ORION: A FULLY DETERMINISTIC AND INTERPRETABLE PIPELINE FOR VIDEO SCENE GRAPH GENERATION WITH EXPLICIT CAUSAL INFLUENCE SCORING

Riddhiman Rana[1]* **Aryav Semwal**[1] **Yogesh Atluru**[1] **Kevin Zhu**[1] **Cristian Meo**[1] **Shivank Garg**[1]†
[1]Algoverse AI Research
riddhiman.rana@gmail.com, aryavsemwal17@gmail.com, yatluru@gmail.com
kevin@algoverseacademy.com, cristian@algoverseairesearch.org
shivank@algoverseairesearch.org

ABSTRACT

Understanding interaction dynamics in egocentric video remains a fundamental challenge for grounded vision language systems. Existing approaches detect objects and actions but often fail to synthesize them into temporally consistent, interpretable relational representations. We introduce Orion, a fully deterministic and interpretable pipeline that transforms raw perceptual streams into symbolic, queryable knowledge graphs through a process termed *semantic uplift*. Semantic uplift converts low-level detections, embeddings, and tracks into structured entities, relations, and influence-aware events. Orion integrates modular perception components, using DINO-based backbones for object proposals, V-JEPA2 for appearance representation and re-identification, and a lightweight FastVLM-style backend for natural language entity descriptions. Entities and relations are assembled into temporally aligned scene graphs. An explicit Causal Influence Score (CIS) deterministically aggregates temporal, spatial, motion, and semantic evidence to estimate directed influence between entity pairs, enabling transparent and auditable reasoning about interaction patterns. Orion positions semantic uplift as a bridge between low-level vision outputs and high-level relational representations while remaining fully interpretable. Code is available at: https://github.com/riddhimanrana/orion-research/.

1 INTRODUCTION

Understanding causal dynamics in video is essential for building grounded vision-language systems capable of reasoning about complex real-world interactions. Applications such as robotics, augmented reality, and human behavior modeling require more than detecting objects and actions in isolation; they demand representations that integrate these elements into temporally consistent and causally meaningful structures that can support inspection, explanation, and downstream decision-making (Ji et al., 2020).

Video Scene Graph Generation (VidSGG) extends image-based scene graph generation into the temporal domain by modeling object identities and relations across time. Early approaches relied on spatio-temporal transformers and recurrent architectures to aggregate interactions (Cong et al., 2021), while later work incorporated stronger temporal reasoning and long-range dependencies (Feng et al., 2023). Recent benchmarks such as Panoptic Video Scene Graph Generation (PVSG) (Yang et al., 2023) and Action Genome (Ji et al., 2020) significantly increased task difficulty by requiring long-term temporal consistency, compositional reasoning, and robust handling of complex interaction dynamics.

Despite these advances, most VidSGG systems rely on end-to-end learned models that introduce stochasticity and opaque decision-making (Nag et al., 2023; Nguyen et al., 2025). As a result, reasoning about interaction dynamics is often implicit and difficult to inspect, debug, or trust in explainability-critical downstream systems. While such models can capture statistical regularities, they rarely expose the intermediate evidence supporting predicted relations. Scene graph anticipation methods further compound this issue by relying on learned temporal priors that obscure why specific future relations are predicted (Ji et al., 2020; Feng et al., 2023).

[*] Lead Author
[†] Senior Author

We address these limitations by introducing *Orion*, a fully deterministic and modular pipeline for video scene graph generation with explicit causal influence scoring. Orion performs *semantic uplift* by composing pretrained perception components into a transparent pipeline that converts raw video into symbolic, queryable scene graphs. Unlike prior approaches that encode interaction patterns implicitly within neural representations (Cong et al., 2021; Nguyen et al., 2025), Orion introduces an explicit, parameter-free Causal Influence Score (CIS) that aggregates interpretable temporal, spatial, motion, and semantic evidence into directed influence estimates. Rather than performing formal causal inference, CIS provides transparent and auditable proxies for potential interaction pathways.

Our contributions are:

- We introduce Orion, a fully deterministic and modular pipeline for semantic uplift from video to symbolic scene graphs.
- We introduce an explicit Causal Influence Scoring (CIS) formulation for transparent, evidence-based estimation of directed interaction influences.

