ASP-Driven Visual Commonsense: A General Framework for Reasoning about Embodied Interaction in the Wild

Jakob Suchan[1,3] , **Mehul Bhatt**[2,3] , **Julius Monsen**[2,3]
[1]Constructor University, Germany
[2]Örebro University, Sweden
[3]CoDesign Lab EU
info@codesign-lab.org

Abstract

We present a general framework for declaratively grounded visual commonsense (reasoning) about embodied interaction in naturalistic, in-the-wild settings relevant to a range of AI application domains. The core computational capabilities of the framework pertaining visual commonsense are driven by a robust neurosymbolic architecture primarily consisting of: **(1)** answer set programming based modelling of foundational aspects pertaining spatio-temporal dynamics, encompassing space, time, events, action, motion; **(2)** modularly integrated visual computing techniques constituting the neural substrate linking quantitative perceptual features serving as low-level counterparts to high-level semantic characterisations of (inter)active visual commonsense.

Practically, we also present a first open-release of the developed framework with the aim to promote independent extensions and real-world applied KRR. The release comprises: **(a)** demonstrated case-studies in domains such as autonomous driving, psychology and media studies; **(b)** systematic evaluation mechanisms for community benchmarking; and **(c)** supporting material such as tutorials and datasets.

1 Motivation

We present a novel framework for visual commonsense in autonomous systems concerned with (inter)active sensemaking in diverse embodied "in-the-wild" situations of everyday life and work. By interactive sensemaking, interchangeably active sensemaking, we allude to:

- interleaved multimodal perception, interpretation and decision-making requiring coordination of attention, integration of sensory inputs, and dynamic exploration of possible worlds and outcomes, possibly under tight temporal constraints.

As a basic example of active sensemaking, consider the illustration of a sample human activity —"*making a cup of tea*"— in Fig. 1. Here, our notion of embodied active vision is inherent and plays a crucial role, as well as offers a compelling challenge for real-world applications of AI/KR and ML/Vision. The sample of Fig. 1 presents data captured from an egocentric viewpoint with a head-mounted RGB-D capture device. From a commonsense viewpoint, this episode may be represented as a sequence of dynamic

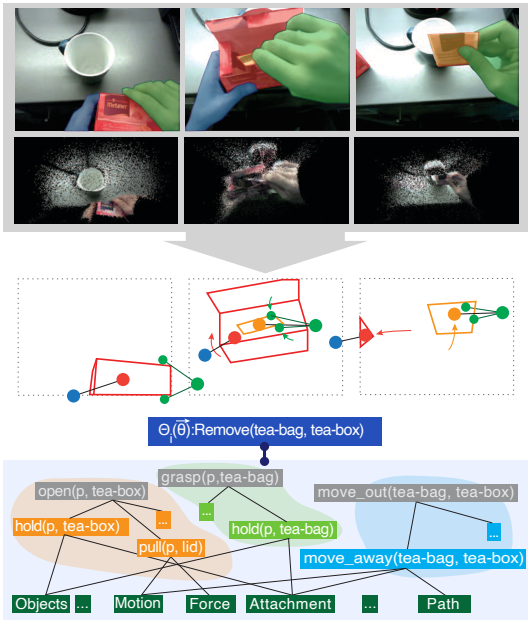


Figure 1: **Relational Grounding of Everyday Interaction:** "Making a cup of tea" (egocentric view from a head-mounted RGB-D capture device)

visuospatial interactions, such as the following:

- opening the tea-box, removing a tea-bag from the box and putting the tea-bag inside a tea-cup filled with water while holding the tea-cup.

Corresponding to such interactions are high-level spatial and temporal relationships between the agent and other involved objects, e.g., involving conceptual representations of contact and containment that hold across specific time-intervals. In this context, manipulation and control actions ($\Theta_1(\vec{\theta})$, ...$\Theta_n(\vec{\theta})$) cause state transitions in the world, which are modelled in a domain-independent manner as changes in the spatio-temporal relations amongst involved domain entities. Our proposed framework enables the seamless maintenance of such dynamic interactive situations in a robust, domain-independent, and elaboration tolerant manner.

Explainable Visual Commonsense. In order to realise a computational model of *active, explainable, visual com-*

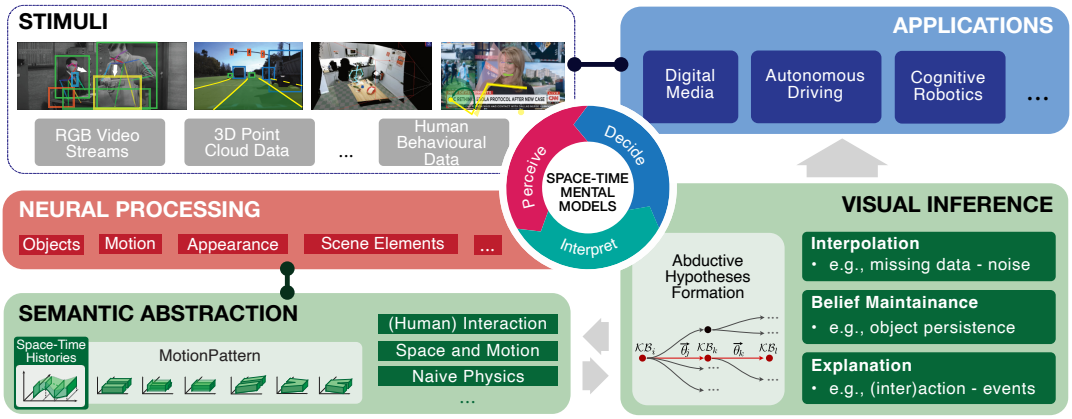


Figure 2: **Neurosymbolic Visual Commonsense:** Integrated Vision & AI foundations for making-sense of embodied multimodal interaction.

monsense reasoning with dynamic visuospatial imagery, we present a systematic, modular, and extensible integration of methods from knowledge representation and reasoning, and computer vision. In central focus of this paper are *answer set programming* (Brewka, Eiter, and Truszczyński 2011; Gebser et al. 2014) based declarative modelling of action/event induced spatio-temporal dynamics and deep learning driven computer vision techniques for the extraction of low-level perceptual features from quantitative stimuli such as video and eye-tracking data (Suchan, Bhatt, and Varadarajan 2019). Our key emphasis here is on developing robust mechanisms for generalised (declarative) neurosymbolic visual commonsense (reasoning) about space, time, events, actions, motion, and spatio-temporal dynamics as relevant to embodied multimodal interaction under *ecologically valid naturalistic settings* of everyday life (Fig. 3). From a practical viewpoint, we showcase the manner in which the developed framework robustly supports general-purpose computational visual commonsense reasoning primarily (but not exclusively) rooted to (neurosymbolic) non-monotonic visual abduction. The presented framework is motivated by and demonstrated in the applied backdrop of areas as diverse as autonomous driving, design of digital visuo-auditory media, and behavioural visual perception research in cognitive psychology.

Emerging AI Regulation. We address the foundations of next-generation computational cognitive vision systems with an emphasis on achieving explainability and human-centred design. We interpret human-centred criteria within the emerging regulatory framework in the European Union (AI HLEG 2019; EU Commission 2019; EU Commission 2021); our particular focus is on the recommendations concerning 'technical robustness', 'explicability', and 'transparency' of future AI systems. In this context, we particularly argue for the need and relevance of methods in knowledge representation and reasoning within next-generation AI systems, and demonstrate this with a systematically implemented novel exemplar of "generalised visual commonsense framework" bridging Vision and AI/KRR techniques and making them available for extensions and applications.

