Universal Scene Graph Generation

Supplementary Material

Overview

The appendix presents more details and additional results not included in the main paper due to page limitation. The list of items included are:

- Specification on Task Definition and Setups in §A;
- Full-version Related Work in §B;
- Limitations and Future Direction in §C;
- Extended Framework Details in §D;
- Detailed Experimental Settings in §E;
- Extended Experimental Results in §F.

A. Specification on Task Definition and Setups

A.1. SG Structure

Here, we provide a detailed description of the nodes and edges in the USG. The USG is formally represented as $\mathcal{G}^{\mathcal{U}} = \{\mathcal{O}, \mathcal{R}\}$, 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. For instance, as illustrated in Fig. 1, the objects node set $\mathcal{O}$ in the USG comprises of textual objects node set $\mathcal{O}^{\mathcal{S}}$ in the TSG and visual objects node set $\mathcal{O}^{\mathcal{I}}$ in the ISG. $\mathcal{R} = \{\mathcal{R}^*, \mathcal{R}^{*\times\diamond}\}$, $*, \diamond \in \{\mathcal{I}, \mathcal{V}, \mathcal{D}, \mathcal{S}\}$ and $* \neq \diamond$. $\mathcal{R}^*$ includes both intra-modality relationships and inter-modality associations $\mathcal{R}^{*\times\diamond}$. We define the existence of inter-modality associations between objects from different modalities if they correspond to the same underlying object described in distinct modalities. For example, as shown in Fig. 1, the textual object "*Peter*" in the TSG should correspond to the visual object "*person*" in the ISG. Similarly, as depicted in Fig. 3, the "*sofa*" in the 3DSG aligns with the "*sofa*" in the ISG. When inter-modality associations exist, the corresponding objects are merged into a unified node, as shown in Fig. 1, with the example of the "*headphones*" This merged node represents the object across multiple modalities, retaining a single category label. Typically, the object name from the textual modality is prioritized for its flexibility and precision in description. Similarly, the relation predicate is preferentially adopted from the TSG, as it often provides a more descriptive and accurate representation. For instance, in Fig. 1, the relationship between "*Peter*" and "*sofa*" in the USG is "*relax on*" derived from the TSG, rather than "*lying*" which might be less descriptive. Despite merging nodes, the segmentation masks from all modalities are preserved. This ensures that each modality's unique contribution to the object's representation is maintained within the USG.

In addition, to parse the USG for scenes derived from video and other modalities, we first establish association relations between nodes from other modalities and the objects in each frame of the VSG. For instance, as illustrated in Fig. 4, the objects "*Peter*", "*sofa*" and "*iPhone*" from the TSG are associated with the objects in every frame of the VSG. To ensure the USG comprehensively represents the scene described by the video and other modalities, the scene from the other modalities is added as the first frame in the USG. The remaining frames correspond to the frame-level scene graph representations from the VSG. This paradigm advances in integrating multimodal information much more seamlessly, enriching the holistic representation of the scene within the USG framework.

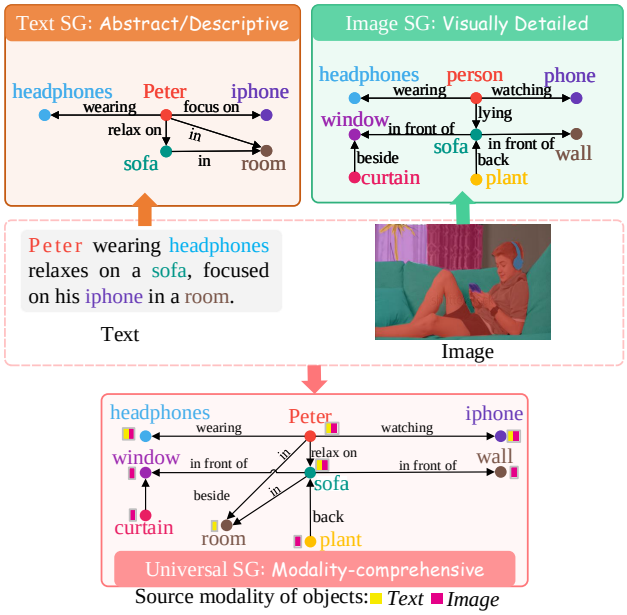

Figure 1. Illustration of USG generated from text and image scenes.

A.2. Combination of All Possible Modalities

Here we provide illustrations of the USG obtained under different modal combinations:

- **Text-Image**: in Fig.1.
- **Text-3D**: in Fig.2.
- **Text-Video**: in Fig.4.
- **Image-3D**: in Fig.3.
- **Text-Image-3D**: in Fig.5.
- **Text-Image-Video-3D (Complete combination)**: in Fig.6 we provide a full illustration of USG obtained from the total 4 modalities, i.e., text, image, video, and 3D.

