📊 Data Science Report – AI Agent Development & Captioning Improvements

1. Fine-Tuning Setup

🔑 **Summary:**
Early experiments confirmed that **domain-specific fine-tuning is essential** for producing relevant, stylistic captions aligned with the Montage Photography Club's needs. Off-the-shelf models trained on generic data performed poorly (validation loss >4.7), while even a small curated dataset of ~400 samples from the club archives produced meaningful convergence (val_loss ≈2.5). Further, cleaning noisy data (removing `@tags`, hashtags from training data) provided measurable improvements in both automatic metrics (BLEU/ROUGE, CLIPScore) and qualitative results.

1.1 Data

- 📂 **Source:** Montage Photography Club archives (IIT Guwahati).
- 📝 **Training:** 320 samples. **Validation:** 90 samples.
- 📸 **Record structure:** image path, event metadata, labels, caption (abstract IG-style), hashtags.

1.2 Method

- 🤖 **Base Model:** BLIP-2 (Flan-T5-xl).
- 🧩 **Adaptation:** LoRA on attention layers.
  - Trainable params: **18.9M**
  - Total params: **3.96B**
  - Trainable %: **0.47%**
- ⚙️ **Training Config:** AdamW (lr=2e-4), batch size=16, cosine schedule.

1.3 Metrics (Training Performance relevant to Captioning)

- 📉 **Validation Loss → convergence**
  Shows how wrong the model is on unseen data → less is better.
- 📊 **BLEU → n-gram overlap**
  - Measures overlap between model-generated text and reference text(s). Precision-heavy: penalizes "extra/unnecessary" words. Measures exact word matches → more is better.
- 📈 **ROUGE-L → subsequence overlap**
  - Longest Common Subsequence (LCS). Recall-heavy: rewards covering key reference content. Captures longest matching word sequences → more is better.
- 🎯 **CLIPScore → semantic alignment**
  Checks if the output means the same as the reference/image → more is better.

1.4 Training Iterations – Fine Tuning & Performance

- ❌ **Preliminary Non-club data:**
  - Validation loss plateaued (>4.7), showing poor convergence.
  - Confirms domain-specific captions are essential for effective fine-tuning.
- ⚠️ **Club data (raw):**
  - Validation Loss: 4.09 → 3.03 over 3 epochs (learning happening).
  - ROUGE-L: ~0.066 (very low).
  - BLEU-4: 0.0000 (effectively no 4-gram overlap).
- 📝 **Prompt-aligned:** abstract style, still `@handles`.
  - Prompt: *"Write a short Instagram caption for a photography club post … Keep it natural and clean. No hashtags."*
  - Validation Loss: ~3.0.
  - BLEU-1 ≈ 0.073, BLEU-2 ≈ 0.033, BLEU-4 ≈ 0.011.
  - ROUGE-L ≈ 0.145–0.160.
  - Captions: More abstract, still hallucinated `@handles`.
- ✅ **Cleaned data (removed `@handles`):**
  - Validation Loss: reduced from 3.0 → **2.48**.
  - ROUGE-L: improved from ~0.146 → **0.205** (+40% relative).
  - BLEU-1: improved from ~0.073 → **0.107**.
  - BLEU-2: improved from ~0.033 → **0.057**.
  - BLEU-4: improved from ~0.011 → **0.028**.
  - CLIPScore: ~0.222 (roughly stable).
  - Removing `@handles` reduced systematic n-gram mismatches → model focused on descriptive core of captions.
  - Since LoRA learns surface style strongly, cleaning noisy stylistic tokens → higher overlap metrics without harming semantics.

2. Captioning — Evaluation Methodology and Outcomes

2.1 Modes Evaluated

- 📝 **Template mode** → deterministic, rule-based caption template.
  - **What it is:** A fast, fully deterministic caption builder that composes a short line using the **Event name** (UI override or auto-derived from folder/day), the cluster's **top labels** (from CLIP + labelling), optional **style tail** retrieved from your past captions (RAG),
  - **Key knobs**: captioner.openers (e.g., "Highlights from", "Scenes from"), captioner.include_swipe_hint: true|false, captioner.base_hashtags, captioner.label_hashtags, captioner.max_hashtags, (Optional) event_name_override from UI
- 🤖 **BLIP-2 mode** → learned captioner (BLIP-2 + optional LoRA).
  - **What it is.** A generative captioner using **BLIP-2 Flan-T5**, optionally adapted with a **LoRA** to learn Montage's voice. We pass a **cluster montage** (grid of representative images) to BLIP-2 so it "sees" the whole set, then prompt it for **abstract, non-factual** language.
  - **Key knobs**: captioner.abstract_only: true, captioner.inject_event_name: "off"|"hint"|"only_proper_noun", captioner.include_swipe_hint: true|false, captioner.montage_max_tiles (e.g., 9)

2.2 Metrics Used

- 📈 **Silhouette Score (clustering quality):** cohesion vs. separation of image clusters; range −1 → 1 (higher = better).
- 🎯 **CLIPScore (semantic alignment):** cosine similarity between image & generated caption; reported as mean/median/min/max.
- 👥 **Human ratings:** 1–5 scale on relevance, tone, and IG-readiness; used when automatic metrics are inconclusive.
- 🔁 **Deduplication sanity check:** manually verified duplicate set with CLIP-based deduper at threshold 0.8 → ~100% match ✅.