2 Space-Time Mental Models in (Inter)Active Vision

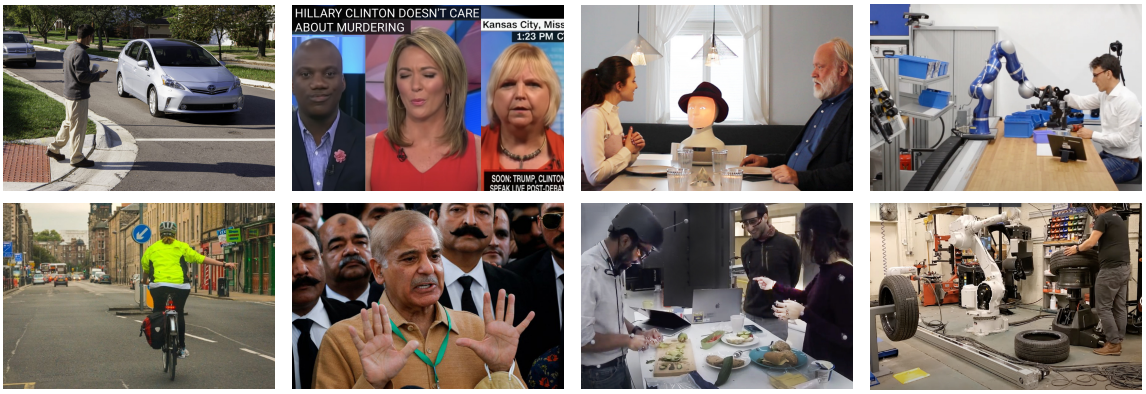
We present a conceptual summary (Fig. 2) of the developed framework for visual commonsense, motivating its overall design and implementation whilst highlighting key functional capabilities vis-a-vis application needs, robust declarative foundations, domain-independent (re)usability, and modular extensibility.

2.1 What's in a Space-Time Mental Model?

Consider a situation where a human is driving a car, navigating complex urban traffic together with other vehicles and vulnerable road users (e.g., pedestrians, cyclists). The human driver's ability to maintain space-time situational awareness of the situation, for instance in an approximately 4-6 sec interval comprising of the immediate past and future constitutes the human driver's space-time mental model. Conceptually akin to this characterisation, we denote the computation of a *space-time mental model* (Bhatt and Suchan 2021) as:

- the ability to semantically interpret and explain the space-time sensory perception of the environment, conceptually making sense of the configuration and dynamics of mutually interacting people, objects, artefacts, and events and actions in space and time both in an active realtime as well as offline setting.

This characterisation of a space-time mental model may be interpreted in a multitude of ways. In the specific context of computational modelling, factors determining our design and implementation choices are influenced by the following aspects: (**1**) Level of ontological expressivity required vis-a-vis domain-independent commonsense characterisations of aspects related to space, time, motion, granularity of human interaction representation etc; (**2**) Scalable algorithmic implementation needs and overall tolerable complexity (e.g., realtime vs. offline processing needs); (**3**) Nature of quantitative input data sources (e.g., video, sound, LIDAR, eye-tracking and other physiological markers); and (**4**) Applica-



(a) (b) (c) (d)

Figure 3: **Embodied multimodal interaction in diverse contexts**: **(a)** A pedestrian establishes joint attention with a driver, and a cyclist's gesture indicates intentions to turn following traffic rules; **(b)** Facial expressions accompany speech during news media discussions or public talks; **(c)** Eye contact and deictic gestures promoting joint attention under social (robotic) collaborative tasks; **(d)** Industrial collaborative tasks with a robotic arm.

tion needs primarily pertaining to types of high-level commonsense reasoning capabilities required.

A detailed characterisation of all design decisions is not essential to the purpose of this paper. Here, we instead focus on a conceptual summary (Fig. 2) of the developed framework vis-a-vis key representational and computational requirements that have been formalised, implemented, and evaluated from the viewpoint of open-source dissemination.

2.2 Ontological and Computational Setup

We highlight two key aspects: (**I**) Ontological modelling of interaction, and (**II**) Key constituents of a foundational formal and computational characterisation of visual commonsense as relevant to a range of visual stimuli / application domains involving diverse forms of multimodal interaction:

I. Modelling Interaction: Embodied multimodal interactions, relevant to diverse applications (e.g., Fig. 3), are characterised based on the relational spatio-temporal structure underlying the respective interaction and the effects on the beliefs about the world (Table 1). Practical actions (e.g. involving pushing/pulling an object, (re)direction of a path) describe the interactions between a person and the environment during an everyday task. Communicative interactions are classified based on the mode of deliverance of the message, as explicit or implicit interactions. Explicit interactions involve a range of modalities such as facial expressions or gestures, e.g. a cyclist's extension of one hand on the side, is a gesture that conveys his intention to turn in the upcoming intersection. Implicit interactions involve a set of modalities as communication tools that require lower effort, such as gaze, body posture or head movements. Visuospatial properties of a scene, such as the visibility of objects/agents, their locations and facing directions, describe the state of the world. A combination of facts and events observed over a longer time interval may lead to *hypotheses* (Sec 3.1) about ongoing interactions, agent's intentions, or the anticipation of near future events. Scene elements are the distinct elements of the physical world obtained from high-level sensing, e.g. a car, pedestrian. Scene elements are categorised based on their type, structure and properties, and are geometrically represented as low-level entities (e.g. bounding boxes) that are involved in spatio-temporal relationships, and that constitute the underlying quantitative features emanating of low-level signal processing. Commonsense spatio-temporal relations and patterns (e.g., *left*, *overlapping*, *during*, *approaching*) offer a human-centered and cognitively adequate formalism for semantic grounding and automated reasoning for everyday (embodied) multimodal interactions (Bhatt, Schultz, and Freksa 2013; Mani and Pustejovsky 2012; Cohn et al. 1997).

II. Domain-Independent Visual Commonsense: Reasoning about space-time mental models is most fundamentally supported by foundational or a meta-characterisation of the basic epistemological phenomena identifiable in diverse scenarios involving modelling of *spatio-temporal dynamics* (Bhatt and Loke 2008; Hazarika 2005). In the proposed framework, this corresponds to the following:

- maintaining **consistent beliefs** respecting (domain-neutral) commonsense criteria, e.g., related to compositionality & indirect effects, space-time continuity, positional changes resulting from motion.
- ability to make **default assumptions**, e.g., pertaining to persistence objects and/or visual and spatial object attributes, e.g., pertaining position, velocity, direction of movement.
- **interpolation** of missing information, e.g., what could be hypothesised about missing information (e.g., moments of visual occlusion); how can this hypothesis support planning an immediate next step?
- object **identity maintenance** at a semantic level, e.g., in the presence of occlusions, missing and noisy quantitative data, low-level errors in visual (mis)detection and tracking.

