INTEGRATION BY PARTS: COMPOSITIONAL MULTIMODAL LEARNING FOR DYNAMIC 3D SCENE GRAPH GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Three-dimensional (3D) scene understanding is essential for autonomous systems operating in dynamic environments. While RGB, LiDAR, and textual semantics provide complementary information, existing dynamic 3D scene graph methods often rely on holistic object representations that are sensitive to occlusion, viewpoint changes, sensor noise, incomplete observations, and cross-modal misalignment. We propose Integration by Parts (IBP), a compositional multimodal framework that represents objects using fine-grained latent components rather than single embeddings. Shared learnable queries *decompose* RGB and LiDAR observations into latent components, which are explicitly *aligned* across modalities and *integrated through recomposition* with global multimodal and textual context to form compositional object representations. A temporal module associates these representations across observations and maintains semantic consistency for spatial-temporal reasoning. Experiments on KITTI-360 demonstrate the effectiveness of IBP for multimodal alignment, temporal reasoning, and robustness to incomplete observations. The code is available at: `https://anonymous.4open.science/r/ibp_anonymized-/`.

1 INTRODUCTION

Three-dimensional (3D) scene understanding is a fundamental capability for intelligent and autonomous systems operating in complex physical environments, with applications spanning autonomous navigation (Li & Cui, 2025; Li et al., 2025), environmental monitoring (Cavaliere et al., 2019; Mendieta & Tabkhi, 2021; Raine & Fischer, 2026), disaster assessment (Wang et al., 2026), precision agriculture (Li et al., 2026), and urban planning (Ro et al., 2025). Across these diverse domains, modern sensing platforms increasingly rely on complementary modalities such as RGB imagery and LiDAR point clouds to jointly capture appearance, semantic, and geometric information, enabling richer scene understanding than individual modalities alone (Kong et al., 2025; Behari et al., 2025; Singh et al., 2025; Sarkar et al., 2023). However, achieving consistent scene understanding from heterogeneous multimodal observations remains challenging in dynamic environments, where objects undergo viewpoint changes, occlusions, partial visibility, sensor noise, and temporal variations. These challenges call for representations that can effectively integrate complementary multimodal evidence while preserving object identity and relational information across time.

Dynamic 3D scene graphs provide a structured representation of such environments by modeling objects, their spatial relationships, and their evolution over time (Feng et al., 2023; Wang et al., 2024b; Heo et al., 2026; Armeni et al., 2019; Kim et al., 2019; Wu et al., 2023). In multimodal settings, different sensing modalities provide complementary information about the same objects. For example, RGB imagery captures rich appearance and semantic cues, whereas LiDAR provides explicit geometric and structural information. Existing approaches, however, predominantly rely on *holistic* object representations, aggregating multimodal information into a single object embedding before fusion and temporal reasoning (Rotondi et al., 2026; Chen et al., 2025; Le et al., 2026). While effective under sufficiently complete observations, such holistic representations may obscure complementary local evidence available across modalities, particularly when objects are occluded or only partially observed. Moreover, as viewpoint and visibility change over time, different portions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of an object may remain observable across successive frames. Consequently, preserving fine grained evidence across modalities and time before object-level fusion may provide a more robust basis for dynamic scene understanding.

Motivated by this observation, we draw inspiration from human scene understanding, where an object can often be recognized even when only partially visible, as informative local evidence can still provide cues about its identity. We propose *Integration by Parts* (IBP), a compositional framework for multimodal dynamic 3D scene graph generation that follows a similar representational principle. Instead of directly collapsing heterogeneous observations into holistic object representations, IBP decomposes RGB and LiDAR observations into latent components and preserves this compositional structure across modalities. Although RGB and LiDAR capture fundamentally different characteristics of an object, IBP organizes their complementary evidence around a shared set of latent components and explicitly aligns corresponding components across the two modalities. The aligned components are subsequently integrated through recomposition with global multimodal and semantic context to form compositional object representations, allowing informative local evidence to contribute when observations are incomplete or vary over time. These representations support object association and spatial-temporal relational reasoning. Thus, rather than treating multimodal fusion as an exclusively object level operation, IBP introduces component level *decomposition*, *alignment*, and *integration through recomposition* prior to temporal association and reasoning.

We evaluate Integration by Parts on KITTI-360 against recent scene understanding and dynamic scene graph methods under a controlled protocol, including evaluation with predicted associations and tracklets. Quantitative and qualitative analyses demonstrate strong crossmodal alignment, reliable temporal association, and improved spatial-temporal relational reasoning. Ablations show that compositional representations particularly benefit temporal reasoning and tracklet coverage, while controlled occlusion experiments demonstrate resilience to moderately incomplete observations. Our main contributions are summarized as follows:

1. We propose *Integration by Parts*, a compositional multimodal framework for dynamic 3D scene graph generation that represents objects through latent components rather than exclusively holistic representations.
2. We introduce component level RGB-LiDAR learning that discovers modality specific evidence within a shared latent component space and explicitly aligns corresponding components before object level fusion.
3. We integrate the resulting compositional representations with cross-frame association and temporal reasoning for dynamic scene graph generation, and demonstrate improved temporal reasoning, tracklet continuity, cross-modal alignment, and robustness under partial observations.

2 Related Works

3D Scene Graphs. 3D scene graphs provide structured object-centric representations of objects and their spatial relations, with early systems such as 3D Scene Graph (Armeni et al., 2019) and SceneGraphFusion (Wu et al., 2021) establishing graph construction from 3D observations. Subsequent approaches increasingly incorporate semantic and linguistic information: VL-SAT (Wang et al., 2023) transfers visual-linguistic knowledge to point-cloud scene graph prediction, while Open3DSG (Koch et al., 2024) and CLIP-driven 3DSSG (Chen et al., 2024) exploit vision-language representations for open-vocabulary reasoning. Recent object-centric representation learning (Heo et al., 2026) further highlights the importance of discriminative object representations. These approaches, however, primarily address static 3D scenes and do not jointly model multimodal component correspondence and temporal object evolution.

Multimodal 3D and RGB-LiDAR Understanding. Cross-modal learning has been widely explored for integrating complementary appearance and geometric information. CAT-Det (Zhang et al., 2022) reduces camera–LiDAR representation discrepancy through contrastive learning, while CLIP^2 (Zeng et al., 2023) aligns image, language, and point-cloud representations in a shared space. CrossOver (Sarkar et al., 2025) extends alignment across multiple 2D/3D modalities, while USG (Wu et al., 2025) combines cross-modal object association with text-centric contrastive learning for universal scene graphs. SAM4D (Xu et al., 2025) further incorporates camera-LiDAR align-

ment with temporal cross-modal memory. Despite this progress, these methods predominantly align or fuse representations at the scene or object level. In contrast, IBP discovers a shared set of latent components and establishes RGB-LiDAR correspondence at the component level before object-level composition and relational reasoning.

Dynamic Scene Graphs. Temporal scene graph generation has primarily been studied in video, including STTran (Cong et al., 2021), which models spatial-temporal context using Transformers, and OED (Wang et al., 2024a), which formulates dynamic scene graph generation as an end-to-end prediction problem. Dynamic 3D graphs have also been explored for autonomous driving (Greve et al., 2024), while recent work addresses robust spatio-temporal scene graph generation (Peddi et al., 2025) and multimodal 3D scene graphs for embodied reasoning (Huang et al., 2026). These methods establish the importance of temporal and multimodal relational reasoning, but do not jointly address component level RGB-LiDAR correspondence, semantic grounding, cross-frame association, and tracklet level spatial-temporal reasoning within a unified framework.

3 Problem Formulation

Given a temporal sequence of multimodal observations $\mathcal{X} = \{X^{(t)}\}_{t=1}^{T}$, the observation at time t is $X^{(t)} = \left(I^{(t)}, P^{(t)}, S^{(t)}\right)$, where $I^{(t)}$ denotes the RGB observation, $P^{(t)}$ the corresponding LiDAR point cloud, and $S^{(t)}$ a caption-derived textual description. Our objective is to construct a dynamic 3D scene graph that captures object semantics, within frame spatial relations, crossframe object correspondence, and temporal evolution. At each time t, the frame level scene graph is $G^{(t)} = \left(V^{(t)}, E_S^{(t)}\right)$, where $V^{(t)}$ denotes the set of object nodes and $E_S^{(t)}$ the spatial relations between objects within the observation. Across consecutive observations, we define the identity association relation $E_A^{(t,t+1)} = \left\{(v_i^{(t)}, v_j^{(t+1)}) \mid v_i^{(t)} \sim_{\text{id}} v_j^{(t+1)}\right\} \subseteq V^{(t)} \times V^{(t+1)}$, where $v_i^{(t)} \sim_{\text{id}} v_j^{(t+1)}$ indicates that the two nodes correspond to the same physical object. These associations form object-centric tracklets that provide temporal context for predicting temporal relation labels r_{ij}^{T} between object pairs, such as *approaching*, *moving away*, and *same motion direction*. The resulting dynamic scene graph is therefore represented as $\mathcal{G} = \left(\{V^{(t)}\}_{t=1}^{T}, \{E_S^{(t)}\}_{t=1}^{T}, \{E_A^{(t,t+1)}\}_{t=1}^{T-1}\right)$, where $E_S^{(t)}$ represents within observation relational edges carrying spatial predicates, while $E_A^{(t,t+1)}$ captures cross observation identity associations. Temporal predicates r_{ij}^{T} are inferred for relational object pairs using their tracklet level temporal context.

4 Methodology

Figure 1 presents the overall architecture of Integration by Parts (IBP), which follows a *decompose–align–recompose–associate–reason* pipeline. First, shared latent queries decompose local RGB and LiDAR evidence into modality specific latent components, while textual semantics provide an object level semantic reference. Second, corresponding RGB and LiDAR components are explicitly aligned to establish finegrained crossmodal correspondence. Third, the aligned components are integrated and recomposed with global multimodal context to form compositional object representations. These representations are then associated across observations to construct object centric tracklets, which provide temporal context for predicting spatial and temporal relations in the dynamic 3D scene graph.

4.1 Compositional Multimodal Representation Learning

For each object i, IBP constructs a compositional representation from complementary RGB, LiDAR, and textual evidence. Rather than collapsing modality specific observations into whole object descriptors before fusion, we preserve local RGB and LiDAR evidence, organize it into shared latent components, and establish crossmodal correspondence before object level aggregation.