2 Related Work

Video scene graph generation (VidSGG) seeks to parse dynamic video content into temporally consistent graphs of objects and their evolving relationships, enabling applications ranging from action understanding to future prediction. Foundational efforts extended static scene graph generation to video by introducing architectures that capture spatio-temporal interactions. The Spatial-Temporal Transformer (STTran) (Cong et al., 2021) and its variants employ transformer layers to aggregate pairwise dependencies across frames. DETR-based approaches, including DSG-DETR (Feng et al., 2023) and DSG-DETR++(Feng et al., 2023), further model long-range dependencies through query propagation and iterative encoder-decoder refinement. Target-adaptive context aggregation methods, such as TRACE (Teng et al., 2021), introduce hierarchical relation trees to efficiently capture spatio-temporal context, improving efficiency and recognition accuracy at the frame level.

Beyond standard scene graph generation (SGG), scene graph anticipation (SGA) aims to predict future object interactions and relations from partial video observations. This setting requires modeling long-term temporal dependencies and reasoning under uncertainty. SceneSayer (Peddi et al., 2024) adopts a continuous-time formulation for SGA, evolving object-centric representations using Neural ODE and SDE dynamics to achieve strong performance on Action Genome. Complementary structural approaches include HIG (Nguyen et al., 2024b), which organizes interactions through multi-level interlacement graphs, and CYCLO (Nguyen et al., 2024a), which employs cyclic graph transformers to capture periodic and overlapping relations. For long-form egocentric video understanding, Egocentric Action Scene Graphs (EASGs) (Rodin et al., 2024) extend scene graph representations to temporally evolving graphs of actions, objects, and relations.

Emerging zero-shot and modality-agnostic approaches seek to mitigate the data scarcity and annotation cost inherent to large-scale video scene graph datasets. SAMJAM (Li et al., 2025) enables zero-shot VidSGG in egocentric kitchen environments by combining Segment Anything Model tracking with multimodal prompting for semantic grounding. Universal Scene Graph (USG) (Wu et al., 2025) proposes a unified representation spanning text, images, video, and 3D, facilitating cross-modal transfer without extensive task-specific supervision.

Standard benchmarks closely mirror the objectives of SGG and SGA. Action Genome (Ji et al., 2020) decomposes Charades actions into spatio-temporal scene graphs under constrained observation, enabling evaluation of future relation prediction. Panoptic Video Scene Graph Generation (PVSG) (Yang et al., 2023) provides panoptic masks and fine-grained temporal annotations across third-person and egocentric videos, supporting holistic and temporally consistent scene graph generation.

While these methods have substantially advanced performance, most rely on supervised end-to-end training or dataset-specific adaptation, entangling relational reasoning within opaque learned parameters. Explicit causal attribution, auditability, and full determinism remain largely unaddressed. Orion departs from this paradigm by assembling off-the-shelf pretrained modules into a fully zero-shot, configuration-driven pipeline. It augments scene graphs with a parameter-free Causal Influence Score (CIS) that provides transparent, multi-modal estimates of directed interaction influence, bridging perceptual grounding and interpretable relational reasoning while achieving competitive accuracy on established benchmarks.

