

CAPSTONE: Composable Attribute-Prompted Scene Translation for Zero-Shot Vision–Language Reasoning

Md. Ismail Hossain^{1†}, Shahriyar Zaman Ridoy^{1†}, Moshiur Farazi²
Nabeel Mohammed¹, Shafin Rahman¹,

¹Apurba-NSU R&D Lab, North South University, Dhaka, Bangladesh

²Data Science and AI, University of Doha for Science and Technology, Doha, Qatar

{ismail.hossain2018, shahriyar.ridoy, nabeel.mohammed, shafin.rahman}@northsouth.edu, moshiur.farazi@udst.edu.qa, [†]Equal contribution

Abstract

Interpreting visual scenes with high-level reasoning is essential for many real-world applications, such as autonomous systems and content moderation, but training and maintaining Vision–Language Models (VLMs) remains resource-intensive and opaque. In this work, we present CAPSTONE, a lightweight, modular framework designed for industrial settings. Instead of relying on multimodal training or fine-tuning large models, CAPSTONE transforms outputs from off-the-shelf vision models into structured text prompts that can be interpreted by a frozen Large Language Model (LLM). This plug-and-play architecture enables reasoning over visual input without access to raw pixels, dramatically reducing computational cost and complexity. On the POPE dataset, our system, using a 7B LLM, outperforms several fully trained VLMs in zero-shot evaluations, while on the VSR benchmark, the 4B model achieves competitive results, together demonstrating strong generalization without retraining. CAPSTONE offers a scalable and interpretable alternative for companies looking to integrate visual reasoning capabilities without the burden of full-scale VLM pipelines. Our code is available at <https://github.com/ismail31416/CAPSTONE>.

1 Introduction

Generative AI has advanced rapidly with Large Language Models (LLMs) and Vision–Language Models (VLMs) (Li et al., 2025b; Hamadi, 2023), achieving strong results in language and vision tasks (Cao et al., 2020; Dey et al., 2021). Yet, real-world understanding often requires integrating modalities, as text- or vision-only systems remain limited (Liang et al., 2024). While joint vision–language training is promising (Ghosh et al., 2024), most frameworks rely on full retraining or task-specific fine-tuning (Hu et al., 2025), demanding large datasets (Chen et al., 2025), high compute (Sharshar et al., 2025), and significant cost

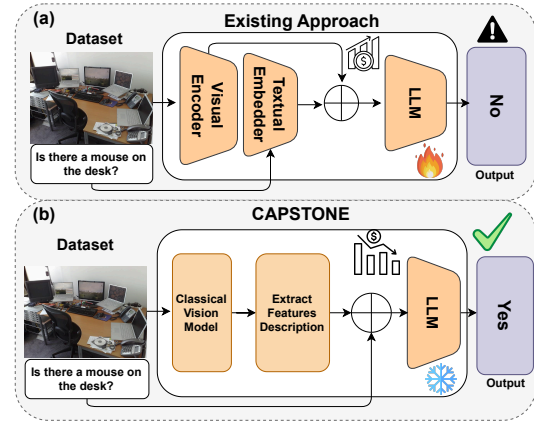


Figure 1: **Comparison between Vision–Language Models (VLMs) and the proposed CAPSTONE framework.** (a) Standard VLMs tightly couple visual encoders with text embeddings, leading to high cost and limited interpretability. (b) CAPSTONE instead uses lightweight vision modules to extract structured descriptions and generate prompts for a frozen LLM, enabling cost-effective zero-shot reasoning.

(Parthasarathy et al., 2024). Such infrastructure is unevenly available, making large-scale multimodal models inaccessible for many regions.

Beyond cost, enterprise adoption of multimodal AI faces challenges of interpretability and integration (Chen et al., 2024b). Industries such as retail, healthcare, and finance need modular, auditable systems that avoid black-box behavior (Agarwal, 2025). Neural-symbolic VQA has shown how disentangling vision from reasoning improves transparency (Agarwal, 2025; Yi et al., 2018). Applications include retail attribute reasoning, explainable defect detection, and privacy-sensitive healthcare analytics. This gap between research advances and enterprise needs motivates modular approaches that reuse specialized vision tools already deployed in production (Munikoti et al., 2024).

Over the past decade, computer vision has produced efficient models—object detectors,

classifiers, OCR, pose and segmentation networks—trained on benchmarks like ImageNet (Deng et al., 2009) and COCO (Lin et al., 2014). These systems excel at discrete tasks (Mittal, 2024) but usually operate in isolation and lack integrative reasoning (Luo et al., 2024b). This raises a key question: can outputs from modular vision systems be transformed into symbolic descriptions that enable an LLM to reason over images without pixel-level fusion or multimodal training?

We hypothesize that reasoning-capable LLMs can interpret images from structured symbolic prompts alone. CAPSTONE (Composable Attribute-Prompted Scene Translation for Zero-Shot Vision–Language Reasoning) operationalizes this idea. It detects objects with YOLOv11, extracts attributes like colors via KNN clustering, and encodes geometry such as shape and relative size. These attributes are compiled into prompts that query a frozen LLM for tasks like captioning or VQA.