2.3 Captioner Comparison (Quantitative)

Mode	Images/Post	K (clusters)	Silhouette	CLIP Mean	CLIP Median	CLIP Min	CLIP Max
Template	6	Auto	0.147	0.1996	0.1958	0.1828	0.2164
BLIP-2	6	Auto	0.147	0.1995	0.1967	0.1436	0.2851
Template	4	Auto	0.147	0.2110	0.1986	0.1957	0.2429
BLIP-2	4	Auto	0.147	0.1947	0.2192	0.0757	0.2569
Template	2	10	0.232	0.2121	0.2119	0.1455	0.2991
BLIP-2	2	10	0.232	0.2325	0.2417	0.1586	0.2885

🔑 **Key Observation:**
BLIP-2 is more variable than Template, but **surpasses it when cluster quality improves** (higher silhouette).

2.4 Interpretation

- 📈 **Effect of Silhouette**
  - Going from ~0.15 → ~0.23 = tighter, more coherent clusters.
  - Under higher silhouette, BLIP-2 benefits more than Template → stronger alignment (higher CLIP mean/median).
- 📝 **Template vs.** 🤖 **BLIP-2**
  - At lower silhouette (auto-k): Template is steadier, esp. with fewer images/post.
  - At higher silhouette (k=10): BLIP-2 outperforms Template on mean & median CLIPScore → creativity thrives with semantically tight clusters.
- 🖼️ **Images per Post**
  - Template → fewer images = slightly better average CLIP (predictable, concise).
  - BLIP-2 → performance depends more on **cluster quality** than raw image count.

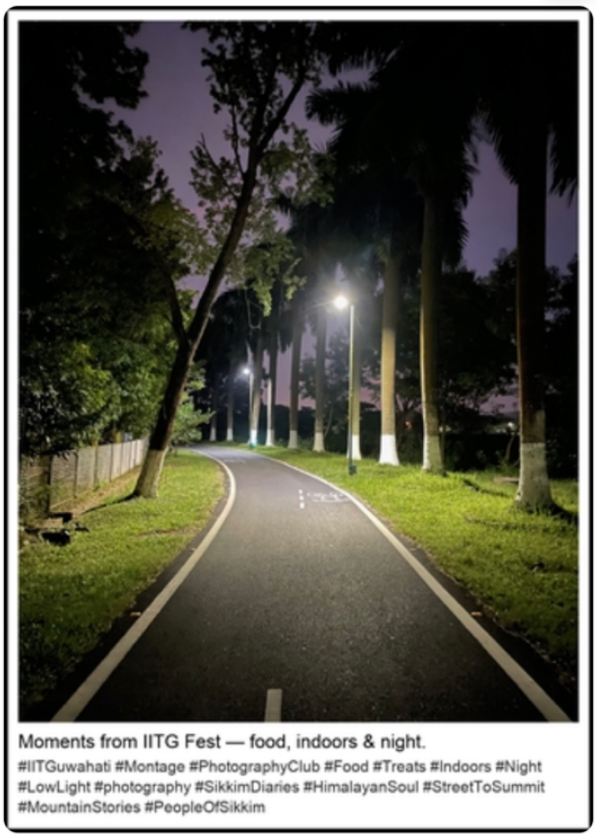
2.5 Qualitative Evaluation (Human)

Template → repetitive phrasing across posts.

