

# Multimodal AI Fusion Architecture Report

**VISOR: AI-Powered Guiding Shield for Vision**

## Executive Summary

VISOR implements a **late fusion multimodal architecture** that combines complementary vision models (YOLO object detection + BLIP image captioning) through a language model reasoner to generate unified, context-aware scene descriptions optimized for assistive technology. The system demonstrates significant improvements in descriptive accuracy and naturalness compared to individual model outputs.

## 1. Fusion Architecture

### 1.1 Overview

The fusion pipeline operates at the **output level (late fusion)**, where independent vision models process the same input image, and their outputs are combined via a reasoning layer:

```
Input Image → YOLOv8 (Object Detection) → Output: Bounding boxes, class labels, confidence scores → BLIP (Image Captioning) → Output: Natural language scene description → BLIP VQA (Question Answering) → Output: Answer to user queries ↓ FUSION LAYER ↓ FLAN-T5 / Gemini 2.5 Flash (Reasoning) → Synthesized Narrative (fused output)
```

### 1.2 Technical Implementation

**Fusion Function:** `generate_narrative()` in `backend.py`

**Inputs:** YOLO detections (top 6 objects) + BLIP caption

**Process:** Detection summarization → Prompt construction → Language model synthesis

**Output:** Single concise sentence (<25 words) optimized for text-to-speech

### 1.3 Why Late Fusion?

**Advantages:** Modularity, interpretability, efficiency, flexibility

**Trade-offs:** Slightly higher latency (mitigated by lightweight reasoner)

## 2. Model Performance Metrics

### 2.1 YOLO Detection Models (coco128 validation)

Model	Parameters (M)	Inference (ms/img)	Precision (P)	Recall (R)	mAP50	mAP50-95
yolov8n	~3.2	~118	0.64	0.537	0.605	0.446
yolov8s	~11.2	~244	0.797	0.664	0.760	0.589
yolov8m	~25.9	~482	0.712	0.730	0.784	0.614

**Analysis:** yolov8n selected for MVP (lowest latency: 118ms). yolov8s shows best precision-recall balance. yolov8m achieves highest mAP50-95 (0.614) but 4x slower.

## 2.2 Available Visualizations

Graphs available in `run/yolov8{n,s,m}/`:

- `BoxPR_curve.png`: Precision-Recall curve
- `BoxF1_curve.png`: F1-score vs. confidence threshold
- `BoxP_curve.png` / `BoxR_curve.png`: Precision/Recall curves
- `confusion_matrix.png`: Per-class confusion matrix
- `val_batch*.jpg`: Sample predictions vs. ground truth

## 2.3 Fusion Pipeline Performance

**Quantitative Results** (20 COCO val images):

- **Object Coverage**: Fused 54.08% vs BLIP 32.75% (+65.1% improvement) ✓
- **ROUGE-L**: Fused 0.7484 (maintains semantic quality)
- **METEOR**: Fused 0.7374 (good semantic matching)
- Lower BLEU expected (synthesizes novel text, doesn't copy BLIP verbatim)

Aspect	BLIP Caption Alone	YOLO Detections Alone	Fused Narrative
Object Specificity	General ("a person")	Specific ("person 93%")	Contextual ("person seated at desk")
Spatial Relationships	Limited	None	Inferred ("person with laptop")
Naturalness	Good	Poor (list format)	Excellent (sentences)
TTS Optimization	Fair	Poor	Excellent (concise, natural)

## 3. Integration of Gemini 2.5 Flash

The integration of **Gemini 2.5 Flash** alongside FLAN-T5 provides:

- Improved context understanding and reasoning capabilities
- Better instruction following for user-specific prompts
- Reduced hallucinations through superior grounding

The fusion layer supports multiple reasoner backends (FLAN-T5-small or Gemini 2.5 Flash) via `REASONER_CKPT` configuration.

## 4. Use Case: Assistive Technology

**Requirements:** Real-time performance (<500ms), high accuracy, natural speech output, reliability

**Fusion Benefits:**

1. Comprehensive descriptions combining 'what' (caption) with 'where/what exactly' (detections)
2. Reduced ambiguity through multiple complementary signals
3. Natural speech output optimized for text-to-speech
4. Context-aware reasoning that infers spatial relationships

## 5. Conclusion

The multimodal fusion architecture in VISOR successfully combines object detection and image captioning to produce superior scene descriptions. Key achievements:

- ✓ Higher descriptive accuracy than individual models
- ✓ Natural, TTS-optimized outputs for assistive applications
- ✓ Modular, extensible design for future enhancements
- ✓ Real-time performance suitable for mobile/edge devices

The integration of Gemini 2.5 Flash further enhances reasoning quality, demonstrating the flexibility of the fusion approach.

*Report Generated: 2025 | Models: YOLOv8 (n/s/m), BLIP-base, FLAN-T5-small, Gemini 2.5 Flash | Dataset: COCO128 validation | Graphs: run/yolov8{n,s,m}/*