Supported by foundation or meta-level axiomatisation of the fundamental epistemological aspects, the key focus of the

Embodied Multimodal Interaction Mechanisms		
Practical Action	Object / Environment Interactions - Auditory cues - Motion Paths	enters(P,Q), crossing(P,Q), passing_behind(P,Q), hides_behind(P,Q), approaching(P,Q), opening(P,Q), removing(P,Q), holding(P,Q), touching(P,Q), ...
Explicit Interaction	Eye Contact - Facial Expressions - Gestures - Speech - Nodding	joint_attention(P,Q), monitoring_attention(P,Q), gesture(P, Gesture), hand_sign(P, Sign), auditory_cue(Source, Cue), ...
Implicit Interaction	Body Posture / Positioning - Head Movement - Gaze - Intonation - Behavioural changes	pose(P,Pose), turn_head(P, Direction), speed_up(P), maintain_steady_speed(P), slow_down(P), detect(P,Q), track(P,Q), ...
Facts / Beliefs (Fluents)		
Scene Properties	visibility: hidden(P), partially_hidden(P), occluded_by(P, Q), , ...; attention: looking_at(P, Q), attentive(P), ...; location: on(P, Q), in(P, Q), next_to(P, Q), ...	
Scene Elements		
Types	object: dynamic / static — (range of domain-independent and dependent categories and instances)	
Structure & Properties	human: body-parts (hands, face, ...), body pose, facing direction, gaze direction, ... objects: orientation, parts, ...	
Spatio-Temporal Characterisations		
Domains	Mereotopology, Incidence, Orientation, Distance, Size, Motion, ...	
Relations	topology / position: inside, outside, overlapping, connected, left, right, in front, behind, on top, touching; direction: facing towards, facing away, same direction, opposite direction; moving: towards, away, parallel; ...	
Entities	bounding boxes, polygons, line-segments, points, oriented-points, motion trajectories, time-points, time intervals, ...	

Table 1: **Ontological Structure of Modelling Embodied Multimodal Interactions**. Only a select indicative sample is included to illustrate key high-level categories.

developed framework is on a hybrid architecture for systematically computing robust visual explanation(s) encompassing hypothesis formation, belief revision, and default reasoning with visual data (for real-time as well as for offline processing). In other words, the framework supports visuospatial abduction with space, events, actions, and motion practically implemented within answer set programming. Sections 3–4 will further elaborate on the formalisation and practical (applied) relevance of formalising foundational visual commonsense as a a domain-independent theory that is usable in diverse scenarios involving embodied interaction.

3 Neurosymbolic Visual Commonsense: A General Framework

We present a formal and computational characterisation of a theory of spatio-temporal dynamics in the backdrop of our notion of space-time mental models in Sec 2. Furthermore, we also summarise the technical implementation of the framework in view of its intended applications, and further technical extensions. Our developmental goal in this paper is to integrate recent advances in knowledge-centric computational cognitive vision research aimed at developing formal models for the processing and semantic interpretation of dynamic visuo-spatial imagery with cognitively rooted structured characterisations of commonsense knowledge pertaining to space and motion (Suchan, Bhatt, and Varadarajan 2021; Suchan, Bhatt, and Varadarajan 2019).

3.1 Spatio-Temporal Dynamics

Visual commonsense reasoning about embodied interaction requires a high-level representation of objects, and their spatio-temporal dynamics, e.g., respective motions & mutual interactions in space-time amongst other aspects. For the purposes of this paper, key foundational ontological primitives in this respect are (Table 2):

- Σ_{st} corresponds to primitives for representing space, time, motion and scene-level relational spatiotemporal structure
- Σ_{dyn} corresponds to the domain-independent commonsense theory for representing and reasoning about change.

Let $\Sigma \equiv_{def} \Sigma_{st} <\mathcal{O}, \mathcal{E}, \mathcal{T}, \mathcal{MT}, \mathcal{R}> \cup \Sigma_{dyn} <\Phi, \Theta>$ denote the background theory of space, action, events, motion, and change as follows (Table 2):

• **Domain Objects** ($\mathcal{O}$). The high-level, domain-dependent visual elements in the scene, e.g., road-side stakeholders such as *people*, *cars*, *cyclists*, constitute domain objects. Domain objects are denoted by $\mathcal{O} = \{o_1, ..., o_n\}$; elements in $\mathcal{O}$ are geometrically interpreted as *spatial entities*.

• **Spatial Entities** ($\mathcal{E}$). Spatial entities correspond to abstractions of domain objects by way of *points*, *line-segments* or (axis-aligned) *rectangles* based on their spatial properties (and a particular reasoning task at hand). Spatial entities are denoted by $\mathcal{E} = \{\varepsilon_1, ..., \varepsilon_n\}$.

• **Time** ($\mathcal{T}$). The temporal dimension is represented by **time points**, denoted as $\mathcal{T} = \{t_1, ..., t_n\}$.

• **Motion Tracks** ($\mathcal{MT}$). Motion-tracks represent the spacetime motion trajectories (e.g., bottom of Fig. 2; (Hazarika 2005; Bennett et al. 2000; Schultz et al. 2018)) of abstract spatial entities ($\mathcal{E}$) corresponding to domain object ($\mathcal{O}$) of interest. $\mathcal{MT}_{o_i} = (\varepsilon_{t_s}, ..., \varepsilon_{t_e})$ represents the **motion track** of a single object o_i, where t_s and t_e denote the start and end time of the track and ε_{t_s} to ε_{t_e} denotes the spatial entity ($\mathcal{E}$) —e.g., the *axis-aligned bounding box*— corresponding to the object o_i at time points t_s to t_e.

• **Spatio-Temporal Relationships** ($\mathcal{R}$). The spatial configuration of the scene and changes thereof are characterised based on the spatio-temporal relationships ($\mathcal{R}$; Table 1) between abstract representations ($\mathcal{E}$) of the domain objects ($\mathcal{O}$). For the running and demo examples of this paper,

SPACETIME AND MOTION	REPRESENTATION	
Space-Temporal Primitives (Σ_{st})		
Domain Objects	$\mathcal{O} = \{o_1, ..., o_n\}$	e.g., *cars, people, cyclists*
Spatial Entities	$\mathcal{E} = \{\varepsilon_1, ..., \varepsilon_n\}$	*points, line-segments, rectangles*
Time	$\mathcal{T} = \{t_1, ..., t_n\}$	*time-points, time-intervals*
Motion	$\mathcal{MT}_{o_i} = (\varepsilon_{t_s}, ..., \varepsilon_{t_e})$	*motion tracks / space-time histories*
Spatio-Temporal Relationships	$\mathcal{R}$	*e.g., topology, orientation, distance*
Spatio-Temporal Dynamics (Σ_{dyn})		
Fluents	$\Phi = \{\phi_1, ..., \phi_n\}$	*e.g., visibility, hidden_by, clipped*
Events	$\Theta = \{\theta_1, ..., \theta_n\}$	*e.g., hides_behind, missing_detections*
Problem Specification		
Visual Observations	$\mathcal{VO}_t = \{obs_1, ..., obs_n\}$	*e.g., $\mathcal{E}$ corresponding to object detections*
Predictions	$\mathcal{P}_t = \{p_{trk_1}, ..., p_{trk_n}\}$	*e.g., $\mathcal{E}$ for predicted track*
Matching Likelihood	$\mathcal{ML}_t = \{ml_{trk_1,obs_1}, ..., ml_{trk_n,obs_m}\}$	*e.g., IoU between tracks and detections*
Hypothesis		
Assignments	$\mathcal{H}^{assign}$	*abduced assignments*
Events	$\mathcal{H}^{events} = \{\theta_1, ..., \theta_n\}$	*abduced event sequence*
Explanations	$\mathcal{EXP} \leftarrow < \mathcal{H}^{events}, \mathcal{MT} >$	*scene dynamics; abduced events and corresponding motion tracks*

Table 2: **Commonsense – Space – Motion:** Ontological and Representational Setup

positional relations on axis-aligned rectangles based on the Rectangle Algebra (RA) (Balbiani, Condotta, and del Cerro 1999) suffice; RA uses the relations of Interval Algebra (IA) (Allen 1983) $\mathcal{R}_{\mathsf{IA}} \equiv \{$before, after, during, contains, starts, started_by, finishes, finished_by, overlaps, overlapped_by, meets, met_by, equal$\}$ to relate two objects by the *interval relations* projected along each modelled dimension separately (e.g., horizontal and vertical dimensions).