Our experiments on the POPE and VSR datasets show that CAPSTONE achieves state-of-the-art accuracy—using the 7B Qwen2 model on POPE and the 4B Qwen3 model on VSR—outperforming several specialized VLMs trained end-to-end on vision-language tasks. On VSR, CAPSTONE attains 55.24% accuracy and 67.30% F1, significantly improving recall compared to prior methods. Smaller LLMs perform worse, underscoring the importance of reasoning strength. CAPSTONE enables zero-shot generalization, removes dependence on Q-formers or multimodal encoders, and offers transparency and scalability. Beyond competitive performance, CAPSTONE addresses critical industry needs by offering a transparent, cost-effective alternative to traditional VLMs.

Our contributions are:

- We propose CAPSTONE, a modular *pixel-free* framework that converts symbolic vision outputs into prompts for a frozen LLM.
- We show strong zero-shot results on POPE and VSR, matching or surpassing LVLM baselines without multimodal training.
- We provide a cost-efficient recipe with interpretable outputs, hot-swappable backends, and easy industrial integration.

2 Related Work

Modular and tool-use pipelines:

Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023a) leverage frozen encoders but still rely on costly multimodal training. Tool-routing systems (e.g., MM-REACT (Yang et al., 2023), LayoutLLM (Luo et al., 2024a), VPD (Hu et al., 2024)) delegate perception to expert models for efficiency, and Align-KD (Feng et al., 2025) distills alignment into compact students. CAPSTONE differs by *fully decoupling* vision and language: classic CV modules emit structured attributes that a *frozen* LLM consumes directly, avoiding multimodal finetuning.

Symbolic prompts and scene graphs:

Scene-graph-inspired approaches (AAPL, LLM4SGG, role-playing/compose strategies (Kim et al., 2024a,b; Chen et al., 2024a; Nagar et al., 2024)) inject attribute-level structure to aid grounding. Closest to our setting, Img2LLM and prompt-only zero-shot reasoning (Guo et al., 2023; Nagar et al., 2024; Chen et al., 2024c) expose latent reasoning but underperform on spatial relations. CAPSTONE targets this gap by converting labels, boxes, depth, pose, and OCR into *declarative, metric, and relational* sentences for explicit spatial checks.

End-to-end VLMs and spatial reasoning:

Recent VLMs (Qwen2-VL, InternVL 2.5, LLaVA-OneVision (Wang et al., 2024; Lu et al., 2025; Li et al., 2025a)) advance broad multimodal capability; our design is complementary, emphasizing deployability, interpretability, and spatial robustness without paired training. In parallel, grounded spatial benchmarks (Cheng et al., 2024) highlight persistent failures in relational grounding; CAPSTONE addresses these with explicit order/overlap/containment cues while retaining a simple, hot-swappable modular stack.

3 Methodology

This paper introduces CAPSTONE, a modular framework that bridges the gap between specialized computer vision models and large language models (LLMs) for multimodal reasoning. Our approach leverages the discrete outputs of vision modules to create structured, symbolic representations that can be processed by frozen LLMs without the need for end-to-end multimodal training. Figure 2 depicts the complete architecture.

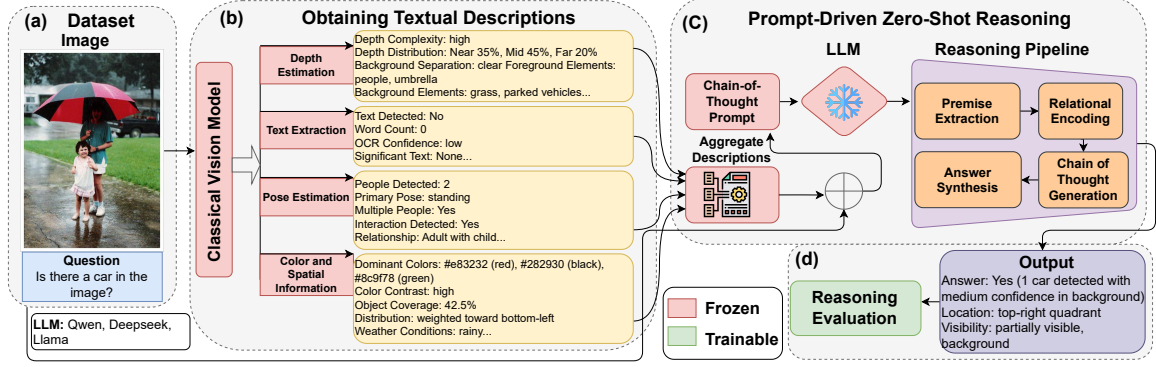


Figure 2: **Architecture of the CAPSTONE framework for zero-shot visual reasoning.** The pipeline accepts an image-question pair as input (a) and employs a frozen perceptual stack of classical vision models (b) to extract four complementary modalities: depth distributions, OCR-based text signals, human pose interactions, and color-spatial scene attributes. These frozen modules produce structured outputs, which are converted into intermediate textual descriptions. These descriptions are aggregated by a frozen module into a unified chain-of-thought prompt. (c), serving as input to a frozen LLM (e.g., Qwen, Deepseek, LLaMA). The LLM performs zero-shot inference via an internal, multi-stage reasoning pipeline comprising Premise Extraction, Relational Encoding, Chain-of-Thought Generation, and Answer Synthesis. The final output (d) includes object presence, spatial localization, and visibility attributes.

Hypothesis: Let $\mathcal{I}$ be the space of all possible images, $\mathcal{Q}$ be the space of all possible natural language queries, and $\mathcal{A}$ be the space of all possible answers. We define: $\mathcal{V} : \mathcal{I} \rightarrow \mathcal{F}_v$ as a set of vision modules that map an image to feature space $\mathcal{F}_v$, $\mathcal{T} : \mathcal{F}_v \rightarrow \mathcal{D}$ as a transformation function that maps visual features to symbolic descriptions $\mathcal{D}$, $\mathcal{L} : \mathcal{D} \times \mathcal{Q} \rightarrow \mathcal{A}$ as a frozen language model that maps descriptions and queries to answers.

