# Universal Scene Graph Generation

Shengqiong Wu, Hao Fei*, Tat-seng Chua
National University of Singapore
swu@u.nus.edu, haofei37@nus.edu.sg, dcscts@nus.edu.sg

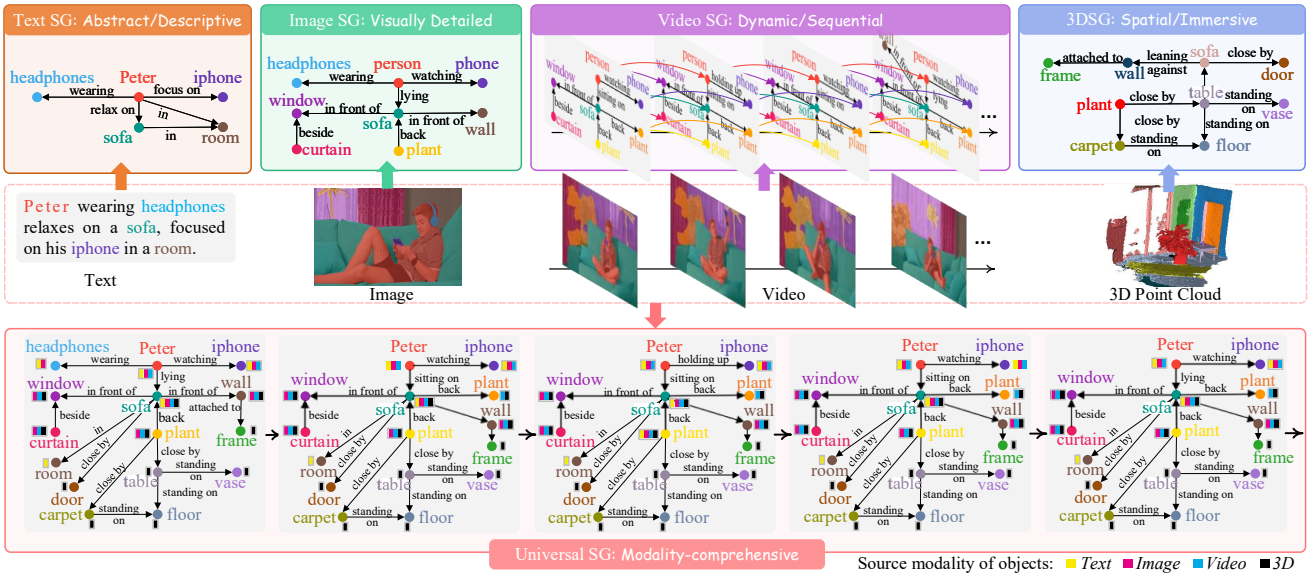



Figure 1. Illustrations of SGs (top) of single modalities in text, image, video, and 3D, and our proposed Universal SG (bottom). Note that the USG instance shown here is under the combination of four complete modalities, while practically any modality can be absent freely. Also, the temporal coreference edges are omitted for visual clarity (a full version is given in the Appendix).

## Abstract

*Scene graph (SG) representations can neatly and efficiently describe scene semantics, which has driven sustained intensive research in SG generation. In the real world, multiple modalities often coexist, with different types, such as images, text, video, and 3D data, expressing distinct characteristics. Unfortunately, current SG research is largely confined to single-modality scene modeling, preventing the full utilization of the complementary strengths of different modality SG representations in depicting holistic scene semantics. To this end, we introduce Universal SG (USG), a novel representation capable of fully characterizing comprehensive semantic scenes from any given combination of modality inputs, encompassing modality-invariant and modality-specific scenes. Further, we tailor a niche-targeting USG parser, USG-Par, which effectively addresses two key bottlenecks of cross-modal object alignment and out-of-domain challenges. We design the USG-Par with modular architecture for end-to-end USG generation, in which we devise an object associator to relieve the modality gap for cross-modal object alignment. Further, we propose a text-centric scene contrasting learning mechanism to mitigate domain imbalances by aligning multimodal objects and relations with textual SGs. Through extensive experiments, we demonstrate that USG offers a stronger capability for expressing scene semantics than standalone SGs, and also that our USG-Par achieves higher efficacy and performance. The project page is https://sqwu.top/USG/.*

## 1. Introduction

Scene understanding is a fundamental topic in computer vision and artificial intelligence, aiming to comprehend and interpret scenes in a manner akin to human perception. Within scene understanding, SG generation [21, 50, 61] stands as a pivotal task, seeking to identify and classify all constituent objects in a given scene, along with their attributes and interrelationships. This process constructs a semantic graph representation that facilitates a comprehensive understanding of the specific scene. SGs are widely applied in various real-world applications [3, 9, 10, 33, 36, 43], such as au-

*Corresponding author.

tonomous driving [14, 49, 65], robot navigation [39, 52], augmented reality [20, 44], etc. Consequently, SG generation has garnered significant research attention in past decades [7, 8, 15, 24, 30, 45].

Humans perceive the world through a multitude of sensory modalities, acquiring information via different channels to form a complete perception of their environment. Thus, beyond images, scenes can be represented through various other modalities as well, including text, video, and 3D formats. Correspondingly, current research has focused on constructing scene graphs across these different modalities, such as generations of Image SG (ISG) [21, 45, 61], Textual SG (TSG) [6, 26], Video SG (VSG) [7, 40, 41], and 3D SG (3DSG) [46, 51, 53]. Moreover, due to the inherent nature of each modality, SGs in each modality possess distinct capabilities, as illustrated in Fig. 1. Specifically:

- **ISG** – Images provide concrete visual details, enabling SGs to precisely describe object locations, sizes, and visual attributes, but typically cannot directly convey temporal sequences or dynamic changes.
- **TSG** – Compared to the concreteness of vision, text can flexibly describe abstract entities, actions, events, and abstract relationships between objects and entities that are not visually explicit, but usually lack specific visual and spatial details.
- **VSG** – Building upon static images, videos are more adept at expressing dynamic events, actions, and temporal changes.
- **3DSG** On top of 2D vision, 3D scenes can further model objects and their spatial relationships, sizes, orientations, and other 3-dimensional spatial attributes.

It is evident that SG representations across different modalities provide complementary insights for describing a whole scene semantics. Conversely, this implies that relying solely on one single modality's SG cannot offer a comprehensive scene representation. In practical applications, an ideal process involves users providing inputs in any single modality or even multiple combinations and the system simultaneously extracting modality-specific and modality-shared scene information to derive a unified SG representation for a comprehensive scene understanding. Unfortunately, current research [4, 8, 24, 48, 50] communities study SGs for different modalities separately, and a universal representation encompassing all modalities does not yet exist. To bridge this gap, this paper proposes a Universal SG (USG) representation capable of characterizing any combination of modalities. As shown in Fig. 1, USG can incorporate all the characteristics of individual SGs, capturing a comprehensive semantic scene from any given combination of input modalities.

Yet, realizing such a universal SG generation presents two non-trivial challenges from the methodological perspective. First, regarding **model architecture**, the current community has most largely explored singleton-modality SG generation methods [17, 46, 57, 61, 64]. To achieve USG parsing, a direct approach would be to use a pipeline paradigm, i.e., first parsing each modality's SG separately, and then merging the individual SGs into one USG [11, 33]. The most challenging issue here is merging identical objects across multiple modalities while retaining modality-specific scene information. However, because the independent SG parsers for certain modalities operate in isolation, it leads to critical problems: i), complementary semantic information across modalities may be overlooked; ii), due to differences in feature spaces across modalities, it becomes difficult to precisely align identical objects across modalities, resulting in a final USG that is neither concise nor effective. Second, regarding **data**, the lack of annotated USG data is a significant hurdle. Since manual annotation is labor-intensive, a feasible solution is to leverage the existing abundance of single-modality SG annotation datasets to learn USG, i.e., through joint training on various singleton SG datasets. Unfortunately, significant domain divergence exists among different modality data. For example, 3DSG data might focus solely on static indoor scenes, VSGs are mostly biased towards action-rich scenes, while only TSGs hold for general domains. Consequently, the resulting USG parser inevitably suffers from scene biases, thereby limiting its effectiveness.