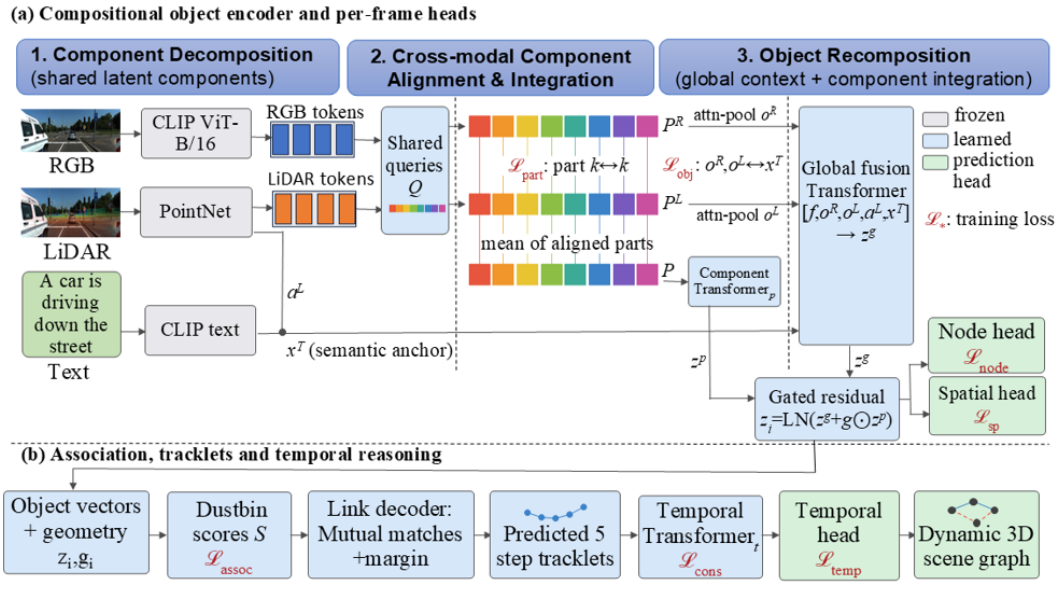


Figure 1: Overview of Integration by Parts. (a) The compositional object encoder first decomposes RGB and LiDAR observations into shared latent components, aligns and integrates corresponding components across modalities, and recomposes them with global multimodal context to form the object representation zi. (b) The resulting object representations are used for association, tracklet construction, and temporal reasoning to build a dynamic 3D scene graph G.

4.1.1 Modality specific feature encoding

For each object i, we extract complementary appearance, geometric, and semantic features from RGB, LiDAR, and text, respectively. The RGB object crop is encoded using CLIP ViT-B/16 (Radford et al., 2021), producing patch level tokens $H_i^R \in \mathbb{R}^{196\times768}$, where 196 corresponds to 14×14 image patches and 768 is the CLIP feature dimension. LiDAR points within the corresponding 3D bounding box are processed by a PointNet-based encoder (Qi et al., 2017), yielding local geometric tokens $H_i^L \in \mathbb{R}^{128\times512}$ and a global max-pooled representation $a_i^L \in \mathbb{R}^{512}$. An object centric caption generated by frozen BLIP (Li et al., 2022) is encoded using the CLIP text encoder to obtain the semantic representation $x_i^T \in \mathbb{R}^{512}$. Since the textual description is derived from RGB and does not provide reliable component level grounding, x_i^T serves as an object level semantic reference, while latent component discovery operates on the local RGB and LiDAR tokens. Further preprocessing and encoder details are provided in the Appendix A.2.

4.1.2 Latent Component Discovery

Directly pooling each modality into a single object vector can suppress localized evidence, particularly under occlusion and partial observation. IBP therefore represents each object using multiple latent components before global fusion. The RGB and LiDAR token sequences are first projected into a common d-dimensional space and augmented with positional and modality embeddings, $\widetilde{X}_i^m = X_i^m W^m + E^m$, where $m \in \{R, L\}$.

We introduce K shared learnable queries $Q \in \mathbb{R}^{K\times d}$, where each query acts as a latent component slot that independently aggregates evidence from the RGB and LiDAR token sequences. Sharing the query bank provides a common latent indexing mechanism across modalities, while subsequent crossmodal alignment encourages corresponding indices to encode compatible object evidence. Importantly, no semantic part labels or predefined physical correspondences are assigned to these components.

For each modality, the shared queries attend to its projected token sequence through multihead cross attention,

$$P_i^m = \mathrm{MHA}\left(Q, \widetilde{X}_i^m, \widetilde{X}_i^m\right) = [p_{i1}^m, \ldots, p_{iK}^m] \in \mathbb{R}^{K\times d}, \text{ where } m \in \{R, L\}. \tag{1}$$

Thus, P_i^R and P_i^L provide K modality specific realizations associated with the same latent component indices. Their crossmodal correspondence is not assumed a priori, rather it is learned through the alignment objective introduced in section 4.1.3

Component Pooling While the complete component sets are retained for subsequent component aware processing, we also obtain an object level summary for each modality through attention pooling, $\alpha_{ik}^m = \mathrm{softmax}_k(w^\top p_{ik}^m)$, and $o_i^m = \sum_{k=1}^K \alpha_{ik}^m p_{ik}^m \in \mathbb{R}^d$. The pooled representations o_i^R and o_i^L provide modality level object summaries, while P_i^R and P_i^L preserve the finegrained evidence required for component level alignment and subsequent compositional refinement.

4.1.3 Crossmodal Component Alignment

Shared queries establish a common component organization but do not by themselves guarantee correspondence between RGB and LiDAR components. We therefore explicitly align the component representations. Corresponding RGB and LiDAR components of the same object are treated as positive pairs and non-corresponding components as negatives through a symmetric contrastive objective,

$$\mathcal{L}_{\mathrm{part}} = \mathcal{L}_{\mathrm{NCE}}(P^R, P^L). \tag{2}$$

This encourages corresponding appearance and geometric components to occupy nearby locations in the shared latent space while retaining modality specific evidence.

At the object level, the pooled representations o_i^R and o_i^L are aligned with the textual representation, which serves as a common semantic anchor, through the objective,

$$\mathcal{L}_{\mathrm{obj}} = \frac{1}{2}\left[\mathcal{L}_{\mathrm{NCE}}(o^R, x^T) + \mathcal{L}_{\mathrm{NCE}}(o^L, x^T)\right]. \tag{3}$$

Thus, RGB and LiDAR are aligned at the component level, while text provides semantic consistency at the object level. The detailed contrastive formulation is provided in the Appendix A.2.

4.1.4 Compositional Object Representation

Having decomposed the RGB and LiDAR observations into aligned latent components, IBP recomposes their complementary evidence into a unified object representation. We preserve two levels of information during this process: global multimodal context provides a holistic description of the object, while the aligned components retain finegrained crossmodal evidence. To obtain the global representation, the pooled RGB and LiDAR descriptors, global LiDAR feature, and textual representation are jointly contextualized using a learnable fusion token f:

$$z_i^g = \mathrm{Transformer}_{\mathrm{global}}\left([f, o_i^R, o_i^L, a_i^L, x_i^T]\right)_0, \tag{4}$$

where the subscript 0 denotes the output corresponding to the prepended fusion token. Thus, z_i^g summarizes the global appearance, geometric, and semantic context of object i.

In parallel, the aligned RGB and LiDAR components are integrated and contextualized using a component level Transformer given as

$$P_i = \mathrm{Transformer}_{\mathrm{part}}\left(\frac{P_i^R + P_i^L}{2}\right), \qquad z_i^p = \sum_{k=1}^K \beta_{ik} P_{ik}, \tag{5}$$

where β_{ik} are learned attention weights over the refined components. The resulting z_i^p summarizes the aligned component level evidence retained during decomposition and crossmodal alignment.

Finally, IBP recomposes the global and component level representations through gated residual integration:

$$g_i = \sigma(W_g[z_i^g \| z_i^p]), \qquad z_i = \mathrm{LN}(z_i^g + g_i \odot z_i^p), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, $\odot$ denotes elementwise multiplication, and LN denotes layer normalization. The gate adaptively controls the contribution of component level evidence to the global representation. The resulting z_i therefore recomposes global multimodal context with aligned finegrained evidence and is used for subsequent cross observation association and relational reasoning.

4.2 Temporal Association and Tracklet Construction

Dynamic scene graph reasoning requires observations of the same object to be linked across time despite viewpoint changes, camera motion, occlusion, and partial visibility. IBP therefore associates the compositional object representations across consecutive observations using explicit geometric cues and constructs object centric tracklets for subsequent temporal reasoning.

4.2.1 Geometry aware object association

Let $Z^{(t)} = [z_1^{(t)}, \ldots, z_{N_t}^{(t)}]$ and $Z^{(t+1)} = [z_1^{(t+1)}, \ldots, z_{N_{t+1}}^{(t+1)}]$ denote the compositional object representations in two consecutive observations. For each candidate pair (i, j), we construct

$$a_{ij} = [z_i^{(t)} \| z_j^{(t+1)} \| (z_i^{(t)} - z_j^{(t+1)}) \| (z_i^{(t)} \odot z_j^{(t+1)}) \| g_{ij}^a], \qquad s_{ij} = h_a(a_{ij}), \tag{7}$$

where $g_{ij}^a \in \mathbb{R}^{16}$ provides geometric and observation level consistency cues and $h_a(\cdot)$ is an MLP that produces the association score s_{ij}. Thus, the association jointly captures object specific information, cross observation variation, feature compatibility, and geometric consistency.

The score matrix is augmented with dustbin scores (Sarlin et al., 2020) to accommodate unmatched observations, where $s_{i\varnothing}$ and $s_{\varnothing j}$ denote the scores for leaving source object i and target object j unmatched, respectively. The augmented scores are normalized using log-domain Sinkhorn iterations (Cuturi, 2013) to obtain approximately one-to-one assignment scores. Mutual assignments are retained as cross observation identity associations, while dustbin assignments represent unmatched observations. Further details of the geometric descriptor and Sinkhorn normalization are provided in the Appendix A.2.