• **Dynamics / Fluents and Events**. The set of **fluents** $\Phi = \{\phi_1, ..., \phi_n\}$ and **events** $\Theta = \{\theta_1, ..., \theta_n\}$ respectively characterise the dynamic properties of the objects in the scene and high-level abducibles (e.g., passing_behind, approaching, touching; Sec 2.2, Table 1). For reasoning about dynamics (with $<\Phi, \Theta>$), we use the epistemic generalisation of the event calculus (Kowalski and Sergot 1989) as per the formalisation in (Ma et al. 2014; Miller, Morgenstern, and Patkos 2013); in particular, for examples of this paper, the Functional Event Calculus (FEC) fragment of Ma et al. (2014) suffices. Main axioms relevant for this paper pertain to occurs-at(θ, t) denoting that an event occurred at time t and holds-at(ϕ, v, t) denoting that v holds for a fluent ϕ at time t. It it worth noting that in so far as the approach to reason about changes is concerned, our modular framework is by no means limited to the specific approach being utilised. In principle, any method capable of modelling *dynamic spatial systems* (Bhatt and Loke 2008) encompassing *space, actions, and change* (Bhatt 2012; Bhatt et al. 2011) is usable; basic considerations guiding choice of an action theory pertain to expressivity, modular elaboration tolerance, and support for basic epistemological aspects such as *frame* and *ramification* (Shanahan 1997). For instance, other epistemic settings for abductive inference with ASP too may be utilised (Eppe and Bhatt 2015a; Eppe and Bhatt 2015b).

• **Problem Specification** $< \mathcal{VO}_t, \mathcal{P}_t, \mathcal{ML}_t >$. The abduction for each time point is given by the visual observations ($\mathcal{VO}_t$) consisting of spatial entities $\mathcal{E}$, i.e., bounding boxes for the detected objects, spatial entities $\mathcal{E}$ of object detections; the predicted locations ($\mathcal{P}_t$) for each track at time point t given as spatial entities $\mathcal{E}$; and the matching likelihood ($\mathcal{ML}_t$), i.e., based on the Intersection over Union (IoU) between detected objects and tracks, providing an estimate of how likely a detection belongs to a track.

• **Hypothesis Abduction** Abduced hypothesis consist of assignments ($\mathcal{H}^{assign}$) of detections to tracks and high-level events ($\mathcal{H}^{events}$) explaining object motion, e.g., occlusion of an object, caused by the object passing behind an other object. The online abduction results in abduced visuospatial dynamics ($\mathcal{EXP}$) consisting of motion tracks ($\mathcal{MT}$) (generated using the abduced assignments in $\mathcal{H}^{assign}$) and the events ($\mathcal{H}^{events}$) explaining the motion tracks. Following perception as logical abduction most directly in the sense of Shanahan (2005), we define the task of abducing visual explanations as finding an association ($\mathcal{H}_t^{assign}$) of observed scene elements ($\mathcal{VO}_t$) to the motion tracks of objects ($\mathcal{MT}$) given by the predictions $\mathcal{P}_t$, together with a high-level explanation ($\mathcal{H}_t^{events}$), such that $[\mathcal{H}_t^{assign} \wedge \mathcal{H}_t^{events}]$ is consistent with the background knowledge and the previously abduced event sequence $\mathcal{H}^{events}$, and entails the perceived scene given by $< \mathcal{VO}_t, \mathcal{P}_t, \mathcal{ML}_t >$:

$$\Sigma \wedge \mathcal{H}^{events} \wedge [\mathcal{H}_t^{assign} \wedge \mathcal{H}_t^{events}] \models \mathcal{VO}_t \wedge \mathcal{P}_t \wedge \mathcal{ML}_t$$

where $\mathcal{H}_t^{assign}$ consists of the assignment of detections to object tracks[1], and $\mathcal{H}_t^{events}$ consists of the high-level *events* Θ explaining the assignments.

ASP encoding of the formal framework is included in the supplementary. Select fragments from example applications are included in Sec 4.

3.2 Technical Design and Implementation

The visual commonsense framework is available as a modularly engineered platform that seamlessly integrates with standard APIs: Python bindings, ROS, and vanilla docker

[1]In this setup, assignment of detections to object tracks is one amongst a range of possible assignments. In principle, this assignment could include any arbitrary visuospatial feature. For brevity, the present narration uses object-track assignment as one example.

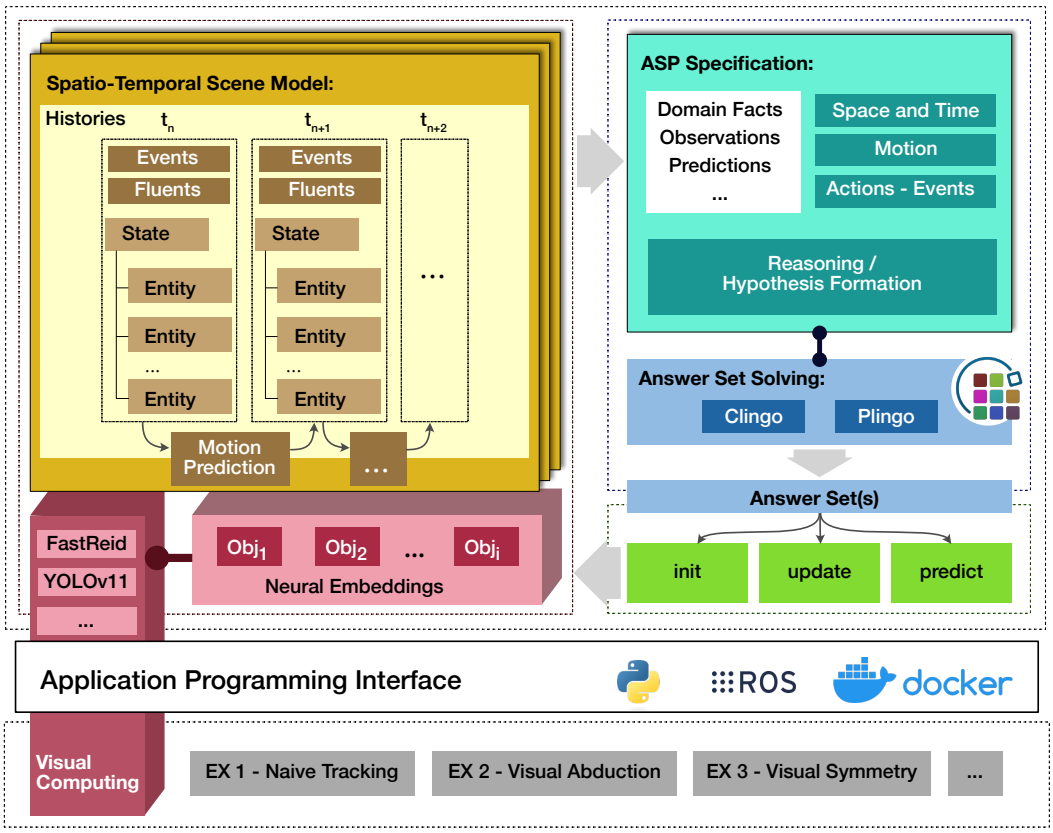


Figure 4: **Technical Architecture of Visual Commonsense Framework**. Central emphasis is on modularity, extensibility, and seamless integration as part of large-scale hybrid, embodied AI systems.

for light-weight use and/or development (Sec 4). A brief summary of the architecture of the framework follows:

- **Scene Model.** This module maintains (one or many) representations of the scene in terms of time series data characterising the scene state over time, including dynamic and static scene attributes and scene elements, which are characterised by their geometric extend, and can also have dynamic and static scene attributes. Furthermore, the mental model maintains sequences of events and corresponding fluents holding at specific time points. For interacting with this space-time scene model, it provides functions for initialising a new scene, adding / updating scene elements, and for predicting future positions of scene elements based on constant velocity using Kalman filters (Kalman 1960). To integrate with the ASP driven abstractions, the module generates ASP specifications describing the scene.