To address these challenges, we present a USG Parser (termed **USG-Par**), capable of end-to-end scene parsing from any modality inputs, outputting a USG representation. Technically, USG-Par works sequentially through 5 main modules: Step ❶, modality-specific encoders encode inputs from different modalities; Step ❷, we employ Mask2Former [5] as a shared mask decoder to generate representations for scene objects; Step ❸, an object associator is devised to determine whether objects from different modalities are identical. Specifically, to eliminate the modality gap, objects are transformed into their respective modality-specific feature spaces before object association and alignment; Step ❹, a relation proposal constructor generates the most feasible relation pairs by modeling object-level interactions; Step ❺, a relation decoder finally predicts the final relations among different objects based on the selected pairs from the previous step. During model training, to combat the issue of data domain imbalance, we propose a *text-centric scene contrasting learning* mechanism, where considering that the text modality is scene-unbiased in the general domain, we align objects and relations from various modalities to the objects and relations in the text space of TSG.

Experimental results demonstrate that USG provides a more powerful and comprehensive scene representation compared to standalone SGs of individual modalities. Furthermore, extensive experimental results on various SG benchmark datasets indicate that, i), the proposed USG-Par achieves significant performance improvements in single-modality scene graph parsing; and ii), in multi-modality SG

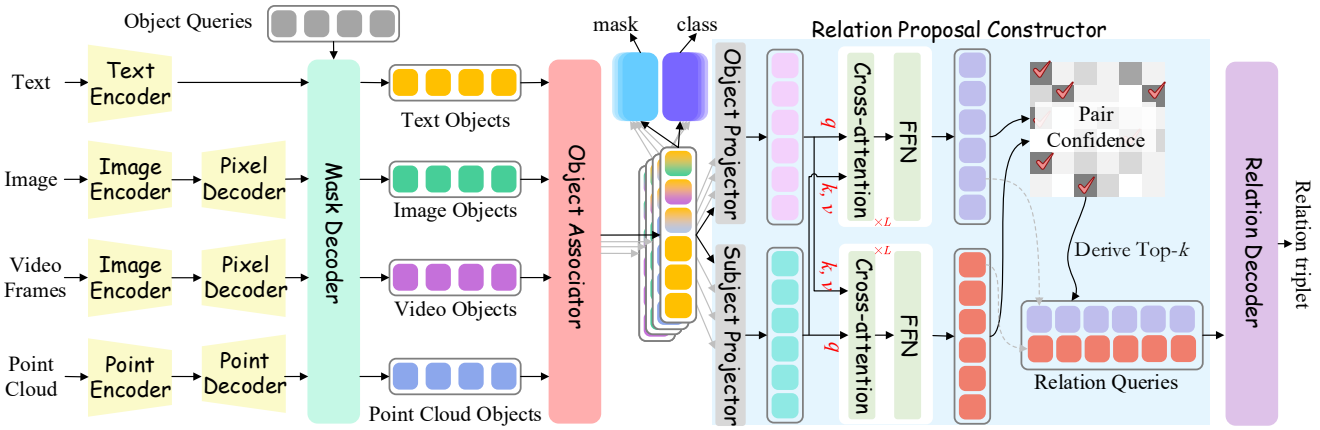



Figure 2. Overview of USG-Par architecture. It mainly consists of five modules, including modality-specific encoders, shared mask decoder, object associator, relation proposal constructor, and relation decoder.

parsing, USG-Par accurately constructs associations between objects from different modalities, achieving better performance than pipeline approaches. Further, we show that USG-Par effectively handles scene parsing for unseen scene domains and unseen modality combinations, thanks to the text-centric scene contrasting learning for weak supervision.

In summary, this work makes two primary contributions to the community. **First**, we are the first to propose a Universal SG representation for holistic semantic scene understanding. **Second**, we introduce a novel USG generator, USG-Par, which effectively addresses cross-modal object alignment and out-of-domain challenges simultaneously.

## 2. Universal Scene Graph

### 2.1. Preliminary

The concept of the SG is initially introduced in [19] as a visually grounded graph structure representing object instances within an image, where edges depict pairwise relationships between these objects. Formally, given an image $\mathcal{I} \in \mathbb{R}^{H \times W}$, ISG generation task is defined as:

$$F(\mathcal{G}^{\mathcal{I}}|\mathcal{I}) = F(\{\mathcal{O}, \mathcal{R}\}|\mathcal{I}), \quad (1)$$

where $\mathcal{O}$ is the set of objects, with each object node $o_i$ characterized by a bounding box or a mask segmentation $m_i$ and an associated category label $c_i^o \in \mathbb{C}^{\mathcal{O}}$, *i.e.* $o_i = (c_i^o, m_i)$. $\mathcal{R}$ is the set of relations, with each directed edge $r_{i,j}$ between two objects (subject and object) described by a predicate $c_{i,j}^r \in \mathbb{C}^{\mathcal{R}}$, i.e., $r_{i,j} = (o_i, c_{i,j}^r, o_j)$. $\mathbb{C}^{\mathcal{O}}$ and $\mathbb{C}^{\mathcal{R}}$ means the object and predicate classes.

Recently, by incorporating temporal dimensions, VSG [21] generation has been introduced:

$$F(\mathcal{G}^{\mathcal{V}}|\mathcal{V}) = F(\{\mathcal{G}_t^{\mathcal{V}}\}_{t=1}^T|\mathcal{V}) = F(\{\mathcal{O}_t, \mathcal{R}_t\}_{t=1}^T|\mathcal{V}), \quad (2)$$

where $\mathcal{V} \in \mathbb{R}^{T \times H \times W}$ is the input video, and $\mathcal{G}_t^{\mathcal{V}} = \{\mathcal{O}_t, \mathcal{R}_t\}$ denotes a ISG at $t$-th frame.

To enable spatially immersive understanding, methods for generating 3DSG [46] are then developed:

$$F(\mathcal{G}^{\mathcal{D}}|\mathcal{D}) = F(\{\mathcal{O}, \mathcal{R}\}|\mathcal{D}), \quad (3)$$

where $\mathcal{D} \in \mathbb{R}^{P \times 6}$ is the input 3D point clouds, with $P$ standing for the number of point clouds of interest and 6 representing xyz coordinates plus RGB values. Similar to ISG, each 3D object $o_i$ in the object set $\mathcal{O}$ is identified by an instance segmentation $m_i \in \{0,1\}^P$ and a category label $c_i^o$

Lastly, text, as a highly abstract and flexible modality for scene description, has motivated research on generating SGs from textual inputs [38], formulated as:

$$F(\mathcal{G}^{\mathcal{S}}|\mathcal{S}) = F(\{\mathcal{O}, \mathcal{R}\}|\mathcal{S}), \quad (4)$$

where $\mathcal{S} \in \mathbb{R}^L$ is a given text. In TSGs, each object node is defined solely by its category label.

