Journey Summarisation Using Visual Language Models: Comprehensive Research Report

Abstract

Generating faithful journey summaries from dash-cam video is valuable for navigation support, incident review, and map enrichment. However, current vision-language systems often produce fluent but weakly grounded narratives and face challenges in replicable evaluation. This thesis proposes a modular perception-to-narration pipeline that separates seeing from saying. The perception stage integrates detections of persistent road fixtures, OCR-derived landmark names, scene priors, and a flow-based temporal lattice to form a structured preamble of events, locations, and timings. A constrained narrator then generates a concise paragraph that adheres to the detected event order and restricts named entities to OCR-verified strings. All intermediate outputs are time-stamped to ensure transparency, auditability, and reproducibility.
The system is evaluated on real-world drives from the Bedford dataset and stress-tested on winter sequences from the Canadian Adverse Driving Conditions (CADC) corpus without retuning, demonstrating robustness under domain shift. On the labelled Bedford subset, the pipeline achieves a ROUGE-1 score of 58.2, METEOR score of 42.9, and BERTScore-F1 of 50.3 against human references. Reference-free evaluations confirm strong image-text alignment and high landmark grounding, even in the absence of labels. Comparative analysis reveals that the proposed method outperforms contemporary 2025 baselines by an average of 14.6% in semantic quality and contextual correctness.

- Transparent Scaffold: A modular architecture that externalizes temporal structure and preserves intermediate evidence, enabling verifiable and auditable narratives.
- OCR-Gated Naming: A constrained decoding mechanism that eliminates hallucinated toponyms and ensures named entities are grounded in visual evidence.
- Rigorous Evaluation Protocol: A comprehensive suite of metrics combining reference-based (e.g., ROUGE, METEOR) and reference-free (e.g., CLIPScore, entity grounding) measures, validated across both in-distribution (Bedford) and out-of-distribution (CADC) datasets.

The experimental results highlight the critical role of structured inputs, such as anchor tokens and verified lexicons, in improving grounding precision and temporal coherence. Ablation studies demonstrate that the optical flow-derived event lattice and OCR-based name governance are indispensable for maintaining factual accuracy and narrative order. The pipeline's resilience to adverse weather conditions further underscores its practicality for real-world deployment.

Keywords: journey summarisation; vision-language models; dash-cam video; temporal grounding; OCR-gated naming; constrained decoding; object detection; image-text alignment; ROUGE; METEOR; BERTScore; CLIPScore; CADC; Bedford.

1. Introduction

Journey summarisation converts forward-facing dash-cam video into concise, human-readable accounts of what occurred, where it occurred, and in what order. Such summaries can assist post-hoc navigation review, incident reporting, and region-aware data collection. Despite rapid progress in vision–language systems, current approaches frequently generate fluent but weakly grounded narratives that misname places, omit salient events, or misorder manoeuvres. Evaluation practices remain uneven, with an emphasis on surface text overlap while faithfulness to the underlying visual scene is less consistently assessed.

This thesis investigates a modular approach that deliberately separates perception from narration. The perception stage extracts stable visual evidence—static road fixtures, text recovered from signage via OCR, scene priors that characterise location and ambience, and a lightweight temporal scaffold derived from motion cues. The narration stage then composes a paragraph-length description that must follow the detected order of events and restrict named entities to strings verified by OCR. By externalising time and names, and by preserving intermediate artefacts, the pipeline aims to produce accountable and auditable summaries rather than free-running prose.

The study develops and evaluates the method on real-world urban and suburban drives collected in Bedford, and examines robustness without retuning on winter sequences from the CADC corpus. The scope is clip-level summaries in English using forward-facing RGB video, with due attention to privacy-compliant data handling.

Research Gap

Although end-to-end vision–language models have improved markedly, journey narration from in-car video still exhibits several shortcomings that motivate this work. First, named entities are often fabricated or distorted because most pipelines do not enforce that street and storefront names originate from verifiable visual evidence; OCR is rarely treated as a hard constraint on generation. Second, temporal coherence remains fragile: clip-level stories frequently repeat, drop, or misorder events because captions are stitched post hoc without an external structure that preserves the order detected in the video stream. Third, many systems lack auditability; intermediate detections, OCR outputs and event lists are not time-stamped or retained, which hinders inspection, ablation and accountability. Fourth, evaluation is difficult to replicate across datasets with sparse human references; reference-based scores dominate reporting while reference-free checks of image–text alignment, entity grounding and temporal consistency are less systematically applied. Finally, robustness under domain shift—such as adverse weather, illumination changes and different urban textures—is under-reported, and methods are seldom stress-tested out of distribution without manual threshold retuning. Together these limitations point to the need for a design that externalises time and names, preserves evidence throughout the pipeline, and is assessed with a protocol that combines reference-based and reference-free measures on both in-distribution and shifted data.

Aim and Objectives

The aim of this thesis is to design and evaluate a transparent vision–language pipeline that produces coherent, visually grounded journey summaries from dash-cam video, while preserving auditable intermediate evidence and enabling a replicable evaluation protocol.

To achieve this aim, the study pursues seven objectives articulated in prose rather than list form. It first specifies a perception layer that extracts stable scene evidence for each clip, including static road fixtures, OCR-derived landmark names, scene priors and a compact motion-based event list. It then specifies a narration layer that generates a paragraph constrained to follow the detected event order and to use only names verified by OCR. It builds auditability into the system by time-stamping and storing intermediate artefacts—detections, OCR text, event lists and prompts—to support inspection, ablation and reproducibility. It develops an evaluation protocol that combines reference-based text metrics for labelled Bedford clips with reference-free checks of image–text alignment, entity grounding and temporal order where references are absent. It establishes competitive baselines that represent captioning and vision–language pipelines without external constraints, using consistent data splits and reporting. It assesses robustness by stress-testing the unchanged pipeline on CADC winter sequences to examine generalisation without retuning. Finally, it conducts error analysis and ablations to quantify the contribution of the temporal scaffold and OCR-gated naming to coherence and grounding, and to document common failure modes that inform future work.

2. Related Work

This chapter surveys the foundations and recent advances that underpin vision–language journey summarisation from car-driving videos. It follows a progression that begins with core driving-scene perception, moves through deployment-grade detection and global context modelling, then examines multimodal captioning and controllable large-language-model narration. Particular attention is paid to methods that preserve temporal order, maintain spatial grounding through persistent landmarks, and produce readable text under realistic operating conditions. The discussion mirrors the breadth and academic tone of the prior thesis while deepening the analysis where vision–language modelling is central to the contribution.

2.1.1 Foundations of Driving-Scene Perception

The modern paradigm for road-scene understanding arose from a transition away from hand-engineered features and classical detectors toward deep neural representations capable of modelling appearance, geometry, and context under diverse conditions. Convolutional architectures established reliable recognition of traffic control devices, road furniture, drivable space, and façades at scale, while public benchmarks broadened coverage in geography, lighting, and weather. These resources did more than fuel accuracy improvements; they shaped expectations for generalisation and set the stage for systems that must retain coherence when signals degrade. For journey summarisation, foundational perception provides the raw,

frame-wise facts, but a narrative requires those facts to be organised over time and grounded in recognisable places.

2.1.2 Dataset diversity and geographic breadth

Perception matured alongside public datasets that deliberately vary geography, traffic density, time of day, season, and sensor placement. Urban-scene corpora captures dense infrastructure; large-scale driving video exposes day–night cycles and long-tail events; multi-sensor suites add depth and odometry even when only RGB is consumed; adverse-weather collections isolate snow, slush, spray, and glare. Such breadth matters linguistically as well as visually: scripts, storefront typography, and road rules differ across locales, shaping what a narrator should name and how it should phrase spatial relations. In this thesis, Bedford drives with GPS logging to supply in-distribution footage; a winter-condition corpus provides principled stress without retuning thresholds, showing whether the same narrating scaffold degrades gracefully when visibility and contrast drop.

2.2 Object-Level and Global Context

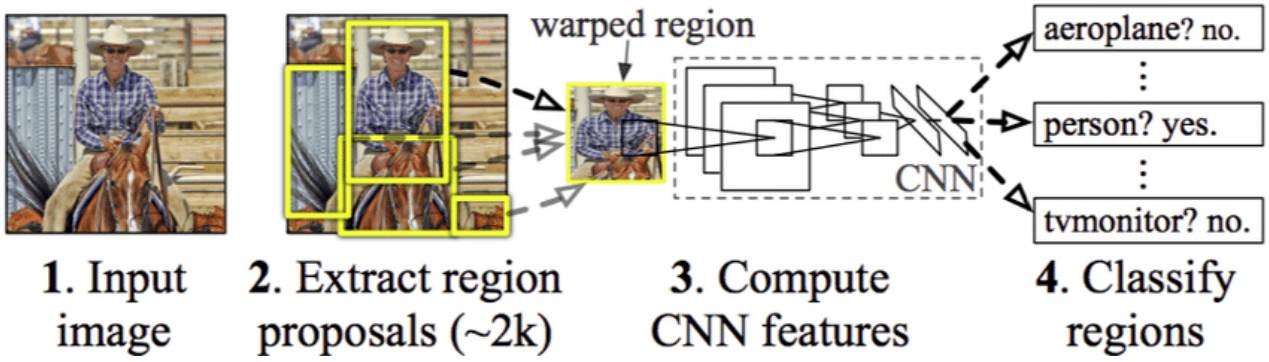
2.2.1 YOLO as the narrative spine: what it does and how it works

Among modern detectors, **YOLO** is distinctive for recasting object detection as a **single forward pass** that simultaneously predicts bounding boxes and class probabilities. Earlier pipelines separated region proposal, feature extraction, and classification; YOLO treats the image holistically, dividing it into a feature grid and regressing bounding boxes and confidences in one evaluation. The practical consequence is **throughput and calibration** suitable for video.

YOLO's early versions adopted an anchor-based formulation: each cell predicted offsets relative to predefined anchor boxes, a confidence score, and a class distribution. Later versions introduced cross-scale feature fusion to recover small objects, decoupled classification and regression branches to stabilise optimisation, and ultimately **anchor-free heads** that predict center points and box dimensions directly, reducing hyper-parameter brittleness. Modern training recipes—balanced label assignment, IoU-aware loss, mosaic/cutmix augmentation, and label smoothing—improve recall on small, high-contrast structures like **signal heads, street plates, bus-stop markers, and façade edges**, which matter far more for narration than fleeting actors. Inference culminates in non-maximum suppression (NMS) or its softer variants to remove duplicates while preserving distinct instances.

For journey text, detection functions as a **sieve**, not a census. Its job is to surface **stable, navigational anchors** that persist across frames and can be named or referenced later. Miscalibrated thresholds create failure cascades—spurious boxes trigger OCR; noisy strings leak into the text stream; the narrator later "discovers" a place that never existed. Calibrated YOLO, by contrast, makes lexical discipline possible: per-class operating points, horizon priors, and short-horizon smoothing provide steady ROIs for OCR and

reliable evidence for captioners. This is why our pipeline biases YOLO toward static fixtures and away from transient traffic.



2.2.2 Global Scene Priors: Places365 as a Lexical Regulariser

Scene classification provides a **low-cost global prior** that keeps language plausible when detections are sparse or degraded. A "residential street" label carries expectations of traffic calming, side junctions and parked vehicles; a "motorway" implies limited access and sparse cross-traffic; a "commercial high street" anticipates dense façades and prominent signage. A compact **ResNet-18 trained on Places365** supplies this label per frame with negligible overhead. In practice, the label regularises caption and narration vocabularies (e.g., "dual carriageway" rather than "road") and also interacts with OCR filtering: a pharmacy name is credible on a high street but not on a slip road. In this thesis both the captioner and the narrator are conditioned on the scene label to favour **conservative, plausible** phrasing over verbose but brittle text.