- **ASP Encoding** This module provides the ASP characterisations for doing inference over the space-time scene model. In particular, it implements the Space and Motion primitives, including external predicates for defining the spatial context for (geometric) spatial reasoning in ASP, and provides event based reasoning using Event Calculus (Kowalski and Sergot 1989). These elements are then used in the examples to characterise the domain. Together with the scene specification obtained from the scene model module (Fig. 4) these formalisations are used in the Solving step to find Answer Sets using the Clingo Python API [2]. The resulting Answer Set is then getting parsed to extract events and fluents, which can then be used to update the scene model according to these events.

- **Neural Embeddings.** Scene elements are obtained using state-of-the-art visual processing for obtaining geometrical abstractions of objects in the scene (e.g., bounding boxes, line-segments, etc).[3] Furthermore, the scene elements may also be represented in terms of neural feature representations that facilitate feature level analysis of scene elements, e.g. for estimating visual similarity based on the cosine similarity of the neural features as provided by (He et al. 2020). Such neural processing is made available in the ASP specification using neural predicates as proposed by NeurASP (Yang, Ishay, and Lee 2020).

[2] Clingo Python API. potassco.org/clingo/python-api/current

[3] It suffices to mention that low-level visual computing foundations are driven by deep learning based computer vision techniques, e.g., for visual feature detection, tracking supporting extraction and analyses of scene elements such as people, body-structure and objects in the scene, and object and scene motion (He et al. 2016; Ren et al. 2017; Deng et al. 2020; Baltrusaitis et al. 2018; Bergmann, Meinhardt, and Leal-Taixé 2019; Bewley et al. 2016; Jocher and Qiu 2024). Additional information may be obtained in the supplementary material.

We posit that this level and manner of generalised visual commonsense functionality relevant to diverse practical contexts is a first of its kind system that is now available for direct integration within AI applications, as well as for further technical extensions from the viewpoint of its core commonsense reasoning capabilities (Sec 6).

4 Applied Examples: Visual Commonsense in the Wild

We illustrate the gist of neurosymbolic visual commonsense through diverse but synergistic examples: Examples 1 and 2 covering complementary tracking and event abduction tasks, and Example 3 presents explainable visuospatial symmetry derivation task. Please note that the complete examples are rather elaborate; hence, here we present small relevant snippets aimed at communicating the key aspects.

EX 1. Naive Tracking: This example shows basic object tracking, using the presented framework. The application consists of two parts: the python part maintaining the geometric model of the scene, and the ASP program characterising the logic of the problem.

1.1. Python Setup. To setup the scene model within python, the scene is initialised as follows:

```
global scene
scene = dvsg.Scene()
```

Next, update positions of scene objects:

```
pos2D = dvsg.spatial_entities.Point2D(pos.x, pos.y)
if not id in scene.objects:
  static_attributes = {"type": type }
  scene.objects[id] =
  dvsg.Object(timestamp, id, pos2D, static_attributes)
dynamic_attributes = { "width": width, "height": height }
scene.objects[id].update_pos(timestamp, pos2D)
scene.objects[id].update_attributes(timestamp, dynamic_attributes)
```

Predicting object positions:

```
if id in scene.objects:
  scene.objects[id].predict_pos(curr_time)
```

1.2. Abducing Assignments for Moving Object Tracking. Within ASP, object tracking corresponds to the problem of assigning observed objects to tracks within the mental model; this is characterised using choice rules to generate possible assignments, together with 'starting' and 'ending' tracks respectively:

```
1{occurs_at(assign(Trk, Det), T):
    position(obj(detection(Det)), Det_Pos, curr_time(T);
    occurs_at(end(Trk), T): curr_time(T) } 1
:- position(obj(track(Trk)), Trk_Pos).

1{occurs_at(assign(Trk, Det), T): position(obj(track(Trk)),
    Trk_Pos), curr_time(T);
    occurs_at(new_track(Det), T): curr_time(T)} 1
:- position(obj(detection(Det)), Det_Pos).
```

The most promising assignment is then obtained by using the built-in optimisation of CLINGO (Gebser et al. 2014).

```
#minimize {Int_Dist@1: occurs_at(assign(Trk, Det), T),
  position(obj(track(Trk)), position2D(X1, Y1)),
  position(obj(detection(Det)), position2D(X2, Y2)),
  Dist = @distance2d_(X1,Y1,X2,Y2),  Int_Dist = @to_int_(Dist)}.
```

Figure 5: **Naive Tracking Results:** Example scene from the PathTrack dataset.

1.3. Tracking Result. The solving step results in a model containing the assignments, starting, and ending tracks; the outcomes get stored in the event sequence of the mental model component. As next step, we search within the events of the current time point and update the scene model according to the abduced events, i.e, we (**a**) update tracks with the new detections assigned to them; (**b**) initialise new tracks with the respective detections; and (**c**) delete ending tracks:

```
# update tracks
if event.arguments[0].name == "assign":
trk_id_str =  "trk(" + str(event.arguments[0].arguments[0]) + ")"
det_id_str =
  "det(" + str(event.arguments[0].arguments[1]) + ")"
  update_object(frame_nr, trk_id_str,
    detections[det_id_str].static_attributes["type"],
    detections[det_id_str].position2D, ...)

# new tracks
if event.arguments[0].name == "new_track":
  trk_id_str = "trk(" + str(trk_id) + ")"
  new_trk_det_id =
  "det(" + str(event.arguments[0].arguments[0]) + ")"
  update_object(frame_nr, trk_id_str,
    detections[new_trk_det_id].static_attributes["type"],
    detections[new_trk_det_id].position2D, ...)
  track_id += 1

# end tracks
if event.arguments[0].name == "end":
  trk_id_str = "trk(" + str(event.arguments[0].arguments[0]) + ")"
  dvsg.Scene.objects.pop(trk_id_str)
```

For the example scene from the PathTrack dataset (Manen et al. 2017), this simple tracking results in a sequence of motion tracks associated with the individuals in the scene, as depicted in Fig. 5.