4.2.2 Tracklet construction and recovery

Predicted associations are chained into object centric tracklets over five consecutive key samples (approximately 2 seconds). For object i,

$$Z_i = [z_{i,0}, \ldots, z_{i,4}], \qquad M_i = [m_{i,0}, \ldots, m_{i,4}], \quad m_{i,t} \in \{0, 1\}, \tag{8}$$

where M_i indicates valid observations at each temporal position, and tracklets are restricted to individual sequence/clip boundaries.

During training, persistent groundtruth identities are used only to construct tracklets and association targets, and are never supplied as model inputs. Inference uses predicted associations exclusively. To recover associations rejected by mutual matching, an unmatched pair (i, j) is reconsidered when its association score exceeds both corresponding dustbin scores, $\min(s_{ij} - s_{i\varnothing},\, s_{ij} - s_{\varnothing j}) \geq 0$. Valid candidates are ranked by this margin and greedily selected under one-to-one matching. Recovery can reconnect fragmented tracklets across adjacent keyframes, but cannot recover unobserved objects or bridge gaps without an adjacent candidate.

4.3 Spatial-Temporal Reasoning

Given the object tracklets from Section 4.2, IBP contextualizes the compositional representations using a temporal Transformer:

$$H_i = \mathrm{LN}(\mathrm{Transformer}_t(Z_i + E_{\mathrm{pos}}, M_i)) = [h_{i,0}, \ldots, h_{i,4}], \tag{9}$$

where E_{pos} denotes temporal positional embeddings and M_i masks missing observations. Temporal consistency across valid adjacent observations is encouraged by the following objective:

$$\mathcal{L}_{\mathrm{cons}} = \mathrm{mean}_{(i,t)\in\mathcal{V}}[1 - \cos(h_{i,t}, h_{i,t+1})]. \tag{10}$$

A cosine consistency loss serves as an auxiliary regularizer to preserve object level semantic consistency across adjacent valid observations, while allowing the temporal Transformer to model their evolving context. For relational reasoning, the refined components are summarized as $\hat{P}_i = \frac{1}{K}\sum_k P_{ik}$ and combined pairwise as

$$p_{ij} = \mathrm{MLP}[\hat{P}_i \| \hat{P}_j \| (\hat{P}_i - \hat{P}_j) \| (\hat{P}_i \odot \hat{P}_j)]. \tag{11}$$

Spatial prediction combines p_{ij} with z_i, z_j and geometric cues g_{ij}^s, while temporal prediction combines it with $h_{i,m}, h_{j,m}$ and motion cues g_{ij}^t. Separate MLP heads predict spatial and temporal predicates, and node categories are predicted from z_i. Detailed geometric descriptors are provided in the Appendix A.2.

4.4 Learning objectives and training strategy

In addition to the representation objectives $\mathcal{L}_{\text{part}}$, $\mathcal{L}_{\text{obj}}$, and $\mathcal{L}_{\text{cons}}$, IBP uses $\mathcal{L}_{\text{assoc}}$ for cross observation association, $\mathcal{L}_{\text{node}}$ is class weighted cross-entropy for object classification, and $\mathcal{L}_{\text{sp}}$ and $\mathcal{L}_{\text{temp}}$ uses weighted focal binary crossentropy for spatial and temporal relation prediction. The complete objective is

$$\mathcal{L} = \lambda_{\text{part}}\mathcal{L}_{\text{part}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{assoc}}\mathcal{L}_{\text{assoc}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{node}}\mathcal{L}_{\text{node}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}. \quad (12)$$

Training proceeds in three stages: (1) learning compositional representations using the component and object level alignment objectives (2) jointly learning association and scene graph prediction using teacher-forced tracklets constructed from persistent ground-truth identities, and (3) adapting relation reasoning to predicted tracklets while keeping the representation and association modules fixed, reducing the train–inference mismatch. Detailed loss formulations, weighting parameters, and optimization settings are provided in the Appendix A.3.1.

5 Experimental Setup

We evaluate IBP on KITTI-360 (Liao et al., 2022), which provides synchronized RGB-LiDAR observations, 3D semantic and instance annotations, vehicle poses, and persistent object identities suitable for dynamic multimodal scene graph evaluation. We use a sequence level train/validation/test split to prevent temporal leakage and apply identical observations, relation annotations, candidate pairs, and evaluation protocols across all methods. Spatial and temporal relation prediction is evaluated using macro-F1 (M-F1), mean Average Precision (mAP), and Recall@K (R@K). For predicted tracklets, we additionally report association accuracy (As. Acc.) and tracklet coverage (T.C.).

We compare against DSG-DETR (Chen et al., 2025), TEMPURA (Nag et al., 2023), and UNO (Le et al., 2026) for dynamic scene-graph and temporal representation learning, and Scene-LLM (Fu et al., 2025) for multimodal 3D representation and semantic alignment. Since these methods were developed under different modalities and settings, we adapt them to the common KITTI-360 inputs and prediction targets.

We consider two settings: (1) *teacher-forced PredCLS*, where all methods receive ground truth object pairs and five step tracklets, isolating relation prediction from association errors, and (2) *predicted tracklets*, where IBP constructs crossframe associations and tracklets, evaluating the complete pipeline. Dataset statistics, data splits and annotation details, baseline adaptations, implementation settings, computational costs, and additional evaluation metrics are provided in the Appendix A.1.3. IBP contains 16.45M trainable parameters (excluding frozen pretrained encoders). Detailed training and inference costs are reported in the Appendix A.3.3.

6 Results

6.1 Controlled Baseline Comparison

Table 1 compares the methods under teacher-forced PredCLS, isolating relation prediction from association errors. IBP achieves the strongest temporal performance, with M-F1/mAP of 94.14/98.95 and R@20/R@50 of 82.05/98.92. Compared with DSG-DETR, the strongest competing temporal baseline in M-F1, IBP improves temporal M-F1 by 2.95% and mAP by 1.38%. Spatial classification remains stronger for DSG-DETR and TEMPURA, while spatial retrieval is nearly identical among the three methods (R@20 $\approx$29.1 and R@50 $\approx$54.7). Thus, the primary advantage of IBP lies in temporal relational reasoning rather than frame level spatial prediction.

6.2 Evaluation under predicted tracklets

We next evaluate the complete IBP pipeline using model predicted associations and tracklets rather than groundtruth identities. Across three random seeds (Table 2), IBP obtains 98.76 ± 0.08 association accuracy and 84.88 ± 0.58 tracklet coverage. Despite association and tracklet errors, it achieves temporal M-F1/mAP of 84.93 ± 0.59/85.95 ± 0.51 and combined M-F1/mAP of 90.50 ± 0.23/93.69 ± 0.22. The low variance across seeds further indicates stable end-to-end behavior.

Table 1: Comparison of relation prediction performance (%) under the teacher-forced PredCLS setting. **Bold** and underlined values indicate the highest and second-highest performance, respectively.

Method	Temporal Relations				Spatial Relations			
	M-F1	mAP	R@20	R@50	M-F1	mAP	R@20	R@50
Our Model	**94.14**	**98.95**	**82.05**	**98.92**	92.45	98.29	29.05	54.74
DSG-DETR	91.19	97.57	81.24	98.34	94.59	99.21	29.06	54.73
TEMPURA	90.83	98.52	81.90	98.67	**96.12**	**99.49**	29.07	**54.78**
Scene-LLM	47.76	44.42	57.38	83.43	52.24	55.66	20.94	39.83
UNO	45.11	40.41	42.35	82.43	20.22	50.08	**30.01**	33.59

Table 2: Performance (%) across three random seeds. **Bold** marks the best seed per metric, underline marks the second best.

Seed	Spatial						Temporal						Association		Combined	
	M-F1	mAP	R@20	mR@20	R@50	mR@50	M-F1	mAP	R@20	mR@20	R@50	mR@50	As. Acc.	T.C.	M-F1	mAP
42	93.76	98.31	29.05	29.65	**54.76**	**55.59**	**85.57**	**86.27**	67.92	66.06	82.68	82.96	**98.82**	84.36	**90.69**	93.80
43	93.75	98.28	**29.06**	**29.83**	54.75	55.45	84.40	85.36	69.10	67.06	83.56	83.80	98.67	84.78	90.24	93.44
44	**94.03**	**98.41**	29.05	29.46	54.74	55.02	84.81	86.22	**69.65**	**68.51**	**84.26**	**84.95**	98.79	**85.50**	90.57	**93.84**
Mean	93.85	98.33	29.05	29.65	54.75	55.35	84.93	85.95	68.89	67.21	83.50	83.90	98.76	84.88	90.50	93.69
SD	0.16	0.07	0.01	0.19	0.01	0.30	0.59	0.51	0.88	1.23	0.79	1.00	0.08	0.58	0.23	0.22

6.3 Component and Temporal Ablations

Table 4 evaluates the contributions of compositional and temporal components. Removing component level representation and alignment has little effect on spatial prediction, but reduces temporal M-F1/mAP from 85.57/86.27 to 81.04/78.19 and tracklet coverage from 84.36 to 78.91, indicating that compositional representations primarily benefit temporal reasoning and cross observation continuity. Temporal modeling has a larger effect. Removing temporal context (T.Co.), motion geometry (M.G.), and consistency learning reduces temporal M-F1/mAP to 50.08/49.58. Retaining only the temporal Transformer (T.T.) improves these to 52.50/55.76, but remains substantially below the complete model, demonstrating the complementary roles of compositional and temporal modeling. We additionally vary the number of latent components K, the corresponding results are reported in Appendix A.5. We use $K = 8$ in the final model, which provides the strongest temporal retrieval performance while remaining competitive on the remaining metrics.

6.4 Robustness to Incomplete Observations

We further evaluate robustness using node classification accuracy under progressive RGB and LiDAR degradation in (Table 3). Compared with the whole-object ablation, IBP generally preserves higher accuracy, with its advantage increasing as degradation becomes more severe. At 60% degradation, IBP improves accuracy by 2.22, 3.75, and 2.68% under RGB-only, LiDAR-only, and joint degradation, respectively. Under joint degradation, the whole-object representation is slightly better at 20% (+1.22%), but IBP becomes stronger at 40% and 60%. These results indicate that compositional representations better preserve discriminative object information under substantially incomplete observations.

