Eurographics Conference on Visualization (EuroVis) 2026
N. Andrienko, M. Chen, and B. Wang
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 3

URBANCLIPATLAS: A Visual Analytics Framework for Event and Scene Retrieval in Urban Videos

Joel Perca[1], Luis Sante[1], Juanpablo Heredia[1], Joao Rulff [1,2], Claudio Silva[2] and Jorge Poco[1]

[1]Fundação Getulio Vargas, Brazil
[2]New York University, USA

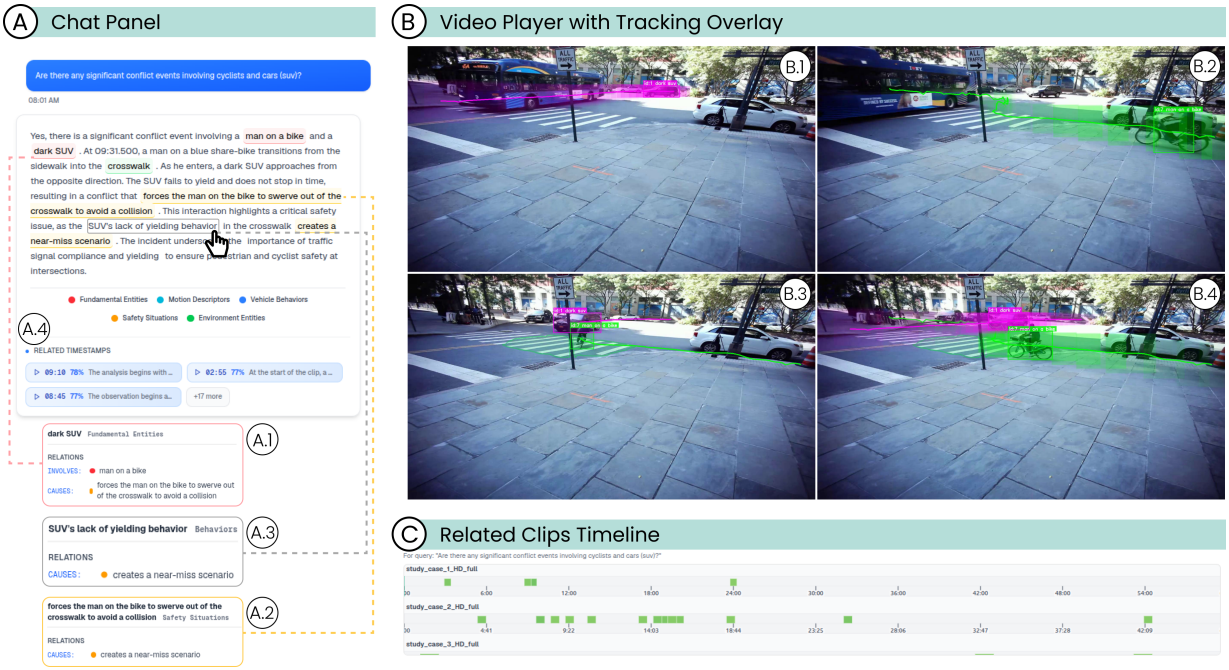

Figure 1: URBANCLIPATLAS *interface.* *(A) The* Chat Panel *displays the user's query, the RAG-generated narrative answer, and entity-level tooltips linked to the knowledge graph. (B) The* Video Player with tracking overlays *shows the current frame with dynamic entities and highlighted static layout elements. (C) The* Related Clips Timeline *summarizes retrieved clips across videos, with cells encoded by their semantic relevance to the query.*

Abstract

Extracting actionable insights from long-duration urban videos is often labor-intensive: analysts must manually sift through raw footage to pinpoint target events or uncover broader behavioral trends. In this work, we present URBANCLIPATLAS*, a visual analytics system for exploring long urban videos recorded at street intersections.* URBANCLIPATLAS *combines retrieval-augmented generation (RAG), taxonomy-aware entity extraction, and video grounding to support event retrieval and interpretation. The system segments extended recordings into short clips, generates textual descriptions with a vision–language model, and indexes them for semantic retrieval. A knowledge graph maps entities and relations from LLM answers onto a domain-specific taxonomy and aligns them with detected objects and trajectories to support visual grounding and verification.* URBANCLIPATLAS *supports scene retrieval through an augmented chat-based interface and improves scene interpretation by tightly aligning textual outputs with video evidence. This design strengthens the connection between textual reasoning and visual evidence, reducing the effort required to validate model outputs and refine hypotheses. We demonstrate the usefulness of* URBANCLIPATLAS *on the StreetAware dataset through two case studies involving hazardous scenarios and crossing dynamics at street intersections.* URBANCLIPATLAS *helps analysts reason about safety- and mobility-related patterns across large urban video collections.*

CCS Concepts

• ***Human-centered computing*** → *Visualization; Visualization systems and tools;*

1. Introduction

Understanding the multifaceted dynamics of urban environments is fundamental to fostering resilient, safe, and equitable cities. Street intersections are key nodes where diverse traffic flows converge, occupying only a small fraction of the street network yet playing a disproportionate role in safety and efficiency. The National Association of City Transportation Officials (NACTO) calls them "one of the most critical and most complicated elements in roadway design," and the US Department of Transportation (USDOT) reports that intersections account for a substantial share of traffic fatalities, with economic and societal costs reaching billions of dollars [Fed24, BMZ*15].

To optimize these locations, urban planners and transportation engineers rely on geometric and operational interventions—such as curb extensions, raised medians, and signal timing changes [SMMSL20]. Yet the intended effect of a design often differs from its real-world outcome. Intersections are not just static infrastructure but dynamic environments shaped by human behavior, where pedestrians, cyclists, and vehicles negotiate shared space. Assessing safety and efficiency thus requires granular observations of how these agents actually move and interact.

New sensing infrastructures now enable extended recording of intersection activity. Unlike earlier approaches based on aerial imagery [CZYW17, BHM*25] or GPS traces [DHZ*18, AH20], modern systems capture rich, high-definition video directly from street level [PRB*23]. This street-view perspective provides nuanced information about vulnerable road users, fine-grained behavioral cues, and subtle interaction dynamics [RPH*24].

Despite these advances, domain experts still lack scalable tools to retrieve and analyze specific events without laborious manual inspection. Existing automated approaches typically rely on heavily supervised models and a limited vocabulary of predefined events [ASM*21, AE24], which restricts exploratory analysis and rapid adaptation to new policy questions [HERMCP25]. Vision–language models (VLMs) offer a promising alternative: their zero-shot capabilities and natural-language interfaces allow analysts to describe scenarios in plain text and retrieve matching video segments. However, interpreting VLM outputs in complex urban scenes remains challenging. Analysts must relate long textual descriptions to visual counterparts involving many entities and intricate spatial relationships, a cognitively demanding process that is amplified when working with long-duration recordings.