### 2.2. USG Definition

In contrast to existing methods that focus solely on single-modality SG generation, we define our USG generation task as being able to handle any single-modality SG generation as well as any combination of modalities. Given a set of input data in various modalities (e.g., image $\mathcal{I}$, video $\mathcal{V}$, 3D point cloud $\mathcal{D}$ and text $\mathcal{S}$), USG generation is formulated as:

$$F(\mathcal{G}^{\mathcal{U}}|\{\mathcal{I}, \mathcal{V}, \mathcal{D}, \mathcal{S}\}) = F(\{\mathcal{O}, \mathcal{R}\}|\{\mathcal{I}, \mathcal{V}, \mathcal{D}, \mathcal{S}\}), \quad (5)$$

where $\mathcal{O} = \{\mathcal{O}^*\}$, $* \in \{\mathcal{I}, \mathcal{V}, \mathcal{D}, \mathcal{S}\}$ represents the set of objects across all modalities. Each node involves a category label $c_i^o \in \mathbb{C}^{\mathcal{O}}$ and a segmentation mask $m_i$. Additionally, for objects extracted from textual descriptions, we construct a positional binary mask to indicate their locations. $\mathcal{R} = \{\mathcal{R}^*, \mathcal{R}^{*\times\diamond}\}$, $*, \diamond \in \{\mathcal{I}, \mathcal{V}, \mathcal{D}, \mathcal{S}\}$ and $* \neq \diamond$. $\mathcal{R}^*$ includes both intra-modality relationships and inter-modality associations $\mathcal{R}^{*\times\diamond}$. For example, in Fig. 1, the text and image describe the same scene; thus, the textual object "*Peter*" in the TSG should correspond to the visual object "*person*" in the ISG. Similarly, the "*sofa*" in the 3DSG corresponds to the object "*sofa*" in the ISG.

## 3. Methodology

Our model consists of five main modules, as shown in Fig. 2. **First**, we first extract the modality-specific features with a modality-specific backbone. **Second**, we employ a shared

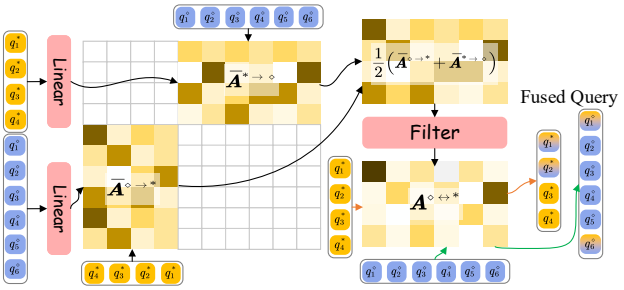



Figure 3. Illustration of the object associator for establishing associations between different modalities.

mask decoder to extract object queries for various modalities. These object queries are then fed into the modality-specific object detection head to obtain the category label and tracked positions of the corresponding objects. **Third**, the object queries are input into the object associator, which determines the association relationships between objects across modalities. **Fourth**, a relation proposal constructor is utilized to retrieve the most confidential subject-object pairs. **Finally**, a relation decoder is employed to decode the final predicate prediction between the subjects and objects.

### 3.1. Modality-specific Encoder

To encode each modality, we propose using specialized encoders for each one: **1) Text Encoder.** We employ OpenCLIP [34] to encode the input text $\mathcal{S}$ to obtain the text contextualized features, $\boldsymbol{H}^{\mathcal{S}}$. **2) Image Encoder and Pixel Decoder.** We adopt the frozen CLIP-ConvNeXt [29] as the backbone image encoder to model the given image/video inputs, yielding frozen feature $\bar{\boldsymbol{H}}^{\mathcal{I}/\mathcal{V}}$. The pixel decoder, adapted from Mask2Former, consists of multi-stage deformable attention layers that transform the frozen features $\bar{\boldsymbol{H}}^{\mathcal{I}/\mathcal{V}}$, into the fused multi-scale feature $\{\boldsymbol{H}_i^{\mathcal{I}/\mathcal{V}}\}_{i=1}^{3}$, with the same channel dimension, where $i$ is the layer index, and $i=3$ corresponds to the highest-resolution feature. **3) Point Encoder and Point Decoder.** We employ Point-BERT [60] as the point encoder to encode the input point cloud $\mathcal{D}$, generating the super-point features, $\bar{\boldsymbol{H}}^{\mathcal{D}}$. The point decoder is designed to propagate the super-point features to each point hierarchically, producing multi-scale point features $\{\boldsymbol{H}^{\mathcal{D}}\}_{i=1}^{3}$, where $i=3$ denotes point clouds features with the original number points. All features are projected into a common $d$-dimensional space using a linear layer.

### 3.2. Shared Mask Decoder

Objects across different modalities can provide complementary information, facilitating cross-modal learning. Therefore, we employ a shared mask decoder framework to implicitly integrate these cross-modal complementary features. Following [5], we utilize multi-scale features and a cascaded decoder to perform masked cross-attention between modality-specific features $\boldsymbol{H}^{*}$ and the corresponding object query features $\boldsymbol{X}_l^{*} \in \mathbb{R}^{N_q^{*} \times d}, * \in \{\mathcal{I}, \mathcal{V}, \mathcal{D}, \mathcal{S}\}$ as follows:

$$\boldsymbol{X}_l^{*} = \text{softmax}(\boldsymbol{M}_{l-1}^{*} + \boldsymbol{Q}_{l-1}^{*}\boldsymbol{K}_{l-1}^{*\top})\boldsymbol{V}_{l-1}^{*} + \boldsymbol{X}_{l-1}^{*}, \quad (6)$$

where $N_q^{*}$ is the number of queries and $l$ is the layer index. $\boldsymbol{M}_{l-1}^{*}$ is the binarized output of the resized mask prediction from the previous stage. $\boldsymbol{X}_0^{*}$ denotes input object query features to the mask decoder. $\boldsymbol{Q}_{l-1}^{*} = F_q(\boldsymbol{X}_{l-1}^{*})$, while $\boldsymbol{K}_{l-1}^{*} = F_k(\boldsymbol{H}^{*})$ and $\boldsymbol{V}_{l-1}^{*} = F_v(\boldsymbol{H}^{*})$. Here, $F_q(\cdot)$, $F_k(\cdot)$ and $F_v(\cdot)$ are linear transformations as typically applied in attention mechanisms. In practice, for image, video, and 3D data, $\boldsymbol{H}^{*}$ is sampled from the multi-scale feature output $\{\boldsymbol{H}^{\mathcal{I}/\mathcal{V}/\mathcal{D}}\}_{i=1}^{3}$, while for text, we employ $\boldsymbol{H}^{\mathcal{S}}$ across different scales. In addition, for video data, to effectively capture the temporal information across frames, we incorporate a transformer-based temporal encoder $F_{temp}$ to model the temporal relationships between objects. After $L^{mask}$ layers, we obtain the refined object queries $\boldsymbol{Q}^{*} = \{\boldsymbol{q}_i^{*}\}_{i=1}^{N_q^{*}}$.

### 3.3. Object Associator

The biggest challenge in USG generation would be accurately merging identical objects across multiple modalities. To establish robust associations and bridge the modality gap, we propose projecting objects into each other's feature spaces using a transformation layer before determining association relationships:

$$\begin{aligned} \bar{\boldsymbol{A}}^{*\to\diamond} &= \cos(F_{*\to\diamond}(\boldsymbol{Q}^{*}), \boldsymbol{Q}^{\diamond}), \\ \bar{\boldsymbol{A}}^{\diamond\to*} &= \cos(F_{\diamond\to*}(\boldsymbol{Q}^{\diamond}), \boldsymbol{Q}^{*}), \\ \bar{\boldsymbol{A}}^{*\leftrightarrow\diamond} &= (\bar{\boldsymbol{A}}^{*\to\diamond} + \bar{\boldsymbol{A}}^{\diamond\to*})/2\,, \end{aligned} \quad (7)$$

where $F_{*\to\diamond}(\cdot)$ is a linear transformation of $\boldsymbol{Q}^{*}$, similar to $F_{\diamond\to*}(\cdot)$. We then design a filtering module to further refine and learn feasible sparse association pairs. In practice, we employ a CNN-based architecture, which leverages local details while efficiently filtering out redundant noise. The output is a refined association matrix, denoted as $\boldsymbol{A}^{*\leftrightarrow\diamond}$. In Fig. 3 we illustrate the above process in detail.