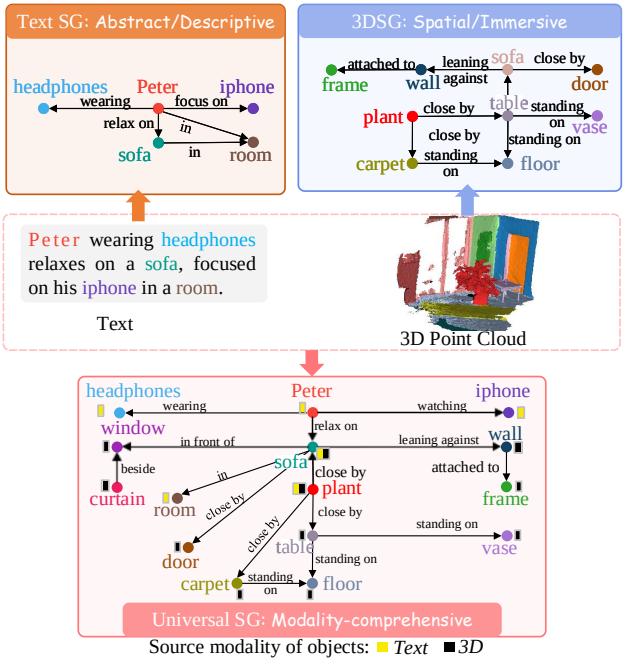

Figure 2. Illustration of USG generated from text and 3D scenes.

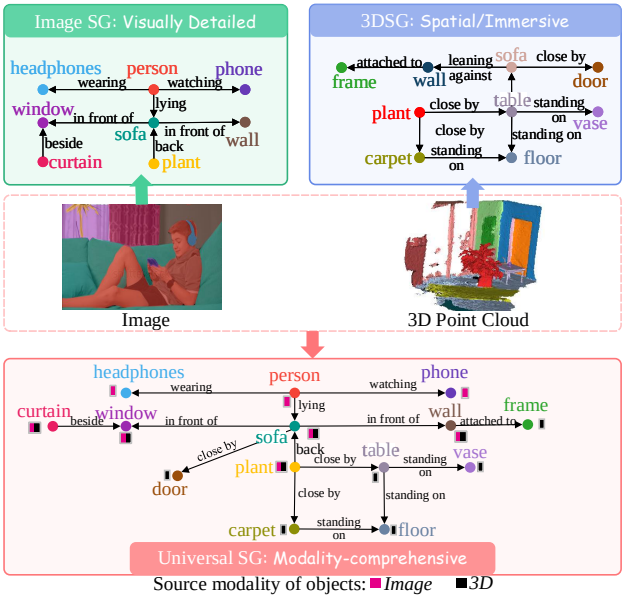

Figure 3. Illustration of USG generated from image and 3D scenes.

B. Full-version Related Work

B.1. SG Representation and Definition

Research on SG generation [22, 28, 52, 64] has long been a significant focus within the relevant community, aiming to deeply understand environmental scenes by not only recognizing individual objects but also the semantic relationships between them. Over decades of development, SGs have garnered substantial research attention and efforts, where various definitions of SG representations under different modalities and settings are developed [18, 22, 23, 43, 52]. Initially, centered on the static vision, researchers pioneered image SG [20, 22, 64], where nodes represent objects and edges denote the relationships between them. Subsequently, textual SGs [28, 43] are proposed to acknowledge that the textual modality can also convey a complete scene. Later studies extended SG representations to other data modalities, including video [22] and 3D [52, 56], and even to more settings such as panoptic SG [59, 60] and ego-view SG [42], etc. SGs can accurately capture the semantics of a scene while filtering out undesired visual information. Moreover, different modalities possess distinct characteristics, allowing SGs to model semantic scenes with subtly different traits. Thus, SGs have been widely applied to various downstream tasks [43, 50, 55, 66].

As previously emphasized, the definitions of SGs across differing modalities result in varied features and strengths. While almost all existing SG research is confined to modeling within a single modality, we recognize 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 [4, 33, 57, 67], where an increasing number of applications require the ability to understand and process multimodal information. Therefore, this paper explores a novel USG representation for the first time.

B.2. SG Generation Methods

Historically, SG generation methods can be broadly categorized into two main groups: two-stage methodsand one-stage methods. The two-stage methods [24, 25, 30, 34, 48, 58, 64, 68] involve training separate object detection and relation prediction models sequentially. Typically, these methods rely on off-the-shelf object detectors, such as Faster R-CNN [40], to detect N object queries. Subsequently, features such as appearance, spatial information, labels, depth, and masks are extracted for all possible combinations of detected objects. These features are then fed into the relation prediction model to infer relationships between each object pair. Despite achieving high relation extraction performance, the inherent limitations of the pipeline approaches, particularly the separate training of components, lead to significant model complexity.