There is thus a need for visual analytics techniques that structure and enrich VLM outputs, connect them to higher-level semantic models of intersection events, and ground these semantics in video to support sensemaking. To address this need, we introduce URBANCLIPATLAS, a visual analytics framework for event and scene retrieval in long-duration urban intersection videos. URBANCLIPATLAS combines a retrieval-augmented generation (RAG) architecture with taxonomy-aware entity extraction and smart video analysis to support both event retrieval and interpretation. The system segments long recordings into clips, generates textual descriptions with a VLM, and indexes them for semantic retrieval. At query time, natural-language questions are enriched to better match the indexed descriptions; relevant clips are retrieved; and entities mentioned in language-model responses are extracted and organized into a knowledge graph guided by a domain-specific taxonomy of intersection events. Object detection and multi-object tracking then align these entities with visual instances, providing grounding that links semantic descriptions to video evidence. In contrast to existing video RAG systems that operate mainly at the text-index level, our approach integrates retrieval, taxonomy-guided entity alignment, and video grounding within an interactive visual analytics framework. This framework supports output-level auditing of black-box models through an augmented narrative view that presents model-generated answers alongside supporting video snippets, highlighted entities, and trajectories, reducing the effort required to interpret model outputs. In summary, this work makes the following contributions:

- **RAG-based video processing pipeline.** We propose a video RAG pipeline that segments long-duration recordings into clips, generates textual descriptions with a vision–language model, indexes them for semantic search, and performs prompt enrichment at query time to support interactive identification of salient video segments.
- **Taxonomy-aware entity alignment and visual grounding.** We introduce a knowledge-graph augmentation strategy that extracts entities from VLM responses, aligns them with detected visual objects, and organizes them according to a domain-specific taxonomy of urban intersection events, enabling consistent text–video grounding and higher-level reasoning.
- **Visual analytics system and evaluation.** We implement URBANCLIPATLAS, an open-source visual analytics system that integrates these components into an augmented-narrative interface, bridging textual descriptions with visual evidence and supporting human-in-the-loop verification of model outputs. We demonstrate its usefulness through two case studies on the StreetAware dataset and report insights from structured interviews with urban specialists on usability and deployment potential.

2. Related Work

Our work lies at the intersection of RAG, video understanding, and visual analytics for urban monitoring.

Retrieval-Augmented Generation for Video Understanding. RAG enhances the factual accuracy and knowledge scope of large language models (LLMs) by incorporating external knowledge sources [JKBH25]. Extending RAG to video introduces additional challenges, including the cost of converting long videos into text and the loss of multimodal cues such as motion and spatial context when relying solely on textual representations [RXX*25]. These issues become particularly pronounced for long-duration corpora, where purely textual indexing often fails to preserve event-level structure [ADUC24, RXX*25]. To address these limitations, several systems adopt incremental pipelines that delay expensive processing until necessary. Some enrich representations only for retrieved segments [ADUC24], while others maintain evolving knowledge structures to support low-latency retrieval as new content arrives [SRC25]. Graph-based indices and hierarchical retrieval strategies [GXY*25], as well as cascaded pipelines combining lightweight models with VLM refinement [ADSUC24], further reduce computational costs.

Another line of work strengthens retrieval by preserving multimodal signals rather than relying solely on text. Approaches augment VLM inputs with auxiliary channels such as OCR, ASR, or object detections [LZY*24], apply question-aware frame sampling [TYL*25], or jointly reason across video, audio, and text [MPCK*25]. While these techniques improve retrieval relevance, they still struggle to capture long-range temporal dependencies. More recent approaches construct structured representations across multiple videos, treating collections as interconnected knowledge spaces. Graph-based grounding and multimodal encoders enable reasoning across entities and events at the corpus level [RXX*25, JKBH25, ADC24]. However, these strategies introduce challenges related to entity canonicalization, graph maintenance, and multimodal alignment at scale.

Retrieval-oriented approaches are particularly promising for urban analytics, where scenes evolve rapidly and contextual grounding is essential [ADC24, MPCK*25]. In contrast to these general-purpose retrieval methods, URBANCLIPATLAS incorporates a domain-specific taxonomy and an explicit grounding loop to ensure that retrieved narratives remain aligned with the semantic requirements of urban safety analysis and can be audited against the underlying video evidence.

Object Detection and Tracking for Grounding. Grounding language in video requires linking textual descriptions to entities in space and time through detection, tracking, and spatial reasoning. Conventional closed-vocabulary detectors often struggle in open environments and fail to maintain consistent identities when objects are occluded or change appearance [JCL*24]. These issues are amplified in dense urban scenes involving heterogeneous agents. Open-vocabulary detection addresses this limitation by aligning visual features with textual embeddings. YOLO-World [CSG*24] enables real-time detection of unseen categories, while OV-STAD [WGQ*24] combines open-vocabulary detection with action understanding. Despite these advances, maintaining temporally stable identities across frames remains challenging.

Recent work also emphasizes spatial reasoning. Jiang et al. [JCL*24] propose a pipeline that links entities, attributes, and relations across frames, supporting more temporally consistent alignments between visual content and generated text. SpatialBot [CPY*25] leverages a VLM to reason about distances, directions, and geometric relations, enabling explicit spatial understanding. Together, these systems point toward grounding models that can both identify objects and situate them in their environment. These capabilities are vital for interpretable, context-aware video analysis, but they do not by themselves constrain how VLM or LLM outputs are used. In URBANCLIPATLAS, we close this loop by combining detection, multi-object tracking, and a specialized taxonomy into a grounding layer that explicitly checks model-generated claims against concrete visual evidence in urban intersection videos, exposing mismatches to the analyst.

Visual Analytics for Video Data. Visual video analytics integrates computer vision with interactive visualization to support human reasoning over complex video data [SH19]. Prior systems demonstrate the value of linking automated analysis to visual interfaces that enable analysts to inspect and guide algorithmic results. Several systems combine visual features with textual information. Wu and Qu [WQ20] explore storytelling patterns in TED talks using transcripts and video features, while Chen et al. [CBH*20] integrate trajectory analysis with spatial views to study traffic behavior. Motion Browser [CNC*20] similarly employs coordinated views to analyze motion patterns in clinical data. Recent visual analytics systems have also explored interactive analysis of spatiotemporal phenomena through coordinated structural and temporal views, including association-rule-based exploration of categorical data and visual analysis of evolving geographic regions [DSP*25, NDP25].

Although these systems effectively expose spatiotemporal patterns, they typically rely on predefined features and offer limited support for natural-language queries and generative explanations. Moreover, high-level descriptions are rarely grounded to concrete entities in the video. To address these gaps, URBANCLIPATLAS introduces a visual analytics framework that couples a RAG-based video pipeline with taxonomy-aware entity alignment and explicit grounding. By anchoring model-generated explanations directly to retrieved clips and verifiable entities in the scene, the augmented narrative interface turns otherwise opaque VLM/LLM outputs into auditable, evidence-linked insights for urban safety analysis.

3. System Overview

Prior work on RAG-based video retrieval and visual video analytics has highlighted persistent challenges in (1) semantically searching long-duration videos, (2) exposing entity-level spatiotemporal context, and (3) supporting explanation and validation of algorithmic outputs directly in video frames (*e.g.*, [RXX*25, ADC24, SH19, WQ20, CBH*20]). Guided by these observations and formative discussions with urban traffic analysts, we derive a set of design goals and corresponding technical tasks for URBANCLIPATLAS. We then describe how these goals are operationalized through the three-stage workflow illustrated in Fig. 2.

3.1. Design Goals

We conceptualize URBANCLIPATLAS as a visual analytics system that sits between raw video streams and domain experts' questions about safety and behavior at intersections. The following design goals capture the capabilities needed to analyze large collections of urban videos.

G1 Text-based semantic access to long-duration videos. Enable analysts to formulate information needs in natural language and retrieve semantically relevant video segments, avoiding manual inspection of hours of footage.

G2 Entity-centric spatiotemporal context. Provide representations that expose the locations of entities (*e.g.*, vehicles and pedestrians) in the scene and how their states and interactions evolve over time, supporting reasoning about trajectories, conflicts, and usage patterns.

G3 Transparent, visually grounded explanations. Couple model-generated narratives with explicit visual evidence so that analysts can quickly understand, verify, and critique system outputs rather than treating them as black-box predictions, thereby supporting output-level auditing of VLM/LLM behavior in real scenes.