### 3.4. Modality-specific Object Detection Head

Detecting objects involves predicting the segmentation mask and category label (containing a "no object" label) from each object query. To achieve simultaneous object detection across different modalities while retaining modality-specific scene information, we employ modality-specific heads. The design is consistent across modalities; here, we illustrate the process in the image modality as an example. Upon establishing association relationships between objects across modalities, we fuse the object query via $\boldsymbol{q}_i^{\mathcal{I}} = \boldsymbol{q}_i^{\mathcal{I}} + \sum_j \boldsymbol{A}_{i,j}^{\mathcal{I}\leftrightarrow*}(\boldsymbol{q}_j^{*}), * \in \{\mathcal{V}, \mathcal{D}, \mathcal{S}\}$, allowing the incorporation of complementary information from other modalities to enrich the object embeddings of the current modality. Then, a category classifier is applied on the fused query features $\boldsymbol{q}_i^{\mathcal{I}}$ to yield category label probability predictions $\bar{c}_i^{o,\mathcal{I}}$ for each segment. For mask prediction, each binary mask prediction $\bar{m}_i^{\mathcal{I}} \in [0,1]^{H\times W}$ is obtained by computing the dot product between the fused query features and per-pixel embeddings

$\boldsymbol{H}_3^{\mathcal{I}}$, i.e., $\bar{m}_i^{\mathcal{I}} = \text{sigmoid}(\text{MLP}(\boldsymbol{Q}^{\mathcal{I}}) \cdot \boldsymbol{H}_3^{\mathcal{I}\top})$.

### 3.5. Relation Proposal Constructor

We employ a subject and object projector, implemented as an MLP, to generate subject and object embeddings $\boldsymbol{E}^{obj}$, $\boldsymbol{E}^{sub}$, respectively. we omit the modality superscript for simplicity. A straightforward approach would involve calculating relationship embeddings by combining embeddings for all possible subject-object pairs and subsequently classifying the relationship predicates. However, such an exhaustive pairwise computation is computationally infeasible. Moreover, intuitively, improving recall of relevant pairs correlates with enhanced relationship recall, suggesting that focusing on the most promising object pairs could increase computational efficiency and overall performance. To this end, we introduce a Relation Proposal Constructor (RPC) to selectively identify promising object pairs, as shown in Fig. 2. Specifically, we design a two-way relation-aware cross-attention mechanism $F_{\text{CA}}(q, k, v)$, to iteratively refine subject and object features as follows:

$$\begin{aligned} \boldsymbol{X}_l^{sub} &= F_{\text{CA}}^{obj \to sub}(\boldsymbol{X}_{l-1}^{sub}, \boldsymbol{X}_{l-1}^{obj}, \boldsymbol{X}_{l-1}^{obj}), \\ \boldsymbol{X}_l^{obj} &= F_{\text{CA}}^{sub \to obj}(\boldsymbol{X}_{l-1}^{obj}, \boldsymbol{X}_{l-1}^{sub}, \boldsymbol{X}_{l-1}^{sub}), \end{aligned} \tag{8}$$

where $l$ denotes the layer index, and $\boldsymbol{X}_0^{sub} = \boldsymbol{E}^{sub}$, $\boldsymbol{X}_0^{obj} = \boldsymbol{E}^{obj}$. Following $L^{RPC}$ layers of interaction, we compute the cosine similarity between the refined subject and object embeddings, resulting in a Pair Confidence Matrix $\boldsymbol{C} = \cos(\boldsymbol{X}_L^{sub}, \boldsymbol{X}_L^{obj})$. We then perform a top-$k$ selection on the $\boldsymbol{C}$, with the top-$k$ indices used to retrieve the corresponding subject and object queries from $\boldsymbol{X}_L^{sub}$,$\boldsymbol{X}_L^{obj}$, which are denoted as $\boldsymbol{Q}^{sub}$ and $\boldsymbol{Q}^{obj}$ respectively.

### 3.6. Relation Detector

After retrieving the potential subject and object queries from RPC, they are concatenated together along the length dimension to construct relationship queries:

$$\boldsymbol{Q}^{rel} = [\boldsymbol{Q}^{sub} + \boldsymbol{E}^{sub}; \boldsymbol{Q}^{obj} + \boldsymbol{E}^{obj}], \tag{9}$$

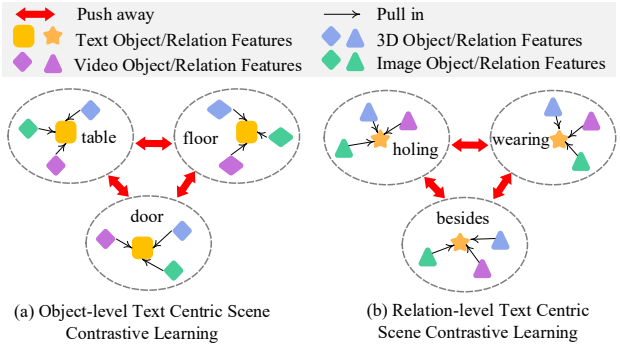
where $[;]$ means the concatenation. Then, we employ a transformer-based relation decoder applying cross-attention with keys and values from contextualized input features and self-attention on queries to predict the final relation:

$$\boldsymbol{X}_l^{rel} = \text{F}_{CA}^{rel}(\boldsymbol{X}_{l-1}^{rel}, \boldsymbol{H}, \boldsymbol{H}), \tag{10}$$

where $\boldsymbol{X}_0^{rel} = \boldsymbol{Q}^{rel}$ is the relationship query features into the relation decoder. Since relationship analysis primarily focuses on examining semantic information, we leverage the contextualized representations generated by each modality-specific encoder as $\boldsymbol{H}$. After applying $L^{rel}$ transformer layer, we apply a relationship classifier on the refined relationship queries $\boldsymbol{X}_L^{rel}$ to predict the final relationship probabilities.

### 3.7. Training with Domain-balancing Strategy

This section elaborates on the training objectives and strategies to optimize our system.




(a) Object-level Text Centric Scene Contrastive Learning

(b) Relation-level Text Centric Scene Contrastive Learning

Figure 4. Illustration of the object-level and relation-level text-centric scene contrasting learning mechanism.

**Object Detection Loss.** During training, we first apply Hungarian matching between the predicted and ground-truth entity masks to assign object queries to entities in text, video, image, and 3D modalities. This assignment is then used to supervise the mask predictions and category label classifications. We employ a sigmoid Cross-Entropy (CE) loss $L_{cls}^{o}$ as in [31], computed between the ground truth object classes and logits obtained by computing the inner product of the object queries with the text embeddings of the category names. Moreover, following [50], a binary CE loss $L_{ce}$ and Dice loss $L_{dice}$ are leveraged for segmentation.

$$\mathcal{L}_{obj} = \lambda_{cls} L_{cls}^{o} + \lambda_{ce} L_{ce} + \lambda_{dice} L_{dice}, \tag{11}$$

where $\lambda_{cls}, \lambda_{ce}, \lambda_{dice}$ are the parameter weights.

**Object Association Loss.** To optimize the object associator, we take the ground-truth association matrix, which is a binary matrix, as the supervised signal. Due to the sparsity of the association matrix, we utilize a weighted binary CE loss, $\mathcal{L}_{ass}$, to ensure stable training by significantly increasing the weight of the positive entries in the matrix.