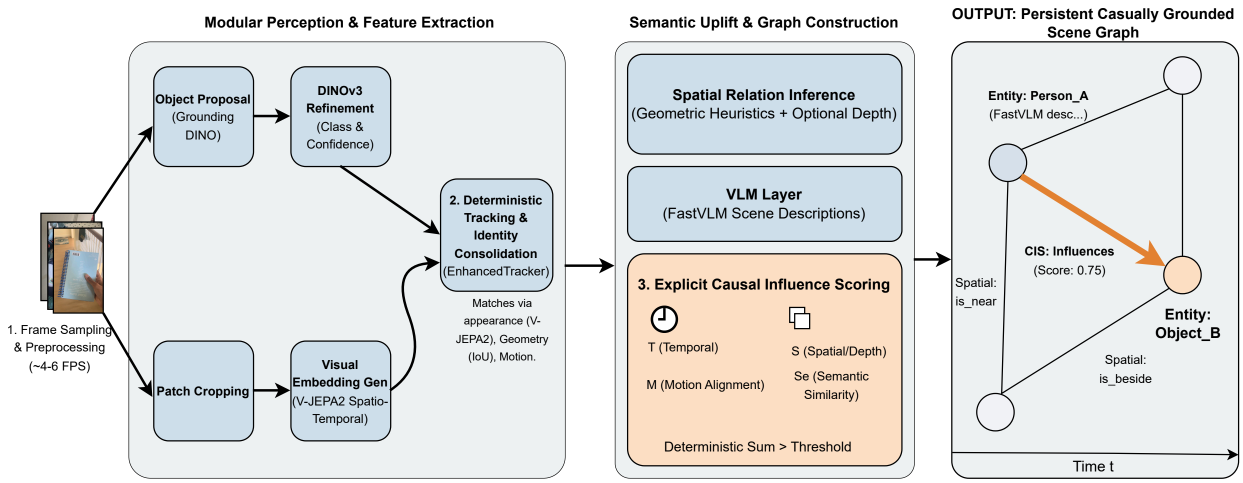

Modular Perception & Feature Extraction
1. Frame Sampling & Preprocessing (~4-6 FPS)
Object Proposal (Grounding DINO)
DINOv3 Refinement (Class & Confidence)
2. Deterministic Tracking & Identity Consolidation (EnhancedTracker)
Matches via appearance (V-JEPA2), Geometry (IoU), Motion.
Patch Cropping
Visual Embedding Gen (V-JEPA2 Spatio-Temporal)
Semantic Uplift & Graph Construction
Spatial Relation Inference (Geometric Heuristics + Optional Depth)
VLM Layer (FastVLM Scene Descriptions)
3. Explicit Causal Influence Scoring
T (Temporal)
S (Spatial/Depth)
M (Motion Alignment)
Se (Semantic Similarity)
Deterministic Sum > Threshold
OUTPUT: Persistent Casually Grounded Scene Graph
Entity: Person_A (FastVLM desc...)
CIS: Influences (Score: 0.75)
Spatial: is_near
Entity: Object_B
Spatial: is_beside
Time t

Figure 1: Overview of Orion: Raw egocentric video frames are processed through modular, deterministic stages to produce temporally aligned scene graphs with explicit Causal Influence Scores (CIS) quantifying directed entity influences.

3 Method

Orion processes input videos through a series of modular and fully deterministic stages: frame sampling and preprocessing, object detection and feature extraction, temporal tracking, and scene-graph construction from tracked entities. This yields temporally aligned per-frame scene graphs, optionally augmented with Causal Influence Scoring (CIS) that assigns directed influence edges (e.g., influences, grasps, moves with) using temporal co-occurrence, spatial proximity (with depth gating), motion alignment, and semantic compatibility. Figure 1 shows an overview of Orion.

3.1 Frame Sampling and Object Detection / Proposal Generation

The input video is sampled at 5 FPS. Detection uses a two-stage propose → refine pipeline: GroundingDINO (Liu et al., 2024) generates bounding-box proposals, and DINOv3 (Siméoni et al., 2025) refines labels and confidences using full-frame context (detailed refinement process in appendix).

3.2 Visual Embedding Generation

Appearance representations and re-identification use V-JEPA2 (Assran et al., 2025) to generate normalized, video-aware (spatio-temporal) embeddings per tracked detection, supporting robust tracking under occlusion and viewpoint changes.

3.3 Causal Influence Scoring (CIS)

For every ordered entity pair (a, b) whose observations fall within a temporal window of 30 frames, Orion computes a deterministic Causal Influence Score (CIS) estimating directed influence $a \to b$:

$$\text{CIS}(a \to b) = w_t T + w_s S + w_m M + w_{se} S_{se}, \tag{1}$$

where T, S, M, and S_e capture temporal proximity (exponential decay over time difference), spatial proximity (normalized centroid distance), motion alignment (velocity cosine similarity), and semantic compatibility (embedding cosine similarity), respectively; full definitions are provided in the appendix. Default weights are $w_t = 0.30$, $w_s = 0.44$, $w_m = 0.21$, $w_{se} = 0.05$; a directed link is retained when CIS ≥ 0.50.

4 Experiments

We evaluate Orion on standard benchmarks for video scene graph generation (SGG) and anticipation (SGA): PVSG for SGG and Action Genome for SGA.

Method	R@10	R@20	R@50
STTran+ (Cong et al., 2021)	14.9	22.6	42.9
DSG-DETR+ (Feng et al., 2023)	15.2	23.1	43.3
STTran++ (Cong et al., 2021)	16.6	29.1	51.5
DSG-DETR++ (Feng et al., 2023)	17.4	30.5	51.9
SceneSayer ODE (Peddi et al., 2024)	26.4	36.6	49.8
SceneSayer SDE (Peddi et al., 2024)	28.4	38.6	51.4
HyperGraph (Nguyen et al., 2025)	29.2	39.3	52.1
HyperGLM (Nguyen et al., 2025)	30.0	40.5	53.5
Orion (ours)	18.4	32.7	47.7

Table 1a: Comparison (%) on Action Genome (SGA) with $F = 0.5$. Recall@10, @20, @50 (No Constraint).