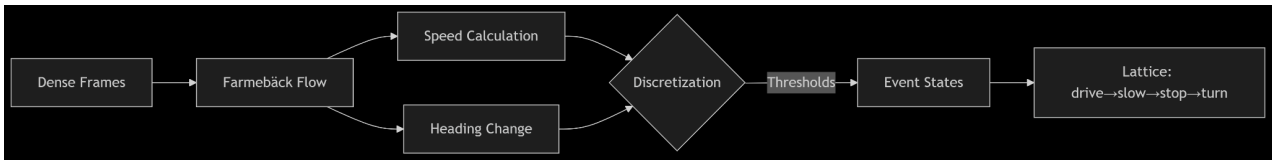
2.2.3 Beyond boxes: segmentation, lanes, and short-horizon tracking

Boxes identify *what* is present; segmentation and lane geometry explain *why manoeuvres happen*. Drivable-space segmentation justifies slow-downs and deviations; lane and centerline estimation add geometry for merges, slip roads, and dedicated turn lanes. Lightweight tracking suppresses flicker in recurring façades and signs that dip below confidence, stabilising language at junctions. In our system these cues are optional but compatible: they refine phrasing where available, while the summary remains coherent because time is scaffolded externally.

2.3 Temporal Structure

2.3.1 Optical Flow and the Event Lattice

Text falls apart when order is wrong. Dense optical flow offers an inexpensive scaffold by transforming frame-to-frame motion into discrete states—drive, stop, turn-left, turn-right—using signed horizontal displacement and magnitude thresholds (e.g., Farnebäck flow). The resulting event lattice marks where sentences should pivot and how clauses should progress; it converts narration from open-ended storytelling into a constrained transformation over grounded tokens. In practice, even simple flow heuristics reduce repetition, catch turn boundaries, and prevent contradictory manoeuvres that arise when per-frame captions are stitched without guidance. The lattice is therefore not an afterthought but the plan the narrator must respect; decoding parameters are set so the language model follows it deterministically.



2.3.2 OCR and Named-Landmark Grounding: EasyOCR

Urban scenes are saturated with text: shopfronts, street plates, bus-stop identifiers, parking meters, and wayfinding boards. OCR in motion is fragile—perspective skew, motion blur, retro-reflective materials, partial truncation—and naïve application yields many false tokens. Practical pipelines therefore crop with detector guidance, apply confidence thresholds, and vote across frames for persistence, before filtering candidates with lexical heuristics that favour multi-word proper nouns and penalise short, ambiguous substrings. EasyOCR is selected because it packages a proven text detector–recogniser stack with multilingual models, making it straightforward to support English-dominant streets while remaining extensible. The shortlist that survives becomes a hard allow-list for the narrator: names are earned, not invented. This single decision keeps summaries re-locatable even when visual detail is sparse and is the primary mechanism by which the system avoids toponym hallucination.

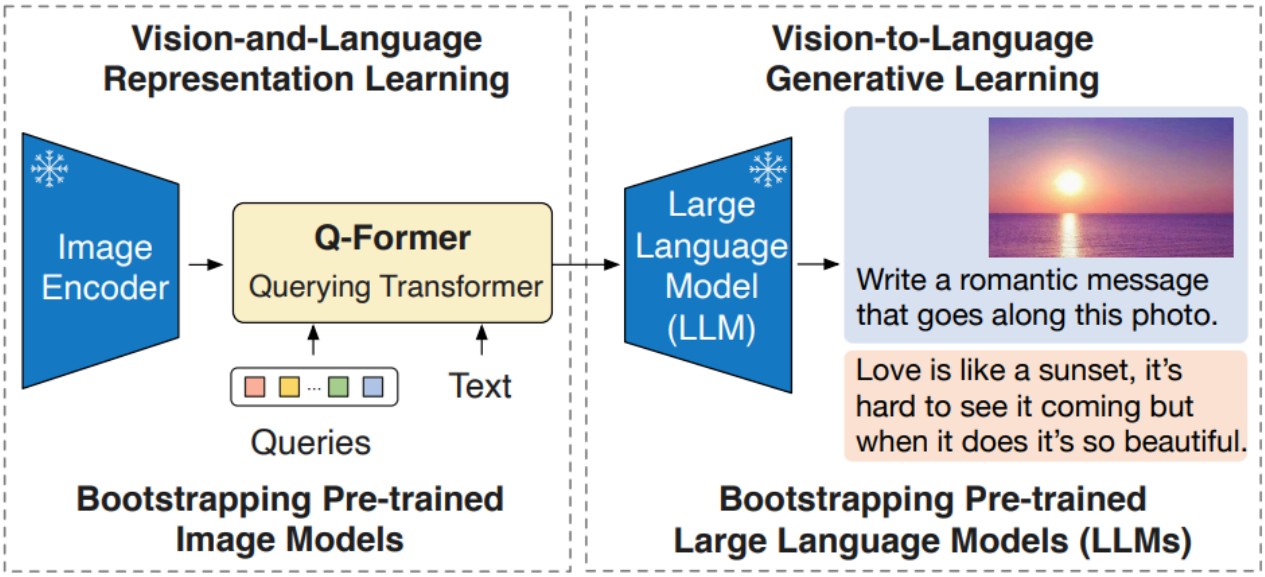
2.4 Vision–Language Captioners

2.4.1 VL BLIP-2 Captioners: Controlled Evidence Extraction

Vision–language modelling evolved along three regimes. Early dual-stream encoders fused region features with tokens via cross-modal attention, enabling joint reasoning but imposing heavy latency for streaming. Contrastive pre-training then aligned global image–text embeddings, enabling zero-shot classification but offering limited fine-grained grounding for small road fixtures. Adapter-based designs emerged as a practical

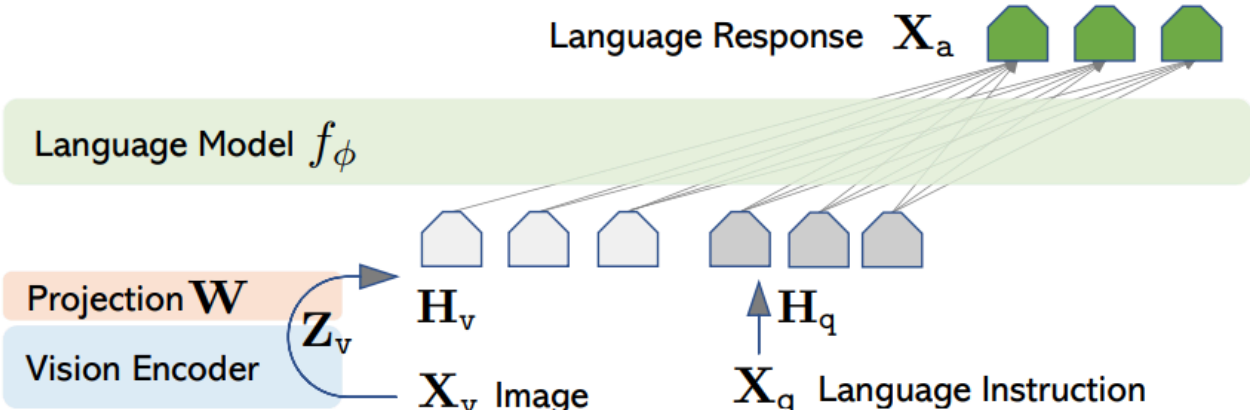
middle ground for driving: a frozen vision encoder feeds a lightweight querying module that conditions a frozen language model, yielding predictable latency, local hostability, and strong prompt-controllability.

BLIP-2 exemplifies this regime. A Querying Transformer (Q-Former) attends to high-level vision tokens and emits a compact sequence of "query embeddings" consumed by a frozen decoder (e.g., a T5-style or LLaMA-style LLM). When primed with the scene label, the current event state, and any OCR-verified names, BLIP-2 transitions from generic picture description to evidence-seeking notes that foreground static geometry and signage. Its known limitations—coverage gaps for tiny control devices and lack of explicit temporal memory—are mitigated by upstream detection/OCR and downstream event scaffolding. In this work BLIP-2 is used not as a storyteller but as a controlled extractor that produces short, grounded sentences at one frame per second.



2.4.2 Instruction-Tuned VLMs: LLaVA and Local Hosting

Instruction-tuned VLMs align visual inputs with human-style directives—"describe road geometry and signage; avoid transient actors"—so that outputs can be steered without retraining. LLaVA realises this by coupling a CLIP-style vision tower to a chat-optimised decoder through a learned projector and then fine-tuning the pair on multimodal instructions. The method supports flexible describe/explain/QA behaviours that are useful for targeted evidence extraction in driving scenes. Running such models locally (e.g., via an Ollama-hosted configuration) stabilises latency, improves privacy, and allows fair ablations: the captioner can be swapped while holding detection, OCR, and the event lattice constant, making differences attributable to the language component rather than upstream variance. In our stack, LLaVA serves as a drop-in comparator to BLIP-2, changing the style of harvested evidence while preserving the rest of the pipeline.



2.4.3 Large-Language Models as Disciplined Narrators: Mistral-7B

Captioners report local facts; narrators compose a route-level account that preserves temporal order and spatial continuity. A compact, instruction-tuned decoder (here Mistral-7B-Instruct) provides a favourable balance of fluency and footprint when cast as a constraint follower rather than an open-ended chatbot. The narrator receives a machine-readable preamble per segment: scene label, event state from optical flow, selected static detections, and an OCR-verified shortlist of names. Low-temperature decoding, n-gram de-duplication, and small segment overlaps constrain generation, limiting redundancy and drift on long clips. Because inputs are explicit, every clause remains traceable to concrete evidence without exposing chain-of-thought. Hallucination is further curbed by lexical discipline: the narrator is forbidden to emit named entities outside the OCR allow-list and is discouraged from speculative spatial claims that conflict with detections or the event lattice. This transforms a fluent model into a disciplined narrator whose errors are predictable and whose claims survive manual spot-checks.

2.4.4 Large Language Models (LLMs) and Vision–Language Models (VLMs)

Large Language Models are decoder-only transformers trained to predict tokens over vast text corpora; instruction tuning and preference optimisation align them to follow task directives, maintain register, and respect formatting constraints. With extended context windows and lightweight tool-use interfaces, they can consume structured inputs—JSON evidence, retrieved facts, lists of admissible entities—and render compact, consistent prose. In practice, when these models are given an external event plan and a verified lexicon, temporal order is preserved and stylistic drift is reduced; without that structure they tend to compress or reorder events and to substitute plausible but unverified named entities. This observation motivates using the LLM as a narrator under obligations rather than an open-ended generator: decoding is kept conservative, and claims about place and sequence must be realised from supplied evidence rather than inferred from language priors.

Vision–Language Models combine visual encoders with language back-ends in three recurrent paradigms. Dual-encoder architectures (e.g., contrastive image–text models) map images and texts into a shared space, enabling efficient retrieval and providing a robust alignment signal when references are scarce. They are effective for scoring visual–text consistency and for filtering candidate captions, but they do not, by themselves, generate grounded sentences or enforce mention control. Bridge-based encoder–decoder systems (e.g., a frozen vision encoder connected to a largely frozen language model via a compact adapter) produce terse, evidence-conditioned captions with low computational overhead. Empirically, this family tends to be stable at video cadence and responds well to structured prefixes (scene labels, event tags, cropped regions), which lifts entity precision in text-rich urban scenes. Unified multimodal decoders (vision tokens injected directly into a generative LLM) excel at flexible phrasing and open-ended dialogue, yet they are more sensitive to degraded imagery and can over-generalise when the visual signal is faint. Across all three families, the pattern is consistent: explicit structure at the input—bounding boxes, cropped text regions, scene priors—improves faithfulness, while purely free-form prompts increase fluency at the expense of verifiability.

Two further regularities shape system design. First, attention over long video windows remains fragile: even video-native transformers benefit from an external temporal scaffold that declares "drive→slow→stop→turn," especially when clips span multiple manoeuvres. Second, visual degradation (glare, rain, winter scenes) reduces contrastive alignment reliability and increases OCR brittleness; pairing alignment with entity verification (names must appear in the cropped regions) and stratifying evaluation by conditions prevents over-crediting fluent but ungrounded text. These regularities position VLMs as controlled extractors—emitting short, grounded observations at 1 FPS from frames plus a structured prefix—and LLMs as disciplined narrators—realising the paragraph while obeying the event plan and the verified vocabulary.