**Relation Classification Loss.** For relation predicate classification, we employ a sigmoid CE loss $\mathcal{L}_{cls}^{r}$ for the predicate classification, similar to object category classification. In addition, to supervise the relation proposal constructor in selecting the most confidential object pairs for further relation classification, we utilize a weighted binary CE loss, $\mathcal{L}_{pair}$, on the pair confidence matrix $\boldsymbol{C}$:

$$\mathcal{L}_{rel} = \mathcal{L}_{cls}^{r} + \mathcal{L}_{pair}. \tag{12}$$

**Text-centric Scene Contrastive Learning.** A significant challenge when leveraging single-modal SG data for USG-Par learning is the domain imbalance across modalities, compounded by the lack of USG data for various modality combinations, which can result in suboptimal performance. To address these, we propose a text-centric scene contrastive learning approach that aligns other modalities with text data, attributed with two key advantages: **1)** TSG data encompass the most diverse and general domain, and **2)** binding information from other modalities into text effectively addresses the scarcity of USG data for certain modality combinations

| Modality | Dataset | #Obj. | #Rela. | #Tri. | #Ins. |
|---|---|---|---|---|---|
| Text | FACTUAL [26] | 4,042 | 1,607 | 40,149 | 40,369 |
| Image | VG [21] | 5,996 | 1,024 | 1,683,231 | 108,077 |
| | PSG [57] | 133 | 56 | 275,371 | 48,749 |
| Video | AG [18] | 36 | 25 | 772,013 | 288,782 |
| | PVSG [59] | 126 | 65 | 4,587 | 400 |
| 3D | 3DDSG [46] | 528 | 39 | 543,956 | 1,335 |
| Multimodal | $\mathcal{S}-\mathcal{I}$ | 6,089 | 1,235 | 1,791,309 | 124,357 |
| | $\mathcal{S}-\mathcal{V}$ | 150 | 132 | 6,751 | 400 |
| | $\mathcal{S}-\mathcal{D}$ | 724 | 257 | 230,865 | 46,173 |
| | $\mathcal{I}-\mathcal{V}$ | 126 | 65 | 4,587 | 400 |
| | $\mathcal{I}-\mathcal{D}$ | 345 | 75 | 7,689 | 4,492 |

Table 1. Statistics of SG datasets. '#Obj.' and '#Rela.' denote the number of the object and relation categories, respectively. '#Tri.' is relation triplets count, and '#Ins.' is instance count.

[13, 68]. Considering the unique characteristics of SG, we design both object-level and relation-level text-centric contrastive learning, as illustrated in Fig. 4. Given text-* pairs as inputs, where * represents image, video, or 3D modalities, we extract textual queries $\boldsymbol{Q}^{\mathcal{T}}$ and other modal queries $\boldsymbol{Q}^{*}, * \in \{\mathcal{I}, \mathcal{V}, \mathcal{D}\}$. Positive targets are constructed when corresponding textual objects are present in other modalities; otherwise, they serve as negative targets. Thus, the object-level text-centric contrastive loss is formulated as:

$$\mathcal{L}^{o}_{cons} = -\sum_{y^{+}} \log \frac{\exp(\mathbf{x}\cdot\mathbf{y}^{+})}{\exp(\mathbf{x}\cdot\mathbf{y}^{+}) + \sum_{y^{-}} \exp(\mathbf{x}\cdot\mathbf{y}^{-})}, \tag{13}$$

where $\mathbf{x}$, $\mathbf{y}^{+}$, and $\mathbf{y}^{-}$ are query embeddings of text-* pairs, their positive targets, and negative targets, sampled from object queries $\boldsymbol{Q}^{\mathcal{T}}$ and $\boldsymbol{Q}^{*}$, respectively, Similarly, we compute the relation-level text-centric contrastive loss $\mathcal{L}^{r}_{cons}$. The total contrastive loss is given by $\mathcal{L}_{cons} = \mathcal{L}^{o}_{cons} + \mathcal{L}^{r}_{cons}$.

**Training Target in Total.** We combine all four loss terms in a linear manner as our final loss function:

$$\mathcal{L} = \alpha\mathcal{L}_{obj} + \beta\mathcal{L}_{ass} + \gamma\mathcal{L}_{rel} + \eta\mathcal{L}_{cons}, \tag{14}$$

where $\alpha, \beta, \gamma$ and $\eta$ are the weights for the loss terms.

## 4. Experimental Settings

**Datasets and Resources.** We conduct experiments on two distinct groups of datasets to comprehensively evaluate our method's SG generation capability in single and multiple modalities. **1) Single modality**, we employ a range of well-established datasets: VG [21] and PSG [57] for images, AG [18] and PVSG[59] for videos, 3DDSG [46] for 3D scenes, and FACTUAL [26] for text-based scenes. Notably, some of these datasets provide only bounding box annotations; thus, we use SAM-2 [35] to generate pseudo-segmentation masks, using the bounding boxes as prompts. **2) Multiple modalities**, on the one hand, we construct multimodal SGs by leveraging paired text-image/video/3D data. We utilize GPT-4o [32] to parse initial TSGs from textual captions, and then link textual and visual objects through label matching. To increase the diversity and richness of textual descriptions in these multimodal pairs, we rephrase and enrich captions, allowing for flexible and partially non-literal associations with visual content. Additionally, for image-video cross-modal SGs, we pair randomly selected frames with temporally non-adjacent video segments. Similarly, we pair 2D image views with corresponding 3D scenes for image-3D cross-modal SGs. Tab. 1 shows the statistics of the datasets, and we provide further details on data construction in the Appendix.

| Method | Backbone | R/mR@20 | R/mR@50 | R/mR@100 |
|---|---|---|---|---|
| IMP [56] | R50 | 16.5 / 6.5 | 18.2 / 7.1 | 18.6 / 7.2 |
| Motifs [61] | R50 | 20.0 / 9.1 | 21.5 / 9.6 | 22.0 / 9.7 |
| VCTree [45] | R50 | 20.6 / 9.7 | 22.1 / 10.2 | 22.5 / 10.2 |
| GPS-Net [27] | R50 | 17.8 / 7.0 | 19.6 / 7.5 | 20.1 / 7.7 |
| PSGTR [57] | R50 | 28.4 / 16.6 | 34.4 / 20.8 | 36.3 / 22.1 |
| PSFormer [57] | R50 | 18.1 / 14.8 | 19.6 / 20.1 | 17.4 / 18.7 |
| HiLo [67] | R50 | 34.1 / 23.7 | 40.7 / 30.3 | 43.0 / 33.1 |
| DSGG [16] | R50 | 32.7 / 30.8 | 42.8 / 38.8 | 50.0 / 43.4 |
| Pair-Net [50] | R50 | 29.6 / 24.7 | 35.6 / 28.5 | 39.6 / 30.6 |
| Pair-Net [50] | Swin-B | 33.3 / 25.4 | 39.3 / 28.2 | 42.4 / 29.7 |
| USG-Par$^{\natural}$ (Ours) | OpenCLIP | 35.7 / 29.9 | 44.6 / 40.9 | 51.3 / 42.7 |
| USG-Par (Ours) | OpenCLIP | **36.9 / 32.1** | **46.4 / 41.7** | **52.4 / 44.6** |

Table 2. Evaluation on the PSG [57] under the SGDet task. $^{\natural}$ means the model is trained solely on the corresponding single-modality dataset, here, PSG. The top baseline results are underlined, and the best overall performance is highlighted in bold. The tables below follow the same format.

| Method | R/mR@20 | R/mR@50 | R/mR@100 |
|---|---|---|---|
| IPS+T+1D Conv. [59] | 2.79 / 1.24 | 2.80 / 1.47 | 3.10 / 1.59 |
| IPS+T+Trans. [59] | 4.02 / 1.75 | 4.41 / 1.86 | 4.88 / 2.03 |
| VPS+1D Conv. [59] | 0.60 / 0.27 | 0.73 / 0.28 | 0.76 / 0.29 |
| VPS+Trans. [59] | 0.75 / 0.36 | 0.91 / 0.39 | 0.94 / 0.40 |
| USG-Par$^{\natural}$ (Ours) | 4.68 / 2.01 | 5.37 / 2.02 | 6.15 / 3.03 |
| USG-Par (Ours) | **5.08 / 2.23** | **6.64 / 2.36** | **7.45 / 3.76** |

Table 3. Evaluation on the PVSG [59].