EX 2. Visual Abduction: Example 1 focussed on naive tracking; building on this, this example proceeds with high-level event abduction as interpreted in our framework (Sec 3.1). Towards this, we introduce events into the ASP specification with the aim to explain perceived visual observations ($\mathcal{VO}$; Table 2), and utilise additional abductive steps aimed at hypothesising event occurrences explaining the adopted assignments in the (naive) tracking step. In particular, below we showcase abduced events explaining disappearance and reappearance of objects by inferring that one or more objects have been hidden by some other object(s). We first define the respective fluents and events:

```
fluent(hidden(Trk)) :- trk(Trk, _).
fluent(hidden_by(Trk1, Trk2)) :- trk(Trk2, _), trk(Trk1, _).

event(hides_behind(Trk1, Trk2)) :- trk(Trk1, _), trk(Trk2, _).
initiates(hides_behind(Trk1, Trk2), hidden(Trk1), T) :-
  trk(Trk1, _), trk(Trk2, _), time(T).
initiates(hides_behind(Trk1, Trk2), hidden_by(Trk1, Trk2), T) :-
  trk(Trk1, _), trk(Trk2, _), time(T).

event(unhides_from_behind(Trk1, Trk2)) :-
  trk(Trk1, _), trk(Trk2, _).
terminates(unhides_from_behind(Trk1, Trk2), hidden(Trk1), T) :-
  trk(Trk1, _), trk(Trk2, _), time(T).
terminates(unhides_from_behind(Trk1, Trk2),
  hidden_by(Trk1, Trk2), T) :-
    trk(Trk1, _), trk(Trk2, _), time(T).
```

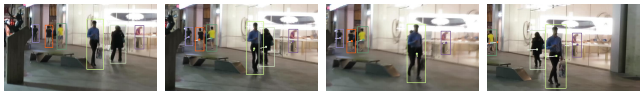


Figure 6: **Visual Abduction Results:** Example scene from the MOT dataset.

Next, we introduce choice rules to generate appropriate explanations:

```
1{occurs_at(hides_behind(Trk, Trk2), curr_time):
    trk(Trk2,_), position(overlapping, Trk, Trk2),
    not holds_at(hidden(Trk), curr_time);
  ...
}1 :- halt(Trk).

1{occurs_at(unhides_from_behind(Trk, Trk2), curr_time):
    trk(Trk2,_), not holds_at(hidden(Trk2), curr_time),
    holds_at(hidden_by(Trk, Trk2), curr_time),
    position(overlapping, Det, Trk2),
    @trk_estimate(Trk, Trk_est), @in_range(Det, Trk_est);
  ...
}1 :- resume(Trk, Det).
```

For the example scene in Fig. 6, the above characterisations are suitable to abduce that the person on the right hides behind another person and reappears afterwards.

```
...
occurs_at(hides_behind(trk_7, trk_2),14)
occurs_at(unhides_from_behind(trk_7, trk_2),27)
...
```

EX 3. Visuospatial Symmetry: This example diverges from Examples 1-2 to illustrate the flexibility of the framework in analysing visual stimuli based on any arbitrary set of requirements / constraints, exemplified for the case of visual symmetry. The example is relevant to domains in media studies and visual perception studies in psychology where the emphasis is on the study of the structure and function of visual stimuli. For the case of finding symmetrical structures in images, one may follow a similar approach as before, without considering the temporal dimension:[4]

3.1. Python Setup. Towards this, we initialise the scene as before and subsequently populate the scene with the detected scene elements based on object detection, in this case, driven by YOLOv11 (Jocher and Qiu 2024):

```
model = YOLO("yolo11x.pt")
results = model.predict(im)

for result in results:
  xywh = result.boxes.xywh
  conf = result.boxes.conf
  names = [result.names[cls.item()] for
    cls in result.boxes.cls.int()]
  for id in range(0, num_dets):
    pos2D = dvsg.spatial_entities.Point2D(int(xywh[id][0]), ... )
    static_attributes = { "class": names[id], ... }
    dynamic_attributes = { "width": int(xywh[id][2]), ... }
    scene.objects[id] =
      dvsg.Object(0, id, pos2D, static_attributes)
    scene.objects[id].update_attributes(0, dynamic_attributes)
```

3.2. Semantically Characterising Symmetry. In the ASP encodings, we characterise the notion of (reflectional) symmetry by taking the divergence from the perfect symmetrical position. Towards this, we define two possibilities of symmetrical placement: (**a**) there is a symmetrical pair of objects having equal size, being equidistant from the centre axis, and having the same appearance; and (**b**) there is a single object on the centre axis. Please note that this is merely a minimal example interpretation for the purposes of below:

```
divergence(sym_pair(ID1, ID2, Class), size(Div_W, Div_H),
  pos(Div_Horizontal, Div_Vertical), Class) :-
  divergence_width(ID1, ID2, Div_W),
  divergence_height(ID1, ID2, Div_H),
  divergence_sym_pos(ID1, ID2, pos(Div_Horizontal, Div_Vertical)),
  divergence_appearance(ID1, ID2, Appear_Div).

divergence(single_box(ID), Div) :- divergence_sym_pos(ID, Div).
```

Next, we obtain the divergence from the above basic configuration based on deviation in the **visual appearance** given by the neural features, and the geometrical divergences in **position** and **size**:

```
divergence_sym_pos(ID, Val) :- v_line(symmetry_axis, X_sym_axis),
  position(ID, position2D(X_box, _)), Val = |X_sym_axis-X_box|.

divergence_sym_pos(ID1, ID2, pos(Div_Horizontal, Div_Vertical)) :-
  v_line(symmetry_axis, X_sym_axis), position(obj(ID1),
  position2D(X_box1, Y_box1)), position(obj(ID2),
  position2D(X_box2, Y_box2)), image_size(W_img, H_img),
  Div_Vertical = |Y_box2 - Y_box1|, Dist1 = X_sym_axis - X_box1,
  Dist2 = X_box2 - X_sym_axis, Div_Horizontal = |Dist2 - Dist1|.

divergence_width(ID1, ID2, Val) :-
  width(ID1, W1), width(ID2, W2), Val = |W1-W2|.

divergence_height(ID1, ID2, Val) :-
  height(ID1, H1), width(ID2, H2), Val = |H1-H2|.

divergence_appearance(ID1, ID2, Div) :-
  position(ID1, _), position(ID2, _),
  @neural(similarity, ID1, ID2, Sim), Div = 100-Sim.
```

Finding symmetrical structures is then done by abducing pairs of symmetrical elements, and single elements:

```
1{ sym_pair(ID, ID2, Class): pred(ID2), class(Class), ID != ID2;
  sym_pair(ID1, ID, Class): pred(ID1), class(Class), ID != ID1;
  single_box(ID); ... } 1 :- pred(ID).

:- sym_pair(ID1, ID2, _), sym_pair(ID2, ID1, _).
:- sym_pair(ID1, _, _), single_box(ID1).
:- sym_pair(_, ID2, _), single_box(ID2).
:- sym_pair(ID1, ID2, _), sym_pair(ID1, ID3, _), ID2 != ID3.
```

Finally, picking matching symmetrical objects is based on optimisation over weighted divergence in symmetry features to minimize the divergence from the symmetrical placement:

```
#minimize {(Div)@1 :
  sym_pair(ID1, ID2, Class),
  divergence(sym_pair(ID1, ID2, Class), size(Div_W, Div_H),
  pos(Div_Vertical, Div_Horizontal), Class_Div, App_Div),
  Div = w1*Div_W + w2*Div_H + w3*Div_Vertical +
  w4*Div_Horizontal+w5*Class_Div}.

#minimize {(Div)@1 : single_box(ID),
  divergence(single_box(ID), Div_Pos), Div = w6*Div_Pos}.
```

3.3. Symmetry Results. For the example image in Fig. 7 from the MS COCO dataset (Lin et al. 2014), where we have 9 scene elements detected[5], we generate the following scene description:

```
position(obj(0), position2D(157, 239)). class(obj(0), elephant).
conf(obj(0), 95). width(obj(0), 200). height(obj(0), 260).

position(obj(1), position2D(359, 223)). class(obj(1), elephant).
conf(obj(1), 94). width(obj(1), 267). height(obj(1), 281).
...

position(obj(8), position2D(622, 205)). class(obj(8), elephant).
conf(obj(8), 59). width(obj(8), 34). height(obj(8), 41).
```

[4]Indeed, symmetry may be interpreted also temporally, but space is limited to present such a characterisation here.