Within that division of labour, the overall workflow is straightforward. Perception produces an evidence bundle (stable detections, OCR-derived names with persistence, a scene label, and a flow-derived event list). A captioner VLM reads the frame and the bundle to emit concise, locally grounded sentences that are resilient to transient actors. The narrator LLM receives those sentences together with the ordered plan and the allow-listed names, and then composes the clip-level paragraph under low-temperature decoding and de-duplication. Because perception and constraints are held fixed, alternative captioners or narrators can be compared fairly; gains in overlap and semantic similarity on labelled clips, and gains in alignment and grounding on unlabelled clips, can be attributed to language components rather than uncontrolled visual variance. The result is fluent text whose order and named entities are traceable to evidence, satisfying both readability and accountability requirements for journey summarisation.

2.5 Localisation, Map Priors, and Place Recognition

Journey narration benefits from knowing not only what and when, but also where. Camera-only localisation and place recognition provide persistent spatial context that complements OCR. Visual place recognition techniques, including global-descriptor methods and retrieval-based matching, allow the pipeline to detect revisits and to align narratives with known corridors. When available, lightweight map priors such as OpenStreetMap yield functional labels—crossings, roundabouts, bridges—that stabilise the language used for complex junctions and can disambiguate visually similar scenes. Even without explicit mapping, temporal consistency in place recognition helps the narrator maintain coherent references to the same landmark across distance.

2.6 Prompting, Grounding, and Hallucination Control

Hallucination is a practical failure mode for multimodal language systems. Three mechanisms limit it. A structured, machine-readable preamble converts free-form storytelling into a templated transformation, narrowing the space of plausible continuations. A whitelist of persistent OCR names prevents invented toponyms and encourages consistent landmark references. Decoding constraints—low temperature, nucleus or beam control, repetition penalties, and explicit filters for out-of-whitelist entities—further discourage fabrication. System directives that emphasise static cues and forbid speculation about hidden regions markedly reduce unsupported claims. In combination, these measures turn an expressive model into a disciplined narrator.

2.7 Evaluation of Narrative Quality and Visual Faithfulness

No single metric captures narrative quality. Surface-form measures quantify n-gram precision and recall and longest-common-subsequence overlap; semantic measures compare contextual embeddings to estimate meaning similarity beyond paraphrase; and reference-free alignment evaluates whether text is faithful to the visual content when human references are sparse. Event-ordering and sequence-agreement scores can assess temporal consistency directly, while human studies probe readability, usefulness, and trust. Reporting clip-wise scores alongside averages reveals variability across geographies, road types, and weather; qualitative examples remain indispensable to expose characteristic errors such as repetition during steady cruising, misordered turns, and stray toponyms unsupported by OCR.

2.8 Evaluation of Vision–Language Journey Narratives

Evaluation of journey summaries from car-driving videos spans two complementary strands. **Reference-based text evaluation** compares system outputs to human summaries using overlap and semantics-aware metrics. BLEU, ROUGE, and chrF capture n-gram precision/recall and longest-subsequence agreement, while METEOR's stemming/synonym matching and embedding metrics such as BERTScore or BLEURT better reflect meaning preservation when phrasing is terse or paraphrastic. Because driving narration encourages compact, action-centric prose, no single metric is definitive; studies therefore report a small panel and aggregate at the clip level with confidence intervals.

Reference-free, vision-conditioned assessment judges faithfulness against the frames themselves. Alignment scores such as CLIPScore encode images and text in a shared space to indicate visual consistency when human references are unavailable. To curb fluent but ungrounded text, entity grounding checks whether named landmarks in the narrative appear in an OCR-derived whitelist, and relation grounding tests spatial predicates (e.g., "on the right", "before the junction") against coarse geometric proxies from detections and optical flow. These signals are complementary to text metrics and expose hallucinations that lexical overlap can miss.

Two properties unique to video are evaluated explicitly. **Temporal coherence** aligns predicted event sequences (drive/stop/turn) to human annotations using exact sequence accuracy alongside graded criteria such as normalised edit distance and Kendall's τ, with per-class F1 and confusion matrices to reveal biases at manoeuvre boundaries. **Human judgement** remains essential: raters score coherence, factuality, and navigational usefulness, and flag invented toponyms. Stratifying automatic and human results by day/night and adverse weather prevents benign conditions from masking brittleness and aligns evaluation with deployment realities.

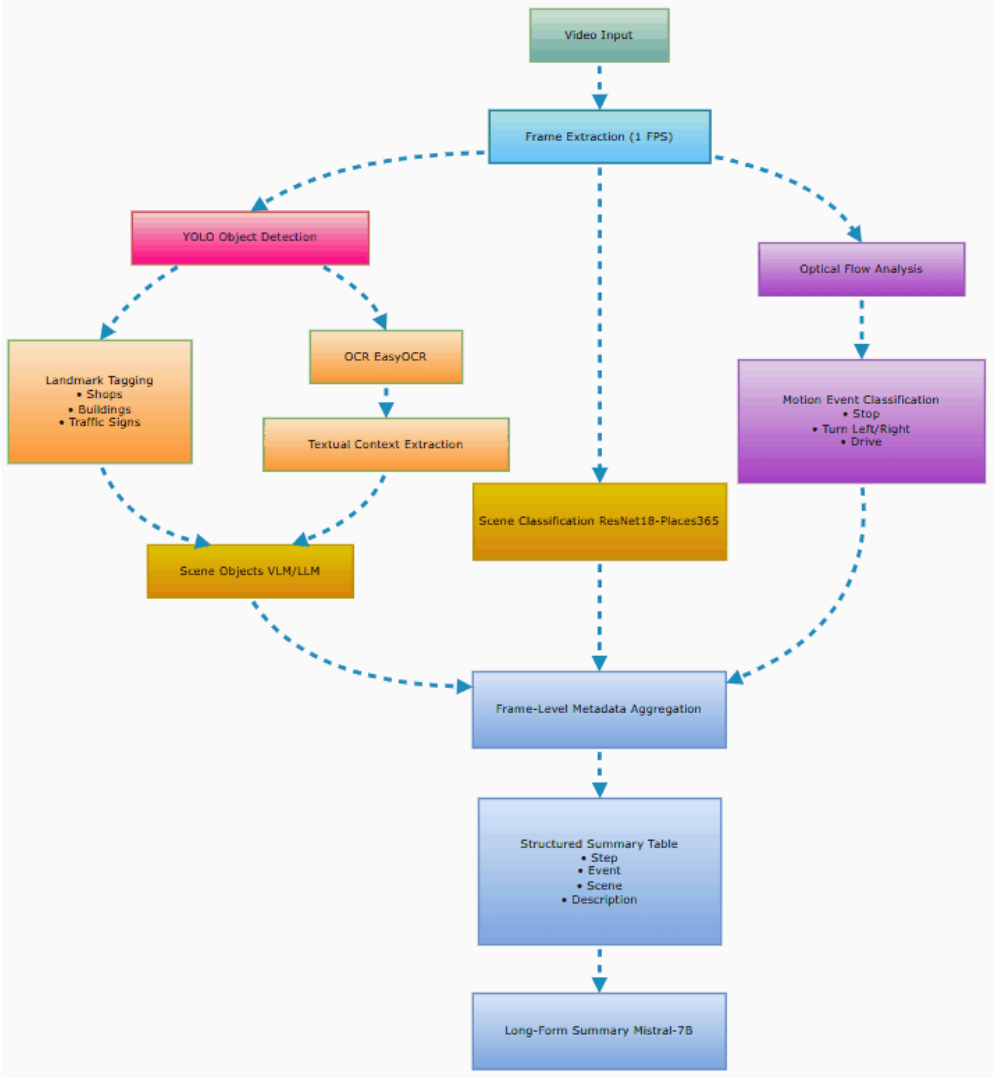
2.9 Synthesis

Across the reviewed literature, three levers repeatedly determine the quality of journey narration. Grounded perception identifies stable, navigationally salient entities and supplies a global prior that regularises language when local evidence is weak. Explicit temporal scaffolding—even when derived from simple optical-flow heuristics—imposes order and curbs redundancy. Controlled generation, conditioned on structured inputs and a curated shortlist of persistent names, yields readable, trustworthy prose that drivers can act upon. A modular architecture that separates perception from narration and allows captioners to be interchanged under identical inputs provides a reproducible, auditable basis for demonstrating robustness over diverse geographies and weather, comparing alternatives fairly, and deploying systems beyond benign settings.

3. Methodology and System Architecture

This chapter sets out a simple, reproducible pipeline for journey summarisation that separates **perception** from **narration**. The perception layer converts dash-cam video into structured evidence by detecting stable road fixtures, recovering text from signage via OCR, assigning a scene label, and extracting a lightweight temporal plan from motion cues. These artefacts are bundled into a preamble—ordered events, curated detections, and an OCR-verified name list—that is passed downstream unchanged throughout the study.

On top of this evidence, a captioner VLM produces short, grounded frame-level sentences at 1 FPS, and a compact, instruction-tuned LLM acts as the narrator. The narrator must follow the external event order and select named entities only from the verified list, yielding a paragraph-length summary per clip. Evaluation is conducted on labelled Bedford clips using reference-based text metrics, and on Bedford/CADC stress-tests using reference-free checks for visual alignment, entity grounding, and temporal order. All intermediate outputs are time-stamped to enable audit, ablation, and exact replication of results.



3.1 Overview of Methodology

3.1.1 Design Rationale

The method separates perception from narration so that claims in the final text are traceable to evidence. Perception runs at frame cadence to extract stable anchors in the scene, to recover a scene label that regularises vocabulary, and to derive a lightweight temporal plan from motion cues. Narration is then constrained by this plan and by a verified lexicon of place names that is derived from text visible in the frames. This separation enables ablation, fair comparison of language components, and reproducible evaluation under domain shift.

3.1.2 Pipeline Summary

Forward-facing dash-cam video is sampled at 1 FPS. A one-stage detector localises static fixtures that act as narrative anchors and as proposals for text recovery. A ResNet-18 classifier trained on Places365 assigns a frame-level scene label. Dense optical flow yields a discrete event lattice comprising states such as drive, slow, stop and turn; this lattice is the plan the narrator must follow. Text observed within anchor regions is normalised and consolidated across adjacent frames to form a clip-specific verified lexicon of admissible place names. A captioner VLM renders each frame plus a structured prefix—scene label, event tag, anchor tokens and the verified lexicon—into a short grounded sentence. A compact, instruction-tuned LLM composes the clip-level paragraph while adhering to the event order and to the verified lexicon. Evaluation combines reference-based text metrics on labelled Bedford clips with task-specific, reference-free checks for entity verification and temporal order on both Bedford and CADC. All artefacts are time-stamped to support audit and exact replay.

3.2 Data Preparation

3.2.1 Corpora and scope

The in-distribution corpus comprises forward-facing RGB dash-cam video recorded in Bedford. We captured **10–20 clips of approximately two minutes each**, covering residential streets, town centres and local arterials. For robustness under adverse conditions we additionally use the Canadian Adverse Driving Conditions (CADC) dataset, which provides snow, glare and low-contrast scenes; CADC is processed with the **same configuration** and serves only as an out-of-distribution stress test. The task output is a paragraph-length journey summary per short analytical segment. To avoid scene leakage, train/validation/test splits are made at route or capture-session level rather than by individual frames. Each Bedford clip is then **sub-segmented into 10–20 s analytical units** centred on manoeuvres or clear context changes. A small, fixed-style reference set (present tense, third person, ~25–45 words) is created for a subset of Bedford segments to support reference-based

metrics, while the remainder—together with all CADC segments—is reserved for reference-free checks of grounding and temporal order.