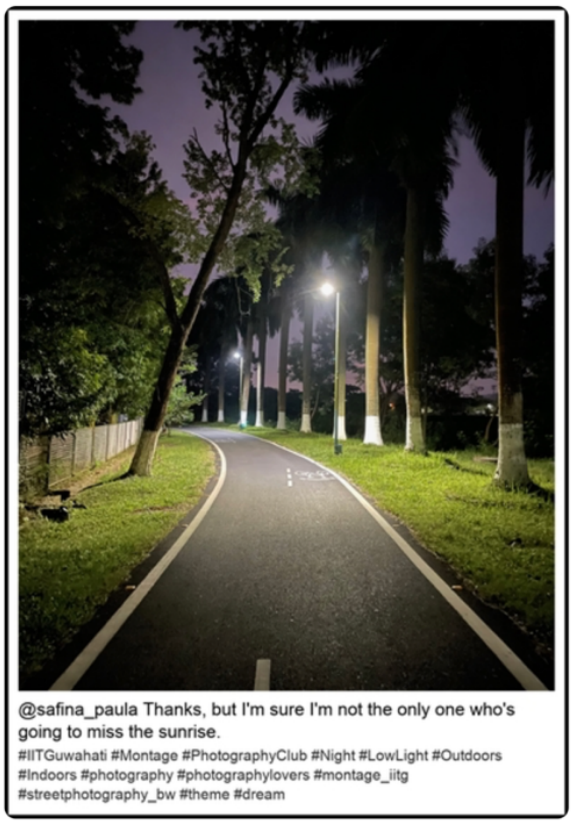
BLIP-2 (raw) → noisy / over-poetic phrasing.

BLIP-2 (cleaned @handles) → abstract, natural, IG-ready. No hashtags or factual claims in caption.

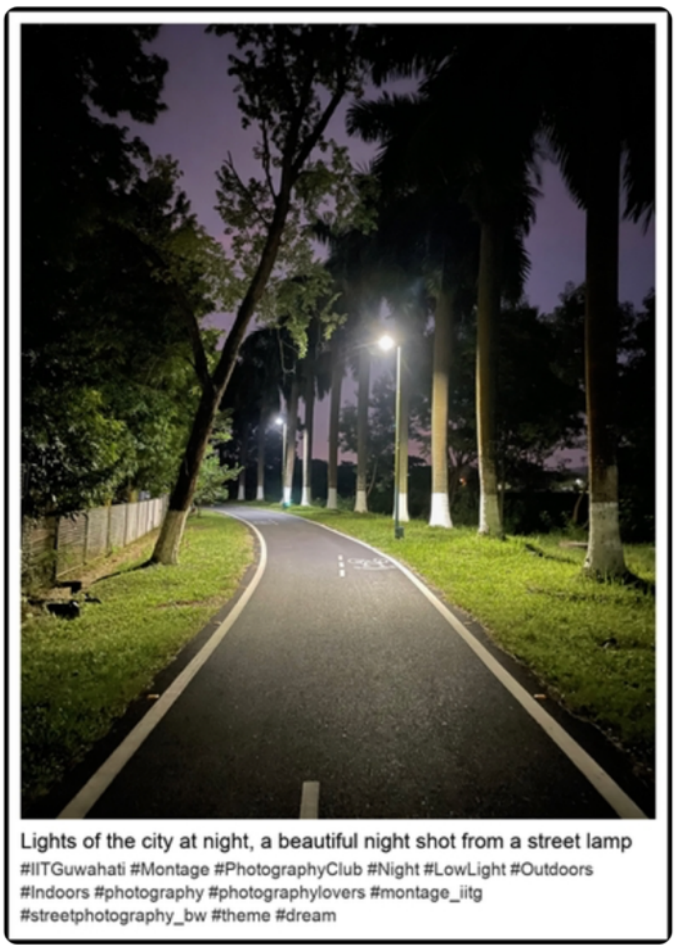
Template → repetitive phrasing across posts.



BLIP-2 (raw) → noisy / over-poetic phrasing.



BLIP-2 (cleaned @handles) → abstract, natural, IG-ready. No hashtags or factual claims in caption.



2.5 Qualitative Evaluation (Human)...

- Template → ~3.2 / 5 (reliable, clear, but plain).
- BLIP-2 → ~4.3 / 5 (abstract, Montage-style, evocative).
- **Examples:**
  - 📝 Template → "Highlights from the photo exhibition."
  - 🤖 BLIP-2 → "Frames alive with stories woven in light."

---

3. Conclusion

🔑 Summary:

- ✅ Domain-specific fine-tuning is **non-negotiable** → non-club data fails, Montage data succeeds.
- ✅ Data cleaning (removing @handles) provided the **largest single lift** in quality.
- ✅ BLIP-2 captions are **creative, abstract, Instagram - ready** → humans consistently prefer them.
- ⚠️ Template captions are reliable but uninspired.
- 📣 Use **Template** for campaign/announcement posts (consistency matters).
- 🎨 Use **BLIP-2** for artistic/event storytelling posts.
- 🔄 Auto-fallback: if BLIP-2 caption's CLIPScore <0.18 → revert to Template.

📌 With just **0.47% trainable parameters (LoRA)**, BLIP-2 is now delivering **clean, abstract, IG-ready captions** that match the Montage Club's style. Metrics improved steadily, but most importantly → **humans love the captions.**