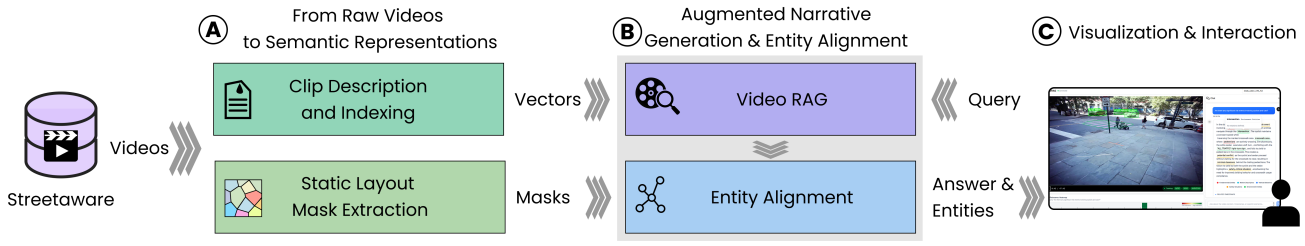

Figure 2: URBANCLIPATLAS *main workflow.* *(A)* Preprocessing*: long videos are segmented into clips, described by a VLM, indexed in a vector store, and paired with static layout masks, yielding semantic and spatial representations for each video. (B)* Augmented Narrative Generation*: at query time, the system enriches the user's question, retrieves relevant clips, and composes narrative answers using the precomputed embeddings, knowledge graph, and masks. (C)* Visualization*: the interface receives user queries, integrates outputs from all components, and supports interactive exploration and validation of retrieved content.*

3.2. Design Tasks

To operationalize these goals in a visual analytics setting, we specify the following technical design tasks for URBANCLIPATLAS:

T1 Natural-language query parsing and semantic retrieval (G1). Implement a text interface that accepts free-form questions, performs query enrichment with domain terminology, encodes text into an embedding space, and retrieves candidate clips from a vector index of video descriptions.

T2 Entity extraction, taxonomy mapping, and knowledge-graph construction (G2, G3). From RAG-generated answers, automatically detect entity mentions such as a car turning right or a pedestrian crossing late, map these mentions to a domain-specific taxonomy of intersection events, and instantiate a knowledge graph that captures entity types, roles, and relationships.

T3 Spatiotemporal grounding via detection and tracking (G2, G3). Run object detection and multi-object tracking over video clips, associate tracked objects with textual entities from the knowledge graph, and maintain persistent identifiers and trajectories across frames and clips.

T4 Linked visual encodings for narrative, entities, and video (G2, G3). Design coordinated views so that the augmented narrative, entity list/graph, and video player share common references. Selecting an entity in the text highlights the corresponding track in the video and vice versa.

T5 Interaction techniques for clip navigation and comparison (G1–G3). Provide mechanisms to navigate between retrieved clips, refine queries, and compare alternative narratives (*e.g.*, multiple queries or parameter settings), supporting iterative exploration of complex behaviors across the video corpus and human-in-the-loop refinement of the system's interpretations.

3.3. System Workflow

To realize these goals, URBANCLIPATLAS transforms raw urban videos into an interactive analysis environment through three integrated stages (Fig. 2). In the *Preprocessing* stage (A), the system segments long-duration street-level recordings, such as those from the StreetAware dataset [PRB*23], into overlapping clips. It then processes these clips offline to obtain two complementary representations: textual descriptions generated by a vision-language model and stored in a vector index for semantic retrieval, and static layout masks that describe the intersection infrastructure, such as crosswalks, sidewalks, and lanes. During *Augmented Narrative Generation* (B), analysts submit natural-language questions. URBANCLIPATLAS enriches these queries with domain terminology, retrieves relevant clips from the semantic index, and uses a RAG pipeline to compose narrative answers. Entities are extracted and mapped to a taxonomy of intersection events, and then instantiated in a knowledge graph that can be aligned with downstream object detection and tracking results. Finally, the *Visualization and Interaction* stage (C) presents these results through an augmented narrative interface that coordinates a video player, a narrative view, and an entity/graph view. Analysts can move smoothly between queries, explanations, and visual evidence to explore and validate safety- and mobility-related patterns across the video corpus.

4. From Raw Videos to Semantic Representations

To support semantic search and retrieval over long urban videos, URBANCLIPATLAS performs an offline preprocessing stage (Fig. 2A) that converts raw intersection footage into structured, queryable representations. This stage comprises two complementary procedures: (1) clip description and indexing for text-based retrieval and (2) static layout mask extraction to contextualize events within the physical infrastructure.

4.1. Clip Description and Indexing

Each input video v has duration T_v seconds (not necessarily the same across videos). To preserve continuity and reduce the risk of missing important interactions at segment boundaries, we divide each video into fixed-length clips ($\tau = 30$ seconds) with a 5-second overlap ($\omega = 5$). For a given video v, the number of clips is

$$m_v = \left\lceil \frac{T_v - \tau}{\tau - \omega} \right\rceil + 1,$$

and the i-th clip $C_{v,i}$ starts at timestamp $t_{v,i} = (i-1)(\tau - \omega)$.

Each segmented clip $C_{v,i}$ is then processed by a VLM to generate a textual description summarizing its visual content and activities. Given a prompt template $\mathcal{P}$, designed to emphasize traffic entities, behaviors, and safety-critical situations, we obtain $D_{v,i} =$

Figure 3: ***Hierarchical taxonomy for knowledge-graph construction.*** *The taxonomy organizes urban-traffic concepts into five top-level categories*—Agents, Motion Descriptors, Individual Behaviors, Safety Situations, *and* Environment Entities—*providing a consistent semantic basis for entity extraction, event classification, and grounding across all components of* URBANCLIPATLAS.

$\mathrm{VLM}(C_{v,i}, \mathcal{P})$. During this stage, we also adjust timestamps when needed to ensure temporal consistency across descriptions by correcting offsets introduced by segmentation or buffering.

To enable semantic similarity search, the textual descriptions are embedded in a continuous vector space. An embedding model E encodes each description $D_{v,i}$ as

$$\mathbf{e}_{v,i} = \mathrm{E}(D_{v,i}), \quad \mathbf{e}_{v,i} \in \mathbb{R}^d,$$

where d denotes the dimensionality of the embedding space. We store these embeddings, together with their metadata (timestamps, clip identifiers, and original descriptions), in a vector database, yielding an indexed collection of clip descriptors that supports efficient semantic retrieval during the query phase.

4.2. Static Layout Mask Extraction

To provide spatial context for retrieved events and to distinguish moving agents from static infrastructure, URBANCLIPATLAS extracts semantic masks of each intersection's layout. The goal is to obtain a clean representation of the environment (*e.g.*, crosswalks, sidewalks, and lanes).

For each video v, we first sample n frames uniformly across its duration, denoted $F_{v,1}, \ldots, F_{v,n}$. Each sampled frame is processed with the open-vocabulary detector YOLO-World [CSG*24] to identify dynamic entities such as vehicles, pedestrians, and cyclists. We selected YOLO-World for its state-of-the-art open-vocabulary detection, which allows us to specify the traffic-related categories of interest. To reduce false positives and enhance precision, we apply a detection confidence threshold of 0.65 across all YOLO-World outputs, based on empirical tuning and used consistently in all our experiments.

Let $N(F_{v,j})$ denote the number of detected objects in frame $F_{v,j}$. We choose as a reference the frame with the lowest detected object count, $F_v^* = \arg\min_{j \in \{1,\ldots,n\}} N(F_{v,j})$, which in practice yields a view of the intersection where static infrastructure is least occluded by traffic. We then apply the semantic segmentation model Mask2Former [CMS*22] to F_v^* to obtain a set of masks $\mathcal{M}_v = \{M_{v,1}, \ldots, M_{v,p}\}$, where each mask corresponds to a specific static component of the environment (*e.g.*, crosswalks, traffic signals, road surfaces, sidewalks, and lane markings). These layout masks serve as spatial reference layers throughout the pipeline, enabling subsequent modules to contextualize detected events relative to the intersection's physical geometry and supporting later tasks such as entity-level grounding and spatial reasoning.

5. Augmented Narrative Generation and Entity Alignment

The answer generation stage is the core of URBANCLIPATLAS (Fig. 2B). It operates on the preprocessed clip index and layout masks and (i) retrieves relevant clips through a video RAG pipeline and (ii) structures and grounds the resulting narrative using a taxonomy-guided knowledge graph.