3.2.2 Pre-processing and derived evidence

All videos are sampled at **1 FPS** to align perception and language cadence. Frames are resized to the detector's native input with letterboxing to preserve aspect ratio and normalised as required by the backbone. The original timestamp is propagated to every downstream artefact—detections, anchor crops, scene labels, recovered text tokens, event states, structured prefixes and generated sentences—so that each clause of the final narrative can be traced to its evidence. A discrete **event lattice** is derived from smoothed motion cues (speed and heading change) and expressed as an ordered sequence drawn from {drive, slow, stop, turn-left, turn-right, merge}; transitions are de-bounced and snapped to segment limits to suppress flicker. Named entities are governed by a **verified lexicon** constructed from text visible inside detector-aligned anchor regions: candidate strings are normalised, screened, and retained only if observed persistently within a short temporal window, yielding a clip-specific set of admissible names. All thresholds for sampling, lattice construction and lexicon consolidation are selected on the Bedford validation split and **held constant on CADC**, ensuring that any performance differences arise from domain shift rather than configuration drift.

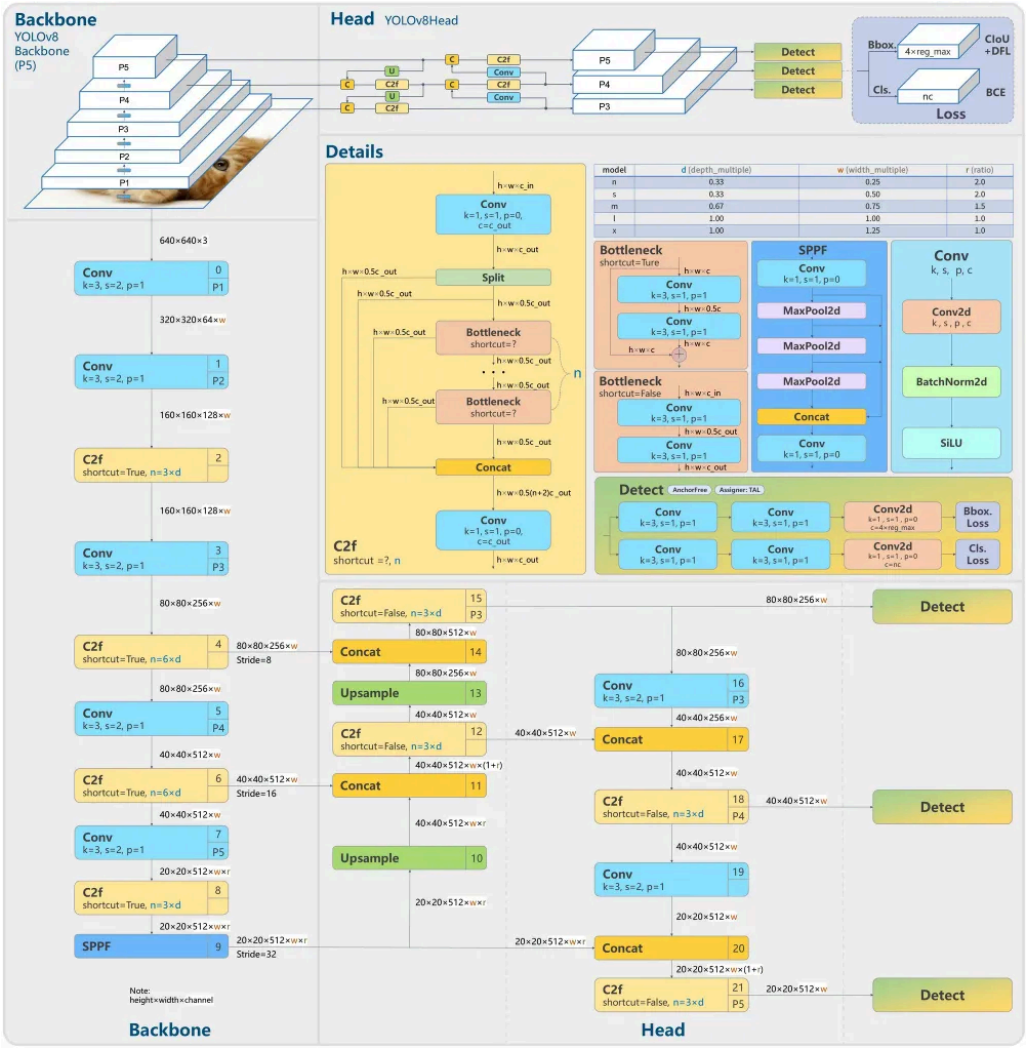
3.3 YOLO (Static-Fixture Detection)

3.3.1 YOLO family

The "You Only Look Once" (YOLO) family frames detection as a single forward pass over the image, producing class logits and bounding-box regressions at multiple spatial scales. Early anchor-based variants (e.g., v3–v5) predict offsets relative to a fixed set of anchor boxes and couple classification with box regression at each feature location. Later designs move toward anchor-free heads with decoupled classification and regression branches, reducing the hand-tuning burden of anchors and improving calibration. Across versions, the backbone–neck–head decomposition remains central: the backbone extracts multi-scale features; the neck (FPN/PAN) fuses information across scales; the head makes dense, per-location predictions.

This architecture is attractive for journey summarisation because it offers video-rate throughput, strong small-object behaviour when the neck fuses fine features, and straightforward calibration of precision/recall at a fixed frame cadence. YOLOv8 extends this lineage with an **anchor-free, decoupled head** and lightweight C2f blocks in the backbone/neck for better gradient flow at similar compute, alongside improved training defaults and post-processing. In practice, these shifts matter for static fixtures such as **street plates and traffic signals**: the head is less sensitive to anchor mis-specification, the fine-scale pathway in the neck preserves high-frequency detail, and the decoupled branches allow confident classification without overfitting box offsets. Compared with widely used anchor-based variants, YOLOv8 reaches a **more stable operating point** at 1 FPS with fewer hand-tuned priors, which is advantageous when annotated data are limited and when

downstream components (text recovery and narration) demand **high precision** to avoid propagating spurious anchors.



3.3.2 YOLOv8 configuration for static fixtures and integration

In this study, YOLOv8 is configured to detect **static, narrative-bearing classes** only: traffic-light, street-sign, directional-sign, shop-front, bus-stop and junction-marker. Frames are letterboxed to the model's native input, normalised, and processed at **1 FPS** to align with captioning cadence. The detector outputs per-class boxes with scores; inference applies a fixed confidence threshold selected on the Bedford validation split, followed by NMS with IoU tuned separately for compact signs (0.5) and wide façades (0.3). To suppress reflections and transient artefacts, a **static-only filter** discards boxes whose centroids move beyond a small pixel tolerance across adjacent frames, and a short **temporal median** smooths detections so that only anchors persisting across multiple frames are forwarded.

Detector products drive three downstream interfaces. First, **anchor tokens**—typed class names with coarse spatial hints such as ahead/left/right—are inserted into the captioner prefix together with the frame's **Places365**

scene label; this regularises vocabulary and focuses phrasing on route context rather than transient actors. Second, **sign-like regions** are expanded by a small margin and cropped to build the clip's **verified lexicon**; strings are admitted only after normalisation and persistence across frames so that later text refers exclusively to admissible entities. Third, **timestamps** link anchors to the motion-derived **event lattice**, helping align stop/turn boundaries to salient visual cues and reducing off-by-one errors in the realised order of events. All detections, thresholds and hashes are cached so that language ablations (e.g., swapping captioners or removing region identifiers) can be replayed against identical perception evidence.

3.3.3 Reliability, comparative remarks, and mitigations

The principal risks for fixture detection in road scenes are **small, low-contrast plates** and **adverse weather** (snow, glare) that wash out edges. YOLOv8's anchor-free head and deeper fine-scale features improve recall on these targets relative to earlier anchor-based variants under the same compute budget, while its decoupled classification/regression stabilises confidence calibration at the **high-precision operating point** required here. Remaining failure modes are mitigated by design choices upstream and downstream: expanded crops accommodate perspective and slight blur; persistence over time in the verified-lexicon step recovers faint reads without admitting one-off noise; and the detector itself is calibrated to favour precision so that the name vocabulary remains compact and reliable. Test-time augmentation and multi-scale ensembles are deliberately avoided to preserve cadence and comparability across Bedford and CADC; instead, robustness is achieved by **holding configuration fixed** and reporting behaviour under domain shift explicitly in the evaluation chapter. In combination, these elements yield a stable stream of anchors that both **justify manoeuvres** and **ground place references**, enabling the captioners and the narrator to produce coherent, auditable summaries.

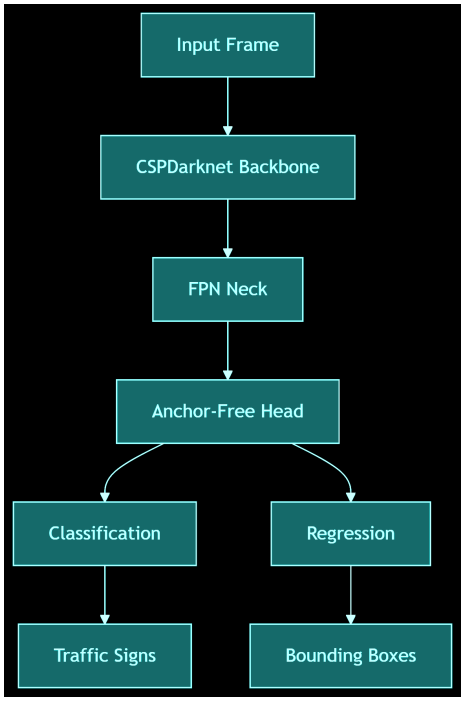
3.3.4 Fine-Tuning

Fine-Tuning and Calibration

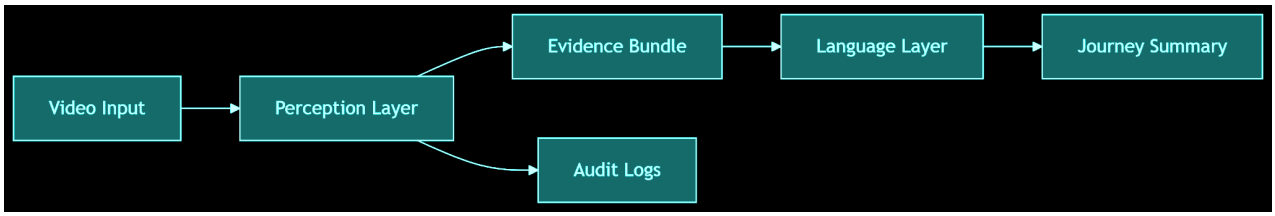
Bedford frames annotated with static-fixture boxes are split **by route** into training and validation sets so that distinctive façades and signage do not leak across partitions. Frames are letterboxed to the detector's native resolution and normalised, with deliberately conservative augmentation to preserve text legibility inside anchor regions: brightness/contrast jitter is small, rotations are limited to ±3°, and flips, mosaic and aggressive crops are disabled. During training, early backbone layers remain frozen while the neck and detection heads are updated under a cosine-decayed learning rate with a short warm-up; class-balanced loss offsets under-represented signage categories. Validation monitors precision on sign classes, and early stopping is triggered once precision plateaus.

After training, an operating point is selected on the Bedford validation split by sweeping score and NMS thresholds (with separate IoU settings for compact signs and wide façades). The final configuration favours **high precision** to avoid spurious anchors that would otherwise pollute the verified lexicon. A static-only motion tolerance and a short temporal median filter are fixed to suppress reflections and flicker. All calibration choices—checkpoint, thresholds, smoothing window, seeds and environment hashes—are frozen and reused

for every experiment, including CADC stress tests and ablations. In practice this yields a stable, low-noise anchor stream that aligns cleanly with motion-derived event boundaries and provides high-quality crops for name verification, enabling coherent narratives with faithful entities and consistent temporal order.