To address this issue, recent research has shifted towards one-stage methods [8, 31, 53, 54], where the object detector and relation extractor are trained in an end-to-end manner. Early studies proposed fully convolutional SG generation models [32] and adopted pixel-based approaches [36]. Following the success of DETR [3], a Transformer-based one-stage object detector, many one-stage SGG studies [13, 17] have adopted similar approaches. These methods effectively model SG generation by introducing object queries or triplet

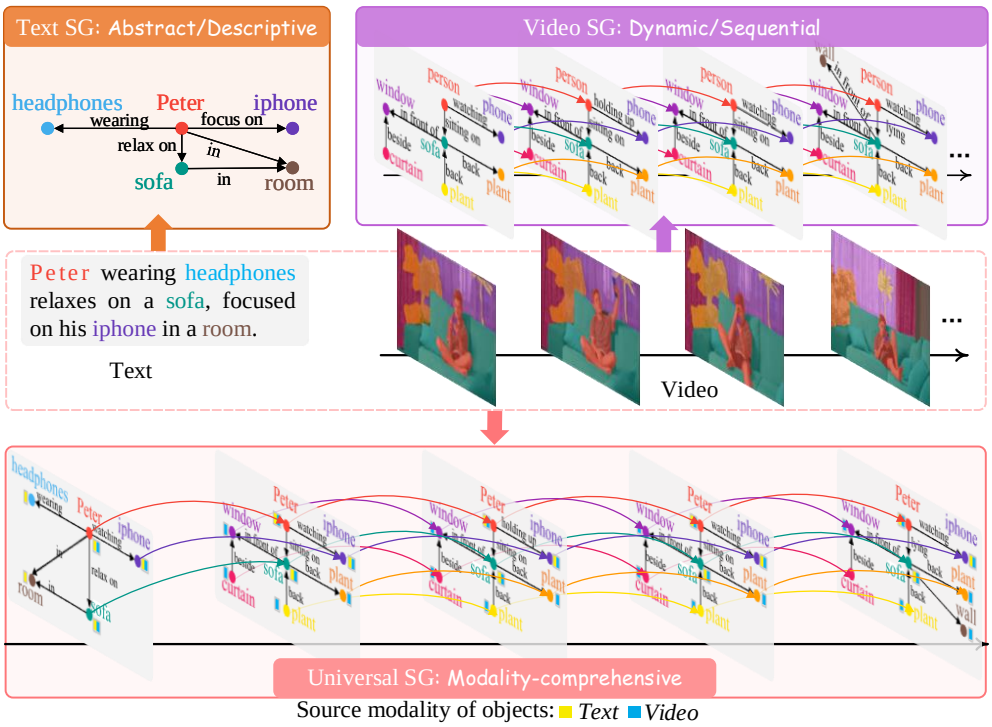

Figure 4. Illustration of USG generated from text and video scenes.

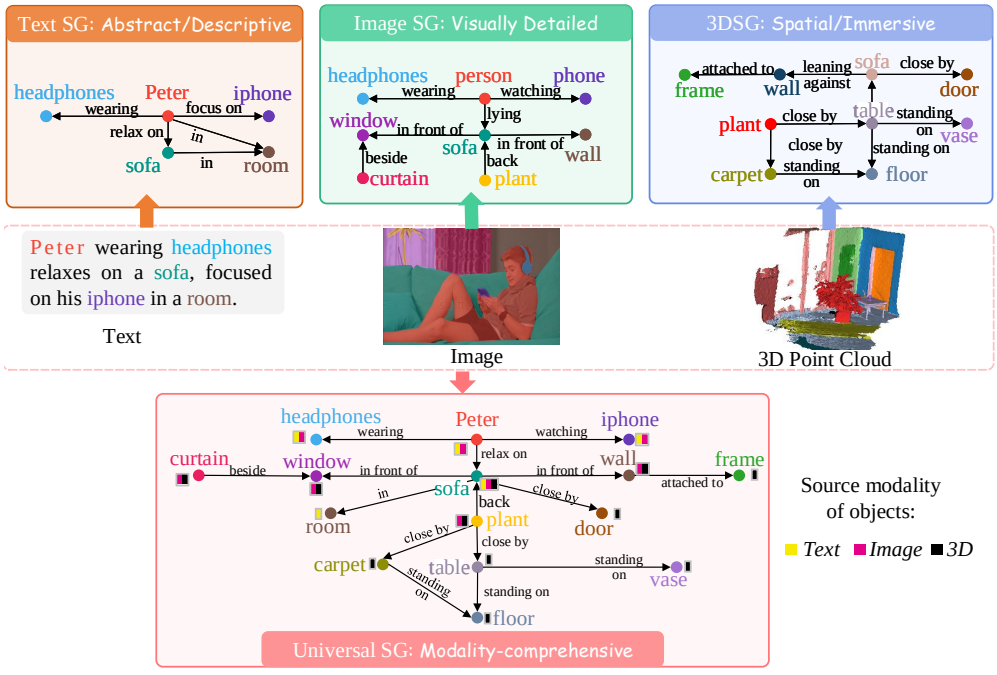

Figure 5. Illustration of USG generated from text, image and 3D scenes.

queries. For instance, RelTR [8] introduced paired subject and object queries, while SGTR [26] proposed compositional queries decoupled into subjects, objects, and predicates. PairNet [54] designed separate relation and object queries, and PSGTR [59] directly introduced triplet queries to detect triplets without relying on an object detector.