5.1. Video RAG

The video RAG component processes a user's natural-language query through three steps: query enrichment, semantic retrieval, and narrative generation. It produces (1) a query-aligned narrative answer and (2) metadata about the selected segments as evidence.

Query Enrichment. Given a user query Q, we first expand it into an enriched query Q' to better match the indexed clip descriptions. A language model augments Q with traffic-specific terminology, paraphrases, and contextual hints drawn from a domain context $\mathcal{K}$ (*e.g.*, intersection events and entities) as $Q' = \mathrm{LLM}(Q, \mathcal{K})$. This improves recall by reducing vocabulary mismatch and also helps filter invalid or out-of-domain questions.

Semantic Retrieval. We embed the enriched query Q' in the same

semantic space as the clip descriptions defined in the preprocessing stage using the embedding model E:

$$\mathbf{q} = \mathrm{E}(Q'), \quad \mathbf{q} \in \mathbb{R}^d.$$

The vector database then computes cosine similarity between $\mathbf{q}$ and all clip embeddings $\mathbf{e}_{v,i}$, ranks clips by similarity, and returns the top-k candidates $\mathcal{R}_k = \{(v,i)\}_{j=1}^{k}$, where each pair (v,i) indexes a retrieved $C_{v,i}$ and its associated $D_{v,i}$.

Narrative Generation. The narrative answer is generated by conditioning an LLM on the original query and the retrieved evidence. In our implementation, we use the highest-ranked clip $C_{v,i}$ together with its description and temporal metadata:

$$A = \mathrm{GenerateNarrative}(Q, C_{v,i}, D_{v,i}, t_{v,i}).$$

The answer A follows domain terminology and explicitly references intersection entities, behaviors, and safety situations. In the next stage, we convert A into a structured representation and align its entities with the video content.

5.2. Taxonomy-Guided Entity Alignment

Raw answers can be long and difficult to interpret. To help analysts understand and validate the generated narratives, URBANCLIPATLAS augments them with a structured, grounded representation. This process has two parts—knowledge graph construction and entity grounding—whose main components are summarized in Fig. 4. The top panel depicts the Knowledge Graph Construction module, and the bottom panel depicts the Entity Grounding Engine.

Knowledge Graph Construction. We represent the answer as a knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ denote entities and edges $\mathcal{E}$ denote relationships. The *Knowledge Graph Construction* module (top part of Fig. 4) performs this transformation. Entity extraction is guided by a fixed crossroad-interaction taxonomy $\mathcal{T}$ inspired by Shirazi and Morris [SM17] and adapted to our setting. Fig. 3 illustrates this taxonomy, which organizes concepts into five top-level categories: agents, motion descriptors, individual behaviors, safety situations, and environment entities, together with examples of each entity type considered.

An LLM-based extractor takes the answer A and outputs entity mentions. Each mention has a class in $\mathcal{T}$, a text span, optional attributes (*e.g.*, *red car*, *turning right*), and a position in the answer. We then perform a light canonicalization step to merge mentions of the same real-world entity (*e.g.*, "a red car" and later "the vehicle" in the same scene) using lexical overlap and LLM-based coreference scoring. The result is a deduplicated set of nodes $\mathcal{V}^*$ with enriched attributes.

Relations are extracted as directed triples (u_s, r, u_o) where $u_s, u_o \in \mathcal{V}^*$ and r is a relation type such as *approaches*, *yields-to*, or *conflicts-with*. The resulting graph $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E})$ provides a compact semantic summary of the narrative that can be visualized and used for grounding.

Entity Grounding Engine. While $\mathcal{G}^*$ captures the semantics of the answer, analysts also need to see where entities appear in the video. The Entity Grounding Engine (bottom part of Fig. 4) attaches visual footprints (bounding boxes or masks) to the entities in $\mathcal{V}^*$ by

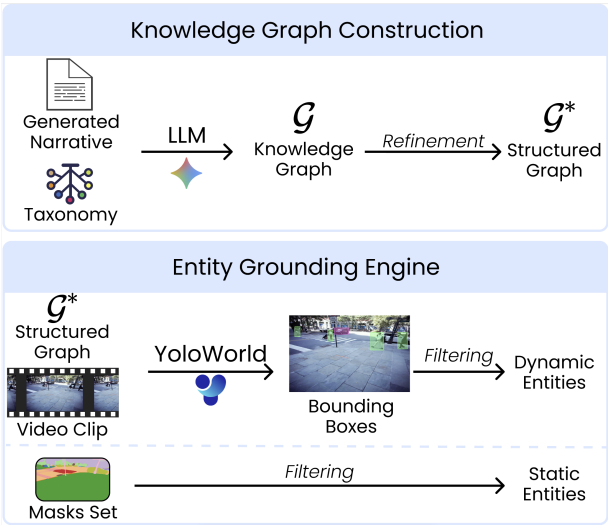

Figure 4: ***Taxonomy-guided entity alignment.*** *The* Knowledge Graph Construction *module uses the generated answer and the fixed taxonomy to build a structured graph $\mathcal{G}^*$ of entities and relations. The* Entity Grounding Engine *then combines $\mathcal{G}^*$, the retrieved clip, and the precomputed masks to localize dynamic entities through detections and tracks, and to select relevant static environment masks for visualization.*

combining YOLO-World detections with the static layout masks from Sec. 4.2.

For *dynamic entities* (*e.g.*, vehicles and pedestrians), we use YOLO-World with the textual description of each relevant entity $u \in \mathcal{V}^*$ as prompt. Given a frame F_t at time t, the detector returns a set of bounding boxes $\mathcal{B}_u = \mathrm{YOLOWorld}(F_t, \mathrm{span}(u))$, where each $b \in \mathcal{B}_u$ encodes position, size, and confidence. Across frames, these detections can be linked into short tracks.

For *static environment entities* (*e.g.*, crosswalks, signals, and lane markings), grounding uses the precomputed layout masks $\mathcal{M}_v$ for the corresponding video. We select only those masks whose class appears in the entity set: $\mathcal{M}_{\text{active}} = \{M \in \mathcal{M}_v \mid \mathrm{class}(M) \in \mathrm{classes}(\mathcal{V}^*)\}$. Restricting to $\mathcal{M}_{\text{active}}$ avoids visual clutter and focuses attention on environmental elements relevant to the current answer.

Together, the knowledge graph and grounding engine connect textual explanations to concrete visual evidence, enabling the interface to highlight entities in both the narrative and the video and to support entity-centric exploration of complex intersection scenes.

6. URBANCLIPATLAS

The URBANCLIPATLAS interface (Fig. 1) is designed as an interactive visual analytics environment centered on an augmented conversational experience. It supports the exploration, retrieval, and validation of specific events in long-duration urban videos by tightly coupling narrative answers, entity-level structures, and video evidence. The interface is organized into three coordinated compo-

nents: the Chat Panel, the Video Player with tracking overlays, and the Related Clips Timeline.

6.1. Chat Panel

The *Chat Panel* (Fig. 1A) is the main entry point for analysts. It receives the user's natural-language query and displays the augmented narrative answer A produced by the video RAG pipeline. The answer describes what happens in the retrieved clip and identifies key actors (*e.g.*, pedestrians, vehicles, and cyclists) as well as environmental elements (*e.g.*, crosswalks, traffic lights, and sidewalks). Using the entity metadata from the knowledge graph, the panel highlights mentions of entities directly in the text. Entities are color-coded according to the taxonomy categories introduced in Fig. 3, with a legend shown at the bottom of the panel (Fig. 1A.4). Clicking or hovering over a highlighted entity links its mention in the text to the corresponding visualization in the video, allowing analysts to directly identify and confirm the entity's role and actions within the event.