**Evaluation Metrics.** We evaluate our methods following three standard evaluation tasks: 1) predicate classification (**PreCls**); 2) scene graph classification (**SGCls**); 3) scene graph detection (**SGDet**). Following previous work [50], we adopt Recall@K (**R@**$K$) and mean Recall@K (**mR@**$K$) as evaluation metrics to measure the fraction of ground truth hit in the top $K$ predictions, where $K \in \{10, 20, 50, 100\}$. For the TSG, we compute the `Set Match` and `SPICE` as in [26] for evaluation.

**Implementation.** We initialize the text and image encoders using OpenCLIP [34]. Following the approach in [5, 25], we design the pixel decoder. For the point encoder, we adopt Point-BERT [60] as the initialization, and for the point decoder, inspired by [60], we implement a hierarchical propagation strategy with distance-based interpolation. The mask decoder follows the design in [5]. We set the number of predefined learnable queries to 100. The object associator is implemented as a 3-layer CNN with a kernel size of $3 \times 3$. The relation decoder comprises a 6-layer transformer with an embedding dimension of 256. During training, we used the AdamW optimizer with an initial learning rate of $10e-4$. More implementation details can refer to the Appendix.

| Method | SGCls | PreCls |
|---|---|---|
| | R@20/50/100 | R@20/50/100 |
| SGPN [47] | 27.0 / 28.8 / 29.0 | 51.9 / 58.0 / 58.5 |
| SGFN [53] | 27.5 / 29.2 / 29.2 | 52.6 / 58.9 / 59.4 |
| EdgeGCN [62] | 28.0 / 29.8 / 29.8 | 54.7 / 60.9 / 61.5 |
| KISGP [63] | 28.5 / 30.0 / 30.1 | 59.3 / 65.0 / 65.3 |
| Feng et al. [12] | - / 31.5 / 31.6 | - / 31.5 / 31.6 |
| VL-SAT [51] | 32.0 / 33.5 / 33.7 | 67.8 / 79.9 / 80.8 |
| CCL-3DDSG [4] | 37.6 / 40.3 / 45.7 | 73.6 / 80.5 / 82.9 |
| USG-Par$^\natural$ (Ours) | 36.6 / 41.4 / 46.2 | 71.9 / 81.0 / 83.4 |
| USG-Par (Ours) | **37.9 / 43.1 / 46.9** | **73.5 / 81.7 / 84.1** |

Table 4. Evaluation results on the 3DDSG [46] dataset.

| Method | Random | | Length | |
|---|---|---|---|---|
| | Set Match | SPICE | Set Match | SPICE |
| SPICE-Parser [1] | 13.00 | 56.15 | 0.94 | 38.04 |
| AMR-SG-T5 [6] | 28.45 | 64.82 | 12.16 | 51.71 |
| CDP-T5 [6] | 46.15 | 73.56 | 26.50 | 61.21 |
| VG-T5 [42] | 11.54 | 47.46 | 2.94 | 42.98 |
| FACTUAL-T5 (pre) [26] | 79.77 | **92.91** | 42.35 | 82.43 |
| FACTUAL-T5 [26] | 79.44 | 92.23 | 38.65 | 80.76 |
| USG-Par$^\natural$ (Ours) | 80.40 | 87.53 | 39.75 | 83.69 |
| USG-Par (Ours) | **82.40** | 88.12 | **43.83** | **84.38** |

Table 5. Performance on the FACTUAL [26] dataset.

| Method | $\mathcal{S}-\mathcal{I}$ | $\mathcal{S}-\mathcal{V}$ | $\mathcal{S}-\mathcal{D}$ | $\mathcal{I}-\mathcal{D}$ | $\mathcal{I}-\mathcal{V}$ |
|---|---|---|---|---|---|
| | 75.4 / 25.4 | 73.3 / 1.9 | 71.1 / 13.3 | 39.1 / 12.6 | 35.4 / 4.2 |
| | [26] + [50] | [26] + [59] | [26] + [4] | [50] + [59] | [50] + [4] |
| USG-Par$^\flat$ | 78.6 / 26.2 | 76.4 / 2.0 | 74.9 / 15.4 | 40.4 / 13.7 | 37.6 / 4.2 |
| USG-Par - $\mathcal{L}_{cons}$ | 79.6 / 29.2 | 79.4 / 2.2 | 77.9 / 17.4 | 43.3 / 16.7 | 41.6 / 7.2 |
| USG-Par | 80.3 / 32.4 | 80.6 / 2.4 | 79.2 / 20.2 | 47.3 / 18.6 | 42.7 / 9.5 |

Table 6. For SGDet task evaluation on multimodal inputs, we apply separate SG parsers per modality as referenced. $^\flat$ means raining on corresponding multimodal data only, and $-\mathcal{L}_{cons}$ denotes training without text-centric scene contrastive loss. We separately report the `Set Match` and `mR@50` scores for text and other modalities.

# 5. Results and Analyses

## 5.1. Main Observations

We compare USG-Par with the existing methods on single and multiple modalities data.

**1) USG Generation in Single Modality.** We present the experimental results for both single-dataset training and joint training across multiple datasets, as shown in Tab. 2 3 4 5. Additional results on other datasets are provided in Appendix. In the single-dataset training setting, our model achieves comparable performance to the best-performing baselines and even slightly surpasses them on certain datasets, such as PVSG. This demonstrates the effectiveness of our model design across different modalities, highlighting its capacity to enhance performance on individual datasets. When comparing our joint-training results with baselines, our model consistently outperforms across all datasets. For example, on PSG, our method shows an average R@K score improvement of 3.2, indicating that joint learning effectively leverages single-modal SG data to boost overall performance.

**2) USG Generation across Multiple Modalities.** To evaluate our model's performance in parsing USG across multiple modalities, we conduct experiments using a collection of pair-wise multimodal datasets. As a baseline, we adopt a pipeline approach, applying the best SG parser for each modality independently and then combining them together. As shown in Tab. 6, our model, trained exclusively on corresponding multimodal data, achieves superior USG generation performance compared to separate SG parsers, demonstrating its ability to leverage cross-modal complementary information to enhance accuracy. Furthermore, joint training on all multimodal datasets consistently achieves the highest performances, underscoring USG-Par's effectiveness in generating USGs across diverse multimodal scenarios.

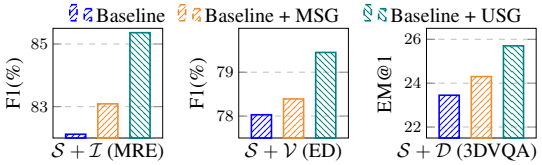



Figure 5. Comparison of multimodal tasks with and without USG integration. Baselines: multimodal relation extraction (MRE) [54], emotion detection (ED) [23], and 3D visual QA (3DVQA) [2].

| LN | Filter | $\mathcal{L}_{ass}$ | $\mathcal{S}-\mathcal{I}/\mathcal{V}/\mathcal{D}$ | $\mathcal{I}-\mathcal{V}$ | $\mathcal{I}-\mathcal{D}$ |
|---|---|---|---|---|---|
| × | ✓ | ✓ | 4.6 / 3.8 / 1.7 | 18.2 | 12.4 |
| ✓ | × | ✓ | 12.5 / 11.7/ 11.1 | 22.3 | 18.2 |
| ✓ | ✓ | × | 10.7 / 10.8 / 11.4 | 22.8 | 19.1 |
| ✓ | ✓ | ✓ | 13.6 / 13.9 / 12.0 | 24.3 | 20.7 |

Table 7. Ablation study of object associator. "LN" means the linear transformation. Association accuracy@5 scores are reported.