Beyond architectural designs, several studies [1, 35, 51, 53] have focused on leveraging modality-specific characteristics to enhance model performance. In the context of VSG generation, modeling spatio-temporal features has garnered

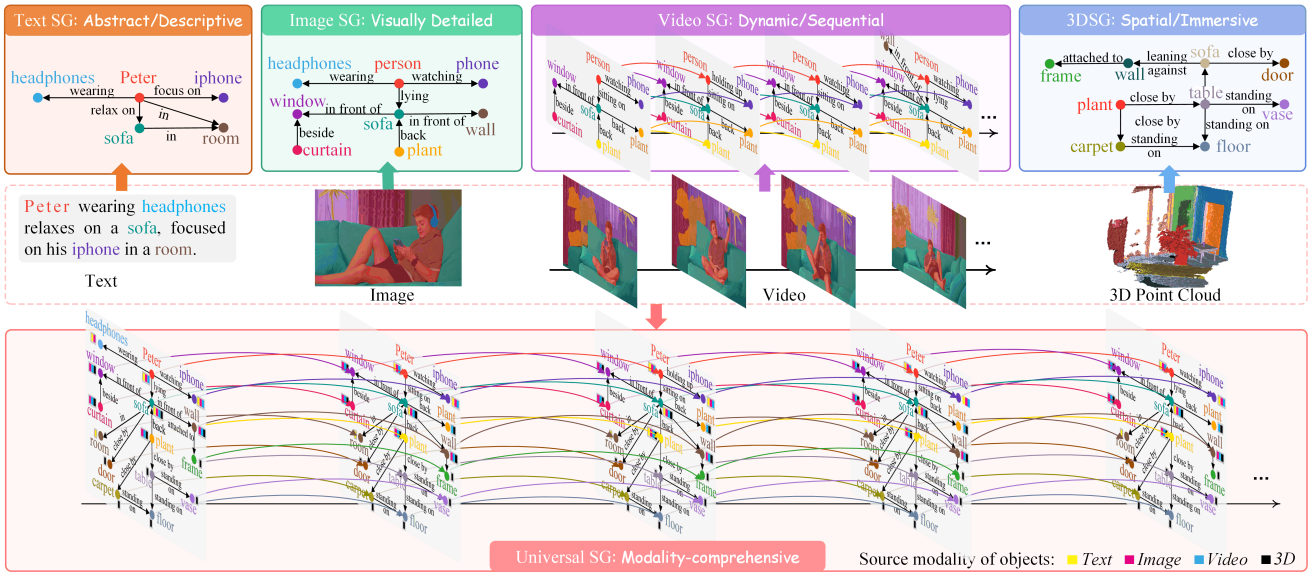

Figure 6. Illustration of USG generated from the text, image, video, and 3D in a stereoscopic viewpoint. This is also the full version of the first illustration shown in the Introduction section of the main article. Best viewed via zooming in.

significant attention. For example, TRACE [51] employs a hierarchical tree structure to aggregate spatial context, and [2] utilizes message passing in a spatio-temporal graph to enhance feature representation. For 3DSG generation, some researches [11, 12] focus on leveraging the spatial layout clues to enhance the 3DSG generation performance.

Additionally, to improve performance, many works are no longer limited to using only visual appearance. External knowledge has been incorporated to further improve SG generation performance [11, 12, 61, 62]. This includes statistical priors [62], such as co-occurrence frequencies, and commonsense knowledge [6, 19, 65] extracted from sources like Wikipedia or ConceptNet [47].

However, existing SG generation methods remain modality-specific, with no approach capable of supporting SG generation across different modalities. This limitation highlights the emergency of developing a universal SG generation method.

C. Limitations and Future Direction

C.1. Potential Limitations

Despite its contributions, this work has several limitations: Firstly, the proposed method faces challenges in associating objects across different modalities in highly complex and densely populated scenes. For instance, distinguishing between multiple similar individuals or matching objects with their textual object names often requires external commonsense knowledge, which is beyond the current scope of the model. Secondly, in video scenes, the method struggles with particular long-term understanding, particularly in object tracking and relation recognition over extended temporal sequences. While our current dataset does not include particularly long videos, such scenarios are common in real-world applications. Addressing this limitation presents a valuable direction for future research.

C.2. Future Work on USG

Going forward on the USG we introduced in this work, we believe the following aspects should be worth exploring.