When hovering over an entity, an interactive tooltip appears (Fig. 1A.1–A.3). This tooltip summarizes the local neighborhood of that node in the knowledge graph, showing related entities and relationships (*e.g.*, *involves* and *causes*). This design follows validation strategies in prior visual analytics systems that couple LLM reasoning with graph structures [CE25], helping users assess how the model connects actors, behaviors, and safety situations.

The *Chat Panel* also surfaces links to additional clips that exhibit similar behaviors or interactions (Fig. 1A.4), together with a confidence indicator reflecting the relevance of the retrieved clip to the user's query. Selecting one of these related clips updates the *Video Player* and the *Related Clips Timeline*. This supports comparison of how a given pattern manifests at different times or in different locations.

6.2. *Video Player* with Tracking Overlays

The *Video Player* (Fig. 1B) provides the primary visual evidence supporting the narrative answer. Its role is to allow analysts to visually confirm and contextualize the retrieved information. The textual explanation, the knowledge graph, and the video frames remain tightly aligned. To support this, the player renders multiple visual layers on top of each frame in response to interactions originating in the Chat Panel or the Timeline.

Dynamic entities. Dynamic actors, such as vehicles and pedestrians, are localized using YOLO-World. This model is prompted with textual descriptions of entities extracted from the answer. They are rendered as colored bounding boxes whose hues follow the taxonomy. Examples of these overlays are shown in Fig. 1B.1, B.2, and B.4. Clicking a dynamic entity mention in the *Chat Panel* jumps the video to the first frame in which that entity is detected and displays its subsequent trajectory.

Static environment entities. Static infrastructure elements (*e.g.*, crosswalks, sidewalks, lanes, and poles) are rendered using the precomputed layout masks from Sec. 4.2. These masks are drawn more subtly than dynamic entities to provide spatial context without overwhelming the scene (Fig. 1B.3 and B.4). Hovering over an environmental entity in the text highlights the corresponding region in the video. This interaction makes explicit how narrative elements relate to physical infrastructure, revealing where and how entities interact with the environment.

All overlays remain synchronized with temporal navigation. Scrubbing, playing back, or jumping via the Timeline keeps the highlighted entities and trajectories aligned with the narrative. This enables analysts to move seamlessly between detailed frame-level inspection and higher-level reasoning about behaviors and safety situations.

6.3. Related Clips Timeline

The *Related Clips Timeline* (Fig. 1C) acts as a temporal overview and navigation aid for the retrieved results. It visualizes clips ranked by semantic relevance to the current query, as computed by the video RAG component (Sec. 5.1).

Each row in the timeline corresponds to a video, and each cell corresponds to a retrieved clip within that video. The horizontal position encodes the clip's temporal location in the source video, while cell appearance encodes its semantic similarity score (clips more relevant to the query are shown more prominently). When the user restricts the search to a single video, the timeline displays a single row. When the search spans the corpus, multiple rows appear, enabling analysts to compare how the queried event or behavior manifests across different viewpoints or intersections.

Interaction with the timeline is tightly coupled to the other views. Clicking a cell in the active video row causes the *Video Player* to seek to the start of that clip. Selecting a cell from another video switches the player to that source, updates the *Chat Panel*, and updates the overlays accordingly. This coordination allows analysts to move smoothly between corpus-level exploration (identifying where relevant events occur) and detailed inspection of individual scenarios.

6.4. Implementation Details

URBANCLIPATLAS is implemented as a Python backend using `FastAPI` for HTTP and WebSocket communication and a `Svelte + TypeScript` frontend with `D3.js` for interactive visualizations. The system integrates multiple foundation models, including `gemini-2.5-pro` for video captioning, `gemini-embedding-001` for semantic indexing, and `gpt-4o` for query enrichment and narrative generation. Object grounding is performed with YOLO-World, while static layout masks are extracted using Mask2Former.

At query time, the end-to-end response latency is typically around 6–7 seconds on a consumer workstation. Additional implementation details, system configuration, prompts used in the LLM pipeline, and performance measurements are provided in the supplementary material. To support reproducibility, the source code, configuration files, and documentation for URBANCLIPATLAS are publicly available at https://visualdslab.com/papers/UrbanClipAtlas/.

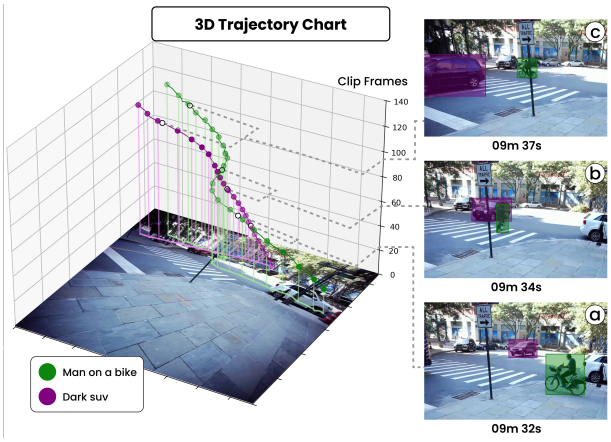

Figure 5: ***Trajectories in Case Study 1.*** *The supporting 3D trajectory chart shows the motion of the* man on a bike *(green) and the* dark SUV *(purple) over time. Insets (a)–(c) show key frames: (a) the onset of the conflict as both approach the crosswalk; (b) the cyclist leaves the crosswalk while the SUV passes; and (c) the cyclist continues along the trajectory after the near miss.*

7. Evaluation

We qualitatively evaluate URBANCLIPATLAS using the *StreetAware* dataset and two case studies. The goal is to illustrate how the system supports event retrieval, multi-entity reasoning, and visual validation of safety-related situations through concrete examples based on real urban videos.

7.1. Dataset: StreetAware

We use the *StreetAware* dataset [PRB*23], which contains high-resolution, synchronized, multimodal street-level videos. These videos were captured at busy urban intersections, with privacy preserved through the removal of audio and the blurring of vehicle license plates and human faces. StreetAware focuses on high-activity crossings in Brooklyn, New York, and includes approximately 8 hours of recordings from three intersections. Unlike traditional automotive datasets captured from a moving ego-vehicle (*e.g.*, KITTI and Cityscapes), StreetAware employs static, multi-angle cameras mounted at intersection corners. This setup provides a stable, wide view of pedestrian–vehicle interactions and the surrounding built environment. It is therefore well suited for analyzing behaviors such as yielding, jaywalking, and near-miss incidents.

7.2. Case Studies

We present two case studies to demonstrate URBANCLIPATLAS's capabilities. Each case study starts from a natural-language query and shows how the *Chat Panel*, *Video Player*, and *Related Clips Timeline* (Fig. 1) work together to surface relevant events and support detailed inspection.

Conflicts between cyclists and cars. In our taxonomy (Fig. 3), a *Conflict* is a *Safety Situation* involving two or more dynamic entities whose trajectories bring them into spatial and temporal proximity, creating a risk of collision without an actual crash [SM17]. This case study examines how URBANCLIPATLAS helps identify and analyze such conflicts between cyclists and vehicles. The analyst issues the query: *"Are there any significant conflict events involving cyclists and cars (SUV)?"* The Video RAG component retrieves a set of relevant clips and generates a narrative answer, which appears in the *Chat Panel* (Fig. 1A). The answer highlights three key entities: *dark SUV*, *man on a bike*, and *crosswalk*, and describes their interactions. The knowledge-graph tooltips (Fig. 1A.1–A.3) expose the main relations:

- *dark SUV* $\xrightarrow{involves}$ *man on a bike*;
- *dark SUV* $\xrightarrow{causes}$ *forces the man on the bike to swerve out of the crosswalk*;
- *forces the man on the bike to swerve out of the crosswalk* $\xrightarrow{causes}$ *creates a near-miss scenario*.