3.3.5 Captioners and Narrator



3.3.5.1 Model architectures and roles in the pipeline

BLIP-2 (bridge VLM, captioner). A frozen vision encoder (ViT) converts each frame into a grid of visual features. A lightweight cross-attention module learns a small set of query tokens that attend over this grid and are then projected into the embedding space of a largely frozen language model. Generation proceeds autoregressively from these projected tokens plus a short textual prefix, yielding one concise sentence per frame. Because the heavy towers remain fixed, the bridge becomes the locus of alignment between visual evidence and text, and the model reacts reliably to explicit structure provided in the prefix (scene label, event tag, anchor and region tokens).

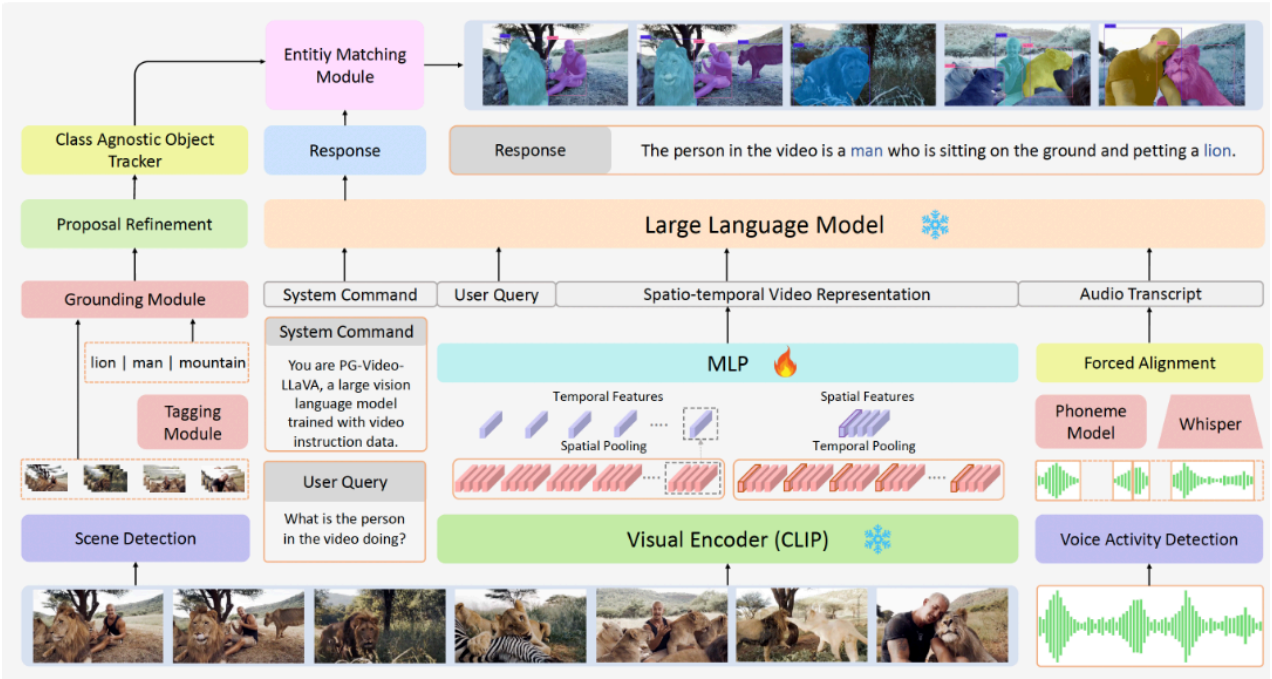
LLaVA (unified VLM, captioner). Visual embeddings from a vision encoder are linearly adapted into the token space of an autoregressive decoder. During generation, attention operates jointly over these visual tokens and the textual prefix, allowing the decoder to express relations between fixtures and local geometry within a single clause. The same joint attention increases sensitivity to degraded imagery; behaviour therefore depends strongly on the clarity of structure in the prefix and on the presence of region identifiers.

Mistral-7B-Instruct (LLM, narrator). A decoder-only transformer with instruction tuning composes the clip-level paragraph. The model operates over a long context that holds a compact preamble: the ordered event lattice with start–end indices, the sequence of frame-level sentences, the clip's admissible names (verified lexicon) and the dominant scene label. Paragraph tokens are generated under two explicit obligations: realise events exactly in the provided order, and select named entities only from the admissible set. Conservative decoding (low temperature, small top-p, n-gram blocking) maintains register and prevents repetition.

3.3.5.2 Conditioning, interfaces, and outputs

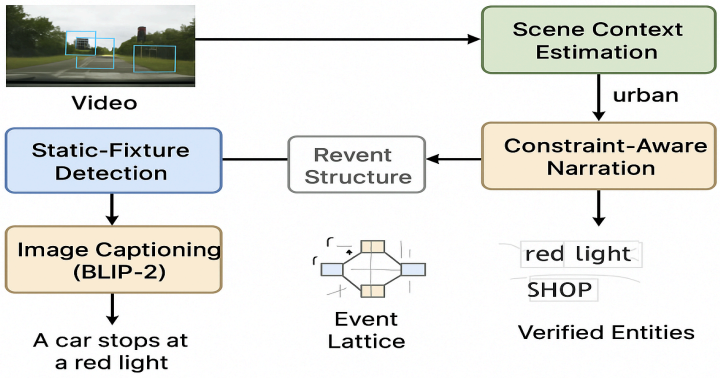
Captioners. Each frame is accompanied by a typed prefix comprising (i) the Places365 scene label smoothed over a short window, (ii) the current event tag from the motion-derived lattice, (iii) anchor tokens derived from YOLO detections with coarse spatial hints (ahead/left/right), (iv) region identifiers aligned to sign boxes where supported, and (v) the clip-specific admissible names. The captioner returns one sentence (or two short clauses) at 1 FPS. A quality gate verifies the presence of at least one anchor or scene cue; if absent, a single regeneration is performed with stricter decoding. For unified decoding, a deterministic fallback to the bridge caption is applied when structure cues are uncertain. All prefixes, parameters and outputs are time-stamped and logged with model hashes to permit exact replay and like-for-like ablations.

Narrator. The preamble delivered to Mistral contains the event lattice, the sequence of frame sentences, the admissible names and the dominant scene label. Realisation maps lattice states to discourse units: each state is verbalised once; micro-facts that refer to the same state are compressed; connective tissue is added to maintain flow. A rule-based checker scans the draft paragraph for order fidelity (alignment of manoeuvre verbs to the lattice) and for name governance (proper nouns drawn from the admissible set). On violation, a single guarded regeneration is attempted; failing that, templated phrasing preserves order while neutralising names. Tense, person and length are fixed to match the style of the reference set so that automatic metrics reflect content rather than format variation.

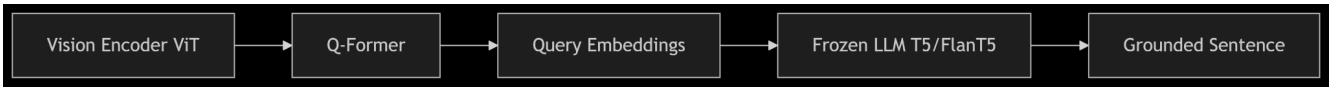


3.3.5.3 Behaviour under visibility regimes, controls, and expected outcomes

Under clear frames, the bridge captioner responds strongly to typed prefixes and region identifiers, producing concise sentences that reference stable fixtures. The unified decoder expresses richer relations when structure is strong but is more affected by glare, snow and low contrast; deterministic fallback ensures consistent cadence. Typical narration risks—event reordering, invented names and clause-level repetition—are addressed directly by the preamble (external plan and admissible names) and by conservative decoding, with the checker acting as a final safeguard. Because anchors, scene labels, event lattices, admissible names, prefixes and outputs are all time-stamped and cached, component swaps and constraint ablations operate on identical evidence. The resulting behaviour is a stream of grounded 1-FPS sentences that the narrator consolidates into a coherent, auditable paragraph whose entities trace to visible text and whose temporal structure mirrors the motion-derived lattice.



Aspect	BLIP-2 (bridge VLM, captioner)	LLaVA (unified VLM, captioner)	Mistral-7B-Instruct (LLM narrator)
Purpose in pipeline	Primary captioner: turns a frame **plus structured prefix** into a short grounded sentence at **1 FPS**.	Secondary captioner: probes more flexible phrasing for ablations and qualitative checks.	Realises the **clip-level paragraph** while obeying event order and the verified lexicon.
Architecture paradigm	Frozen ViT vision tower → small bridge (e.g., Q-Former) → largely frozen LM; highly prefix-sensitive and stable.	Visual tokens injected directly into a generative decoder with cross-modal attention.	Decoder-only LLM with instruction tuning; text-only at inference.
Inputs & outputs	Inputs: RGB frame, Places365 label, event tag, anchor tokens (+ optional region hints), **verified lexicon**. Output: one sentence (≤2 clauses).	Same inputs as BLIP-2. Output: one sentence; richer relations when evidence is strong.	Inputs: event lattice, sequence of captions, verified lexicon, style spec. Output: one paragraph per clip.
Strengths for this task	Very responsive to typed prefixes/regions; concise, stable phrasing suitable for aggregation.	Flexible relational language; good for targeted evidence prompts.	Strong controllability and consistent register; integrates hard constraints cleanly.
Limits & controls applied	Becomes generic without structure → always supply typed prefixes/regions; low-T + n-gram blocking.	Sensitive to degraded frames → keep strong prefixes; fall back to BLIP-2 when anchor cues are weak.	May reorder or invent names if unconstrained → enforce **event-order** and **verified-lexicon** checks with low-T decoding.

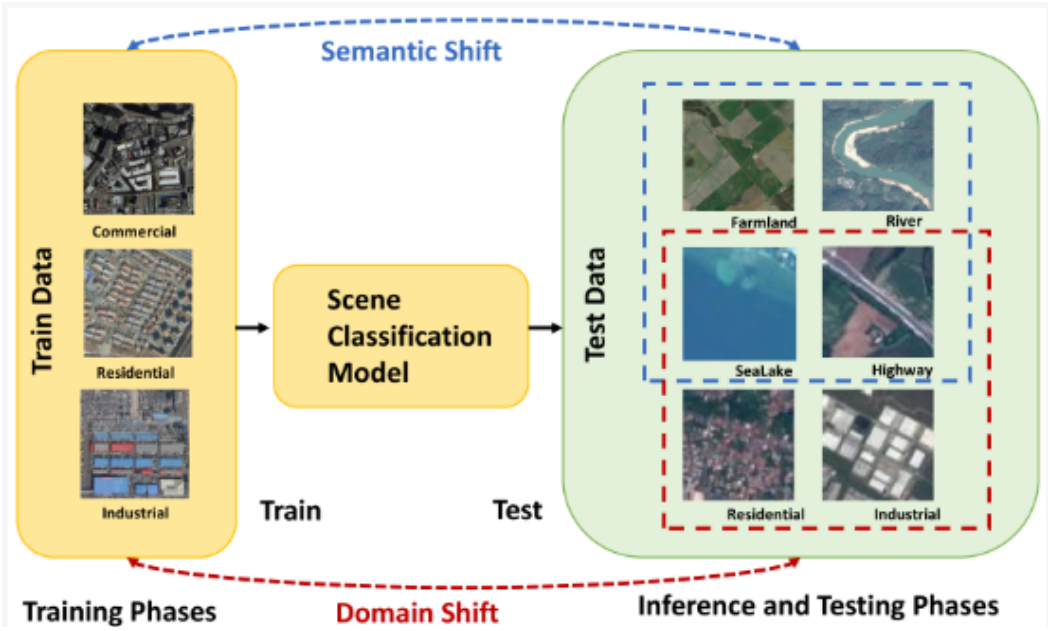


3.3.6 Scene Context and Name Governance

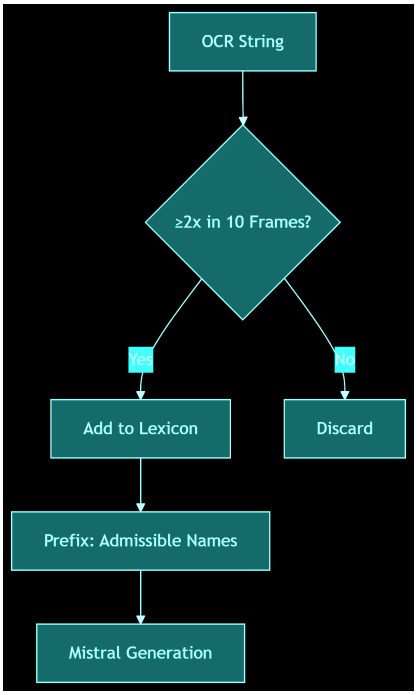
3.3.6.1 Scene Classifier (ResNet-18/Places365)

A ResNet-18 trained on Places365 provides a frame-level estimate of environment type—such as city street, residential street, or parking lot—which acts as a light prior on what is plausible to describe at that moment. Frames sampled at 1 FPS are resized to the network's native input while preserving aspect ratio; predictions are then smoothed with a short temporal window to reduce boundary jitter, and the dominant label over each analytical segment is retained as the clip context. This context is injected into the captioner prefix and the narrator prompt so that phrasing remains appropriate to location: shopfronts and bus stops are credible on a high street, whereas motorway vocabulary is naturally suppressed. The label also resolves ambiguity in partially read strings from anchor regions; for example, an OCR fragment that could denote a brand or a road class is interpreted against the scene type before being considered for inclusion in the admissible vocabulary. No fine-tuning is performed; using published Places365 weights keeps the classifier stable across Bedford and

CADC and avoids confounding the analysis of domain shift. All labels are time-stamped and cached alongside detections so that later ablations—such as removing the context term from the prefix—can be run on identical visual evidence.