First, the USG has significant potential for enhancing the capabilities of multimodal large language models (MLLMs). As a modality-invariant universal representation, the USG facilitates fine-grained alignment across different modalities, including object-level and relation-level correspondences. Inspired by the concept of knowledge masking [49], which focuses on learning more structured knowledge by masking phrases and named entities rather than individual sub-words, USG can inject fine-grained, structured semantic knowledge across modalities into MLLMs. This approach enables the alignment of semantic information at a granular level, fostering a deeper and more precise understanding of cross-modal content. Specifically, pre-training tasks can be designed by masking and predicting various types of nodes in the USG, parsed from multimodal inputs. These nodes may correspond to objects, relationships, or attributes, and their structured representation allows the model to learn modality-invariant features effectively. This strategy not only improves the alignment across modalities but also strengthens the model's reasoning and generalization capabilities.

Beyond MLLMs, the USG can also serve as a foundation for numerous downstream applications. In robotics, USG could facilitate embodied AI tasks, such as planning [15] and navigation [39, 55], by providing a universally structured understanding of dynamic, multimodal environments.

Moreover, in creative applications like content generation, USG could bridge visual and textual modalities to produce contextually coherent and semantically rich outputs, such as image-to-text descriptions or cross-modal story generation.

Looking ahead, the universality of USG should lead to unified multimodal benchmarks, where diverse tasks can be evaluated under a consistent framework. This would drive innovation in creating truly generalizable AI systems capable of reasoning across multiple domains and modalities. Developing such systems could redefine the boundaries of multimodal AI, enabling applications that require deep contextual understanding, such as virtual reality simulations, autonomous systems, and interactive AI agents.

D. Extended Framework Details

In this part, we try to give a more comprehensive picture of our USG generation framework, as an extension to the description in the main article.

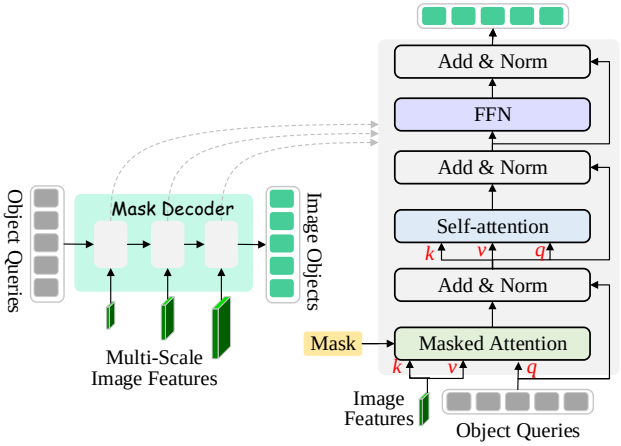

Figure 7. The framework of the mask decoder: multi-scale image features are integrated to refine image object queries, following a similar approach for other modalities such as text, video, and 3D.

D.1. Mask Encoder

As depicted in Fig. 7, the randomly initialized object queries are fed into the mask decoder, in which the multi-scale image features are also injected by masked cross-attention to help refine the object query representations. Specifically, following [7], we 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:

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

where N_q^* is the number of queries and l is the layer index. $\boldsymbol{X}_0^*$ 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. $\boldsymbol{M}_{l-1}^*$ is the binarized output of the resized mask prediction from the previous $l-1$-th Transformer decoder layer:

$$\boldsymbol{M}_{l-1}^*(x,y) = \begin{cases} 0 & \text{if } \boldsymbol{M}_{l-1}^*(x,y) = 1 \\ -\infty & \text{otherwise} \end{cases} \quad (2)$$

Moreover, 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 $\boldsymbol{Q}^* = \{\boldsymbol{q}_i^*\}_{i=1}^{N_q^*}$.

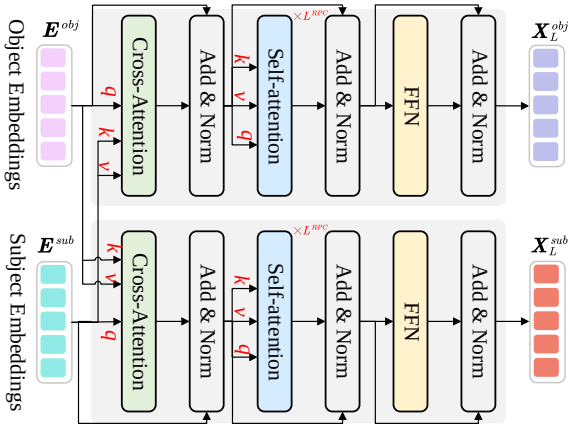

Figure 8. The framework of two-way relation-aware object/subject interaction module.