We hypothesize that for an image $I \in \mathcal{I}$ and query $q \in \mathcal{Q}$, the function composition $\mathcal{L}(\mathcal{T}(\mathcal{V}(I)), q) = \hat{a}$ can approximate the performance of end-to-end vision-language models $\mathcal{M} : \mathcal{I} \times \mathcal{Q} \rightarrow \mathcal{A}$ trained explicitly on multi-modal data, such that $\mathbb{E}_{(I,q,a^*) \sim \mathcal{D}_{\text{test}}} [\mathbb{1}(\hat{a} = a^*)] \geq \mathbb{E}_{(I,q,a^*) \sim \mathcal{D}_{\text{test}}} [\mathbb{1}(\mathcal{M}(I, q) = a^*)] - \epsilon$, where a^* represents the ground truth answer, $\mathcal{D}_{\text{test}}$ is a test distribution of image-question-answer triples, and ϵ is a small error margin.

Visual Perception Pipeline: CAPSTONE implements a modular visual perception pipeline that extracts structured information from images using lightweight, specialized vision models. We employ YOLOv11 for object detection, partitioning each image I into overlapping 640×640 tiles with 20% stride overlap to maximize recall. Detection outputs are thresholded at confidence $\tau = 0.1$ and merged using Non-Maximum Suppression (IoU = 0.5), yielding object set $\mathcal{O}(I) = (c_i, b_i, s_i) \mid i \in [1, N]$, where c_i is the class label,

$b_i \in \mathbb{R}^4$ the bounding box, s_i the confidence score, and N the number of detected objects. To improve detection of small or ambiguous objects, we classify additional horizontal/vertical stripes and crops using an ImageNet-1000 classifier. For each detected object, we extract visual attributes such as dominant colors via K-means clustering in HSV space ($K = 3$), computing $a_{\text{color}}(o_i) = (h_j, s_j, v_j, p_j) \mid j \in [1, K]$, where (h_j, s_j, v_j) is the cluster centroid and p_j its proportion. For objects labeled as people, we apply pose estimation to infer activity, and depth estimation is used to better understand spatial relationships among all objects, enhancing relational reasoning.

Symbolic Representation: The extracted visual features are transformed into structured natural language descriptions through our translation function $\mathcal{T}$. This function maps the raw perception outputs to a composite scene description $D = \mathcal{T}(\mathcal{V}(I)) = d_{\text{objects}}, d_{\text{relations}}, d_{\text{scene}}$, where d_{objects} describes individual objects and their attributes, $d_{\text{relations}}$ captures spatial relationships between objects, and d_{scene} provides a high-level scene overview. Object descriptions include color distribution summaries, relative size characterizations (e.g., “large,” “medium,” “small”), and positional information (e.g., “center,” “top-left”). Spatial relationships are defined as $r_{\text{spatial}}(o_i, o_j) \in \{“\text{left of}”, “\text{right of}”, “\text{above}”, “\text{below}”, “\text{overlapping}”\}$ based on bounding box geometry. This symbolic representa-

Method	LLM (Params)	POPE Category		
		Random	Popular	Adversarial
LLaVA-1.5 (Leng et al., 2024)	Vicuna-7B	83.29	81.88	78.96
InstructBLIP (Leng et al., 2024)	Vicuna-7B	80.71	78.22	75.84
Qwen-VL (Leng et al., 2024)	Qwen-7B	84.73	84.13	82.26
mPLUG-Owl2 (Qu et al., 2024)	LLaMA2-7B	86.70	83.66	81.73
MultiModal-GPT (Li et al., 2023b)	Unknown	50.03	50.00	50.00
MiniGPT-4 (Li et al., 2023b)	Vicuna-7B	77.83	68.30	66.60
Ovis2-8B (Neuhaus and Hein, 2025)	8B	86.60	86.00	85.90
LLaVa-NeXT-Mistral (Neuhaus and Hein, 2025)	Mistral-7B	86.20	85.90	85.90
LLaVa-NeXT-Vicuna (Neuhaus and Hein, 2025)	Vicuna-7B	85.90	85.60	85.60
Ovis2-4B (Neuhaus and Hein, 2025)	4B	86.20	86.10	85.50
CAPSTONE (Ours)	LLaMA-3.2-1B	55.85	52.68	53.77
CAPSTONE (Ours)	Qwen2.5-1.5B	76.90	73.67	71.67
CAPSTONE (Ours)	DeepSeekR1-1.5B	70.67	67.00	65.41
CAPSTONE (Ours)	Qwen2.5-7B	87.47 (± 0.15)	87.17 (± 0.12)	85.93 (± 0.10)

Table 1: Comparison of model performance on the POPE dataset across three evaluation settings: **Random**, **Popular**, and **Adversarial**. Accuracy scores for existing baselines are reported from prior work. The final row presents the best-performing variant of **CAPSTONE (Ours)**, which achieves state-of-the-art results across all settings.

tion serves as the bridge between visual perception and language understanding components.