Method	R@20	R@50	R@100
Transformer (Yang et al., 2023)	4.0	4.4	4.9
HIG (Nguyen et al., 2024b)	4.6	4.9	5.4
CYCLO (Nguyen et al., 2024a)	5.8	6.1	6.7
HyperGraph (Nguyen et al., 2025)	6.5	7.0	7.5
HyperGLM (Nguyen et al., 2025)	7.5	8.1	8.5
Orion (ours)	6.4	7.1	7.6

Table 1b: Comparison (%) on PVSG (SGG) using Recall@20, @50, @100.

Table 1: **Comparative performance of Orion.** We report results on (a) Action Genome for Scene Graph Anticipation (SGA) and (b) PVSG for Video Scene Graph Generation (SGG). Despite its zero-shot nature, Orion remains competitive with supervised baselines.

4.1 Datasets

- **PVSG** (Yang et al., 2023): 400 long videos with panoptic masks and 57 temporal relation classes across third-person and egocentric sources.
- **Action Genome** (Ji et al., 2020): 10K videos with spatio-temporal object and relationship annotations for SGA.

4.2 Setup and Metrics

For PVSG (SGG), we report R@K ($K \in \{20, 50, 100\}$), measuring fraction of ground-truth triplets correctly predicted in top-K ranked by pipeline confidence (volume IoU 0.5 for temporal tube matching).

For Action Genome (SGA), we report R@K ($K \in \{10, 20, 50\}$) under $F = 0.5$ No Constraint setting, quantifying future triplet retrieval in top-K.

Orion is evaluated zero-shot: no training, fine-tuning, or adaptation. We use pretrained components (GroundingDINO (Liu et al., 2024), DINOv3 (Siméoni et al., 2025), V-JEPA2 (Assran et al., 2025), FastVLM (Vasu et al., 2025)) at 5 FPS.

4.3 Results

On Action Genome SGA ($F = 0.5$, No Constraint), Orion achieves R@10 of 18.38%, R@20 of 32.7%, R@50 of 47.73% (Table 1a). These exceed early transformer methods (STTran+, DSG-DETR+) and remain competitive with recent approaches (SceneSayer ODE/SDE, HyperGraph, HyperGLM), despite zero-shot operation versus supervised training. On PVSG SGG, Orion records R@20 of 6.4%, R@50 of 7.1%, R@100 of 7.6% (Table 1b), close to HyperGraph/HyperGLM while surpassing earlier baselines (Transformer, HIG, CYCLO) under stringent volume IoU matching. These results show a modular zero-shot pipeline with explicit CIS can deliver competitive performance while preserving determinism and full interpretability, making Orion suitable for explainability-critical applications like robotics and augmented reality.

5 Conclusion

Orion demonstrates that fully deterministic and transparent video scene graph generation is feasible by combining modern pretrained components in a modular and inspectable pipeline (GroundingDINO (Liu et al., 2024), DINOv3 (Siméoni et al., 2025), V-JEPA2 (Assran et al., 2025), and FastVLM (Vasu et al., 2025)). The explicit Causal Influence Scoring module provides interpretable, evidence-based estimates of directed interaction influences, enabling users to reason about video dynamics in a transparent and auditable manner. This design highlights a practical pathway toward explainable video understanding systems that balance performance, generality, and interpretability.

6 Limitations

Orion is designed to prioritize determinism, modularity, and interpretability, and this design entails several scoped limitations. The proposed Causal Influence Score is a heuristic aggregation of temporal, spatial, motion, and semantic cues, and should be interpreted as an interpretable proxy for interaction influence rather than a formal causal inference mechanism. In addition, the current system primarily relies on 2D bounding-box geometry with limited depth information, which may reduce robustness in highly occluded scenes or complex 3D interactions. Addressing these aspects while preserving transparency remains an important direction for future work.

References

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16372–16382, 2021.

Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5130–5139, 2023.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10236–10247, 2020.

