLIPScore.
- **RAG over Past Captions:**
    - Ensures consistency in tone/hashtags.
    - Avoids over-fitting to rigid templates.
- **BLIP-2 (base model Salesforce/blip2-flan-t5-xl) with LoRA:**
    - Generates abstract, mood-driven captions aligned with club style.
    - LoRA enables style specialization without full finetuning.
    - Per-image batching + common-word fusion prioritizes shared visual cues.
    - Guardrails enforce reliability and stylistic alignment.

---

# Evaluation: Quality & Reliability Metrics

- **CLIPScore:** Measures caption–image alignment (per-image and cluster means).
- **Silhouette Score:** Quick proxy for clustering cohesion/separation.
- **Dedupe Rate:** % of near-duplicates removed.
- **Human Ratings (optional):** Abstractness, tone, Instagram readiness.
- **A/B Protocol:** Compare **template vs BLIP-2/LoRA** on the same clusters; track metrics per event and overall.

- A full **Data Science Report (PDF)** and **Interaction Logs** (prompts + chat history) are generated to document method and outcomes.

**In summary, Montage is a complete, AI agent that automates a real university workflow, integrates a LoRA-tuned model for style-safe captioning, and ships with the metrics and documentation required to evaluate and maintain it.**