6.5 Learned Component Alignment

Figures 2 and 3 examines whether the shared queries and alignment objective establish crossmodal component correspondence. Matched RGB-LiDAR components exhibit substantially higher cosine similarity than non-matched components (0.580 vs. 0.186), producing a clear diagonal structure with a similarity margin of 0.394. This provides direct evidence that corresponding component indices learn compatible cross-modal representations rather than independent modality specific decompositions. Together with the ablation results, these observations indicate that the learned components capture cross modally consistent and task relevant object evidence without requiring part level supervision.

Table 3: Node-classification accuracy (%) under progressive observation degradation. Each entry reports **IBP** / Whole-object.

Degradation	20%	40%	60%
RGB	**76.78**/76.70	**71.95**/71.42	**64.83**/62.61
LiDAR	**79.39**/78.16	**77.01**/75.02	**74.94**/71.19
RGB + LiDAR	74.41/**75.63**	**64.98**/63.68	**49.12**/46.44

Table 4: Ablation study (performance in %) of the proposed framework. A checkmark (✓) indicates that the corresponding component is enabled, while a cross (×) indicates its removal.

Configuration	Parts	T.T.	M.G. T.Co.	Spatial		Temporal		Association		Combined	
				M-F1	mAP	M-F1	mAP	As. Acc.	T.C.	M-F1	mAP
Full Model	✓	✓	✓	93.76	98.31	**85.57**	**86.27**	**98.82**	84.36	**90.69**	**93.80**
Parts / Alignment	×	✓	✓	**93.78**	**98.46**	81.04	78.19	98.50	78.91	89.01	90.86
Temp. Information	✓	×	×	90.24	96.98	50.08	49.58	98.16	**85.99**	75.18	79.20
Temp. Components	✓	✓	×	92.53	97.89	52.50	55.76	98.08	84.22	77.15	82.09

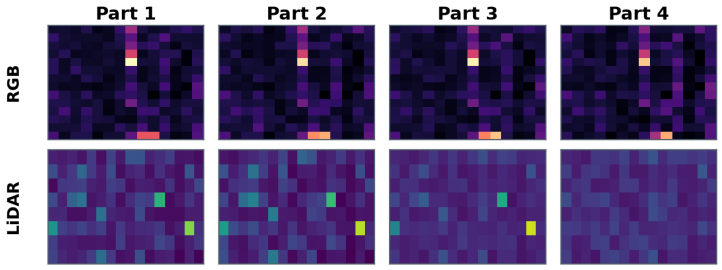


Figure 2: Learned RGB-LiDAR latent parts.

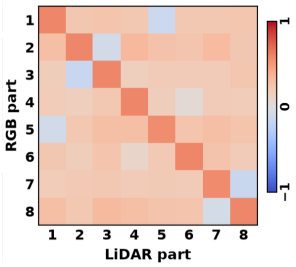


Figure 3: Cosine similarity.

7 Conclusion and Future Work

We presented Integration by Parts (IBP), a compositional multimodal framework for dynamic 3D scene graph generation that decomposes RGB and LiDAR observations into latent components, aligns crossmodal evidence, and recomposes it with global multimodal context for temporal association and relational reasoning. Experiments on KITTI-360 show that the learned components capture crossmodally consistent and task relevant evidence without part level supervision, with compositional representations particularly benefiting temporal reasoning and tracklet construction. Predicted tracklet and experiments on incomplete data further demonstrate effectiveness under association errors and moderately incomplete observations. While the present study considers a fixed component configuration and relation vocabulary on KITTI-360, future work will investigate adaptive component discovery, broader multimodal datasets and relations, robustness to missing or corrupted modalities, and longer-term temporal reasoning.

References

Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5664–5673, 2019.

Nikhil Behari, Aaron Young, Siddharth Somasundaram, Tzofi Klinghoffer, Akshat Dave, and Ramesh Raskar. Blurred lidar for sharper 3d: Robust handheld 3d scanning with diffuse lidar and rgb. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26954–26964, June 2025.

Danilo Cavaliere, Vincenzo Loia, Alessia Saggese, Sabrina Senatore, and Mario Vento. Semantically enhanced uavs to increase the aerial scene understanding. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(3):555–567, 2019. doi: 10.1109/TSMC.2017.2757462.

Lianggangxu Chen, Xuejiao Wang, Jiale Lu, Shaohui Lin, Changbo Wang, and Gaoqi He. Clip-driven open-vocabulary 3d scene graph generation via cross-modality contrastive learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27863–27873. IEEE, 2024.

Mu Chen, Liulei Li, Wenguan Wang, and Yi Yang. Diffvsgg: Diffusion-driven online video scene graph generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29161–29172, 2025. doi: 10.1109/CVPR52734.2025.02715.

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.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

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 (WACV)*, pp. 5130–5139, January 2023.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual reasoning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2195–2206, 2025. doi: 10.1109/WACV61041.2025.00220.

Elias Greve, Martin Büchner, Niclas Vödisch, Wolfram Burgard, and Abhinav Valada. Collaborative dynamic 3d scene graphs for automated driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11118–11124. IEEE, 2024.

KunHo Heo, GiHyun Kim, SuYeon Kim, and MyeongAh Cho. Object-centric representation learning for enhanced 3d semantic scene graph prediction. *Advances in Neural Information Processing Systems*, 38:66654–66685, 2026.

Xun Huang, Shijia Zhao, Yunxiang Wang, Xin Lu, Wanfa Zhang, Rongsheng Qu, Weixin Li, Yunhong Wang, and Chenglu Wen. Msgnav: Unleashing the power of multi-modal 3d scene graph for zero-shot embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37154–37163, 2026.

Ue-Hwan Kim, Jin-Man Park, Taek-Jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE transactions on cybernetics*, 50(12):4921–4933, 2019.

Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14183–14193. IEEE, 2024.

Lingdong Kong, Xiang Xu, Jiawei Ren, Wenwei Zhang, Liang Pan, Kai Chen, Wei Tsang Ooi, and Ziwei Liu. Multi-modal data-efficient 3d scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3748–3765, 2025.

Huy Le, Nhat Chung, Tung Kieu, Jingkang Yang, and Ngan Le. Uno: Unifying one-stage video scene graph generation via object-centric visual representation learning. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2797–2807, 2026. doi: 10.1109/WACV61042.2026.00274.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In *International Conference on Learning Representations*, volume 2025, pp. 29118–29134, 2025.

Qingmei Li, Yang Zhang, Zurong Mai, Yuhang Chen, Henglian Huang, Jiarui Zhang, Zhiwei Zhang, Yibin Wen, Weijia Li, Haohuan Fu, et al. Can large multimodal models understand agricultural scenes? benchmarking with agromind. *Advances in Neural Information Processing Systems*, 38, 2026.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *International Conference on Learning Representations*, volume 2025, pp. 42942–42959, 2025.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

Matias Mendieta and Hisham Tabkhi. Carpe posterum: A convolutional approach for real-time pedestrian path prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11956–11964, 2021.

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

Rohith Peddi, Saurabh Saurabh, Ayush Abhay Shrivastava, Parag Singla, and Vibhav Gogate. Towards unbiased and robust spatio-temporal scene graph generation and anticipation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8648–8657. IEEE, 2025.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Scarlett Raine and Tobias Fischer. Ai-driven marine robotics: emerging trends in underwater perception and ecosystem monitoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 40981–40989, 2026.

Juneyoung Ro, Namwoo Kim, and Yoonjin Yoon. How well do vision-language models understand cities? a comparative study on spatial reasoning from street-view images. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6535–6544. IEEE, 2025.

Dennis Rotondi, Francesco Argenziano, Sebastian Koch, Nathan Hughes, Martin Buechner, Johanna Wald, Lukas Rosenberger Schmid, Daniele Nardi, Abhinav Valada, Liam Paull, et al. 3d scene graphs: Open challenges and future directions. *arXiv preprint arXiv:2606.19383*, 2026.

Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Sgaligner: 3d scene alignment with scene graphs. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21870–21880. IEEE, 2023.

Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Crossover: 3d scene cross-modal alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8985–8994. IEEE, 2025.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946. IEEE, 2020.

Binod Singh, Sayan Deb Sarkar, and Iro Armeni. Sgaligner++: Cross-modal language-aided 3d scene graph alignment. *arXiv preprint arXiv:2509.20401*, 2025.

Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu. Oed: Towards one-stage end-to-end dynamic scene graph generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27938–27947. IEEE, 2024a.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, et al. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. *Advances in Neural Information Processing Systems*, 38, 2026.

Shuo Wang, Yongcai Wang, Zhimin Xu, Yongyu Guo, Wanting Li, Zhe Huang, Xuewei Bai, and Deying Li. Gslamot: A tracklet and query graph-based simultaneous locating, mapping, and multiple object tracking system. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7239–7248, 2024b.

Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, and Lu Sheng. Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21560–21569. IEEE, 2023.

Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14158–14168. IEEE, 2025.

Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7511–7521. IEEE, 2021.

Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Incremental 3d semantic scene graph prediction from rgb sequences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5064–5074, 2023.

Jianyun Xu, Song Wang, Ziqian Ni, Chunyong Hu, Sheng Yang, Jianke Zhu, and Qiang Li. Sam4d: Segment anything in camera and lidar streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28535–28545. IEEE, 2025.

Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clipˆ 2: Contrastive language-image-point pretraining from real-world point cloud data. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15244–15253. IEEE, 2023.

Yanan Zhang, Jiaxin Chen, and Di Huang. Cat-det: Contrastively augmented transformer for multimodal 3d object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 898–907. IEEE, 2022.

A Appendix

A.1 Dataset, Annotation, and Evaluation Protocol

A.1.1 KITTI-360 Dataset and Data Split

We evaluate on KITTI-360 (Liao et al., 2022), a large-scale multimodal urban driving dataset collected in Karlsruhe, Germany. The dataset contains approximately 73.7 km of driving data across 11 continuous sequences, with more than 320K images and 100K LiDAR scans. It provides synchronized perspective and fisheye cameras, Velodyne HDL-64E LiDAR, vehicle poses, calibration parameters, and dense 2D and 3D semantic and instance annotations, including persistent instance identities across observations. These properties enable joint evaluation of frame level relations, cross observation object association, and tracklet level temporal reasoning.