Joshua Li, Fernando Jose Pena Cantu, Emily Yu, Alexander Wong, Yuchen Cui, and Yuhao Chen. Samjam: Zero-shot video scene graph generation for egocentric kitchen videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 467–473, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22803–22813, 2023.

Trong-Thuan Nguyen, Pha Nguyen, Xin Li, Jackson Cothren, Alper Yilmaz, and Khoa Luu. Cyclo: Cyclic graph transformer approach to multi-object relationship modeling in aerial videos. *Advances in Neural Information Processing Systems*, 37:90355–90383, 2024a.

Trong-Thuan Nguyen, Pha Nguyen, and Khoa Luu. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18384–18394, 2024b.

Trong-Thuan Nguyen, Pha Nguyen, Jackson Cothren, Alper Yilmaz, and Khoa Luu. Hyperglm: Hypergraph for video scene graph generation and anticipation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29150–29160, 2025.

Rohith Peddi, Saksham Singh, Saurabh, Parag Singla, and Vibhav Gogate. Towards scene graph anticipation. In *European Conference on Computer Vision*, pp. 159–175. Springer, 2024.

Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18622–18632, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13688–13697, 2021.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, et al. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19769–19780, 2025.

Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14158–14168, 2025.

Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, et al. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18675–18685, 2023.

Appendix

A Implementation Details

A.1 Spatial Relation Inference

Orion augments object tracks with simple, interpretable spatial predicates to form per-frame scene graphs that support downstream querying (e.g., “what is near the cup?”) and provide structured context for later temporal reasoning. For each frame, we infer near, on, and held_by edges from 2D bounding-box geometry (with optional person keypoints when available). Let $b_i = [x_1, y_1, x_2, y_2]$ be the bounding box of entity i, with centroid $c_i \in \mathbb{R}^2$, in an image of size $W_{\text{img}} \times H_{\text{img}}$ and diagonal $D = \sqrt{W_{\text{img}}^2 + H_{\text{img}}^2}$; we define normalized centroid distance $d(i,j) = \|c_i - c_j\|_2 / D$ and the usual $\text{IoU}(b_i, b_j)$. We add $(i, \texttt{near}, j)$ when $d(i,j)$ is below a configurable threshold and $\text{IoU}(b_i, b_j)$ is small to avoid trivial overlap-based matches (defaults $d(i,j) \leq 0.08$ and $\text{IoU}(b_i, b_j) < 0.10$, with a tighter distance threshold of 0.06 for small objects). We add $(i, \texttt{on}, j)$ when i is supported by j, operationalized by sufficient horizontal overlap between boxes and a small vertical gap between the bottom of i and the top of j (normalized by H_{img}), together with a centroid ordering constraint (default horizontal-overlap ≥ 0.30 and vertical-gap ≤ 0.02, with adaptive relaxation for small subjects). Finally, for a non-person entity i and a person entity p, we add $(i, \texttt{held\_by}, p)$ if $\text{IoU}(b_i, b_p) \geq 0.30$ or if the centroid of i lies inside b_p; if hand keypoints are available, we optionally use a hand-to-object proximity test, and we keep at most one held_by edge per object by selecting the best-scoring person in that frame. All predicates and thresholds are configurable; in this work we use the defaults above to obtain stable, easily-audited spatial graphs without requiring a learned relation classifier.

A.2 DINOv3 Refinement & Classification

After initial proposal generation, DINOv3 performs context-aware label refinement. While the proposal stage focuses on generating candidate bounding boxes, this stage leverages full-frame features to re-evaluate each proposal using global scene context. DINOv3 extracts full-frame representations once per frame, pools region features for each proposal, and produces refined class labels, confidence scores, and optional attributes (such as age or gender) obtained either via dedicated linear classification heads or through zero-shot heuristics. This separation enables robust classification without altering box geometry or introducing additional proposals.

A.3 Entity Tracking and Motion Summaries

Stable entity IDs are assigned using the EnhancedTracker / EntityTracker, which combines appearance similarity (computed via V-JEPA2 cosine similarity), geometry (via IoU), and motion heuristics. Each PerceptionEntity maintains a persistent ID along with per-frame observations including bounding box, timestamp, and confidence; it also tracks a prototype embedding, canonical box statistics, temporal bounds defined as $[t_{\text{first}}, t_{\text{last}}]$, and smoothed estimates of velocity and direction.