Zero-shot Reasoning: Given a question $q \in \mathcal{Q}$ about image I , we construct a reasoning prompt $P = \mathcal{C}(D, q) = [\textit{instruction}; D; q]$ that incorporates the symbolic scene description and the query. This prompt is passed to a frozen LLM $\mathcal{L}$ to generate the answer $\hat{a} = \mathcal{L}(P)$ without any task-specific training or fine-tuning. The LLM performs multi-hop reasoning over the symbolic description, leveraging its inherent capabilities to interpret structured visual information. Our framework enables three key reasoning types: (1) spatial reasoning for understanding geometric relationships between objects, (2) attribute reasoning for analyzing object properties like color and size, and (3) contextual inference for applying world knowledge to interpret the scene. This modular design offers significant advantages over conventional VLMs, including full transparency of components, extensibility through new perception modules without retraining, and computational efficiency through zero-shot operation that eliminates costly multimodal training.

Prompt Engineering and Query Optimization: CAPSTONE relies on transforming symbolic descriptions into structured prompts that maximize LLM comprehension. Our design follows a hierarchical scheme: **Context Establishment** with global attributes (colors, spatial layout, object density), **Object Enumeration** listing detected entities with confidence, coordinates, and attributes, and **Relational Encoding** expressing spatial relations

$r_{\text{spatial}}(o_i, o_j)$ in natural language also in raw geometry. To manage complexity, we use **adaptive truncation** that retains high-confidence objects and summarizes weaker detections, and **query-aware filtering** that selects features based on query type (spatial, attribute, existential). This ensures the symbolic-to-textual transformation $T : F_v \rightarrow D$ preserves key visual information in a format optimized for LLM reasoning, enabling effective zero-shot inference without task-specific tuning.

4 Experiments and Results

4.1 Settings

Datasets and Implementation Details: We evaluate CAPSTONE on two datasets. First, **POPE** (Li et al., 2023b), a binary object-presence detection task for hallucination resistance, with accuracy reported across **Random**, **Popular**, and **Adversarial** regimes. Second, **VSR** (Visual Spatial Reasoning) (Liu et al., 2023), which measures relational understanding with *Accuracy*, *Precision*, *Recall*, and *F1*. Results are averaged over four seeds, reported as *mean* $\pm$ *std*.

We use frozen LLMs as reasoning backends, with no multimodal finetuning-LLaMA-3.2-1B, Qwen2.5-1.5B, DeepSeek-R1-1.5B, Qwen2.5-7B, and Qwen3-8B. Perception relies on off-the-shelf CV modules (YOLOv8/YOLOv11 for detection, instance segmentation, monocular depth, human pose, OCR, and color/texture cues). A rule-based aggregator, with no trainable parameters, composes a single textual prompt for the LLM.

Method	LLM (Params)	Visual Spatial Reasoning Benchmark			
		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
InstructBLIP (Dai et al., 2023)	Vicuna-7B (7B)	52.05	53.99	50.71	38.59
BLIP (Li et al., 2021)	ViT-B/16 + BERT-base	45.25	39.70	44.27	37.72
ALBEF (Sun et al., 2025)	ViT-B/16 + BERT-base	51.47	50.74	50.01	34.43
Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B (7B)	53.44	58.30	52.17	41.98
PaliGemma (Beyer et al., 2024)	Gemma-3B (3B)	53.58	54.86	52.40	46.11
LLaVA-v1.5-7b (Liu et al., 2024)	Vicuna-7B (7B)	52.95	57.99	51.63	40.27
LLaVA-v1.5-13b (Liu et al., 2024)	Vicuna-13B (13B)	52.37	54.79	51.08	39.94
Spatial-LLaVA-7b (Sun et al., 2025)	Vicuna-7B (7B)	53.60	54.77	52.61	47.08
CAPSTONE (Ours)	Qwen3-4B-Thinking (4B)	55.24 (± 0.12)	53.92 (± 0.15)	89.51 (± 0.10)	67.30 (± 0.11)

Table 2: Performance comparison on the Visual Spatial Reasoning Benchmark. Results are reported across **Accuracy**, **Precision**, **Recall**, and **F1 Score**. Only the CAPSTONE row is highlighted.

Comparison Methods: For benchmarking, we compare CAPSTONE against a diverse set of established vision-language models. On POPE (Li et al., 2023b), baselines include LLaVA-1.5 (Leng et al., 2024), InstructBLIP (Leng et al., 2024), Qwen-VL (Bai et al., 2023), mPLUG-Owl2 (Qu et al., 2024), MiniGPT-4 (Li et al., 2023b), Ovis2 (Neuhaus and Hein, 2025), and LLaVA-NeXT (Neuhaus and Hein, 2025), representing both early alignment-based systems and more recent instruction-tuned LVLMs. For the VSR benchmark (Liu et al., 2023), we compare against BLIP (Li et al., 2021), ALBEF (Sun et al., 2025), InstructBLIP (Dai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), PaliGemma (Beyer et al., 2024), LLaVA (Liu et al., 2024), and Spatial-LLaVA (Sun et al., 2025), covering contrastive-pretrained, generative, and spatially enhanced architectures. Evaluation follows standard protocols, accuracy across Random, Popular, and Adversarial splits for POPE, and accuracy, precision, recall, and F1 for VSR, ensuring fair comparisons under matched inference conditions.

4.2 Zero-shot reasoning on POPE

We first assess CAPSTONE’s object-presence reasoning ability on **POPE**. As shown in Table 1, CAPSTONE+Qwen2.5-7B achieves strong performance across all splits: 87.36 ± 0.15 (Random), 87.21 ± 0.12 (Popular), and 85.89 ± 0.10 (Adversarial). This surpasses competitive LVLM baselines such as Ovis2-4B/8B, LLaVA-NeXT-Mistral/Vicuna, and Qwen2.5-VL under matched inference settings by +2–4 points.