We use sequences 0000, 0002, 0003, 0004, 0005, and 0010 for training, sequence 0006 for validation, and sequences 0007 and 0009 for testing. The split is performed strictly at the complete driving sequence level, with no random frame, scene, or tracklet level partitioning, thereby preventing temporal leakage across the training, validation, and test sets. The validation sequence is used

for hyperparameter selection and checkpoint selection, while the test sequences are reserved for final evaluation.

The raw sensor streams are recorded at approximately 10 FPS. Each driving sequence is partitioned into non-overlapping clips of 200 raw frames, corresponding to approximately 20 s of driving. Every fifth raw frame is retained as a key sample, resulting in an effective sampling rate of approximately 2 FPS. Five consecutive key samples are grouped into a temporal window spanning approximately 2 s, with windows constrained not to cross clip or driving sequence boundaries.

Our experiments use the left-camera RGB images, Velodyne LiDAR measurements, sensor calibration parameters, vehicle poses, semantic and instance annotations, and 3D bounding-box annotations provided by KITTI-360. Persistent KITTI-360 instance identities are used only to construct supervision signals, association targets, tracklets, and evaluation labels. They are never supplied to the model as input features. All methods are evaluated using the same set of 26 object classes.

Table 5 summarizes the statistics of the processed KITTI-360 data used in our experiments, including the number of key observations, object instances, association links, and tracklets across the training, validation, and test splits.

Table 5: Statistics of the processed KITTI-360 data used in our experiments.

Statistic	**Train**	**Val.**	**Test**
Sequences	6	1	2
Key samples with objects	8,430	1,914	3,330
Object observations (feature rows)	172,700	30,765	59,434
Unique object identities	10,591	2,107	3,969
Category vocabulary	26	26	26
Distinct five-sample GT tracklets	8,410	1,102	2,815

A.1.2 Scene Graph Construction and Relation Annotation

We construct dynamic scene graph annotations from the synchronized KITTI-360 object observations and persistent instance identities. The relation vocabulary contains five spatial and three temporal predicates.

Spatial predicates For an ordered object pair (i, j) within an observation, we consider *left of*, *in front of*, *near*, *overlapping*, and *occluding*. These predicates characterize the relative spatial configuration of the two objects.

Temporal predicates For an object pair represented by their tracklet level observations, we consider *approaching*, *moving away*, and *same motion direction*. Unlike spatial predicates, these relations depend on the evolution of the object pair over the temporal window rather than on an individual observation.

The relation prediction task is multilabel, where more than one predicate may hold for the same object pair. If none of the defined predicates holds, the target is represented by an all-zero multilabel vector and is denoted as *no relation*, and is not treated as an additional predicate class. The *crossing path* relation is retained only for exploratory analysis and is excluded from supervised training and the primary evaluation because of its insufficient number of independent examples.

For directional predicates, inverse relations are defined canonically as

$$\text{right of}(A, B) = \text{left of}(B, A), \tag{13}$$

$$\text{behind}(A, B) = \text{in front of}(B, A), \tag{14}$$

$$\text{occluded by}(A, B) = \text{occluding}(B, A). \tag{15}$$

All compared methods use the same relation vocabulary, prepared object observations, relation identifiers, positive relation annotations, and no-relation candidate pairs. No baseline-specific annotation or separate negative-sample set is constructed.

Table 6 reports the distribution of spatial and temporal relation instances across the dataset splits.

Table 6: Positive spatial and temporal predicate labels in the full-multimodal KITTI-360 relation data.

Relation	Train	Val.	Test
Spatial relations			
Left of	162,293	24,244	55,516
In front of	168,203	25,150	56,916
Near	95,814	17,314	29,706
Overlapping	20,876	4,198	5,526
Occluding	6,853	1,796	1,839
Temporal relations			
Approaching	3,901	439	1,689
Moving away	6,640	530	1,666
Same motion direction	1,454	52	622

A.1.3 Evaluation Protocol and Metrics

We evaluate spatial and temporal relation prediction under a common multilabel evaluation protocol. For each predicate, we compute precision, recall, F1-score (F1), and average precision (AP). Aggregate performance is reported using M-F1 and mAP, which assign equal importance to individual predicate classes. We additionally report R@K and mean Recall@K (mR@K) where applicable.

For association-based evaluation, we report *association accuracy* and *tracklet coverage*. Association accuracy measures the percentage of objects that are correctly matched across two adjacent keyframes, including objects correctly assigned to the no-match (dustbin) option. For association group g, let $n_s^{(g)}$ and $n_t^{(g)}$ denote the numbers of source and target objects, and let $\hat{y}$ and y denote their predicted and ground-truth assignments, respectively. We compute

$$\mathrm{Acc}_{\mathrm{assoc}} = \frac{\sum_g \left[\sum_i \mathbb{I}(\hat{y}_{s,i}^{(g)} = y_{s,i}^{(g)}) + \sum_j \mathbb{I}(\hat{y}_{t,j}^{(g)} = y_{t,j}^{(g)})\right]}{\sum_g (n_s^{(g)} + n_t^{(g)})} \times 100. \tag{16}$$

Predicted assignments are obtained from the row and columnwise maxima of the augmented association matrix. Thus, both correct object matches and correct no-match decisions contribute to the accuracy.

Tracklet coverage measures whether the predicted tracklets provide sufficient temporal observations for evaluating a temporal relation. For candidate r over a five keyframe window, let p_r and q_r be the predicted track IDs of the two objects, and let $m_{r,k} = 1$ when both tracklets contain an observation at keyframe k. The candidate is covered if the two objects have valid, distinct track IDs, are both present in the middle keyframe, and are jointly observed in at least three of the five keyframes:

$$a_r = \mathbb{I}\left(p_r \geq 0,\ q_r \geq 0,\ p_r \neq q_r,\ m_{r,3} = 1,\ \sum_{k=1}^{5} m_{r,k} \geq 3\right). \tag{17}$$

Tracklet coverage is then

$$\mathrm{Coverage}_{\mathrm{tracklet}} = \frac{1}{N_{\mathrm{temp}}} \sum_{r=1}^{N_{\mathrm{temp}}} a_r \times 100, \tag{18}$$

i.e., the percentage of temporal relation candidates for which the predicted tracklets contain sufficient observations for temporal reasoning. For the complete pipeline, we additionally report combined spatial-temporal M-F1 and mAP.

Two evaluation settings are considered. In the *teacher-forced PredCLS* setting, methods receive the same groundtruth object pairs and five step tracklets, thereby isolating relation prediction from errors introduced by object association and tracklet construction. In the *predicted tracklet* setting, IBP constructs cross observation associations and tracklets using its learned association module, providing an end-to-end evaluation of the proposed framework.

For reproducible post-hoc evaluation, we retain the raw, unmodified logits for every prediction together with the corresponding relation identifier, sequence and sample information, subject and object identities, task type, ordered predicate names, object class logits where applicable, and model/checkpoint information. Sigmoid activation is applied to the stored relation logits during evaluation, followed by a fixed decision threshold of 0.5. The same preprocessing, relation targets, candidate pairs, and evaluation settings are used across all compared methods.

A.2 Detailed Model Architecture and Implementation

This section provides additional architectural and implementation details of Integration by Parts (IBP). Following the main paper, the framework consists of five conceptual operations: *decompose*, *align*, *recompose*, *associate*, and *reason*. We first describe the modality specific feature extraction and compositional representation modules, followed by cross observation association and spatial-temporal relation prediction.

A.2.1 Modality Preprocessing and Feature Encoders

For each annotated object, we construct synchronized RGB, LiDAR, and textual representations from the corresponding KITTI-360 observation. The RGB modality is represented using the image region associated with the object annotation. Each object crop is processed by a frozen CLIP ViT-B/16 encoder, from which we retain the 14×14 patch token representation $H_i^R \in \mathbb{R}^{196\times768}$. Retaining the patch tokens rather than immediately pooling them preserves localized appearance evidence for subsequent component discovery.

For the LiDAR modality, points associated with the corresponding 3D object are extracted from the synchronized point cloud. Each object is represented using 128 points with XYZ coordinates and surface-normal information. A pretrained PointNet encoder processes these points to obtain local geometric features $H_i^L \in \mathbb{R}^{128\times512}$. Global max pooling over the PointNet features additionally produces the object level geometric descriptor $a_i^L \in \mathbb{R}^{512}$.

To obtain complementary semantic information, an object centric caption is generated from the RGB observation using a frozen BLIP model. The generated caption is subsequently encoded using the CLIP text encoder, producing $x_i^T \in \mathbb{R}^{512}$. Since this textual representation is derived from the RGB observation rather than from an independent sensing modality, we use it as an object-level semantic reference and do not impose component level correspondence between text and the RGB/LiDAR components.

The modality specific representations are projected to a common embedding dimension $d = 384$ before multimodal processing. The RGB and LiDAR token sequences are augmented with their corresponding positional and modality embeddings before being supplied to the shared query component discovery module.

A.2.2 Latent Component Discovery and Cross-Modal Alignment

IBP uses a shared bank of K learnable queries $Q = \{q_k\}_{k=1}^K$ to organize local RGB and LiDAR evidence into a common set of latent component indices. For modality $m \in \{R, L\}$, the projected token sequence $\widetilde{X}_i^m$ is queried through multi-head cross-attention,

$$P_i^m = \mathrm{MHA}(Q, \widetilde{X}_i^m, \widetilde{X}_i^m) = [p_{i1}^m, \ldots, p_{iK}^m], \tag{19}$$

where $P_i^m \in \mathbb{R}^{K\times d}$. We use $K = 8$ components and $d = 384$ in the primary configuration.

The same query bank is applied to both modalities. Thus, query q_k provides a common latent index for the k^{th} RGB and LiDAR components, while the component representations themselves remain modality specific. Importantly, shared queries alone do not guarantee that p_{ik}^R and p_{ik}^L encode compatible evidence. We therefore explicitly encourage correspondence through a component level contrastive objective.

Before computing contrastive similarity, the representations are ℓ_2-normalized. For two sets of representations $A = \{a_n\}_{n=1}^N$ and $B = \{b_n\}_{n=1}^N$, we define the directional InfoNCE loss as

$$\mathcal{L}_{\mathrm{NCE}}(A \to B) = -\frac{1}{N}\sum_{n=1}^{N} \log \frac{\exp(\mathrm{sim}(a_n, b_n)/\tau)}{\sum_{m=1}^{N} \exp(\mathrm{sim}(a_n, b_m)/\tau)}, \tag{20}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature parameter.