D.2. Two-way Relation-aware Object/Subject Interaction

The detailed framework of the two-way relation-aware object/subject interaction module is demonstrated in Fig. 8. The two inputs are object embeddings and subject embeddings, and then the L^{RPC} layers transformer layers with cross-attention and self-attention mechanisms perform to iteratively refine subject and object features as follows:

$$\begin{aligned} \boldsymbol{X}_l^{sub} &= F_{\text{CA}}^{obj \to sub}(\boldsymbol{X}_{l-1}^{sub}, \boldsymbol{X}_{l-1}^{obj}, \boldsymbol{X}_{l-1}^{obj}), \\ \boldsymbol{X}_l^{obj} &= F_{\text{CA}}^{sub \to obj}(\boldsymbol{X}_{l-1}^{obj}, \boldsymbol{X}_{l-1}^{sub}, \boldsymbol{X}_{l-1}^{sub}), \end{aligned} \quad (3)$$

where l denotes the layer index, and $\boldsymbol{X}_0^{sub} = \boldsymbol{E}^{sub}$, $\boldsymbol{X}_0^{obj} = \boldsymbol{E}^{obj}$. We define the $F_{\text{CA}}(\boldsymbol{X}, \boldsymbol{Y})$ as:

$$F_{\text{CA}}(\boldsymbol{X}, \boldsymbol{Y}) = \text{softmax}(F_q(\boldsymbol{X})^\top \cdot F_k(\boldsymbol{Y})) \cdot F_v(\boldsymbol{Y}), \quad (4)$$

where $F_q(\cdot)$, $F_k(\cdot)$ and $F_v(\cdot)$ are linear transformations as typically applied in attention mechanisms.

D.3. Relation Decoder

As illustrated in Fig. 9, the relation encoder processes the relation queries $\boldsymbol{Q}^{rel}$ alongside contextualized information. Specifically, the initial relation queries are constructed by concatenating the embeddings of selected subject-object pairs. Then, to leverage complementary contextual infor-

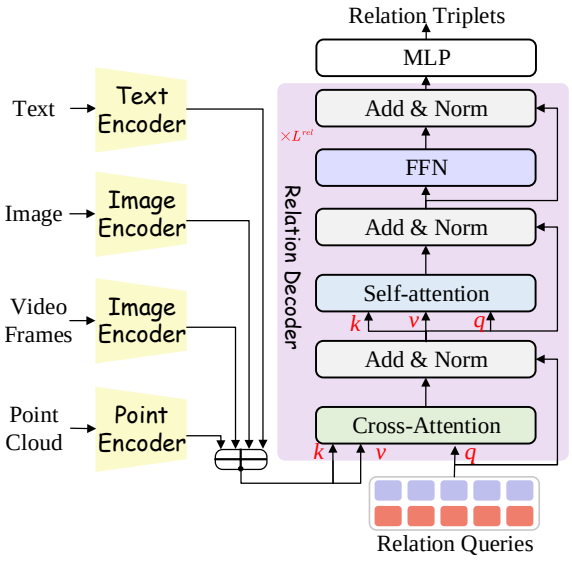

Figure 9. Illustration of relation decoder.

mation from multiple modalities, we propose to fuse the multimodal features to enhance the relation extraction performance:

$$\boldsymbol{H} = [\boldsymbol{H}^{\mathcal{S}}; \bar{\boldsymbol{H}}^{\mathcal{I}}; \bar{\boldsymbol{H}}^{\mathcal{V}}; \bar{\boldsymbol{H}}^{\mathcal{D}}], \tag{5}$$

where $\boldsymbol{H}$ represents the fused features, dependent on the input modalities. For example, given text, image, and 3D inputs, $\boldsymbol{H} = [\boldsymbol{H}^{\mathcal{S}}; \bar{\boldsymbol{H}}^{\mathcal{I}}; \bar{\boldsymbol{H}}^{\mathcal{D}}]$. Then, the fused features are integrated using a cross-attention mechanism to retain critical relational information:

$$\begin{aligned} \boldsymbol{X}_l^{rel} &= \mathrm{F}_{CA}^{rel}(\boldsymbol{X}_{l-1}^{rel}, \boldsymbol{H}, \boldsymbol{H}) \\ &= \mathrm{softmax}(F_q(\boldsymbol{X}_{l-1}^{rel})^\top \cdot F_k(\boldsymbol{H})) \cdot F_v(\boldsymbol{H}), \end{aligned} \tag{6}$$

where $\boldsymbol{X}_0^{rel} = \boldsymbol{Q}^{rel}$ is the initialized relationship query features into the relation decoder.

D.4. Inference

During inference, our framework, developed as a USG parser, supports both single-modality and multimodal input for USG generation. For single-modality USG generation, we first perform object detection, select the most confidential relation proposals, and finally perform the relationship classification. For multimodal USG generation, we introduce an object associator to establish associations between object pairs across different modalities before object detection and relation classification. We leverage Hungarian Assignment to find the associated pairs. Beyond this step, the remaining procedure closely follows that of single-modality SG generation.

For open-vocabulary USG generation, we compute the cosine similarity between each predicted object query embedding and a set of class label embeddings derived from CLIP [38]. The final label for each object is then assigned based on the highest cosine similarity score. Similarly, predicate classes are determined by selecting the label with the closest cosine similarity to the text embeddings of all predicate candidates.

