Video → Perception (Detector + OCR) → World-State: https://arxiv.org/pdf/2311.17406

↓ ↓

Frame Selector RAG Luật VN củng cố suy nghĩ

\ /

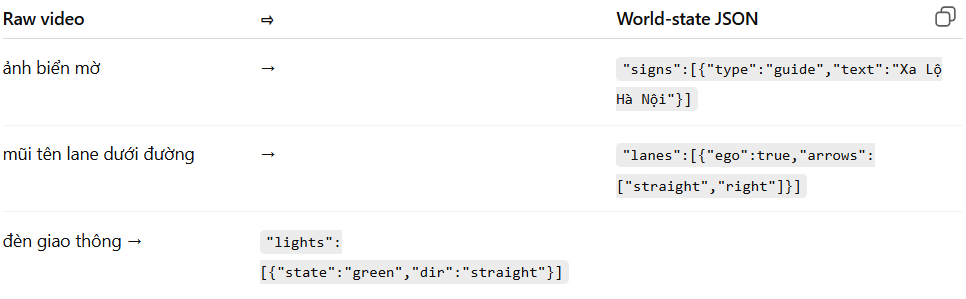
Gom lại thành Structured Chain of Thoughts – Input Model QWEN

→ Qwen-VL + LoRA (hidden-CoT)

↓

Answer + Support frame(s) + **Brief why:**

* Perception vai trò: nhìn ảnh, đọc biển báo,biển hiệu tên đường,đèn tín hiệu
* Đưa Perception này vào World State: World State là một mô hình biểu diễn trạng thái để làm tiền đề đưa vào LLM reasoning cho nó thực hiện được Chain Of Thought không bị nhiễu. Nó như một bản tóm tắt từ video để mô hình tập trung đúng vào những pixel chứa các thông tin mà ta cần quan tâm (thông tin từ Perception).  
  Thay vì để nó nhìn raw pixel và tự suy, ta **đưa cho nó facts đã trích ra**:



Kết quả:  
VLM **không hao token suy pixel** mà tập trung làm **logic giao thông + luật**.

* RAG lúc này đóng vai trò để nó dò luật, quyết định đc là nó lên đi đâu hướng nào
* Frame selector: Top frame trong video chứa thông tin quan trọng để model nhìn vô
* Gom lại thành 1 cái structured CoT
* Thực hiện chain of thought :Facts (nhìn frame) → World-State → Law → Check options → Answer.

Hàm Loss được kết hợp từ **structured Cot** chứa **Facts (nhìn frame) → World-State → Law** + Loss **Answer**