[5]For simplicity of the example, here we are only including object level detections obtained via YOLO (Jocher and Qiu 2024).

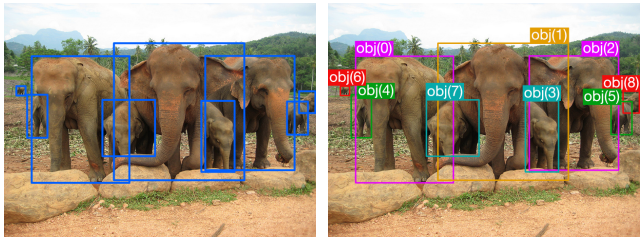


Figure 7: **Symmetry Results:** Example result based on a test image from the MS COCO dataset, depicting symmetry pairs detected from the object configuration.

This scene description then results in the following model providing symmetrical elements of the scene, as visualised in Fig. 7.

```
sym_pair(obj(2),obj(0),elephant) sym_pair(obj(5),obj(4),elephant)
sym_pair(obj(3),obj(7),elephant) sym_pair(obj(6),obj(8),elephant)
single_box(1)
```

This gives the identified symmetry pairs and single boxes. Furthermore, the model provides the divergence of symmetry features from the identified symmetrical structure.

5 Evaluation: Results & Discussion

The presented framework facilitates an easy integration of ASP based reasoning about motion dynamics in visuospatial scenes within common tasks in autonomous vision systems. In the following we are presenting evaluation results for the task of object tracking and in subjective symmetry perception and discuss their specifics.

Multi-Object Tracking. Visual abduction based multi-object tracking (Suchan, Bhatt, and Varadarajan 2021; Suchan, Bhatt, and Varadarajan 2019) has been evaluated on the community established MOT dataset (Milan et al. 2016). Naive Tracking as implemented in **EX 1.** follows the common tracking by detection approach using a simple distance metric without additional methods for enhancing tracking performance, the approach therefore achieves a basic tracking performance comparable to similar approaches, such as SORT (Bewley et al. 2016). In particular, a IOU based version of the approach presented in **EX 1.** reaches a Multi-Object Tracking Accuracy (MOTA) of 41.4% on MOT17, however, due to the simplicity of the method, it reaches an average processing speed of over $200 fps$ on standard hardware (including ASP solving). With the addition of abductive reasoning for occlusion events as presented in **EX 2.**, the tracking performance increases by approx. 5% points, to 46.2%. This increase in tracking performance comes with a drop in processing speed due to the elaborate reasoning. However, the approach is still capable of reaching above realtime performance on challenging real-world scenes of the MOT dataset.

Most recently, the abductive tracking approach of **EX 2.** has been further extended with a *preferential ranking setup* (Monsen, Suchan, and Bhatt 2025). With this approach, including neural appearance features and weight learning further increases the MOTA score to 62.0% on the validation split of the MOT17 dataset.

Method	MOTA (%)	MOTP (%)
Naive Tracking	41.4	88.0
Visual Abduction	46.2	87.9

Table 3: **Multi-Object tracking Results.** Multi-object tracking Accuracy (MOTA) and Precision (MOTP).

Features	CA (%)	Class Prob. Err.
Visual	41.33	0.0572886659
Visual+Objects	54.00	0.0375853705

Table 4: **Subjective Symmetry Perception Results.** Classification accuracy (CA) and mean error.

Visuospatial Symmetry. Semantic characterisations of object level symmetry as presented in **EX 3.** have be applied within a multi-level model of visual symmetry in the context of a human behavioural study focusing on subjective perception of symmetrical structures in real-world images (Suchan et al. 2018). Here, we examined the characterisation of visuospatial symmetry and show that using such semantically founded symmetry characterisations can improve the models' ability to predict the subjective judgment of human participants of symmetrical structures in images (Table 4). In particular, classification of images into four symmetry classes (not_symmetric, somewhat_symmetric, symmetric, and highly_symmetric) increases from 41.33% to 54% by including semantic characterisation of symmetry.

6 Conclusion and Outlook

We have developed a generalised, systematically formalised, declaratively modelled framework for visual commonsense combining diverse techniques in AI and Vision. The developed framework is novel in several ways, primarily centred on its ability to offer domain-independent neurosymbolic reasoning capabilities encompassing space, events, actions, motion within an established non-monotonic setting, and in conjunction with complex quantitative visual data. A secondary motivation behind this work has been to also showcase the value of integrating robust, declarative methods in KRR within large-scale AI systems requiring diverse perceptual and decision-making components.

The developed framework may be either seamlessly used within application domains, or it may be utilised as a research and development platform for research in KR and Vision/ML with the aim to extend the offered visual commonsense reasoning capabilities, e.g., by incorporating causal inference, epistemic reasoning, conceptual reasoning combining semantic knowledge and inference, (neurosymbolic) integration of reasoning and learning, integration of formal argumentation frameworks etc. From an applied viewpoint, presently ongoing work focusses of integration within real-world infrastructure for autonomous driving, e.g., (Suchan and Osterloh 2023).

Dissemination. All relevant materials pertaining to the developed framework (e.g., code, data, documentation) as well as future extensions may be consulted via:

Cognitive Vision., codesign-lab.org/cognitive-vision

Acknowledgements

We acknowledge funding by the Swedish Research Council (Vetenskapsrådet – VR), and the Swedish Foundation for Strategic Research (Stiftelsens för Strategisk Forskning – SSF).

References

Aditya, S.; Yang, Y.; and Baral, C. 2019. Integrating knowledge and reasoning in image understanding. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 6252–6259. ijcai.org.

AI HLEG. 2019. High-level expert group on artificial intelligence: Ethical guidelines for trustworthy ai.

Allen, J. F. 1983. Maintaining knowledge about temporal intervals. *Commun. ACM* 26(11):832–843.

Balbiani, P.; Condotta, J.; and del Cerro, L. F. 1999. A new tractable subclass of the rectangle algebra. In Dean, T., ed., *IJCAI 1999, Sweden*, 442–447. Morgan Kaufmann.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 59–66.

Bennett, B.; Cohn, A. G.; Torrini, P.; and Hazarika, S. M. 2000. A foundation for region-based qualitative geometry. In *Proceedings of the 14th European Conference on Artificial Intelligence*, 204–208.

Bergmann, P.; Meinhardt, T.; and Leal-Taixé, L. 2019. Tracking without bells and whistles. In *The IEEE International Conference on Computer Vision (ICCV)*.

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE Intnl. Conf. on Image Processing (ICIP)*, 3464–3468.

Bhatt, M., and Kersting, K. 2017. Semantic interpretation of multi-modal human-behaviour data - making sense of events, activities, processes. *KI / Artificial Intelligence* 31(4):317–320.

Bhatt, M., and Loke, S. W. 2008. Modelling dynamic spatial systems in the situation calculus. *Spatial Cognition & Computation* 8(1-2):86–130.

Bhatt, M., and Suchan, J. 2021. Artificial visual intelligence: Perceptual commonsense for human-centred cognitive technologies. In *Human-Centered Artificial Intelligence: Advanced Lectures*, 216–242. Springer-Verlag.