E. Detailed Experimental Settings

E.1. Datasets

To evaluate the efficacy of USG-Par, which supports both single-modality and multi-modality scene parsing, we utilize existing single-modality datasets and a manually constructed multimodal dataset.

E.1.1. Single-modal Dataset

The single-modality datasets used in our experiments are categorized into the following four groups based on modality:

Image: **1) Visual Genome (VG)** [22]. We follow the protocols for the widely-used pre-processed subset VG150 [58], which contains the most frequent 150 entities and 50 predicates. The dataset contains approximately 108k images, with 70% for training and 30% for testing. **2) Panoptic Scene Graph (PSG)** [59]. Filtered from COCO [29] and VG datasets [22], the PSG dataset contains 133 object classes, including things, stuff, and 56 relation classes. This dataset has 46k training images and 2k testing images with panoptic segmentation and scene graph annotation. We follow the same data-processing pipelines from [59].

Video: **1) Action Genome (AG)** [18] annotates 234,253 frame scene graphs for sampled frames from around 10K videos, based on Charades dataset [46]. The annotations cover 35 object categories and 25 predicates. The overall predicates consist of three types of predicates: attention, spatial, and contracting. **2) Panoptic Video Scene Graph (PVSG)** [60] consists of 400 videos, including 289 third-person videos from VidOR [44] and 111 egocentric videos from EpicKitchens [10] and Ego4D [14]. Among the videos, 62 videos feature birthday celebrations, while 35 videos center around ceremonies, providing rich content for contextual logic and reasoning.

3D: **3D Scene Graph (3DSG)** [52] includes 1335 3D reconstructed indoor scenes, 528 classes of objects, and 39 types of predicates.

Text: **FACUTAL** [28] is derived from VG [22] dataset, which includes 4,042 classes of objects, 1,607 types of predicates, and 40,369 instances.

E.1.2. Multi-modal Dataset

Here, we show the detailed process for constructing the USG dataset involving two input modalities.

Input Image	Original Caption	Enriched Caption
	A man sitting in a chair holds up a toothbrush while opening a paper bag.	A man sits in a cushioned armchair, holding a blue toothbrush in one hand and an open paper bag in the other. A TV remote rests on the armrest, while a nearby table holds scattered magazines and glasses.
	A workspace with a laptop displaying a green leaf wallpaper, accompanied by a monitor, speakers, a keyboard, and multiple mice on a white desk.	David is sitting in front of the while desk and watching the laptop. Black speakers, a white keyboard, a smartphone on a red mouse pad with a beer logo, and two wireless mice on a white desk, all set in front of white curtains with cables and a black dock behind.
	A girl is playing ball in the park.	A girl plays with a ball in a park surrounded by grass, trees, and a nearby pathway , accompanied with her mother and her dog.
	A baby is playing with a balloon in the living room.	A baby plays with a balloon on a rug in the living room, surrounded by a sofa with cushions, a coffee table holding magazines and a remote, a toy box with stuffed animals and blocks, and a TV paused on a cartoon
	A small rectangular room with two single beds positioned parallel to each other along one wall, separated by a small gap. A desk and chair are located near the corner, with items like books and papers scattered on the desk.	A small rectangular room features two single beds along one wall, separated by a narrow gap, each with neatly tucked white bedding and wooden frames. Between the beds, a small bedside table holds a lamp and a water bottle. Near the corner, a wooden desk with a matching chair is cluttered with books, papers, and a laptop, with a coffee mug and a desk lamp placed on one side. Along the opposite wall, a wardrobe with partially open doors reveals hanging clothes and a suitcase stored at the bottom.
	An outdoor patio area with wooden flooring and bordered by partial fencing. At the center of the patio is a rectangular outdoor table with a glass top, surrounded by matching wicker-style chairs, two of which are pulled slightly away from the table.	An outdoor patio features wooden flooring bordered by low, weathered wooden fencing with vertical slats. At the center of the space stands a rectangular outdoor table with a glass top, supported by a metal or wicker frame. Surrounding the table are four matching wicker-style chairs with cushions. On the table, a small ceramic planter with succulents and a half-filled water glass rest alongside an open book. A potted fern and a taller plant in a terracotta pot sit near the fence on the left, adding greenery to the scene

Figure 10. Examples of constructed Text-Image/Video/3D pair dataset with original caption and enriched caption.