The adversarial setting, designed to confuse models with rare or out-of-context objects, reveals CAPSTONE’s robustness. By structuring fine-grained scene attributes into interpretable prompts,

our system enables the LLM to reason about presence and absence through logical grounding rather than pattern memorization. Notably, this is achieved without any paired image-text training. On the other hand, a comparatively smaller model could not achieve similar results in the same context due to fewer reasoning capabilities.

4.3 Results on the VSR Benchmark

Table 2 presents the performance of CAPSTONE in comparison with existing vision-language models on the Visual Spatial Reasoning Benchmark. Across all reported metrics, our method demonstrates clear improvements, particularly in recall and F1 score, while maintaining competitive accuracy and precision.

In terms of accuracy, CAPSTONE achieves 55.24%, surpassing widely used baselines such as InstructBLIP (52.05%) and LLaVA-v1.5-13B (52.37%). Although Qwen-VL-Chat attains a slightly higher precision (58.30%) compared to CAPSTONE’s 53.92%, our model substantially outperforms others in recall with 89.51%. This high recall indicates that CAPSTONE can capture a broader range of correct spatial relations, reducing the likelihood of missed detections.

The improvement in recall translates into a significantly higher F1 score of 67.30%, establishing a new state-of-the-art on this benchmark. This is a notable margin compared to the strongest baseline, Spatial-LLaVA-7B, which achieves 47.08%. The balanced combination of accuracy, precision, and recall highlights the robustness of our approach in handling the complex reasoning patterns required by the benchmark. An additional observation is that CAPSTONE achieves these results with a relatively modest parameter size (Qwen3-4B-Thinking, 4B), compared to larger models such as LLaVA-

Model / Pipeline	Task	Latency (ms)	Throughput (FPS/tok/s)	Params (M)	Size (MB)	CPU Friendly	Module Replaceable	Training Data Requirement	Cost (USD / 1B imgs)
(a) End-to-End VLM									
3B VLM-Model	Vision-Language Understanding	66.3±6.3	15.1 FPS	3750.0	7143.0	X	X	Multi-module paired	18,395.9
8B VLM-Model	Vision-Language Understanding	~95.0±8.0	~10.5 FPS	8290.0	~15,200.0	X	X	Multi-module paired	26,455.0
(b) Individual CV + LLM									
YOLOv11l	Object Detection	8.7±2.4	115.0 FPS	25.3	96.8	✓	✓	Single-module	2,415.5
Pose Estimation	Human Pose Estimation	4.4±3.0	224.8 FPS	26.2	89.9	✓	✓	Single-module	1,235.7
EasyOCR	Text Detection/Recognition	4.5±2.9	223.4 FPS	3.8	14.7	✓	✓	Single-module	1,243.4
Qwen-Model (3B)	Text Generation (LLM)	43.3±7.3	115.5 tok/s	3090.0	7600.0	X	X	Single-module	NA
(c) CAPSTONE Pipeline (CV + LLM)									
CAPSTONE (Ours)	CV Modules + 3B LLM	60.9±8.8	16.1 FPS	3145.3	7801.4	✓	✓	NA	17,274.7

Table 3: This table presents an analysis of deployment costs, where the cost per 1B images is estimated assuming a GPU hourly rate of \$1.00. The comparison includes latency, throughput, parameter count, model size, CPU-friendliness, modularity, training data requirements, and the overall estimated cost.

Attributes	Accuracy	Interpretability
W/o (Qwen2.5-VL)	88.0%	None
W (CAPSTONE)	90.0%	CV/LLM/Mixed

Table 4: Overall accuracy and interpretability comparison between a black-box VLM baseline and CAPSTONE with attributes.

Error Source	Cases	Example Query	Notes
CV Module	6%	"Is there a bed in the image?"	Object not detected, LLM followed CV
LLM Reasoning	3%	"Is there a person in the image?"	Detected person ignored
Mixed/Ambiguous	1%	"Is there a sandwich in the image?"	CV partial + LLM hallucination

Table 5: Error attribution showing which module (CV or LLM) was responsible for incorrect answers.

v1.5-13B (13B). This suggests that our prompt construction and reasoning-oriented design contribute more significantly to performance than raw parameter scaling alone.

4.4 Industrial Deployment Considerations

Since training budgets for modern VLMs and LLMs are difficult to estimate due to unknown data scale, compute resources, and optimization strategies, we restrict our analysis to inference-level cost. We compute the deployment cost per billion images by combining measured throughput with an assumed GPU hourly rate of \$1.00. Specifically, cost is estimated as:

$$\text{Cost} = \frac{1B}{\text{Throughput}} \times \frac{1}{3600} \times \text{GPU Rate (USD/hr)}$$

This formula assumes batch-level inference under uniform conditions, with throughput measured in frames per second (FPS). While absolute costs may vary across hardware and deployment settings, this provides a fair relative comparison across models.