These relations collectively characterize the situation as a conflict event. The *Video Player* with tracking overlays (Fig. 1B) provides the corresponding visual evidence. The dark SUV and the cyclist are rendered as colored tracks, allowing the analyst to observe how the cyclist changes course to avoid the vehicle. Clicking the relevant entities in the *Chat Panel* centers the playback on the moment of maximum risk and reveals the full trajectories of both actors.

To further support reasoning about motion, we use the 3D trajectory view (an external visualization not implemented directly within URBANCLIPATLAS) shown in Fig. 5. The ground plane shows the reference frame of the intersection, while the vertical axis encodes time. The green trajectory corresponds to the *man on a bike*, and the purple one to the *dark SUV*. Panels (a)–(c) on the right show key frames: (a) when the conflict begins as both approach the crosswalk; (b) the cyclist leaves the crosswalk while the SUV passes; and (c) the cyclist resumes their trajectory. Although the projected paths intersect on the ground plane, the temporal axis reveals that the two entities do not occupy the same space at the same time. This confirms the event as a near miss rather than a collision.

The *Related Clips Timeline* (Fig. 1C) lists additional clips ranked by semantic similarity to the query. Most retrieved clips depict interactions between cyclists and vehicles at varying levels of risk.

Take-away. URBANCLIPATLAS successfully retrieves conflict events that match the analyst's query and explains them through coordinated narrative, structure, and video. The *Chat Panel* and knowledge graph clarify which entities and relations define the conflict, while the *Video Player*, the trajectory visualization, and the *Related Clips Timeline* support inspection of how the event unfolds and how it relates to similar scenarios. Together, these components support both the discovery of near-miss events and the nuanced interpretation of their dynamics.

Illegal crossing triggered by bus occlusion. A key advantage of a chat-based interface is that analysts can iteratively refine their questions, moving from broad descriptions to targeted risk assessments. In this case study, we start with a general request: *"Identify any large vehicle blocking the intersection."* (Fig. 6A.1). We chose this example because multiple large vehicles traverse the intersection throughout the day. In response, URBANCLIPATLAS retrieves a clip in which a coach bus initially occupies part of the intersection. The *Video Player* (Fig. 6A.1) highlights the bus with a bounding

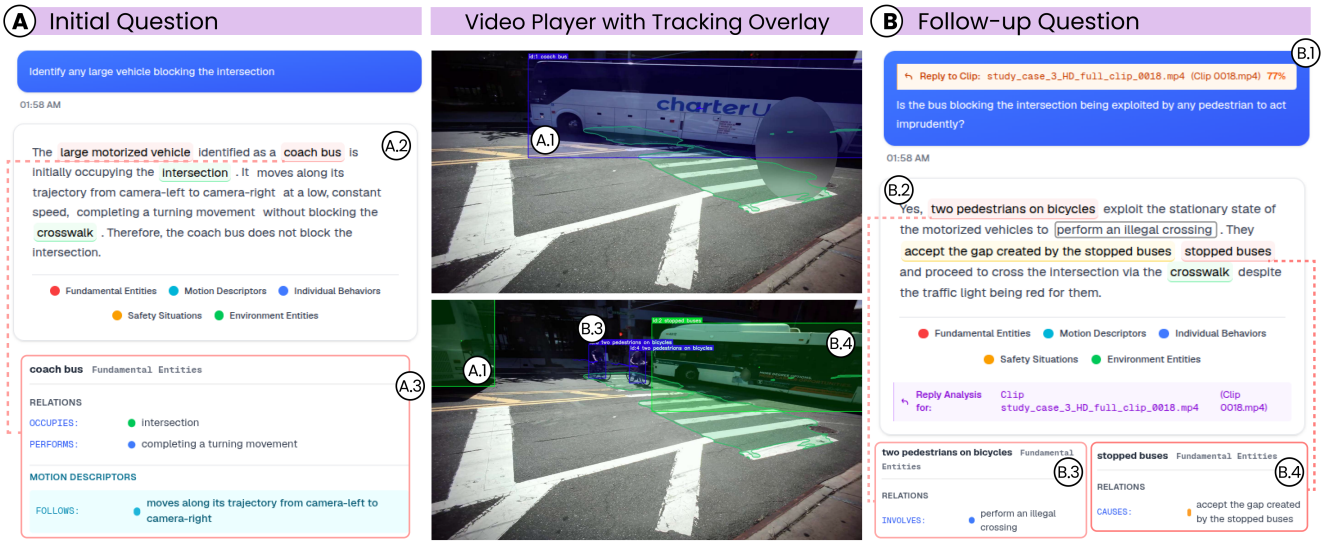

Figure 6: ***Follow-up querying for risk-oriented analysis.*** *(A) Initial prompt asking for large vehicles blocking the intersection: the Chat Panel (A.2–A.3) and Video Player (A.1) highlight a coach bus executing a slow but compliant turning maneuver. (B) Follow-up prompt (B.1) explicitly focuses on safety risks associated with that scene: the updated narrative (B.2) and entity tooltips (B.3–B.4) reveal that a different bus enables an illegal crossing by pedestrians on bicycles.*

box, and the narrative in the *Chat Panel* explains its behavior: *"The coach bus is initially occupying the intersection. It moves along its trajectory from camera-left to camera-right at a low, constant speed, completing a turning movement without blocking the crosswalk."* This description, together with the entity-level breakdown in the tooltip (Fig. 6A.2–A.3), indicates that the bus clears the crosswalk and the intersection without creating an obstruction. At this stage, the system characterizes the vehicle's behavior as compliant.

However, slow turning movements by large vehicles can still create risky situations, particularly when other road users attempt to exploit temporary gaps. This concern motivates a follow-up question that focuses explicitly on risk: *"Is the bus blocking the intersection being exploited by any pedestrian to act imprudently?"* (Fig. 6B.1). Since the query builds on the previous clip, the system maintains context while refocusing the narrative toward safety implications. The new answer (Fig. 6B.2) reveals an emergent pattern: *"Yes, two pedestrians on bicycles exploit the stationary state of the motorized vehicles to perform an illegal crossing. They accept the gap created by the stopped buses and proceed to cross the intersection via the crosswalk despite the traffic light being red for them."* Entity extraction and the knowledge-graph tooltips (Fig. 6B.3–B.4) make this structure explicit by connecting **two pedestrians on bicycles** to the action **perform an illegal crossing** and linking **stopped buses** to the relation **causes: accept the gap created by the stopped buses**. In other words, the buses form a visual shield that encourages risky behavior at the crosswalk.

Importantly, the bus implicated in this second narrative is not the original coach bus from the initial query (Fig. 6A.1), but a different bus that arrives immediately afterward (Fig. 6B.4). This follow-up reframes the scene: the initial query shows compliant bus behavior, but the refined question reveals how a subsequent bus creates conditions for an illegal red-light crossing.

Take-away. This case study illustrates how URBANCLIPATLAS supports semantic retrieval and iterative, risk-focused exploration of long-duration urban videos. By allowing follow-up questions anchored to specific clips, the system helps analysts move from verifying compliant vehicle behavior to uncovering subtle safety risks. Here, an illegal crossing emerges due to bus occlusion, and analysts can identify it without manually re-scrubbing the footage.

8. Discussion and Limitations

Role of the taxonomy and knowledge graph. A curated taxonomy of intersection concepts and its associated knowledge graph are central to URBANCLIPATLAS. By constraining entity classes and relations, the taxonomy focuses the LLM on traffic-relevant semantics (*e.g.*, *yielding*, *illegal crossing*, and *near miss*) and provides a stable scaffold for visualization. Deduplication and canonicalization further reduce noise by merging repeated mentions of the same actor or event. Together, these mechanisms support reasoning that is closer to how traffic engineers and planners interpret interactions than approaches based solely on raw detections.