For component level alignment, corresponding RGB and LiDAR components of the same object and component index form positive pairs, while non-corresponding components form negative pairs. The symmetric component alignment objective is therefore

$$\mathcal{L}_{\mathrm{part}} = \frac{1}{2}\left[\mathcal{L}_{\mathrm{NCE}}(P^R \to P^L) + \mathcal{L}_{\mathrm{NCE}}(P^L \to P^R)\right]. \tag{21}$$

In parallel, the complete component sets are summarized using attention pooling,

$$\alpha_{ik}^m = \frac{\exp(w^\top p_{ik}^m)}{\sum_{l=1}^K \exp(w^\top p_{il}^m)}, \qquad o_i^m = \sum_{k=1}^K \alpha_{ik}^m p_{ik}^m, \tag{22}$$

where $m \in \{R, L\}$. The resulting modality level representations are aligned with the text embedding using the same InfoNCE formulation:

$$\mathcal{L}_{\mathrm{obj}} = \frac{1}{2}\left[\mathcal{L}_{\mathrm{NCE}}(o^R \to x^T) + \mathcal{L}_{\mathrm{NCE}}(o^L \to x^T)\right]. \tag{23}$$

Thus, the same contrastive formulation is used at two different levels: $\mathcal{L}_{\mathrm{part}}$ establishes fine-grained RGB-LiDAR correspondence between latent component indices, whereas $\mathcal{L}_{\mathrm{obj}}$ uses the caption derived text embedding as an object level semantic reference. No semantic part labels or predefined physical part correspondences are used during training.

A.2.3 Compositional Object Representation

Following component alignment, IBP preserves both holistic multimodal context and fine-grained component level evidence. For global fusion, the attention pooled RGB and LiDAR representations o_i^R and o_i^L, the global PointNet feature a_i^L, and the textual representation x_i^T are projected to the common $d = 384$ dimensional space and concatenated with a learnable fusion token. A multimodal Transformer contextualizes these tokens, and the output associated with the fusion token forms the global object representation z_i^g.

In parallel, corresponding RGB and LiDAR components are integrated by averaging their aligned representations and processing the resulting K components with a component level Transformer. Attention pooling over the refined components produces the component representation z_i^p. This branch preserves fine grained evidence that may otherwise be suppressed by direct object level fusion.

The two representations are combined using the gated residual integration described in equation 6 of the main paper. The gate is conditioned jointly on z_i^g and z_i^p and controls the contribution of component level evidence to the global representation on a per feature basis. The resulting representation z_i is used consistently for object classification, cross observation association, and subsequent relational reasoning. For our study, we use 8 attention heads for component attention, 2 layers with 8 heads for the global Transformer, 1 layer with 8 heads for the component Transformer, and 2 layers with 8 heads for the temporal Transformer. All Transformer modules use an FFN dimension of 1536 and a dropout rate of 0.1.

A.2.4 Object Association and Tracklet Construction

For objects observed at consecutive key samples t and $t+1$, IBP predicts whether two observations correspond to the same physical object. For every candidate pair (i, j), the association representation combines the two compositional object embeddings, their difference, their element-wise interaction, and an explicit geometric descriptor, given as

$$a_{ij} = [z_i^{(t)} \| z_j^{(t+1)} \| (z_i^{(t)} - z_j^{(t+1)}) \| (z_i^{(t)} \odot z_j^{(t+1)}) \| g_{ij}^a]. \tag{24}$$

An MLP maps a_{ij} to the association score s_{ij}.

The geometric descriptor $g_{ij}^a \in \mathbb{R}^{16}$ provides explicit geometric and modality-related cues between a pair of observations. It consists of the 3D centre offset $(\Delta x, \Delta y, \Delta z)$, the distance between the

two centres, normalized centre offsets along the three spatial axes, and the differences between their bounding-box dimensions. The descriptor also includes the logarithmic ratios of the corresponding box dimensions and three binary variables indicating the joint availability of RGB, LiDAR, and textual information. Hence, the descriptor contains 16 input-derived quantities covering centre displacement, centre distance, displacement direction, box-size variation, relative box-size change, and modality availability. These quantities are computed directly from the observations and are not learned coefficients. For pairs selected from observations with all modalities present, the three modality indicators are set to one.

To accommodate objects entering or leaving the observation, the pairwise score matrix is augmented with source and target dustbin assignments representing unmatched objects. The augmented matrix is normalized using log-domain Sinkhorn iterations to obtain approximately one-to-one assignments. Cross observation matches are retained only when the source-to-target and target-to-source assignments are mutually consistent. Objects assigned to a dustbin remain unmatched.

The resulting pairwise associations are chained across five consecutive key samples to form object centric tracklets. A binary mask records whether an object is observed at each temporal position. During training, persistent KITTI-360 identities are used only to construct association supervision and teacher-forced tracklets and are never provided as model input. At inference time, tracklets are constructed entirely from predicted associations.

Mutual matching can occasionally reject an otherwise plausible association. For an unmatched candidate (i, j), IBP therefore compares its association score with the corresponding source and target dustbin scores. The candidate is reconsidered when

$$\min(s_{ij} - s_{i\varnothing}, s_{ij} - s_{\varnothing j}) \geq 0. \tag{25}$$

Eligible candidates are ranked by this margin and greedily accepted subject to one-to-one matching. This recovery procedure reconnects fragmented adjacent associations but does not hallucinate missing object observations and cannot bridge a temporal gap when no adjacent candidate exists.

A.2.5 Temporal Contextualization and Relation Prediction

For each predicted or teacher-forced object tracklet, the five compositional object representations are processed by a temporal Transformer after adding temporal positional embeddings. Missing observations are excluded through the corresponding validity mask. The resulting contextualized representations $h_{i,t}$ encode the evolution of an object over the temporal window while retaining its object-level semantic information.

To preserve component information during relation prediction, the refined components are summarized as

$$\hat{P}_i = \frac{1}{K} \sum_{k=1}^{K} P_{ik}, \tag{26}$$

and an ordered object-pair representation is constructed as

$$p_{ij} = \mathrm{MLP}\left([\hat{P}_i \| \hat{P}_j \| (\hat{P}_i - \hat{P}_j) \| (\hat{P}_i \odot \hat{P}_j)] \right). \tag{27}$$

This provides component informed pairwise evidence in addition to the global object representations.

Spatial relation head For objects i and j within the same observation, the spatial relation head combines the compositional object representations, component informed pairwise representation, and within observation geometric descriptor, given as,

$$u^s_{ij} = [z_i \| z_j \| p_{ij} \| g^s_{ij}]. \tag{28}$$

A task-specific MLP maps u^s_{ij} to logits over the five spatial predicates.

Temporal relation head For an object pair (i, j), a temporal Transformer contextualizes each object's five-keyframe tracklet. Let $h_{i,m}$ and $h_{j,m}$ denote the resulting representations at the middle position m. The temporal relation head receives

$$u^t_{ij} = \left[h_{i,m} \| h_{j,m} \| (h_{i,m} - h_{j,m}) \| (h_{i,m} \odot h_{j,m}) \| p_{ij} \| g^t_{ij} \right], \tag{29}$$

where p_{ij} is the component-level pair interaction, g_{ij}^t is the temporal geometric descriptor, $\|$ denotes concatenation, and $\odot$ denotes elementwise multiplication. A separate MLP produces three logits corresponding to *approaching*, *moving away*, and *same motion direction*.

Object categories are predicted independently from z_i using a lightweight classification head over the 26 object classes.

A.2.6 Architecture Summary

Table 7 summarizes the principal architectural settings used in the experiments.

Table 7: Summary of the principal architectural settings of IBP.

Component	**Configuration**
RGB encoder	Frozen CLIP ViT-B/16
RGB local tokens	196×768
LiDAR encoder	Pretrained PointNet
LiDAR points per object	128
LiDAR local tokens	128×512
Text generation	Frozen BLIP
Text encoder	CLIP text encoder
Text embedding	512
Common embedding dimension d	384
Number of latent components K	8
Temporal window	5 key samples ($\sim$2 s)
Component attention heads	8
Global Transformer layers/heads	2/8
Component Transformer layers/heads	1/8
Temporal Transformer layers/heads	2/8
Transformer FFN dimension	1536
Transformer dropout	0.1
Association MLP	$1552 \to 384 \to 1$
Spatial relation MLP	$1936 \to 768 \to 384 \to 5$
Temporal relation MLP	$1944 \to 768 \to 384 \to 3$
Total trainable and non-frozen parameters	16,452,631

A.3 Learning Objectives and Training Details

A.3.1 Learning Objectives

IBP is optimized using objectives corresponding to representation learning, association, temporal consistency, and scene-graph prediction. The component level and object level contrastive objectives $\mathcal{L}_{\text{part}}$ and $\mathcal{L}_{\text{obj}}$ are defined in Appendix A.2.2. Cross observation association is optimized using $\mathcal{L}_{\text{assoc}}$, while temporal consistency is encouraged through

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\mathcal{V}|} \sum_{(i,t)\in\mathcal{V}} \left[1 - \cos(h_{i,t}, h_{i,t+1})\right], \tag{30}$$

where $\mathcal{V}$ contains valid adjacent observations of the same object.

Object classification uses class-weighted cross-entropy $\mathcal{L}_{\text{node}}$, while spatial and temporal relation prediction use weighted focal binary cross-entropy, denoted by $\mathcal{L}_{\text{sp}}$ and $\mathcal{L}_{\text{temp}}$, respectively. The complete objective is

$$\mathcal{L} = \lambda_{\text{part}}\mathcal{L}_{\text{part}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{assoc}}\mathcal{L}_{\text{assoc}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{node}}\mathcal{L}_{\text{node}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}. \tag{31}$$

A.3.2 Three-Stage Training Strategy

Training proceeds in three stages. In *Stage 1*, the compositional representation is learned using the component and object level alignment objectives, establishing RGB-LiDAR component correspondence and object-level semantic alignment. In **Stage 2**, association and scene graph prediction are

learned using teacher-forced tracklets constructed from persistent groundtruth object identities. This allows the relation modules to learn spatial and temporal reasoning without initially being affected by association errors. In **Stage 3**, relation reasoning is adapted to predicted tracklets while the representation and association modules are kept fixed, thereby exposing the relation heads to the tracklet noise encountered at inference time and reducing the train-inference mismatch.

