CapMoo: Automated Behavioral Captioning for Wildlife Engagement

Authors: Supanat Kampapan, Santipab Tongchan and Thitsanapat Siwarattanan
Affiliation: Chulalongkorn University, King Mongkut's University of Technology Thonburi and King Mongkut's Institute of Technology Ladkrabang.
Date: 24th May 2025

Abstract

CapMoo is an end-to-end system that turns live camera feeds of zoo animals into narrated, behavior-aware captions. With Google Gemini Flash 2.5 to deliver real-time text and audio descriptions that highlight both observable actions and inferred motivations. Our observational trials showed that previous models (e.g. Patthama-llm-visions-1.0.0) could detect fine gestures but failed to articulate them in natural, engaging language. CapMoo closes this gap, improving visitor engagement, while minimizing energy consumption through batched inference and mixed-precision encoding.

Introduction & Motivation

It's one thing for an image-captioning model to point out "a tiger is sitting," but quite another to hint at why it might be resting there whether it's conserving energy in the midday heat or keeping an eye on its enclosure. At the Moodeng exhibit, we wanted captions that feel more like a wildlife guide than a photo gallery, so we built CapMoo around Google's Gemini Flash 2.5 API. Rather than training or fine-tuning new models locally, we simply call the Gemini endpoint with carefully crafted prompts that encourage both precise observation and gentle interpretation. This lets us focus on delivering captions that read naturally ("Notice how she gently paws at the water perhaps investigating the ripple patterns") without any of the overhead of data collection or on-site hardware tuning.

System Architecture

Our prototype runs entirely on commodity hardware and cloud services. A standard webcam captures still frames at set intervals; simple software routines discard empty or blurry images. Each remaining frame is packaged with a brief context prompt and sent over HTTPS to the Gemini API. The response—an English sentence describing both action and inferred motivation—is then handed off to an off-the-shelf text-to-speech engine for audio rendering. All captioning and speech synthesis happen in real time, yet no GPUs or specialized edge devices are required, demonstrating how modern LLMs can be slotted into existing exhibits with minimal technical friction.

Observational Evaluation

In informal trials with exhibit staff and a handful of volunteer visitors, CapMoo consistently outperformed the "plain facts" baseline. People remarked that hearing short, story-like narrations ("He seems alert as he shifts his gaze toward the feeder, possibly anticipating his next meal") made them pause longer and ask more questions. Exhibit designers noted that the captions struck a good balance between accuracy and charm, and there were no complaints of robotic or stilted phrasing. These early reactions suggest that, even without local model training or customized hardware, API-driven behavioral captioning paired with TTS can meaningfully enrich the visitor experience.