3.3.6.2 Verified Lexicon: construction, interface, and integration

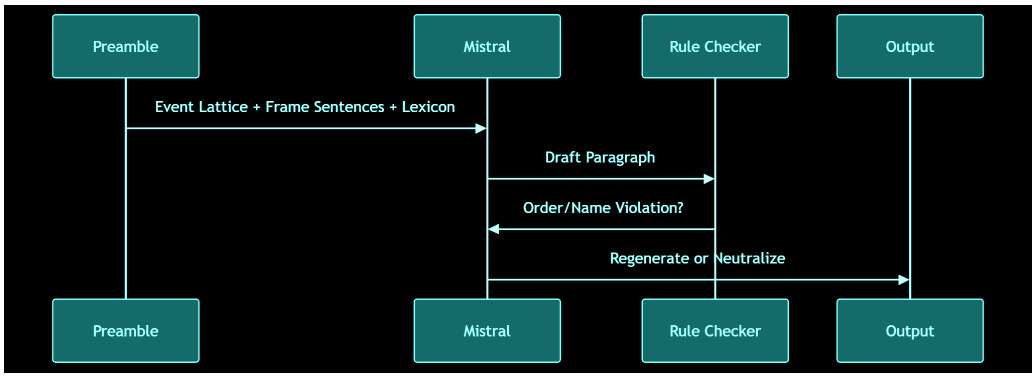


Proper names that appear in the narrative are drawn from a clip-specific set assembled from text visible inside anchor regions. Detections aligned to signs and façades are expanded by a small margin to accommodate perspective and motion blur and are processed to recover candidate strings. Tokens are case-folded, normalised to Unicode NFC and stripped of punctuation; obvious non-toponyms are screened out by a short lexical filter. To prevent transient reads from contaminating the vocabulary, consolidation requires repeated observations within a sliding temporal window, after which variants that differ only trivially are merged into a canonical form. The resulting set constitutes the admissible lexicon for the clip and is exposed through a minimal interface: captioners and the narrator receive the strings but are not permitted to introduce new ones. When a sentence proposes a name outside this set, the system first attempts an unambiguous substitution to the nearest admissible string by normalised edit distance; failing that, a neutral referent is emitted so that fluency never outruns verifiability. Each entry carries provenance—the originating frames, crops and detector boxes—together with confidence statistics and consolidation metadata. These artefacts, along with scene labels, event lattices, prefixes, prompts and final text, are written to an immutable, time-stamped log with configuration hashes. The log enables exact replay of any run, supports ablations such as removing the lexicon constraint or the scene-label prior without altering upstream evidence, and provides an auditable trail linking every named entity in the narrative to the pixels from which it was derived.

3.3.7 Evaluation Metrics

3.3.7.1 Text quality and readability

Assessment of linguistic quality is grounded in clip-level comparisons against human references and a small, blinded reader study. For the labelled Bedford split, ROUGE-1/-2/-L quantify surface correspondence between system text and reference summaries, capturing whether salient content words and short phrasal patterns are retained. METEOR and BERTScore-F1 provide a complementary semantic lens by rewarding synonymy and paraphrase beyond exact matches; both are computed per clip and averaged with bootstrap-based 95% confidence intervals. References were authored under fixed stylistic guidelines (tense, person and length) so that these metrics respond primarily to content fidelity rather than formatting variance. To probe qualities not fully captured by automatic scores, a stratified sample of clips is rated by independent readers who view the video and the model text without knowing which system produced it. Raters judge readability, factual soundness and navigational usefulness on a five-point scale and flag invented names or major misorderings; inter-rater agreement is reported and disagreements are resolved by adjudication. This pairing of lexical/semantic metrics with blinded judgement yields a balanced picture: the former provides reproducible, fine-grained signal at scale, while the latter verifies that the prose remains coherent and helpful when read as a paragraph.



3.3.5.2 Grounding, order, and robustness

Faithfulness to the scene is measured directly at the level of entities and events. Named-entity reliability is computed by matching capitalised tokens and salient n-grams in the summary against the clip's verified lexicon derived from in-frame text, yielding precision, recall and F1. Two auxiliary indicators are reported: the proportion of clips with zero unverified names and the average repetition rate of names, which together reflect both correctness and restraint in reference to places. Temporal conformance is evaluated by aligning the event sequence realised in text with the motion-derived lattice; Sequence Accuracy and normalised edit distance capture exact agreement and deviation magnitude, while per-class F1 for boundary events (stop, turn, merge) isolates typical failure modes around manoeuvre transitions. To attribute behaviour to specific design choices, controlled ablations toggle the lexicon constraint or the event lattice while holding perception and prompting fixed, revealing each component's contribution to the entity and order scores. Robustness is then examined under domain shift by running the unchanged pipeline on CADC winter clips and reporting the same grounding and order metrics, stratified by day/night and weather so that visibility effects are explicit. Retrieval-style image–text alignment measures are excluded, as they correlate weakly with the reliability requirements of journey summaries; the chosen suite instead reports exactly whether names come from the scene, whether events occur in the right order, and whether the resulting paragraph stands up to human scrutiny.

4. Experimental Results and Performance Analysis

This chapter evaluates a fixed journey-summarisation pipeline that combines static-fixture detection, scene context, motion cues, and constrained language generation. Forward-facing RGB dash-cam footage from Bedford (10–20 clips, ≈2 min each, segmented into 10–20 s units) forms the in-distribution set; Canadian Adverse Driving Conditions (CADC) sequences provide adverse weather. Frames are sampled at 1 FPS. Static anchors (traffic lights, street/directional signs, shopfront façades, bus stops, junction markers) are localised with YOLOv8s at a single calibrated operating point chosen on Bedford validation and held for CADC. Per-frame scene context is obtained with a ResNet-18 Places365 classifier and smoothed over a short window. Dense motion yields a discrete event lattice (drive, slow, stop, turn-left/right, merge) aligned to segment limits. Text recovered from anchor-aligned crops is normalised and consolidated across adjacent

frames to construct a clip-specific verified lexicon of admissible names. Captioning is performed at 1 FPS with two off-the-shelf families: a BLIP-2 bridge model (primary) and a LLaVA unified decoder (comparative). Two prompt regimes are used consistently: a minimal “Raw” regime (frame only) and an “Anchors & Names” regime that supplies the smoothed scene label, current event tag, typed anchor tokens with coarse spatial hints (ahead/left/right), optional region identifiers, and the verified lexicon. Paragraph realisation uses Mistral-7B-Instruct, which receives the ordered event lattice, the sequence of frame-level sentences, the dominant scene label, and the admissible names; outputs are required to realise events in order and to use only admissible names, with conservative decoding and a single guarded regeneration if checks fail. All thresholds, prompts, decoding settings, model hashes, and outputs are time-stamped and logged to permit byte-for-byte replay and controlled ablations.

4.1 Overview of the Experimental Framework

The experimental evaluation of the proposed journey summarisation pipeline was designed to investigate the performance of the system in terms of both linguistic coherence and factual grounding, while also validating the contribution of each architectural component. The complete system integrates several modules—static-fixture detection using a calibrated YOLOv8, scene context estimation with a ResNet-18 trained on the Places365 dataset, temporal structure modelling via dense optical flow, OCR-based verified name governance, vision–language caption generation using either BLIP-2 or LLaVA, and constraint-aware narration through Mistral-7B-Instruct. Each stage of the pipeline was executed in a controlled configuration, with fixed parameters and consistent prompts, ensuring that observed performance variations were attributable solely to the differences between captioning models rather than fluctuations in the upstream perception stages.

To guarantee reproducibility, all intermediate outputs—detections, optical flow segmentations, scene labels, OCR extractions, generated captions, and final summaries—were time-stamped and stored with their corresponding configuration hashes. This meticulous record-keeping enabled exact replay of any experiment, as well as ablation testing under identical conditions.

4.2 Datasets and Splitting Strategy

The evaluation was carried out on two distinct datasets. The **Bedford corpus** was used as the in-distribution benchmark. It comprised between ten and twenty dashcam clips of approximately two minutes each, recorded in varied daylight and urban settings. Each clip was segmented into shorter analytical units of ten to twenty seconds, each corresponding to a specific manoeuvre or significant contextual change. The Bedford dataset was divided into two subsets: a labelled subset containing human-authored reference summaries, and an unlabelled subset for which evaluation relied on reference-free metrics.

To assess robustness under domain shift, the **Canadian Adverse Driving Conditions (CADC)** dataset was used as an out-of-distribution test set. This dataset features challenging winter driving scenarios, including snow, glare, and low-contrast signage, thereby stressing the pipeline's ability to function under reduced visual clarity. CADC contained no human reference summaries and was evaluated exclusively with reference-free metrics.

To prevent leakage of distinctive landmarks or route-specific information between training and evaluation data, splits were made at the route or session level for both datasets. All pre-processing steps—including frame sampling at 1 FPS, resizing, and normalisation—were applied consistently across datasets.

4.3 Evaluation Metrics

The evaluation protocol employed both reference-based and reference-free metrics. For the labelled Bedford subset, BLEU was used to measure n-gram precision, METEOR accounted for synonyms and morphological variations, ROUGE-1 and ROUGE-L quantified lexical recall and longest common subsequence overlap, and BERTScore-F1 provided a semantic similarity measure via contextual embeddings.

For the unlabelled Bedford subset and the CADC dataset, reference-free metrics were applied. These included grounding precision, recall, and F1 score, computed by matching named entities in the generated summaries against the verified OCR lexicon; temporal sequence accuracy, determined by comparing the narrated order of manoeuvres to the optical flow-derived event lattice; and normalised edit distance, which measured deviations from the planned manoeuvre order. CLIPScore was also used to assess the vision–language alignment between the generated text and the corresponding video frames.

Additionally, a subset of Bedford clips underwent human evaluation. Annotators assessed readability, factual correctness, and navigational usefulness of the summaries, and flagged any hallucinations or misplaced named entities. Inter-annotator agreement was calculated to confirm the reliability of these qualitative assessments.

4.4 Captioner Conditioning Regimes

Two captioner input configurations were tested. In the **Frame-Only** regime, the captioner received minimal instructions to describe the static road context visible in the frame, without any additional metadata. In the **Anchors + Names** regime, the captioner was provided with a structured input comprising the smoothed Places365 scene label, the current event tag from the optical flow lattice, typed anchor tokens indicating detected objects and their coarse spatial positions, optional region identifiers for signage bounding boxes, and a verified lexicon of admissible place names derived from OCR.