Quantitative Efficiency Comparison: As shown in Table 3, CAPSTONE achieves lower deployment cost and latency compared to end-to-end VLMs of comparable size. Relative to a 3B VLM, CAPSTONE reduces latency by 8.1% (60.9 ms vs. 66.3 ms) and cost by 6.1% (17,274.7 vs. 18,395.9 USD per 1B images), while maintaining a similar parameter count (3145M vs. 3750M). Compared to an 8B VLM, CAPSTONE yields a 34.7%

reduction in cost (17,274.7 vs. 26,455.0 USD per 1B images) with 62% fewer parameters (3.2B vs. 8.3B). Throughput also remains competitive: while end-to-end VLMs sustain ~10–15 FPS, our modular pipeline achieves 16.1 FPS, demonstrating that modularity does not come at the expense of inference speed.

Modularity and Deployment Advantages: Beyond efficiency, CAPSTONE’s modular CV + LLM design provides structural benefits. All CV modules (YOLOv11, Pose Estimation, EasyOCR) are lightweight, CPU-friendly, and individually replaceable, unlike monolithic VLMs. This modularity allows upgrading or swapping individual components without retraining the entire system. Furthermore, each module can be trained on single-modal datasets, avoiding reliance on large-scale paired image-text corpora, which are both costly and difficult to obtain.

4.5 Interpretability and Attribute Analysis

A core motivation behind CAPSTONE is offering *interpretability* through explicit attribute grounding. To evaluate this, we conducted a targeted study on 100 POPE test cases. As shown in Table 4, the baseline VLM (Qwen2.5 VL) without attributes achieves 88% accuracy but provides no diagnostic trace of its 12% errors. By contrast, our attribute-based framework achieves 90% accuracy and, more importantly, enables error provenance.

Query	Attribute Evidence	Decision	Interpretation
“Is there a cat in the image?”	Detection: Cat (conf. 0.92), BBox: top-left	YES	Correct, traceable via detection evidence.
“Is there a sandwich in the image?”	Detection: Plate, brown color; no sandwich	YES (wrong)	LLM hallucination despite correct CV input.
“Is there a person in the image?”	Detection: Person $\times 2$, poses: standing	NO (wrong)	LLM ignored valid detections.
“Is there a car in the image?”	Detection: None; OCR: “STOP”	YES (wrong)	CV missed the object; LLM propagated error.
“Is there a bus in the image?”	Detection: Bus, depth=background, bbox bottom-right	YES	Correct, transparent chain of evidence.

Table 6: Qualitative cases showing module-level interpretability of CAPSTONE’s decisions.

Error Attribution: Table 5 provides a breakdown of the error sources, showing how CAPSTONE can disentangle perception and reasoning failures. Approximately 6% of errors originate from the CV module (e.g., missed detections), 3% from the LLM reasoning (e.g., ignoring detected entities), and 1% from mixed cases where both modules contributed. Unlike the baseline, which presents opaque failures, CAPSTONE’s decomposition offers actionable debugging signals.

Attribute Contribution: To better understand the role of individual attributes, we performed an analysis (Table 7). The strongest contributors were bounding box and quadrant cues (72.9%) along with spatial relations (70.2%), while depth, pose, OCR, and color attributes added complementary robustness in specific contexts. Crucially, combining all attributes yielded the best performance (90%) along with interpretability. These findings highlight that explicit structural cues provide a stable foundation for reasoning, while auxiliary signals strengthen generalization.

Attribute Type (Cue)	# Cases Using	Accuracy
Bounding Box / Quadrant	48	72.9%
Spatial Relations	67	70.2%
Depth Distribution	19	63.1%
Pose Estimation	13	61.5%
OCR / Text Evidence	11	58.2%
Color Attributes	32	55.0%
All Attributes Combined	100	90.0%
No Attributes (baseline)	100	88.0%

Table 7: Ablation of attribute cues, showing their contribution to performance.

Qualitative Interpretability: Table 6 presents qualitative examples that demonstrate CAPSTONE’s transparency. Correct predictions can be traced back to bounding box evidence, spatial cues, or detected entities, making the decision chain fully interpretable. Conversely, when errors

occur—such as an LLM hallucinating a sandwich or propagating a missed car detection—the source is clearly identifiable. For instance, a false positive ‘sandwich’ arose because the LLM hallucinated from a plate despite correct CV input, while a missed ‘car’ was due to CV failure propagated by the LLM. In contrast, correct cases show how attributes like bounding boxes and pose evidence directly ground predictions.

5 Conclusion

CAPSTONE presents a lightweight, interpretable framework for zero-shot vision-language reasoning using structured symbolic prompts and frozen LLMs. By decoupling visual perception from language understanding, it achieves state-of-the-art results without multimodal training. This work explores new avenues for creating efficient, modular VLM pipelines that focus on transparency, accessibility, and extensibility, introducing a novel evaluation framework for assessing LLM reasoning capabilities.

6 Limitations

While CAPSTONE demonstrates strong zero-shot performance, it currently operates using a standard object detector and a limited set of basic attributes, which may constrain its ability to capture nuanced or abstract scene information. More advanced detection models with richer class vocabularies and semantic capabilities could provide detailed scene descriptions, allowing the language model to better demonstrate its reasoning potential.