Integration challenges and semantic drift. Relying on multiple heterogeneous—and largely black-box—foundation models introduces cascading sources of error. Miscaptioned clips may degrade retrieval quality, noisy extractions can propagate into the knowledge graph, and imperfect detections can weaken links between textual entities and visual evidence. We therefore treat VLM-generated descriptions as hypotheses rather than authoritative interpretations. Analysts validate them through visual grounding by comparing textual claims with detected objects and trajectories and by synchronizing video playback. However, entity extraction remains vulnerable: if the LLM diverges from the taxonomy or coreference resolution fails, the graph may drift from the scene (*e.g.*, by

splitting a single vehicle into multiple nodes or conflating distinct actors). The interface mitigates these effects by co-presenting text, structure, and video, allowing analysts to interrogate inconsistencies and verify outputs against the visual evidence.

Bridging narratives and grounded evidence. Although VLMs and LLMs can generate fine-grained descriptions (*e.g.*, *"a pedestrian pushing a stroller"*), open-vocabulary detectors such as YOLO-World typically operate at broader categorical levels and may fail to reliably ground these attributes. Across our experiments, we observed three common failure modes: (1) *grounding misses*, where described entities are not localized or tracked; (2) *entity misalignment*, where behaviors are attributed to incorrect visual instances; and (3) *semantic over-interpretation*, where the LLM labels ambiguous interactions as conflicts or safety threats. Examples are provided in Appx. B. Currently, entities that cannot be localized remain as ungrounded nodes in the knowledge graph, which may bias interpretation if they are not clearly surfaced. Exposing such uncertainties and low-confidence entities is therefore important for maintaining transparency and supporting analyst validation.

Expert feedback. As part of our evaluation, we conducted semi-structured interviews with two domain experts (an architect and an urban-design scholar), each with more than 10 years of experience and active use of tools for urban data exploration. The interviews focused on (i) perceived usefulness relative to current practice and (ii) opportunities for community engagement toward a larger-scale user study. Both experts highlighted the flexibility of chat-based retrieval as a key advantage over interfaces constrained to fixed UI filters (*e.g.*, *"now you can exactly look for something and type it, instead of being limited by the filters that you have."*). They also valued grounding narratives in video to focus attention, while noting that recognizing actions in crowded scenes remains challenging and could become a practical bottleneck. Suggested extensions included an anomaly gallery and curated query starters to guide exploration. Finally, both identified transportation engineers and practitioners in safety and microbehavioral studies as primary beneficiaries, and pointed to partners in academia and public agencies who could support broader adoption and evaluation.

Generality, scalability, and evaluation scope. The current prototype is tuned to the StreetAware dataset and to intersection-focused analysis. Deploying URBANCLIPATLAS in other cities or camera setups would require adapting the taxonomy, revalidating detector performance, and recalibrating prompts. While our preprocessing pipeline handles several hours of video, scaling to city-wide deployments with hundreds of cameras would require more efficient indexing and incremental processing. Our evaluation relies on in-depth case studies rather than a controlled user study, illustrating how experts might use the system but not yet quantifying gains in task performance, trust, or error detection.

Ethical considerations. Acknowledging the fine line between surveillance-driven technology and systems designed to promote public safety and resilient urban infrastructure, we incorporate safeguards to reduce discrimination and misuse. At the data level, the StreetAware videos are anonymized, with faces and license plates blurred and audio excluded. At the system level, the architecture supports query safeguards that reject profiling requests or other misuse (*e.g.*, gender- or race-related queries), and the taxonomy is intentionally restricted to traffic-safety concepts, excluding personal attributes.

Limitations. A key limitation of URBANCLIPATLAS is its reliance on proprietary, cloud-hosted LLMs and VLMs, whose internal reasoning is not fully interpretable. Our design therefore links each generated claim to its originating clips, detected objects, and taxonomy-constrained entities, enabling analysts to audit outputs within a human-in-the-loop workflow. This reliance limits reproducibility, raises costs and privacy concerns, and ties latency to network and API conditions. On the visual analytics side, we currently treat model outputs as point estimates, which can obscure uncertainty in captioning, retrieval, and grounding.

9. Conclusions and Future Work

URBANCLIPATLAS introduces a visual analytics workflow that combines semantic retrieval, entity-centric reasoning, and grounded visual exploration for long-duration urban videos. By integrating lightweight video RAG with taxonomy-aware augmentation and synchronized visual overlays, the system supports traffic analysis tasks in which analysts must locate, explain, and validate complex intersection events. Our case studies show that this integrated approach surfaces relevant interactions more efficiently and strengthens the connection between narrative explanations and observable evidence.

Building on these results, we plan to extend URBANCLIPATLAS in three directions: **Grounding and uncertainty** by propagating uncertainty from captioning, retrieval, and detection into the interface and flagging ungrounded or low-confidence entities; **Scalability and model independence** by scaling indexing and preprocessing to larger camera networks while reducing dependence on proprietary cloud models through open or locally deployable alternatives; and **Interaction and multimodality** by incorporating modalities such as audio or sensor signals and enabling analysts to correct entity labels, merge or split nodes, and flag false positives to improve robustness over time. We also plan to strengthen privacy protection and misuse prevention through more robust query filtering, anonymization pipelines, and audit mechanisms.

Acknowledgments

The authors acknowledge the use of LLM-based tools (ChatGPT and Grammarly) as writing assistants for text polishing. This work was supported by the National Council for Scientific and Technological Development (CNPq, grants #311144/2022-5, #132348/2025-0, and #132349/2025-6), the Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro (FAPERJ, grant #E-26/210.585/2025), the São Paulo Research Foundation (FAPESP, grants #2021/07012-0 and #2023/04868-7), and the School of Applied Mathematics at Fundação Getulio Vargas. This research was also partially supported by the National Science Foundation (NSF, Award #OAC-2411221) and by Connected Cities for Smart Mobility toward Accessible and Resilient Transportation (C2SMART), a Tier 1 University Transportation Center funded by the U.S. Department of Transportation (USDOT, contract #69A3551747124).

References

[ADC24] Arefeen M. A., Debnath B., Chakradhar S.: TrafficLens: Multi-camera traffic video analysis using LLMs. In *IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)* (Sept. 2024), pp. 3974–3981. doi:10.1109/ITSC58415.2024.10920144. 3

[ADSUC24] Arefeen M. A., Debnath B., Sarwar Uddin M. Y., Chakradhar S.: ViTA: An efficient video-to-text algorithm using VLM for RAG-based video analysis system. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (June 2024), pp. 2266–2274. doi:10.1109/CVPRW63382.2024.00232. 2

[ADUC24] Arefeen M. A., Debnath B., Uddin M. Y. S., Chakradhar S.: iRAG: Advancing RAG for videos with an incremental approach. 4341–4348. arXiv:2404.12309. doi:10.1145/3627673.3680088. 2

[AE24] Adewopo V. A., Elsayed N.: Smart city transportation: Deep learning ensemble approach for traffic accident detection. *IEEE Access 12* (2024), 59134–59147. doi:10.1109/ACCESS.2024.3387972. 2

[AH20] Alkaissi Z. A., Hussain R. Y.: Delay time analysis and modelling of signalised intersections using global positioning system (GPS) receivers. *IOP Conference Series: Materials Science and Engineering 671*, 1 (Jan. 2020), 012110. doi:10.1088/1757-899X/671/1/012110. 2

[ASM*21] Aboah A., Shoman M., Mandal V., Davami S., Adu-Gyamfi Y., Sharma A.: A vision-based system for traffic anomaly detection using deep learning and decision trees. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Los Alamitos, CA, USA, June 2021), IEEE Computer Society, pp. 4202–4207. doi:10.1109/CVPRW53098.2021.00475. 2

[BHM*25] Bahmanyar R., Hellekes J., Mühlhaus M., Gstaiger V., Kurz F.: Traffic pattern analysis at urban intersections through vehicle detection in aerial imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences X-G-2025* (2025), 151–158. doi:10.5194/isprs-annals-X-G-2025-151-2025. 2