## 5.2. Ablations and Discussions

Taking one step further, here we give more discussions and in-depth analyses to reveal how the system advances.

**1) Probing Advantage of USG over Singleton SG Representations.** Fig. 5 compares the performance of multimodal tasks with and without USG integration. Across tasks—multimodal relation extraction (MRE), emotion detection (ED), and 3D visual QA (3DVQA)—the incorporation of USG consistently improves results over baselines. Additionally, USG shows superior performance compared to MSG, which is constructed through a pipeline approach, highlighting the advantages of USG representations.

**2) The Necessity of Each Component of Object Associator.** Tab. 7 presents an ablation study on each component of the object associator. Firstly, removing the transformation linear layer—computing cosine similarity directly without modality-specific transformations—leads to the lowest performance, as the model struggles to effectively associate objects across modalities. Incorporating the filter further enhances performance by excluding low-confidence pairs. Finally, adding a supervised signal to the association matrix significantly improves guidance for both the filter learner and the linear layer, enabling more precise performance.

**3) The Necessity of Each Component of RPC.** Tab. 8 presents ablation studies on the effectiveness of each com-

| Proj. | RAC | $\mathcal{L}_{pair}$ | PSG | PVSG | 3DDSG | FACTUAL |
|---|---|---|---|---|---|---|
| | | | R/mR@50 | R/mR@50 | R/mR@50 | Set Match |
| ✓ | ✓ | × | 38.4 / 31.4 | 4.4 / 1.8 | 18.9 / 14.0 | 79.1 |
| ✓ | × | ✓ | 32.5 / 26.5 | 2.3 / 1.1 | 17.9 / 7.2 | 76.3 |
| × | ✓ | ✓ | 36.5 / 28.5 | 4.2 / 1.5 | 18.6 / 14.9 | 78.9 |
| × | × | ✓ | 2.5 / 1.9 | 0.5 / 0.4 | 1.6 / 1.2 | 64.5 |
| ✓ | ✓ | ✓ | 44.6 / 40.9 | 5.4 / 2.3 | 21.8 / 15.4 | 80.4 |

Table 8. Ablation study of the RPC. "Proj." denotes the subject/object projector, and "RAC" is the two-way relation-aware cross-attention module. The evaluation is exclusively performed on the corresponding dataset.

| Architecture | PSG | PVSG | 3DDSG | FACTUAL |
|---|---|---|---|---|
| | R/mR@50 | R/mR@50 | R/mR@50 | Set Match |
| MLP | 34.1 / 20.7 | 3.6 / 1.0 | 12.1 / 7.3 | 61.2 |
| w/o $F_{CA}^{rel}$ | 39.6 / 28.4 | 4.5 / 1.6 | 16.7 / 10.0 | 73.5 |
| Ours | 44.6 / 40.9 | 5.4 / 2.3 | 21.8 / 15.4 | 80.4 |

Table 9. Different Architectures for relation decoder. "w/o $F_{CA}^{rel}$" denotes removing the cross-attention layers.

ponent of RPC. Our findings indicate all three components contribute to the performance. Notably, removing the projection and RAC layers leads to model divergence, resulting in no correct predictions, as object queries lack essential pairwise information and contain only category details. Additionally, removing the pair loss, which encodes critical information on pair distributions to support pair proposal matrix learning, also degrades performance.

**4) The Architecture of Relation Decoder.** We evaluate different architectures for the relation decoder. The results are shown in Tab. 9. We find that the transformer-based architecture outperforms the MLP-based approach. Additionally, integrating pairwise information with contextualized input through cross-attention further improves performance by preserving more input details.

**5) The Impact of Text-centric Scene Contrastive Learning.** In Tab. 6, we compare the model equipped with and without contrastive learning. The results indicate that applying contrastive learning yields consistent improvements in USG generation across all multimodal datasets. This suggests that contrastive learning effectively mitigates modality gaps, leading to enhanced overall performance.

### 5.3. Qualitative Case Study with Visualization

Finally, we visualize a USG of both image and text inputs from our system. As in Fig. 6, the pipeline approach, which applies separate SG parsing of each modality and then combines them, often leads to incorrect associations, such as an erroneous relation between "person" and "Jumbo". Conversely, we find that USG can offer a more comprehensive scene representation by accurately aligning cross-modal objects and integrating information from both modalities. For instance, the USG correctly identifies "Peter" as the person holding the bottle and "Jumbo" as *fed elephant* while also effectively integrating other visual elements, e.g., *trees* and *dirt*. we provide more visualizations in the Appendix.

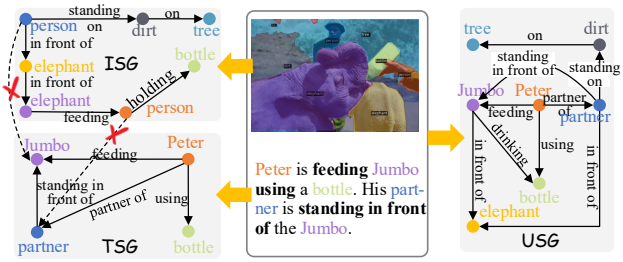



Figure 6. The USG derived from image and text and wrong association built between ISG and TSG by the pipeline method.

## 6. Related Work

Over decades, SGs have garnered substantial research attention [21, 26, 46, 61], where various definitions of SG representations under different modalities and settings are developed [18, 21, 22, 38, 46], including image SG [19, 21, 61], textual SGs [26, 38] video [21], 3D [46, 53], and even more settings such as panoptic SG [57–59] and ego-view SG [37], etc. SGs can accurately capture the semantics of a scene while filtering out undesired visual information. Thus, SGs have been widely applied to various downstream tasks [38, 44, 52, 65]. While almost all existing SG research is confined to modeling within a single modality, we realize that real-world scenarios necessitate a universal SG representation capable of expressing information from various modalities through a unified cross-modal perspective. This need is particularly pressing with the development of multimodal generalist and agent communities [3, 28, 55, 66], where an increasing number of applications require the ability to understand and process multimodal information. Therefore, this paper for the first time explores a novel USG representation. Despite the existence of various SG generation methods for different SG types, there should currently not be a specialized approach for universally parsing SGs across modalities. Specifically, a unified model architecture is required for both modeling modality-invariant SG information and efficiently preserving the complementary modality-specific scene.

## 7. Conclusion

This paper presents a Universal Scene Graph (USG), a novel representation that characterizes comprehensive semantic scenes from any combination of modality inputs, encompassing both modality-invariant and modality-specific aspects. To generate USG effectively, we develop USG-Par, a niche-targeting parser for end-to-end USG generation. USG-Par addresses the critical challenges of cross-modal object alignment and out-of-domain generalization by incorporating an object associator that bridges modality gaps and a text-centric scene contrasting learning mechanism that mitigates domain imbalances. Through extensive experiments, we demonstrate that USG provides a more powerful and comprehensive semantic scene representation compared to standalone SGs. Also USG-Par achieves superior efficacy, offering a strong benchmark method for USG.