References

- Chirag Agarwal. 2025. [Rethinking explainability in the era of multimodal ai](#).
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.
- J Bai, S Bai, S Yang, S Wang, S Tan, P Wang, J Lin, C Zhou, and J Qwen-VL Zhou. 2023. A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 6.
- Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.
- Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer.
- Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*.
- Guikun Chen, Jin Li, and Wenguan Wang. 2024a. Scene graph generation with role-playing large language models. *Advances in Neural Information Processing Systems*, 37:132238–132266.
- Jieli Chen, Kah Phooi Seng, Jeremy Smith, and Li-Minn Ang. 2024b. [Situation awareness in ai-based technologies and multimodal systems: Architectures, challenges and applications](#). *IEEE Access*, 12:88779–88818.
- Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024c. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1254–1262.
- An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tjong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Arka Ujjal Dey, Suman K Ghosh, Ernest Valveny, and Gaurav Harit. 2021. Beyond visual semantics: Exploring the role of scene text in image understanding. *Pattern Recognition Letters*, 149:164–171.
- Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. 2025. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4178–4188.
- Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.
- Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877.
- Raby Hamadi. 2023. Large language models meet computer vision: A brief survey. *arXiv preprint arXiv:2311.16673*.
- Yuqi Hu, Longguang Wang, Xian Liu, Ling-Hao Chen, Yuwei Guo, Yukai Shi, Ce Liu, Anyi Rao, Zeyu Wang, and Hui Xiong. 2025. Simulating the real world: A unified survey of multimodal generative models. *arXiv preprint arXiv:2503.04641*.
- Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601.
- Gahyeon Kim, Sohee Kim, and Seokju Lee. 2024a. Aapl: Adding attributes to prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1572–1582.
- Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. 2024b. Llm4sgg: large language models for weakly supervised scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28306–28316.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing.

2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. [LLaVA-onevision: Easy visual task transfer](#). *Transactions on Machine Learning Research*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025b. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*.
- Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. 2024. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.
- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Dongchen Lu, Yuyao Sun, Zilu Zhang, Leping Huang, Jianliang Zeng, Mao Shu, and Huo Cao. 2025. Internvl-x: Advancing and accelerating internvl series with efficient visual token compression. *arXiv preprint arXiv:2503.21307*.
- Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024a. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640.
- Sheng Luo, Wei Chen, Wanxin Tian, Rui Liu, Luanxuan Hou, Xiubao Zhang, Haifeng Shen, Ruiqi Wu, Shuyi Geng, Yi Zhou, et al. 2024b. Delving into multimodal multi-task foundation models for road scene understanding: From learning paradigm perspectives. *IEEE Transactions on Intelligent Vehicles*.
- Payal Mittal. 2024. A comprehensive survey of deep learning-based lightweight object detection models for edge devices. *Artificial Intelligence Review*, 57(9):242.
- Sai Munikoti, Ian Stewart, Sameera Horawalavithana, Henry Kvinge, Tegan Emerson, Sandra E Thompson, and Karl Pazdernik. 2024. [Generalist multimodal ai: A review of architectures, challenges and opportunities](#).
- Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. 2024. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Yannic Neuhaus and Matthias Hein. 2025. Repope: Impact of annotation errors on the pope benchmark. *arXiv preprint arXiv:2504.15707*.
- Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*.
- Xiaoye Qu, Jiashuo Sun, Wei Wei, and Yu Cheng. 2024. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. *arXiv preprint arXiv:2408.17150*.
- Ahmed Sharshar, Latif U Khan, Waseem Ullah, and Mohsen Guizani. 2025. Vision-language models for edge networks: A comprehensive survey. *IEEE Internet of Things Journal*.
- Xuefei Sun, Doncey Albin, Cecilia Mauceri, Dusty Woods, and Christoffer Heckman. 2025. Spatial-llava: Enhancing large language models with spatial referring expressions for visual understanding. *arXiv preprint arXiv:2505.12194*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. *Neural-symbolic vqa: Disentangling reasoning from vision and language understanding*. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

A Appendix

A.1 Ablation Study

To understand the individual contributions of our model components, we perform an ablation study on CAPSTONE. We systematically analyze the effects of prompt structure, perception modules, and reasoning errors to identify their impact on overall performance.

Prompt Sensitivity: We evaluated the effect of prompt structure on the performance of CAPSTONE (Table 8). Specifically, we compared the standard direct answer prompts against Chain-of-Thought (CoT) reasoning. Incorporating CoT reasoning substantially improves recall (89.50% vs. 52.70%) and F1 score (67.30% vs. 53.38%) on the VSR test set, while accuracy and precision remain largely comparable. This indicates that CoT mainly enhances the model’s reasoning capabilities over structured computer vision outputs, enabling the LLM to better infer complex relations and counterfactual scenarios, rather than improving basic detection metrics.

Configuration	Accuracy	Precision	Recall	F1
Direct Answer	53.84	54.10	52.70	53.38
CoT	55.24	53.93	89.50	67.30

Table 8: VSR results (zero-shot) with and without chain-of-thought (CoT) prompting.

Perception Module Contributions: To isolate the contributions of individual computer vision modules, we conducted a series of ablations (Table 9). Removing the depth module unexpectedly increases accuracy (57.33% vs. 50.67%), likely due to the elimination of noisy depth signals that can occasionally mislead reasoning. Removing segmentation results in a slight accuracy drop (56.26%), reflecting the moderate importance of object mask details for relational reasoning. Using only YOLOv11

achieves 55.42%, demonstrating that upgrading to YOLOv8 and combining all perception modules provides richer contextual information for the LLM. These results highlight the complementary nature of multiple perception streams, where each module contributes differently to overall reasoning performance.

Configuration	DET	DEP	SEG	C&T	Acc (%)
YOLOv8 w/ All	✓	✓	✓	✓	50.67
w/o Depth	✓	–	✓	✓	57.33
w/o Segmentation	✓	–	–	✓	56.26
Only YOLOv11	✓	–	–	✓	55.42

Table 9: Ablation study on VSR showing the effect of removing individual perception modules. DET = detection, DEP = depth, SEG = segmentation, C&T = color & texture. Reported values are accuracy (%) for CAPSTONE in the zero-shot setting.