[BMZ*15] Blincoe L. J., Miller T. R., Zaloshnja E., Lawrence B., et al.: *The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised)*. Tech. rep., U.S. Department of Transportation, National Highway Traffic Safety Administration, 2015. URL: https://rosap.ntl.bts.gov/view/dot/78697. 2

[CBH*20] Chen K., Banerjee T., Huang X., Rangarajan A., Ranka S.: A visual analytics system for processed videos from traffic intersections. In *Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2020)* (2020), Berns K., Helfert M., Gusikhin O., (Eds.), SCITEPRESS, pp. 68–77. doi:10.5220/0009422300680077. 3

[CE25] Coscia A., Endert A.: VisPile: A visual analytics system for analyzing multiple text documents with large language models and knowledge graphs, 2025. URL: https://arxiv.org/abs/2510.09605, arXiv:2510.09605. 7

[CMS*22] Cheng B., Misra I., Schwing A. G., Kirillov A., Girdhar R.: Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 1280–1289. doi:10.1109/CVPR52688.2022.00135. 5

[CNC*20] Chan G. Y.-Y., Nonato L. G., Chu A., Raghavan P., Aluru V., Silva C. T.: Motion browser: Visualizing and understanding complex upper limb movement under obstetrical brachial plexus injuries. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (Jan. 2020), 981–990. doi:10.1109/TVCG.2019.2934280. 3

[CPY*25] Cai W., Ponomarenko I., Yuan J., Li X., Yang W., Dong H., Zhao B.: SpatialBot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)* (2025), pp. 9490–9498. doi:10.1109/ICRA55743.2025.11128671. 3

[CSG*24] Cheng T., Song L., Ge Y., Liu W., Wang X., Shan Y.: YOLO-World: Real-time open-vocabulary object detection, Feb. 2024. arXiv:2401.17270. doi:10.48550/arXiv.2401.17270. 3, 5

[CZYW17] Chen P., Zeng W., Yu G., Wang Y.: Surrogate safety analysis of pedestrian-vehicle conflict at intersections using unmanned aerial vehicle videos. *Journal of Advanced Transportation 2017*, 1 (2017), 5202150. doi:10.1155/2017/5202150. 2

[DHZ*18] Deng M., Huang J., Zhang Y., Liu H., Tang L., Tang J., Yang X.: Generating urban road intersection models from low-frequency gps trajectory data. *International Journal of Geographical Information Science 32*, 12 (2018), 2337–2361. doi:10.1080/13658816.2018.1510124. 2

[DSP*25] Diaz M., Sante L., Perca J., da Silva J. V., Ferreira N., Poco J.: STRive: An association rule-based system for the exploration of spatiotemporal categorical data. *Computers & Graphics 132* (2025), 104410. doi:10.1016/j.cag.2025.104410. 3

[Fed24] Federal Highway Administration: About intersection safety. https://highways.dot.gov/safety/intersection-safety/about, 2024. Last updated July 26, 2024; accessed November 28, 2025. 2

[GXY*25] Guo Z., Xia L., Yu Y., Ao T., Huang C.: LightRAG: Simple and fast retrieval-augmented generation, Apr. 2025. arXiv:2410.05779. doi:10.48550/arXiv.2410.05779. 2

[HERMCP25] Heredia J., Estrada-Rayme L., Matos-Cangalaya J., Poco J.: Interactive exploration and explanation of spatio-temporal anomalies with graph-llm integration. In *2025 38th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (2025), pp. 1–6. doi:10.1109/SIBGRAPI67909.2025.11223398. 2

[JCL*24] Jiang W., Cheng Y., Liu L., Fang Y., Peng Y., Liu Y.: Comprehensive visual grounding for video description. *Proceedings of the AAAI Conference on Artificial Intelligence 38*, 3 (Mar. 2024), 2552–2560. doi:10.1609/aaai.v38i3.28032. 3

[JKBH25] Jeong S., Kim K., Baek J., Hwang S. J.: VideoRAG: Retrieval-augmented generation over video corpus, May 2025. arXiv:2501.05874. doi:10.48550/arXiv.2501.05874. 2, 3

[LZY*24] Luo Y., Zheng X., Yang X., Li G., Lin H., Huang J., Ji J., Chao F., Luo J., Ji R.: Video-RAG: Visually-aligned retrieval-augmented long video comprehension, Dec. 2024. arXiv:2411.13093. doi:10.48550/arXiv.2411.13093. 3

[MPCK*25] Mao M., Perez-Cabarcas M. M., Kallakuri U., Waytowich N. R., Lin X., Mohsenin T.: Multi-RAG: A multimodal retrieval-augmented generation system for adaptive video understanding, June 2025. arXiv:2505.23990. doi:10.48550/arXiv.2505.23990. 3

[NDP25] Nunes A. L., Diaz M., Poco J.: MineTracker: Visual analytics for spatiotemporal analysis of mining areas in the brazilian amazon. In *2025 38th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (2025), pp. 1–6. doi:10.1109/SIBGRAPI67909.2025.11223361. 3

[PRB*23] Piadyk Y., Rulff J., Brewer E., Hosseini M., Ozbay K., Sankaradas M., Chakradhar S., Silva C.: StreetAware: A high-resolution synchronized multimodal urban scene dataset. *Sensors 23*, 7 (Jan. 2023), 3710. doi:10.3390/s23073710. 2, 4, 8

[RPH*24] Rulff J., Pereira G., Hosseini M., Lage M., Silva C.: Towards data-informed interventions: Opportunities and challenges of street-level multimodal sensing, 2024. URL: https://arxiv.org/abs/2410.22092, arXiv:2410.22092. 2

[RXX*25] Ren X., Xu L., Xia L., Wang S., Yin D., Huang C.: VideoRAG: Retrieval-augmented generation with extreme long-context videos, 2025. URL: https://arxiv.org/abs/2502.01549, arXiv:2502.01549. 2, 3

[SH19] SCHÖNING J., HEIDEMANN G.: Visual video analytics for interactive video content analysis. In *Advances in Information and Communication Networks* (Cham, 2019), Arai K., Kapoor S., Bhatia R., (Eds.), Springer International Publishing, pp. 346–360. doi:10.1007/978-3-030-03402-3_23. 3

[SM17] SHIRAZI M. S., MORRIS B. T.: Looking at intersections: A survey of intersection monitoring, behavior and safety analysis of recent studies. *IEEE Transactions on Intelligent Transportation Systems 18*, 1 (Jan. 2017), 4–24. doi:10.1109/TITS.2016.2568920. 6, 8

[SMMSL20] STIPANCIC J., MIRANDA-MORENO L., STRAUSS J., LABBE A.: Pedestrian safety at signalized intersections: Modelling spatial effects of exposure, geometry and signalization on a large urban network. *Accident Analysis & Prevention 134* (2020), 105265. doi:10.1016/j.aap.2019.105265. 2

[SRC25] SANKARADAS M., RAJENDRAN R. K., CHAKRADHAR S. T.: StreamingRAG: Real-time contextual retrieval and generation framework, Jan. 2025. arXiv:2501.14101. doi:10.48550/arXiv.2501.14101. 2

[TYL*25] TAN X., YE Y., LUO Y., WAN Q., LIU F., CAI Z.: RAG-Adapter: A plug-and-play RAG-enhanced framework for long video understanding, Mar. 2025. arXiv:2503.08576. doi:10.48550/arXiv.2503.08576. 3

[WGQ*24] WU T., GE S., QIN J., WU G., WANG L.: Open-vocabulary spatio-temporal action detection, 2024. URL: https://arxiv.org/abs/2405.10832, arXiv:2405.10832. 3

[WQ20] WU A., QU H.: Multimodal analysis of video collections: Visual exploration of presentation techniques in TED talks. *IEEE Transactions on Visualization and Computer Graphics 26*, 7 (2020), 2429–2442. doi:10.1109/TVCG.2018.2889081. 3