A.4 Causal Influence Scoring

The components are defined as follows. Temporal proximity is captured by

$$T = \exp\left(-\frac{\Delta t}{\tau}\right), \tag{2}$$

where $\Delta t = |t_a - t_b|$ is the absolute time difference (in seconds) between the timestamps t_a and t_b of the two observations, and $\tau > 0$ is a decay constant.

Spatial proximity is given by

$$S = 1 - \min(1, d_{\text{norm}}), \tag{3}$$

where $d_{\text{norm}} = \|c_a - c_b\|_2 / D \in [0, 1]$ is the normalized Euclidean distance between the 2D bounding-box centroids $c_a, c_b \in \mathbb{R}^2$ (in pixel coordinates), and $D = \sqrt{H_{\text{img}}^2 + W_{\text{img}}^2}$ is the image diagonal.

Motion alignment is measured via

$$M = \max(0, \cos(v_a, v_b)) \,, \tag{4}$$

where $v_a, v_b \in \mathbb{R}^2$ are the per-entity 2D velocity vectors (pixels/frame), estimated from consecutive centroid differences via the upstream tracker.

Finally, semantic compatibility is

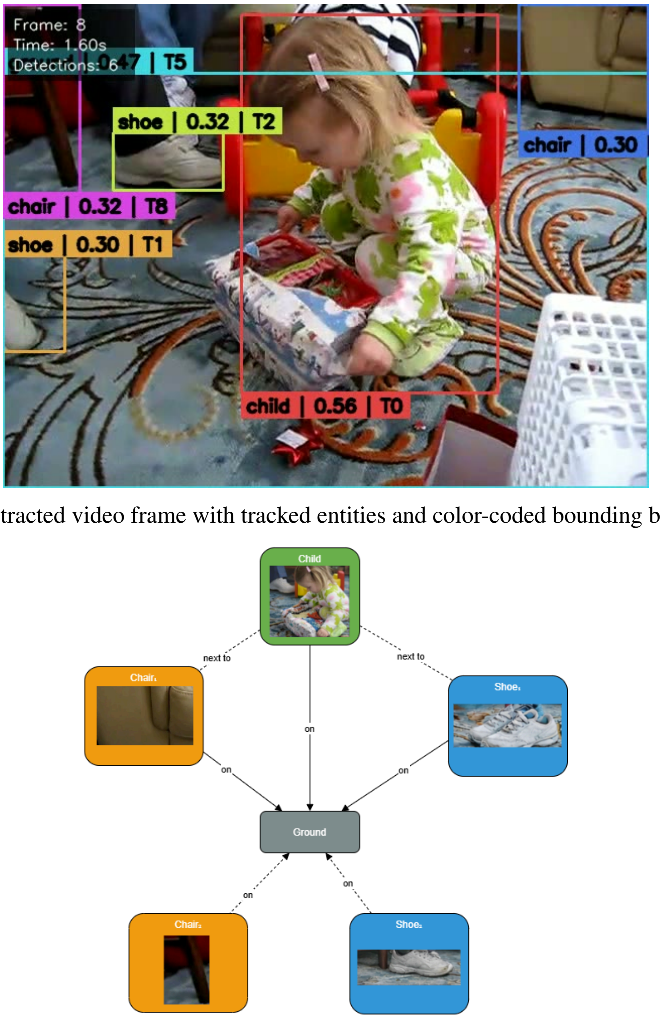
$$S_{se} = \max(0, \cos(e_a, e_b)) \,, \tag{5}$$

where $e_a, e_b \in \mathbb{R}^d$ are unit-normalized prototype embeddings associated with the entities (i.e., $\|e_a\|_2 = \|e_b\|_2 = 1$), and $\cos(\cdot, \cdot)$ denotes cosine similarity:

$$\cos(x, y) = \frac{x^\top y}{\|x\|_2 \|y\|_2}. \tag{6}$$

B Qualitative Example

Figure 2 shows a representative frame from the PVSG dataset together with Orion's predicted entities and relations, illustrating the interpretability of the generated scene graph.



(a) Extracted video frame with tracked entities and color-coded bounding boxes.

(b) Predicted scene graph. Directed edges represent causal influence (CIS > 0.5), illustrating the deterministic semantic uplift from detections to relational interaction.

Figure 2: **Qualitative example of Orion's interpretability.** (top) A frame from PVSG dataset and (bottom) the resulting scene graph. Note how the directed edges and CIS scoring allow for auditable reasoning of the interaction between the subject and the object.