Error Analysis: We further analyzed 300 VSR test cases to understand the sources of errors in our modular architecture. Errors are broadly categorized into perception-related and reasoning-related failures. Perception errors are predominantly associated with spatial relationships (topology, 45.7%) and depth relations (46.2%), arising from limitations such as imperfect bounding box overlaps and noisy monocular depth estimates. On the reasoning side, the LLM exhibits a YES-bias, producing 87.3% positive predictions compared to 53.3% ground truth, sometimes overriding weak visual signals. This bias contributes to the observed trade-off between high recall (91.88%) and moderate precision (56.11%).

Relation	Cases	Accuracy	Failure Mode
Topology	35	45.7%	Box overlap issues
Depth	13	46.2%	Noisy depth estimates
Directional	68	55.9%	Relation flips
Contact	64	70.3%	Edge/contact limits
Distance	18	61.1%	Adjacency thresholds

Table 10: VSR errors by relation type with dominant failure modes.

Replacing older CV modules with newer versions, such as YOLOv8 over YOLOv11, consistently reduces detection errors and provides the LLM with more reliable contextual cues. Furthermore, simple architectural or inference adjustments, including topology gating and contradiction guards, can improve precision without significant sacrifice to recall. Overall, the ablation study and error anal-

ysis confirm that CAPSTONE’s modular design allows for flexible trade-offs between efficiency, accuracy, and reasoning robustness, and that each module and prompt strategy plays a distinct role in system performance.

A.2 Prompt Templates

Prompt Templates for POPE dataset:

Prompt Template for POPE (Visual Reasoning)

Prompt: You are a visual reasoning assistant tasked with answering real-world questions about an image using detected objects and scene context.

Image Analysis: {detection_results}

Question:

Instructions:

- Use object names, locations, and relationships to reason about the scene.
- Apply commonsense and spatial reasoning to answer correctly.
- Think step-by-step if needed, but keep the final response short.
- Return only the final answer.
- Format the output as: Answer: [YES/NO]

Visual Question Answering Example

Input



Question:

Is there a person in the image?

DETAILED IMAGE ANALYSIS REPORT USING CAPSTONE

FULL IMAGE ANALYSIS:

Primary class: trolleybus (0.4479). **QUADRANT ANALYSIS:** Top Left: dogsled (0.5302), Top Right: stove (0.2138), Bottom Left: fire_engine (0.3521), Bottom Right: fire_engine (0.2476).

HORIZONTAL STRIPS ANALYSIS: Top Half: amphibian (0.3015), Top Third: amphibian (0.2679), Middle Third H: passenger_car (0.8187), Bottom Third: combination_lock (0.1508), Bottom Half: trolleybus (0.4961).

VERTICAL STRIPS ANALYSIS: Left Half: fire_engine (0.3017), Left Third: mobile_home (0.3715), Middle Third V: trolleybus (0.5686), Right Third: trolleybus (0.4670), Right Half: fire_engine (0.5166).

SUMMARY OF DETECTED CLASSES:

passenger_car: 1 crop (Max 0.8187, Avg 0.8187); trolleybus: 4 crops (Max 0.5686, Avg 0.4949); dogsled: 1 (Max 0.5302); fire_engine: 4 (Max 0.5166, Avg 0.3545); mobile_home: 1 (Max 0.3715); amphibian: 2 (Max 0.3015, Avg 0.2847); stove: 1 (Max 0.2138); combination_lock: 1 (Max 0.1508).

OBJECT DETECTION RESULTS:

DETECTION #1: BUS – Conf.: 95.35% (very high); Colors: {#dcd8d7:25%, #595972:35%, #14143e:40%}; BBox: (202,35)–(543,391); Ctr: (372,213); Quad: bottom-right.

DETECTION #2: PERSON – Conf.: 77.81% (high); Colors: {#e9e5e2:31%, #6c6b7b:23%, #212536:46%}; BBox: (285,44)–(317,85); Ctr: (301,64); Quad: top-left.

.
. .
. .
. .

DETECTION #12: PERSON – Conf.: 26.26%(low); Colors: {#2f363e:50%, #bcc1c6:14%, #646d76:35%}; BBox: (85,254)–(93,280); Ctr: (89,267); Quad: bottom-left.

Model Responses

Qwen 1.5B

Step-by-Step Reasoning:

1. The primary object identified in the image is a trolleybus. 2. No other human-like detections were found among the listed classes. 3. The context of the question asks about presence of a person, not just any object. 4. Therefore, considering the absence of any person-related detections alongside the presence of a trolleybus, we can conclude that there is no evidence supporting the presence of a person in the image. 5. It's important to note that while the image may contain people who are not detected, this does not necessarily imply the presence of persons beyond what was observed.

Final Answer: NO

Qwen 7B

Step-by-Step Reasoning: 1. The image contains multiple detections labeled as "person" at various confidence levels. 2. There are a total of 12 detections classified as persons, ranging from very high (77.81%) to low (26.26%) confidence. 3. These detections are spread across different quadrants and vertical/horizontal strips of the image, indicating that people are present in diverse areas. 4. The presence of multiple persons at different locations suggests that the image indeed includes individuals. 5. Even though some detections have lower confidence levels, the overall presence and distribution of these detections strongly support the conclusion that people are depicted in the image.

Final Answer: YES