## References


[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016. 7

[2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, pages 19107–19117, 2022. 7

[3] Shivam Chandhok. Scenegpt: A language model for 3d scene understanding. *CoRR*, abs/2408.06926, 2024. 1, 8

[4] 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 *CVPR*, pages 27863–27873, 2024. 2, 7

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1280–1289, 2022. 2, 4, 6

[6] Woo Suk Choi, Yu-Jung Heo, Dharani Punithan, and Byoung-Tak Zhang. Scene graph parsing via abstract meaning representation in pre-trained language models. In *Workshop on DLG4NLP*, pages 30–35, 2022. 2, 7

[7] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *ICCV*, pages 16352–16362, 2021. 2

[8] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11169–11183, 2023. 2

[9] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *CVPR*, 2024. 1

[10] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *ICML*, 2024. 1

[11] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *CoRR*, abs/2406.19255, 2024. 2

[12] Mingtao Feng, Haoran Hou, Liang Zhang, Zijie Wu, Yulan Guo, and Ajmal Mian. 3d spatial multimodal knowledge accumulation for scene graph prediction in point cloud. In *CVPR*, pages 9182–9191, 2023. 7

[13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 6

[14] Elias Greve, Martin Büchner, Niclas Vödisch, Wolfram Burgard, and Abhinav Valada. Collaborative dynamic 3d scene graphs for automated driving. In *ICRA*, pages 11118–11124, 2024. 2

[15] Peng Hao, Xiaobing Wang, Yingying Jiang, Hanchao Jia, and Xiaoshuai Hao. BCTR: bidirectional conditioning transformer for scene graph generation. *CoRR*, abs/2407.18715, 2024. 2

[16] Zeeshan Hayder and Xuming He. DSGG: dense relation transformer for an end-to-end scene graph generation. In *CVPR*, pages 28317–28326, 2024. 6

[17] Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. EGTR: extracting graph from transformer for scene graph generation. In *CVPR*, pages 24229–24238, 2024. 2

[18] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10233–10244, 2020. 6, 8

[19] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678, 2015. 3, 8

[20] Denis Kalkofen, Erick Méndez, and Dieter Schmalstieg. Comprehensible visualization for augmented reality. *IEEE Trans. Vis. Comput. Graph.*, 15(2):193–204, 2009. 2

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 1, 2, 3, 6, 8

[22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset V4. *Int. J. Comput. Vis.*, 128(7):1956–1981, 2020. 8

[23] Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024. 7

[24] Rongjie Li, Songyang Zhang, and Xuming He. SGTR: end-to-end scene graph generation with transformer. In *CVPR*, pages 19464–19474, 2022. 2

[25] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *CVPR*, pages 27948–27959, 2024. 6

[26] Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of ACL*, pages 6377–6390, 2023. 2, 6, 7, 8

[27] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, pages 3743–3752, 2020. 6

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 8

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CVPR*, 2022. 4

[30] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016. 2

[31] Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil

Houlsby. Simple open-vocabulary object detection. In *ECCV*, pages 728–755, 2022. 5

[32] OpenAI. Gpt-4 technical report. https://openai.com/research/gpt-4, 2023. https://openai.com/research/gpt-4. 6

[33] Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *CVPR*, pages 14354–14363, 2024. 1, 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 6

[35] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *CoRR*, abs/2408.00714, 2024. 6

[36] Sonia Raychaudhuri, Tommaso Campari, Unnat Jain, Manolis Savva, and Angel X. Chang. Reduce, reuse, recycle: Modular multi-object navigation. *CoRR*, abs/2304.03696, 2023. 1

[37] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *CVPR*, pages 18622–18632, 2024. 8

[38] Sebastian Schuster, Ranjay Krishna, Angel X. Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *EMNLP Workshop*, pages 70–80, 2015. 3, 8

[39] Zachary Seymour, Niluthpol Chowdhury Mithun, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Graphmapper: Efficient visual navigation by scene graph generation. In *ICPR*, pages 4146–4153, 2022. 2

[40] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM MM*, 2017. 2

[41] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, pages 279–287, 2019. 2

[42] Sahand Sharifzadeh, Sina Moayed Baharlou, Martin Schmitt, Hinrich Schütze, and Volker Tresp. Improving scene graph classification by exploiting knowledge from texts. In *AAAI*, pages 2189–2197, 2022. 7

[43] Guibao Shen, Luozhou Wang, Jiantao Lin, Wenhang Ge, Chaozhe Zhang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Guangyong Chen, Yijun Li, and Ying-Cong Chen. Sg-adapter: Enhancing text-to-image generation with scene graph guidance. *CoRR*, abs/2405.15321, 2024. 1

[44] Tomu Tahara, Takashi Seno, Gaku Narita, and Tomoya Ishikawa. Retargetable AR: context-aware augmented reality in indoor scenes based on 3d scene graph. In *ISMAR*, pages 249–255, 2020. 2, 8

[45] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019. 2, 6

[46] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *CVPR*, pages 3960–3969, 2020. 2, 3, 6, 7, 8

[47] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *CVPR*, pages 3960–3969, 2020. 7

[48] Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu. OED: towards one-stage end-to-end dynamic scene graph generation. In *CVPR*, pages 27938–27947, 2024. 2

[49] Junyao Wang, Arnav Vaibhav Malawade, Junhong Zhou, Shih-Yuan Yu, and Mohammad Abdullah Al Faruque. RS2G: data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In *WACV*, pages 7478–7487, 2024. 2

[50] Jinghao Wang, Zhengyu Wen, Xiangtai Li, Zujin Guo, Jingkang Yang, and Ziwei Liu. Pair then relation: Pair-net for panoptic scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2024. 1, 2, 5, 6, 7

[51] 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 *CVPR*, pages 21560–21569, 2023. 2, 7

[52] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. *CoRR*, abs/2403.17846, 2024. 2, 8

[53] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from RGB-D sequences. In *CVPR*, pages 7515–7525, 2021. 2, 7, 8

[54] Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *ACL*, pages 14734–14751, 2023. 7

[55] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. In *ICML*, 2024. 8

[56] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 3097–3106, 2017. 6

[57] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, pages 178–196, 2022. 2, 6, 8

[58] Jingkang Yang, Jun CEN, Wenxuan Peng, Shuai Liu, Fangzhou Hong, Xiangtai Li, Kaiyang Zhou, Qifeng Chen, and Ziwei Liu. 4d panoptic scene graph generation. In *NeurIPS*, 2023.

[59] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *CVPR*, pages 18675–18685, 2023. 6, 7, 8

[60] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, pages 19291–19300, 2022. 4, 6

[61] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 1, 2, 6, 8

[62] Chaoyi Zhang, Jianhui Yu, Yang Song, and Weidong Cai. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *CVPR*, pages 9705–9715, 2021. 7

[63] Shoulong Zhang, Shuai Li, Aimin Hao, and Hong Qin. Knowledge-inspired 3d scene graph prediction in point cloud. In *NeurIPS*, pages 18620–18632, 2021. 7

[64] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang Wen Chen. End-to-end video scene graph generation with temporal propagation transformer. *IEEE Trans. Multim.*, 26:1613–1625, 2024. 2

[65] Yunpeng Zhang, Deheng Qian, Ding Li, Yifeng Pan, Yong Chen, Zhenbao Liang, Zhiyao Zhang, Shurui Zhang, Hongxu Li, Maolei Fu, Yun Ye, Zhujin Liang, Yi Shan, and Dalong Du. Graphad: Interaction scene graph for end-to-end autonomous driving. *CoRR*, abs/2403.19098, 2024. 2, 8

[66] Zhuosheng Zhang and Aston Zhang. You only look at screens: Multimodal chain-of-action agents. In *ACL*, pages 3132–3149, 2024. 8

[67] Zijian Zhou, Miaojing Shi, and Holger Caesar. Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. In *ICCV*, pages 21580–21591, 2023. 6

[68] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024. 6