Ground-truth persistent identities are used only to construct supervision and teacher-forced tracklets during training; inference relies exclusively on predicted associations.

A.3.3 Computational Complexity and Parameters

The computational cost of the proposed IBP framework was evaluated in terms of training and inference time. The average training time per epoch was 4 min 06 s for Stage 1, 8 min 46 s for Stage 2, and 7 min 40 s for Stage 3. The first epoch of Stage 1 incurred additional overhead, while the remaining epochs averaged approximately 3 min 31 s. Across all three stages, the logged training loop time was 8 h 23 min 38 s. For the held-out sequences 0007 and 0009, evaluation required 3 min 49 s for object encoding, 11 min 51 s for spatial prediction, and 2 min 12 s for temporal prediction, resulting in approximately 17 min 52 s in total. These measurements include evaluation loop data loading but exclude raw feature extraction, tracklet generation, and final metric aggregation. Hence, they should not be interpreted as per-frame latency. The IBP network contains exactly $16{,}452{,}631$ trainable parameters, excluding the frozen pretrained CLIP, PointNet, and BLIP encoders. During the final Stage 3 optimization, 8,034,546 parameters are updated. The overall parameter count is given in Table 7.

A.3.4 Optimization and Hyperparameters

Table 8 summarizes the principal training hyperparameters. Unless otherwise stated, the same configuration is used throughout the reported IBP experiments.

Table 8: Training and optimization settings used for IBP.

Setting	**Value**
Optimizer	AdamW
Learning rate (Stages 1/2/3)	$10^{-4}/10^{-4}/5 \times 10^{-5}$
Stage 2 object-encoder learning rate	10^{-5}
Weight decay	10^{-4}
Batch size (Stages 1/2/3)	$32/16/16$
Training epochs (Stage 1)	40
Training epochs (Stage 2)	30 (1000 steps/epoch)
Training epochs (Stage 3)	10 (1000 steps/epoch)
Contrastive temperature τ	0.07
Focal-loss γ	2.0
λ_{div} (Stages 1–2)	0.1
λ_{part} (Stages 1–2)	1.0
λ_{obj} (Stages 1–2)	1.0
λ_{assoc} (Stage 2)	1.0
λ_{cons} (Stages 2–3)	0.2
λ_{node} (Stages 1–2 / Stage 3)	$1.0/0.5$
λ_{sp} (Stages 2–3)	1.0
λ_{temp} (Stages 2–3)	1.0
Random seed	42

The proposed IBP model was implemented in PyTorch 2.6.0. The principal IBP training runs were conducted on NVIDIA H100 PCIe GPUs, while selected ablations used NVIDIA A100 GPUs.

A.4 Adaptation of the Baselines to KITTI-360 Data

Since the existing KITTI-360 dataset does not provide the dynamic predicates considered in our study, a direct comparison with prior methods is not possible. We therefore adapt the selected

baseline methods, originally developed and evaluated on other datasets, to our extended KITTI-360 setting. The adaptation is performed while preserving the core design of each baseline as closely as possible, with modifications limited to the input representation, label definitions, and other dataset-specific requirements needed for evaluation on our data. The details of these adaptations are described below.

- **DSG-DETR:** DSG-DETR is adapted to KITTI-360 by replacing the original Action Genome detector-based inputs with frozen whole-object RGB, LiDAR, text, and category representations, together with predefined ordered object pairs. The original spatial, contact, and attention prediction heads are replaced by dedicated spatial and temporal relation classifiers for the KITTI-360 predicates. The temporal module retains the original $1{+}3$ transformer structure, while five-step relation sequences are processed using positional encoding and attention masks. This adaptation operates within the PredCLS protocol and avoids introducing detector, ROI, or other Action Genome-specific components.
- **Scene-LLM:** Since no public implementation is available for this method, we reimplemented it following the methodology described in the original paper. For KITTI-360, object representations are obtained by ROI pooling frozen ViT-B features of dimension 768 at the groundtruth bounding box locations, using pre-extracted per object features and therefore avoiding an additional object detection stage. These representations are subsequently processed through the VQ-VAE, spatial clustering, and LLM based reasoning stages described in the original method. The implementation uses separate prediction heads for the five spatial and three temporal predicates, with *crossing_path* omitted from the supervised target set.
- **UNO:** We use the authors' official implementation and introduce a KITTI-360 data loader to map our relation annotations to the input format expected by UNO. The eight predicates considered in our study are assigned to UNO's "contacting" relation head, while the remaining two heads, which are designed for person-object interactions in Action Genome, are retained as inactive placeholders to preserve the original architectural structure. Although UNO normally performs object discovery through learned slot attention, we modify the input interface to use the same groundtruth object instances provided to Scene-LLM, thereby enabling a controlled PredCLS comparison under identical object inputs.
- **TEMPURA:** *Adapted Multimodal TEMPURA-K360 (PredCLS)* modifies only the input representation of TEMPURA for the multimodal KITTI-360 setting, while preserving its spatial-temporal encoder, temporal decoder, memory guided debiasing, and GMM based uncertainty estimation. Instead of the original Faster R-CNN features, union-region representations, 2D box geometry, and GloVe embeddings, each object is represented using pre-extracted CLIP RGB, PointNet LiDAR, and CLIP text features projected into a common space and fused into a whole-object embedding. For each object pair, the two object embeddings, their feature difference, and geometric cues, including relative position, distance, size, IoU, depth, and motion geometry, are combined for relation prediction. The resulting representations are processed by the unchanged TEMPURA architecture using five-frame temporal context, followed by its memory-based training and GMM uncertainty estimation. No IBP-specific part tokens, part interactions, or part-aware relation features are introduced, providing a controlled comparison between multimodal whole-object reasoning and the proposed part-aware representation.

A.5 Additional Experimental Results

Table 9: Performance (%) under increasing RGB and LiDAR occlusion across sequences using teacher-forced PredCLS. Best performance at each occlusion level is shown in **bold**. T-denotes temporal and S-denotes spatial performance.

Occlusion→	0%				20%				40%				60%			
Method↓	T-F1	S-F1	T-mAP	S-mAP	T-F1	S-F1	T-mAP	S-mAP	T-F1	S-F1	T-mAP	S-mAP	T-F1	S-F1	T-mAP	S-mAP
Ours	**93.33**	93.75	**98.58**	98.31	**92.70**	93.58	98.29	98.16	**91.59**	93.42	**98.31**	98.11	89.90	93.17	**97.97**	97.97
DSG-DETR	91.19	94.59	97.57	99.21	90.63	94.65	97.28	99.22	90.86	94.66	97.35	99.22	**90.72**	94.81	97.04	99.23
TEMPURA	90.83	**96.12**	98.52	**99.49**	90.67	**96.07**	**98.32**	**99.47**	89.58	**96.00**	97.84	**99.46**	89.32	**95.99**	97.79	**99.46**

A.5.1 Robustness Comparison under Incomplete Observations

We additionally compare IBP with DSG-DETR and TEMPURA under controlled observation degradation. RGB and LiDAR inputs are progressively degraded at 20%, 40%, and 60% levels, and all methods are evaluated under the same perturbed observations and evaluation protocol. The 0% setting corresponds to the original observations without degradation. Table 9 reports spatial and temporal M-F1 and mAP at each degradation level.

Table 9 shows that IBP maintains strong temporal performance under moderate degradation, achieving the highest temporal F1 at both 20% and 40%. At 60%, DSG-DETR slightly exceeds IBP in temporal F1 (90.72 vs. 89.90), while IBP retains the highest temporal mAP (97.97). Spatial performance remains comparatively stable across degradation levels. Overall, IBP's temporal advantage persists under moderately incomplete observations but diminishes under severe information loss.

A.5.2 Sensitivity to the Number of Latent Components

We study the sensitivity of IBP to the number of latent components K in Table 10. Spatial performance remains largely stable across different values of K, whereas temporal reasoning benefits from a richer component representation. Increasing K from 3 to 5–8 consistently improves temporal and combined performance, indicating that a coarse decomposition provides insufficient representational granularity. No single value dominates all metrics: $K = 7$ achieves the highest temporal M-F1/mAP and combined mAP, while $K = 8$ provides the strongest temporal retrieval performance, particularly at R@50. Overall, the results indicate that IBP is not highly sensitive to a specific choice of K within the 5–8 range, while very small K is less effective for temporal reasoning.

Table 10: Sensitivity to the Number of Latent Components (k) used in the proposed framework (performance in %). Best results are shown in bold, while the second-best results are underlined.

Parts (k)	Spatial					Temporal				Association		Combined	
	M-F1	mAP	R@20	R@50	R@100	M-F1	mAP	R@20	R@50	As. Acc.	T.C.	M-F1	mAP
3	93.50	98.39	**29.06**	**54.75**	**76.81**	83.24	84.03	68.72	81.54	98.57	83.13	89.65	93.01
5	**94.07**	98.29	<u>29.04</u>	**54.75**	<u>76.77</u>	<u>85.76</u>	<u>86.24</u>	69.83	83.58	**98.87**	<u>84.91</u>	**90.96**	<u>93.77</u>
7	93.62	**98.41**	29.05	54.75	76.78	**86.11**	**86.71**	<u>70.20</u>	<u>83.96</u>	98.67	**85.13**	<u>90.80</u>	**94.02**
8	<u>93.76</u>	98.31	<u>29.04</u>	<u>54.74</u>	76.78	85.57	86.27	**73.57**	**91.20**	<u>98.82</u>	84.36	90.69	93.80

A.6 Additional Visualization Results

To further examine the proposed hypothesis, we present additional qualitative results in Figures 4 and 5, which illustrate a fully correct prediction and a representative non-perfect prediction from scene 0007, respectively. In the successful case (frame 2775), IBP correctly recovers all eight ground-truth relations involving a large pole and two cars, achieving precision, recall, and F1 of 1.00. The predicted graph includes directional relations such as *left of* and *in front of*, along with reciprocal *near* relations between the two cars, indicating that the model can jointly represent lateral ordering, depth ordering, and proximity within a densely connected scene graph. The second example (frame 1805) represents a typical imperfect prediction in which all five ground-truth relations are identified, giving a recall of 1.00, while one additional *in front of* relation from 8cf2 to 4599 is incorrectly introduced, resulting in a precision of 0.83 and an F1 score of 0.91. Here, 8cf2 is a small pole that is incorrectly classified as a traffic sign, with the two categories sharing similar visual characteristics due to their thin and distant appearance along the roadside. Their comparable positions and depths may also make the depth-ordering relation less distinctive. These examples indicate that prediction errors are more often associated with additional incorrect edges than with missing relations, particularly around small or visually ambiguous objects; meanwhile, relations involving clearly distinguishable objects remain largely intact. In the visualization, green, red, and grey denote correct, incorrect, and missed relations, respectively, while the RGB bounding boxes represent the projections of the corresponding 3D annotations.

