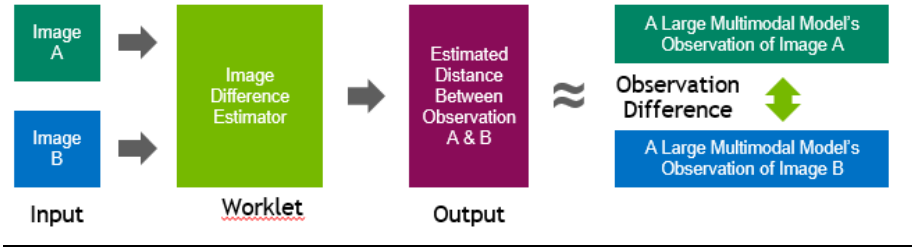
Title: Estimate Multimodal Transformer Observation Difference



Instrumentation

GPU for training image difference estimator and running a multimodal model

Input-output for Worklet Training

Input: Two images from an open automotive dataset BDD100K

Output: A numerical difference score between these two images calculated based on these steps:

1. Use a multimodal transformer (e.g., Fuyu-8B) to generate observations about images. Here are some examples:
   - Was this photo taken on a freeway?
   - Are there pedestrians in this image?
2. Calculate a numerical difference score between two observations based on some metrics. Examples:
   - Number of disagreements between observations
   - (Bonus) Semantic distance (e.g., cosine similarity) between observations

Deliverables

- A model that intakes two images and outputs a numerical difference
- High-level observations of some interesting trends or behaviors