Bhatt, M.; Guesgen, H. W.; Wölfl, S.; and Hazarika, S. M. 2011. Qualitative spatial and temporal reasoning: Emerging applications, trends, and directions. *Spatial Cognition & Computation* 11(1):1–14.

Bhatt, M.; Schultz, C.; and Freksa, C. 2013. The 'Space' in Spatial Assistance Systems: Conception, Formalisation and Computation. In Tenbrink, T.; Wiener, J.; and Claramunt, C., eds., *Representing space in cognition: Interrelations of behavior, language, and formal models. Series: Explorations in Language and Space*. 978-0-19-967991-1, Oxford University Press.

Bhatt, M. 2012. Reasoning about Space, Actions and Change: A Paradigm for Applications of Spatial Reasoning. In *Qualitative Spatial Representation and Reasoning: Trends and Future Directions*. IGI Global, USA.

Blythe, J.; Hobbs, J. R.; Domingos, P.; Kate, R. J.; and Mooney, R. J. 2011. Implementing weighted abduction in markov logic. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11. USA: Association for Computational Linguistics.

Brewka, G.; Eiter, T.; and Truszczyński, M. 2011. Answer set programming at a glance. *Commun. ACM* 54(12):92.

Cohn, A.; Bennett, B.; Gooday, J.; and Gotts, N. 1997. Representing and reasoning with qualitative spatial relations about regions. In Stock, O., ed., *Spatial and Temporal Reasoning*. Dordrecht: Kluwer Academic Publishers. 97–134.

Davis, E., and Marcus, G. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* 58(9):92–103.

Davis, E. 2017. Logical formalizations of commonsense reasoning: A survey. *J. Artif. Intell. Res.* 59:651–723.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*.

Eiter, T.; Higuera, N.; Oetsch, J.; and Pritz, M. 2022. A neuro-symbolic ASP pipeline for visual question answering. *Theory Pract. Log. Program.* 22(5):739–754.

Eppe, M., and Bhatt, M. 2015a. Approximate postdictive reasoning with answer set programming. *J. Appl. Log.* 13(4):676–719.

Eppe, M., and Bhatt, M. 2015b. A history based approximate epistemic action theory for efficient postdictive reasoning. *J. Appl. Log.* 13(4):720–769.

EU Commission. 2019. Communication: Building trust in human centric artificial intelligence.

EU Commission. 2021. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

Gebser, M.; Kaminski, R.; König, A.; and Schaub, T. 2011. Advances in gringo series 3. In *LPNMR*, volume 6645 of *Lecture Notes in Computer Science*, 345–351. Springer.

Gebser, M.; Kaminski, R.; Kaufmann, B.; and Schaub, T. 2012. *Answer Set Solving in Practice*. Morgan & Claypool.

Gebser, M.; Kaminski, R.; Kaufmann, B.; and Schaub, T. 2014. Clingo = ASP + control: Preliminary report. *CoRR* abs/1405.3694.

Hazarika, S. M. 2005. *Qualitative spatial change: space-time histories and continuity*. Ph.D. Dissertation, The University of Leeds.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

He, L.; Liao, X.; Liu, W.; Liu, X.; Cheng, P.; and Mei, T. 2020. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*.

Jocher, G., and Qiu, J. 2024. Ultralytics yolo11.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82(1):35–45.

Kowalski, R., and Sergot, M. 1989. *A Logic-Based Calculus of Events*. Berlin, Heidelberg: Springer-Verlag. 23?51.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755, Springer.

Ma, J.; Miller, R.; Morgenstern, L.; and Patkos, T. 2014. An epistemic event calculus for asp-based reasoning about knowledge of the past, present and future. In *LPAR: 19th Intl. Conf. on Logic for Programming, Artificial Intelligence and Reasoning*, volume 26 of *EPiC Series in Computing*, 75–87. EasyChair.

Manen, S.; Gygli, M.; Dai, D.; and Gool, L. V. 2017. Pathtrack: Fast trajectory annotation with path supervision. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 290–299.

Mani, I., and Pustejovsky, J. 2012. *Interpreting Motion - Grounded Representations for Spatial Language*, volume 5 of *Explorations in language and space*. Oxford Uni. Press.

Milan, A.; Leal-Taixé, L.; Reid, I. D.; Roth, S.; and Schindler, K. 2016. MOT16: A benchmark for multi-object tracking. *CoRR* abs/1603.00831.

Miller, R.; Morgenstern, L.; and Patkos, T. 2013. Reasoning about knowledge and action in an epistemic event calculus. In *COMMONSENSE 2013*.

Monsen, J.; Suchan, J.; and Bhatt, M. 2025. Probabilistic Answer Set Programming Driven Ranking of Dynamic Space-Time Belief Models. In *The 9th International Joint Conference, RuleML+RR 2025*, Lecture Notes in Computer Science. Springer.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(6):1137–1149.

Schaub, T., and Woltran, S. 2018. Special issue on answer set programming. *KI* 32(2-3):101–103.

Schultz, C. P. L.; Bhatt, M.; Suchan, J.; and Walega, P. A. 2018. Answer Set Programming Modulo Space-Time. In Benzmüller, C.; Ricca, F.; Parent, X.; and Roman, D., eds., *Rules and Reasoning - Second International Joint Conference, RuleML+RR 2018, Luxembourg, September 18-21, 2018, Proceedings*, volume 11092 of *Lecture Notes in Computer Science*, 318–326. Springer.

Shanahan, M. 1997. *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. Cambridge, MA, USA: MIT Press.

Shanahan, M. 2005. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science* 29(1):103–134.

Suchan, J., and Osterloh, J.-P. 2023. Assessing drivers' situation awareness in semi-autonomous vehicles: Asp based characterisations of driving dynamics for modelling scene interpretation and projection. *Electronic Proceedings in Theoretical Computer Science* 385:300–313.

Suchan, J.; Bhatt, M.; Vardarajan, S.; Amirshahi, S. A.; and Yu, S. 2018. Semantic Analysis of (Reflectional) Visual Symmetry: A Human-Centred Computational Model for Declarative Explainability. *Advances in Cognitive Systems* 6:65–84.

Suchan, J.; Bhatt, M.; and Varadarajan, S. 2019. Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving. In Kraus, S., ed., *Proc. of 25th Intnl. Joint Conference on Artificial Intelligence, IJCAI 2019*.

Suchan, J.; Bhatt, M.; and Varadarajan, S. 2021. Commonsense visual sensemaking for autonomous driving - on generalised neurosymbolic online abduction integrating vision and semantics. *Artif. Intell.* 299:103522.

Tan, M., and Le, Q. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K., and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6105–6114. Long Beach, California, USA: PMLR.

Tu, K.; Meng, M.; Lee, M. W.; Choe, T. E.; and Zhu, S. C. 2014. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*.

Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B. B. G.; Geiger, A.; and Leibe, B. 2019. Mots: Multi-object tracking and segmentation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Z.; Ishay, A.; and Lee, J. 2020. Neurasp: Embracing neural networks into answer set programming. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 1755–1762. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yu, H.; Siddharth, N.; Barbu, A.; and Siskind, J. M. 2015. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. *J. Artif. Intell. Res. (JAIR)* 52:601–713.

Zeng, W.; Luo, W.; Suo, S.; Sadat, A.; Yang, B.; Casas, S.; and Urtasun, R. 2019. End-to-end interpretable neural motion planner. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.