We also examine the resulting representations through a qualitative temporal relation example from IBP-K360 (drive 0007, frames 2650–2670 with stride 5), shown in Figure 6. In this sequence, two vehicles move ahead of the ego vehicle in the same direction, while their separation increases from 17.6,m at t_1 to 25.2,m at t_5, corresponding to a total increase of 7.6,m or approximately

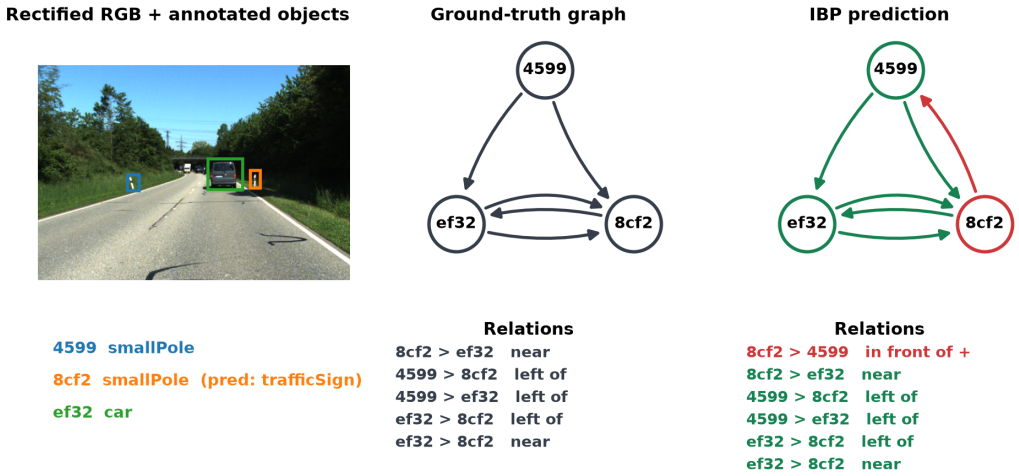


Figure 4: Median (non-perfect) prediction by IBP.

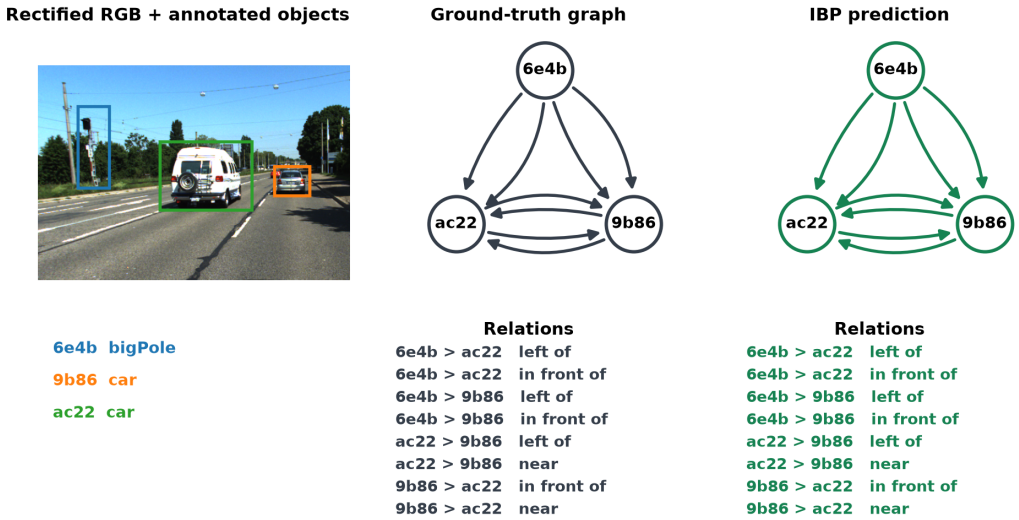


Figure 5: Successful prediction by IBP.

1.9,m per step. Based on this temporal evolution, the model assigns probabilities of 0.99 to *moving away* and *same motion direction*, and 0.00 to *approaching*, correctly matching all three ground-truth predicates and yielding precision, recall, and F1 scores of 1.00. The predictions are well separated from the 0.5 decision boundary, indicating consistent confidence across the inferred relations. Such predicates depend on changes in relative position and heading over time and therefore cannot be reliably determined from a single frame. The observed progression across the tracklet illustrates the temporal cues exploited by the motion-geometry and temporal-consistency mechanisms, in line with the performance degradation observed in Table 4 when these components are removed. Additionally, Figure 7 presents latent representations for eight different RGB–LiDAR parts, four of which are included in the main manuscript. The remaining visualizations further demonstrate the alignment between the RGB and LiDAR components produced by the proposed framework across different parts.

AI-Assisted Writing

We used AI-based language tools solely to improve the grammar, clarity, and readability of the manuscript. The authors reviewed and verified the final content and remain responsible for all technical claims and conclusions.

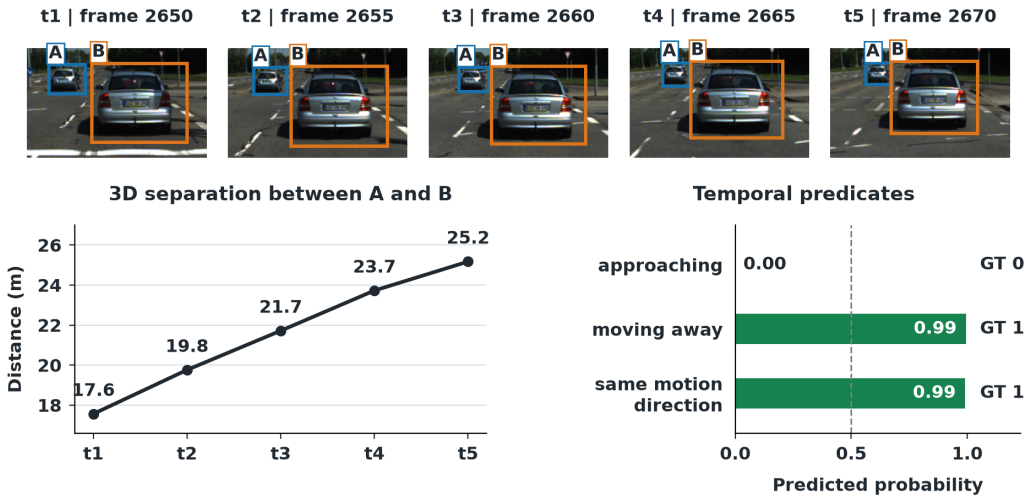


Figure 6: Temporal tracklet predictions over 5 frames.

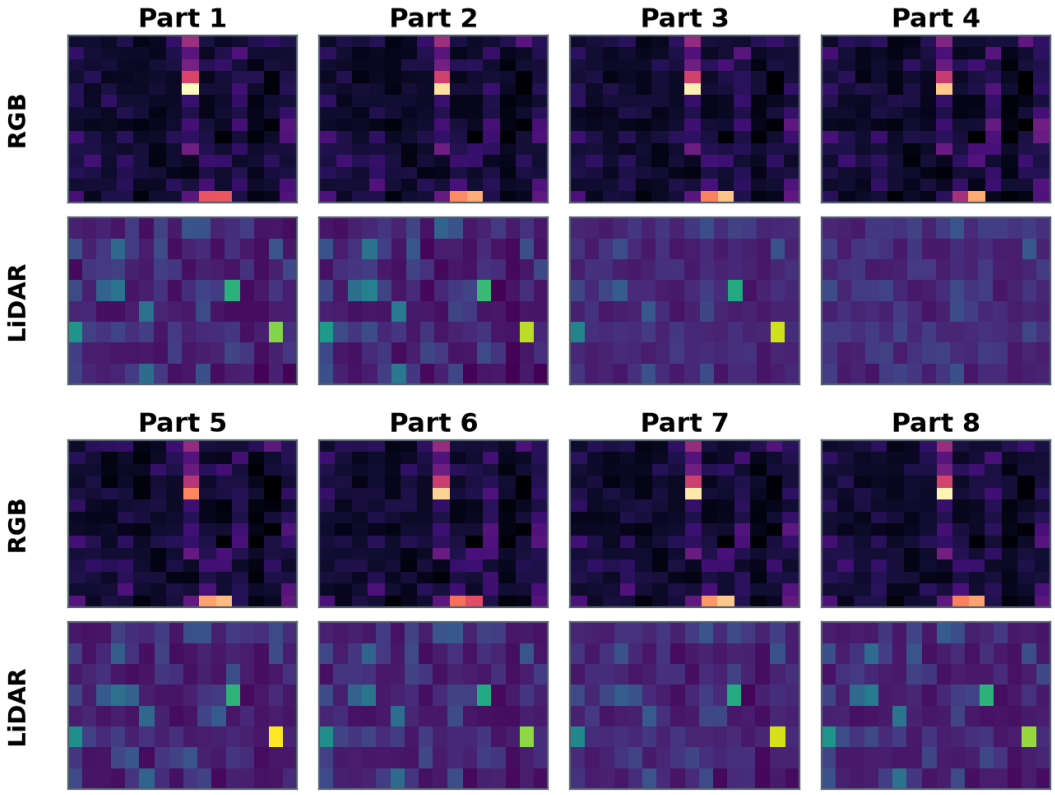


Figure 7: Learned RGB-LiDAR latent parts.