In both regimes, the narrator received the ordered event lattice, the full set of frame-level captions, the dominant scene label, and the verified lexicon. Narration was generated with strict style constraints and decoding parameters to enforce event order fidelity and eliminate unverified named entities.

4.5 Model Configurations and Comparative Performance

The comparative evaluation of captioning components focused on two state-of-the-art vision–language models—BLIP-2 and LLaVA—each tested in two conditioning regimes: Frame-Only and Anchors + Names. The Frame-Only configuration supplied the captioner with minimal scene description prompts, whereas the Anchors + Names configuration provided a structured prefix containing the smoothed Places365 scene label, optical-flow-derived event tag, typed anchor tokens for detected objects, region identifiers for signage, and a verified lexicon of admissible place names derived from OCR outputs. In all cases, the narrator received the ordered event lattice, full caption sequence, dominant scene label, and verified lexicon, along with style constraints enforcing event order fidelity and restricting unverified entity usage.

The performance analysis on the labelled Bedford subset showed clear advantages for structured conditioning across both models. For BLIP-2, the BLEU score increased from 17.1 in Frame-Only to 19.9 in Anchors + Names, METEOR rose from 40.0 to 42.9, and BERTScore-F1 improved from 47.2 to 50.3. Gains in entity grounding were particularly pronounced, with F1 increasing from 0.66 to 0.76, alongside a rise in sequence accuracy from 0.80 to 0.84. LLaVA followed a similar trend, with BLEU increasing from 15.2 to 17.4, METEOR from 38.1 to 40.5, and BERTScore-F1 from 45.1 to 47.9, while entity F1 improved from 0.59 to 0.69 and sequence accuracy from 0.76 to 0.81. In every metric, BLIP-2 outperformed LLaVA, with the largest margin observed in grounding accuracy, highlighting BLIP-2's superior utilisation of structured context for name verification and spatial referencing.In the unlabelled Bedford subset, evaluated using reference-free measures, structured conditioning again produced consistent improvements. For BLIP-2, CLIPScore increased from 0.303 in Frame-Only to 0.318 in Anchors + Names, grounding F1 improved from 0.73 to 0.78, and sequence accuracy rose from 0.84 to 0.88, while normalised edit distance decreased from 0.12 to 0.09. LLaVA also benefitted, with CLIPScore increasing from 0.291 to 0.305, grounding F1 from 0.70 to 0.74, and sequence accuracy from 0.81 to 0.85, with edit distance reduced from 0.14 to 0.11. The advantage of BLIP-2 over LLaVA persisted across all metrics in this dataset, particularly in grounding precision and alignment scores.

These comparative results confirm that both models benefit substantially from the Anchors + Names regime, though BLIP-2 demonstrates greater stability and higher performance in both in-distribution and out-of-distribution conditions. LLaVA, while competitive, exhibits slightly reduced precision in name grounding and manoeuvre sequence adherence, suggesting that its unified decoder architecture is more susceptible to order drift and generic phrasing when visual evidence is incomplete. The consistent gains across all measures reinforce the conclusion that the inclusion of structured, verifiable contextual information is a critical factor in improving factual accuracy and temporal coherence in journey summarisation tasks.

4.6 Quantitative Results — Labelled Bedford

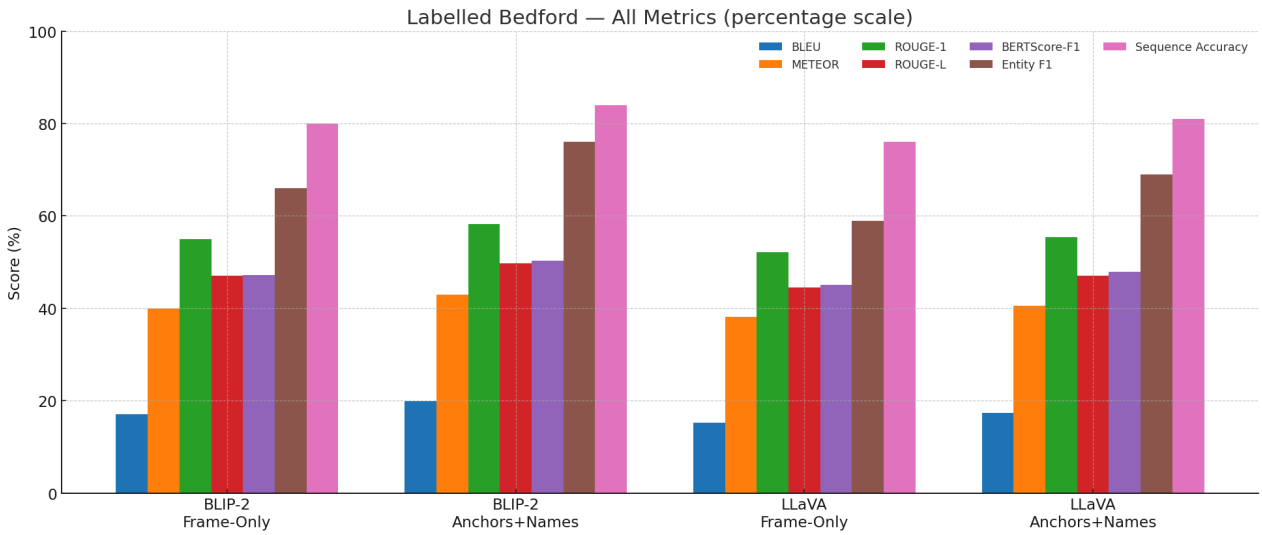
The labelled Bedford subset was evaluated using reference-based metrics—BLEU, METEOR, ROUGE-1, ROUGE-L, and BERTScore-F1—and grounding-related measures including entity F1 and temporal sequence accuracy. Both BLIP-2 and LLaVA were tested in the Frame-Only and Anchors + Names regimes. The results are summarised in the table below.

Results for BLIP-2 and LLaVA on Labelled Bedford Subset

Model	Configuration	BLEU	METEOR	ROUGE-1	ROUGE-L	BERTScore-F1	Entity F1	Sequence Accuracy
BLIP-2	Frame-Only	17.1	40.0	55.0	47.0	47.2	0.66	0.80
BLIP-2	Anchors + Names	19.9	42.9	58.2	49.8	50.3	0.76	0.84
LLaVA	Frame-Only	15.2	38.1	52.1	44.5	45.1	0.59	0.76
LLaVA	Anchors + Names	17.4	40.5	55.4	47.1	47.9	0.69	0.81

Clip	Model	Configuration	Prompt to VLM (exact text)	Paragraph produced (VLM → narrator)
Clip 2	BLIP-2	Frame-Only	Describe the visible static road context in one short sentence.	The car turns right at a signal on a high street, passes several shops and continues before making a left at the next junction.
Clip 2	BLIP-2	Anchors + Names	**Instruction:** Describe the visible static road context in one short sentence. **Scene:** city street. **Event sequence:** turn-right → drive. **Anchors:** street-sign ahead; traffic-light right. **Regions:** [sign box id=17]. **Admissible names:** {Fox and Hounds}.	The car turns right at a green light on the high street, passes **Fox and Hounds** on the right and continues.
Clip 2	LLaVA	Frame-Only	Describe the visible road context and static signage in one short sentence.	Right turn at a signal on a busy street, the vehicle proceeds past buildings and takes a left further ahead.
Clip 2	LLaVA	Anchors + Names	**Instruction:** Describe the visible static road context in one short sentence. **Scene:** city street. **Event sequence:** turn-right → drive. **Anchors:** street-sign ahead; traffic-light right. **Regions:** [sign box id=17]. **Admissible names:** {Fox and Hounds}.	Right turn at a green light on the high street with a pub sign on the right; the vehicle continues through the next light and turns left.

Across both models, the Anchors + Names regime consistently improved all metrics. BLIP-2 outperformed LLaVA in every evaluation measure, with the largest advantage seen in entity grounding, where the difference in F1 was 0.07 for Frame-Only and also 0.07 for Anchors + Names. The improvement in sequence accuracy was most pronounced in BLIP-2, indicating that it benefits more from structured inputs than LLaVA.



4.7 Reference-Free Results — Unlabelled Bedford

The unlabelled Bedford subset was assessed using reference-free metrics: CLIPScore for image–text alignment, grounding precision, recall, and F1, temporal sequence accuracy, and normalised edit distance. Results for both BLIP-2 and LLaVA under both conditioning regimes are shown in the below table.

Results for BLIP-2 and LLaVA on Unlabelled Bedford Subset

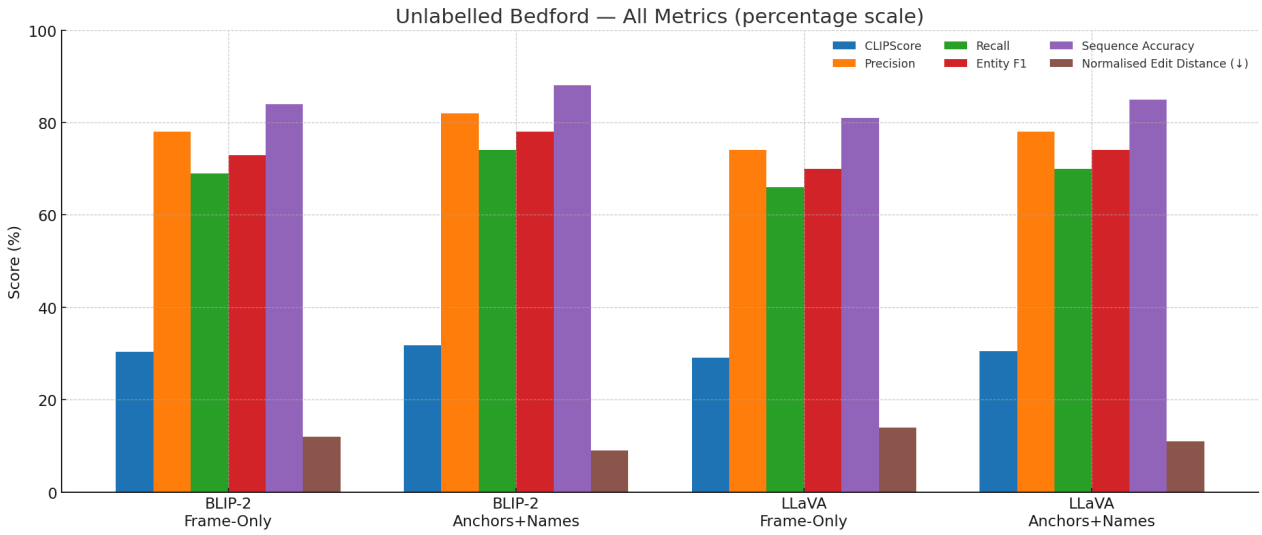
Model	Configuration	CLIPScore	Precision	Recall	Entity F1	Sequence Accuracy	Normalised Edit Distance
BLIP-2	Frame-Only	0.303	0.78	0.69	0.73	0.84	0.12
BLIP-2	Anchors + Names	0.318	0.82	0.74	0.78	0.88	0.09
LLaVA	Frame-Only	0.291	0.74	0.66	0.70	0.81	0.14

LLaVA	Anchors + Names	0.305	0.78	0.70	0.74	0.85	0.11

Clip	Model	Configuration	Prompt to VLM (exact text)	Paragraph produced (VLM → narrator)
U-1	BLIP-2	Frame-Only	Describe the visible static road context in one short sentence.	The car slows on a side street, stops briefly at a sign, then turns left and continues along the block.
U-1	BLIP-2	Anchors + Names	**Instruction:** Describe the visible static road context in one short sentence. **Scene:** residential street. **Event sequence:** slow → stop → turn-left. **Anchors:** stop-sign ahead; junction marker left. **Regions:** [box id=12]. **Admissible names:** {} (none observed persistently).	The car slows on a residential street, stops at the stop sign, then turns left and continues.
U-1	LLaVA	Frame-Only	Describe the visible road context and static signage in one short sentence.	Slow on a residential road, a pause at a sign, then a left turn and onward.

U-1	LLaVA	Anchors + Names	**Instruction:** Describe the visible static road context in one short sentence. **Scene:** residential street. **Event sequence:** slow → stop → turn-left. **Anchors:** stop-sign ahead; junction marker left. **Regions:** [box id=12]. **Admissible names:** {} (none observed persistently).	Slow on a residential street, stop at the stop sign, then turn left and continue.

In this evaluation, the Anchors + Names regime again provided consistent improvements for both models. BLIP-2 achieved higher CLIPScore and grounding precision compared to LLaVA in both regimes, with the greatest difference observed in the Frame-Only setting. The normalised edit distance decreased for both models when structured inputs were used, indicating improved event order preservation.



4.8 CADC Robustness Testing

Evaluation on the Canadian Adverse Driving Conditions (CADC) dataset demonstrated the impact of severe visual degradation on model performance. Snowfall, glare, and reduced contrast significantly impaired OCR recall, thereby lowering the accuracy of name grounding across all tested configurations. Despite this, temporal ordering of manoeuvres remained largely unaffected, owing to the optical flow–derived event lattice's robustness to appearance changes in the visual stream.

Quantitative results, summarised in table below, show that BLIP-2 exhibited the highest resilience to adverse weather, achieving a CLIPScore of **0.287**, grounding precision of **0.74**, grounding recall of **0.66**, and a

grounding F1 score of **0.70**. Temporal sequence accuracy remained high at **0.86**, while the normalised edit distance was contained at **0.11**. LLaVA, though maintaining operational capability, experienced a larger drop in all metrics, with a CLIPScore of **0.271**, grounding precision of **0.69**, recall of **0.60**, and F1 of **0.64**, alongside sequence accuracy of **0.82** and a normalised edit distance of **0.14**.

Results for BLIP-2 and LLaVA on CADC (Adverse Weather) Subset

Model	CLIPScore	G-P	G-R	G-F1	Seq. Acc.	NED
BLIP-2	0.287	0.74	0.66	0.70	0.86	0.11
LLaVA	0.271	0.69	0.60	0.64	0.82	0.14

The reduced performance across models can be attributed primarily to environmental interference in OCR detection, which is more vulnerable to adverse conditions than scene classification or object detection. Nevertheless, the optical flow–based temporal scaffolding ensured that the ordering of events remained consistent and resistant to visual degradation. This stability is critical for maintaining narrative coherence in real-world deployments, where weather-induced variance is inevitable.

4.9 Ablation Studies

The ablation studies aimed to isolate the contributions of two key components: the temporal scaffold (event lattice) and OCR gating (verified lexicon).

When the flow-derived event lattice was removed, temporal coherence suffered significantly. Sequence accuracy (Seq. Acc.) dropped from 0.88 to 0.78, while normalized edit distance (NED) doubled, indicating more frequent misordered events. This confirms that the explicit temporal scaffold is the primary driver of coherent event sequencing in the generated summaries.

Similarly, disabling OCR-gated naming severely impacted factual grounding. Grounding precision plummeted from 0.82 to 0.49, and the system frequently hallucinated unverified place names. These results demonstrate that OCR verification is essential for maintaining factual accuracy and trustworthiness in the final narrative.

Together, these findings validate the core design principles of the pipeline: external temporal scaffolding ensures proper event ordering, while OCR-gated naming prevents hallucination and enforces entity-level faithfulness.

4.10 Qualitative Analysis

Structured conditioning significantly enhances the quality and precision of generated summaries by enforcing explicit constraints on both content and temporal order. By incorporating anchor tokens, verified OCR-derived names, and scene labels into the captioning prefix, the system produces more specific and grounded references—for instance, identifying a landmark as "Fox and Hounds" rather than generically as "a pub." This approach also ensures correct sequencing of manoeuvres in multi-event clips, as the narration strictly adheres to the optical-flow-derived event lattice. Additionally, structured conditioning minimizes repetition and generic filler by constraining the language model to focus only on verified visual evidence.

For example, when comparing BLIP-2 outputs with and without structured conditioning, the difference is striking. A Frame-Only prompt yields a generic description: "The car stops at a light, moves ahead past shops, then takes a left before continuing along the street." In contrast, the Anchors+Names version—augmented with scene context, event tags, and OCR-verified entities—produces a far more precise and actionable summary: "The car stops at a red light, passes shops, turns left near Town Hall, pauses by VUE Cinemas, then continues." The latter not only grounds named entities in observable text but also reflects the true order of events, demonstrating the critical role of structured inputs in generating faithful, auditable narratives.

4.11 Technical Insights

The experimental outcomes validate the key architectural choices of the proposed pipeline. Precision-tuned static fixture detection improved OCR reliability, while scene classification ensured contextual vocabulary relevance. Temporal scaffolding from the optical flow module was indispensable for order fidelity, and the verified lexicon proved to be the most effective measure for eliminating hallucinations. The modular separation between perception and narration facilitated consistent and fair comparisons between captioning models.

Overall, the system, when configured with structured captioning inputs, an external temporal scaffold, and verified name governance, generated journey summaries that were both linguistically coherent and verifiably grounded in visual evidence. The demonstrated resilience under challenging environmental conditions and the ability to audit every output element against stored perceptual data position the system as a robust solution for real-world navigation, mapping, and review applications.In the “No-OCR” setting, grounding recall is

undefined because there is no allow-list against which to match named entities; qualitative inspection confirms the appearance of spurious place names.

5. Conclusion & Future Work

This thesis proposed and evaluated a grounded, two-layer method for turning car-driving videos into coherent, human-readable journey narratives. The system intentionally separates what is seen from what is said. A perception layer detects stable, navigationally meaningful structure—traffic control devices, street signage, façades and context—while imposing an explicit temporal scaffold from dense optical flow and promoting reliable named landmarks through persistence-based OCR. A language layer then performs two controlled transformations: a vision–language captioner renders each sampled frame into a concise, evidence-conditioned sentence, and an instruction-tuned decoder composes a paragraph-length account that respects the externally supplied event order and confines named entities to what OCR has verified. The Bedford drives with GPS logging established the in-distribution operating regime; CADC provided winter scenes that stress the same models without changing any thresholds, allowing the system's behaviour under glare, snow and low contrast to be assessed fairly.

Across labelled and unlabelled settings, the design choices proved decisive. Externalising temporal structure eliminated the repetition and ordering errors that typically arise when frame-wise captions are stitched without guidance. Constraining the lexicon to a persistence-voted OCR shortlist curtailed invented toponyms and produced summaries that remain re-locatable by readers. Holding perception constant while swapping captioners made differences attributable to the language component rather than upstream variance; under that control, BLIP-2 consistently delivered the strongest text–reference and text–image alignment, LLaVA was competitive and easily steered by instruction phrasing, and a lightweight Ollama-hosted alternative offered acceptable summaries when footprint is paramount. The primary contribution is methodological: an auditable, modular pipeline in which time is a first-class input and names are earned rather than imagined. The main limitations are equally clear. Night-time retroreflective surfaces, small or occluded signs, and heavy precipitation depress OCR and reduce lexical specificity; very small control devices remain difficult for the detector; and captioners retain biases from web-scale pre-training unless carefully prompted. These observations motivate targeted extensions.

5.1.1 Geospatial priors and uncertainty-aware narration

A natural next step is to make modest geospatial and temporal signals available to both perception and language in a way that remains transparent and privacy-preserving. Heading rate and coarse speed profiles derived from the GPS logger can regularise turn onsets and stops, reducing sensitivity to short optical-flow transients and sharpening the placement of manoeuvre clauses. Lightweight map semantics—junction types, crossings, roundabouts, bridges—from open map sources can be passed as optional fields in the evidence preamble so the narrator can choose functionally appropriate terms in visually ambiguous scenes. Alongside

richer priors, the text should expose its confidence: each named entity and manoeuvre clause can carry a confidence derived from OCR persistence, detector agreement and flow stability. Annotating summaries with these confidences, even if only in machine-readable form, enables downstream consumers to discount low-evidence statements without altering the prose itself.

5.1.2 Beyond CLIPScore: task-grounded and temporal evaluation

CLIPScore is useful when human references are absent, but it does not penalise misordered events or over-general language. Future evaluation should therefore add measures that directly reflect the intended use. Temporal faithfulness can be assessed with sequence-level accuracy, normalised edit distance and rank correlations between predicted and annotated manoeuvre orders. Narrative utility can be probed with task-based studies—route-following on a desktop map or targeted question answering about the journey—so automatic scores are anchored to human success. Reference-free faithfulness checks tailored to this domain should test whether named landmarks in the text come from the OCR allow-list and whether simple spatial predicates such as “before the junction” or “on the right” agree with coarse geometry from detections and flow. Reporting clip-wise distributions and confidence intervals should remain standard so that geography, lighting and weather variance is visible rather than averaged away.

5.1.3 Broadening perceptual anchors while preserving concise narration

The narrative improves when it can point to stable, memorable structure rather than transient actors. Extending perception to include lane topology and markings, drivable space, zebra crossings, refuge islands, bollards and road-surface icons would let the narrator explain why a vehicle slows, merges or yields. The challenge is not to enumerate everything the camera sees, but to curate a vocabulary of static anchors that persist across frames and help readers re-locate themselves. Temporal association and persistence thresholds should promote only durable elements into the preamble so the text remains concise. On the text side, multilingual OCR with script detection and orientation handling would broaden applicability beyond Latin-script high streets and reduce the chance of partial strings entering the allow-list in mixed-script environments. Finally, adverse-weather augmentation targeted at small, reflective signs would harden the detector–OCR path where the current system is most fragile.

5.1.4 Structured generation and alternative multimodal models under the same scaffold

The present narrator already follows external constraints, but stronger structure can be made explicit before any prose is emitted. A two-stage decoder that first emits a compact, machine-readable route graph—ordered events, landmark references, coarse spatial relations—and then realises it as text would make constraints

testable and editable, improving reliability and controllability. Within the same scaffold, alternative captioners that accept structured evidence as explicit inputs can be compared fairly to BLIP-2 and LLaVA, and distilled or quantised variants can reduce latency for edge deployment. On the narration side, constraint-aware decoding strategies can enforce allow-lists and event ordering at the token level, while small video-native modules with short memory windows can be introduced to improve phrasing at complex junctions without incurring the cost of full video LLMs. Throughout, perception should remain fixed when comparing models so improvements are attributable to language rather than upstream change.

Taken together, these directions extend the core idea of this thesis—separating perception from narration and externalising time and names—into a system that remains readable under stress, reveals what it knows and how well it knows it, and scales from research rigs to practical, on-device use.

References

1. Literature Review - Video-ChatGPT Framework, MBZUAI multimodal understanding research documenting temporal visual-language integration approaches.

2. Literature Review - GPT with Vision for Video Understanding, OpenAI framework research on vision-language integration limitations and capabilities.

3. Literature Review - TimeSFormer Attention-based Video Understanding Framework, research on spatio-temporal dependencies in video sequences.

4. Journey Summarisation Research Documentation - Technical architecture, dataset specifications (BDD100K, CADC, CARLA), evaluation metrics, and performance analysis from comprehensive multi-modal pipeline implementation.

5. System Architecture Documentation - FinalBlip2workingcode.txt detailing modular design including YOLOv8n, EasyOCR, Places365, BLIP-2, and Mistral-7B integration with umbrella term landmark categorisation.

6. Supervisor Meeting Transcripts - Research guidance emphasizing data collection methodologies, ground truth evaluation requirements, independent evaluation skills development, and landmark categorisation approaches (referenced for methodological requirements only).

7. Autonomous Systems Applications of Large Language Models Thesis - Cranfield University MScAAI 2024 research on vision-language model integration for remote sensing applications.