Text-Image ($\mathcal{S}-\mathcal{I}$). We leverage the three image caption datasets: COCO caption [29], Conceptual (CC) caption [45], and VG [22] caption to build the Text-Image pair-wise SG. Specifically, following [21], we first employ the GPT-4o [37] to extract the triplets from the original caption. Then, we align entities in the triplets with entity classes of interest and align predicates in the triplets with predicate classes of interest. Finally, for the VG dataset with ISG annotation, we link the textual and visual objects through label matching. For the COCO and CC datasets without annotated SG, we ground the extracted triplets over relevant image regions to get localized triplets via state-of-the-art grounding methods, i.e., Grounded SAM [41]. Finally, we led to utilizing 64K images on the COCO caption dataset, 145K images on the CC caption dataset, and 57K images on the VG caption dataset. Furthermore, we utilize GPT-4o to rephrase and enhance captions, aiming to increase the diversity and richness of textual descriptions, guided by the following prompts:

Rephrase and Enrich captions

Input Data: textual captions
Instruction: From the given sentence, the task is to enrich the caption with reasonable scenes. Let's take a few examples to understand how to enrich the captions.

[Example-1]:
Input: A lady and a child near a park bench with kites and ducks flying in the sky and on the ground.
Output: A lady and a child near a park bench surrounded by lush greenery, with colorful kites soaring in the sky,

ducks flying overhead, and a few waddling on the ground. Nearby, a serene pond reflects the vibrant scene, and children can be seen playing in the background.

[Example-2]:
Input: Two men sit on a bench near the sidewalk and one of them talks on a cell phone.
Output: Two men sit on a wooden bench near a bustling sidewalk, shaded by nearby trees. One of them is engaged in a conversation on his cell phone, gesturing slightly with his free hand, while the other man sits calmly, gazing at passersby.
...

We present two examples in the first two rows of Fig. 10. After generating the enriched captions, we apply the aforementioned method used for the original captions to produce the final pairwise text-image SG annotations.

Text-Video ($\mathcal{S}-\mathcal{V}$). To construct the text-video pairwise USG dataset, we select 400 videos from ActivityNet [16], which includes dense caption annotations. Following the procedure for text-image pairs, we first extract triplets and align the entities and predicates with relevant concepts. Finally, we track textual objects in the videos by integrating frame-level object grounding results using Grounded SAM [41]. Additionally, we enrich the video captions to create textual SGs, incorporating partially nonliteral associations with the video content. We depict two examples in the middle two rows of Fig. 10.

Text-3D ($\mathcal{S}-\mathcal{D}$). To construct the text-3D pairwise USG dataset, we use the ScanRefer [5] dataset, which contains 46,173 descriptions of 724 object types across 800 ScanNet [9] scenes. Triplets are extracted from these descriptions using GPT-4o. Since ScanRefer provides object localizations, we directly align textual entities with 3D objects to establish associations between text and 3D data. Additionally, we enrich the textual descriptions to create partially overlapping text-3D USG datasets, enhancing diversity and coverage. Two examples are demonstrated in the last two rows of Fig. 10.

Image-Video ($\mathcal{I}-\mathcal{V}$). To construct the image-video pairwise USG dataset, we utilize the existing PVSG [60] video dataset. Specifically, we select the first frame of each video to construct frame-level ISGs. We then extract temporally non-adjacent video segments as the corresponding pairwise video. The associations between image objects and video objects are derived from the original PVSG annotations, ensuring accurate cross-modal connections.

Image-3D ($\mathcal{I}-\mathcal{D}$). To construct the Image-3D USG dataset, we leverage the existing 3DSG [52] dataset. Specifically, we randomly select 2D image views corresponding to 3D scenes. Using object annotations from the 3DSG dataset, we ground the objects in the selected images to obtain their positional information. The relationships among the detected objects are then derived from the original 3DSG annotations, resulting in complete SG annotations for the 2D image views. The associations between objects in the image and 3D scenes are determined by whether the 3D objects can be successfully grounded in the image. By integrating ISGs, 3DSGs, and association relations, we construct the final Image-3D pairwise USG dataset.

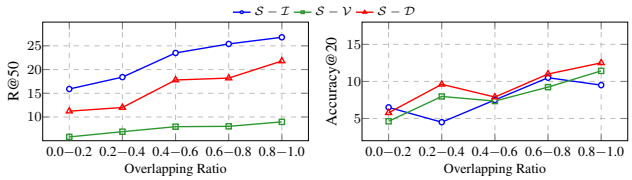

Figure 11. Performance of SGDet and association accuracy scores under varying overlap ratios between the two modalities.

E.2. Implementations

We initialize the text and image encoders using OpenCLIP [38], where the specific version of the image encoder is ConvNext-L. we design the pixel decoder by following the approach in [7, 27]. For the point encoder, we adopt PointBERT [63] as the initialization, and for the point decoder, inspired by [63], we implement a hierarchical propagation strategy with distance-based interpolation. After the encoding, all the features are projected into a 256-dimension using a linear layer. The mask decoder follows the design in [7]. We set the number of predefined learnable queries to 100. The number of layers L^{mask} is set as 9, with 3 transformer layers per scale. The object associator is implemented as a 3-layer CNN with a kernel size of 3×3. For the two-way relation-aware object/subject module, the number of layers, L_{RPC}, is set to 4. 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 $10e-4$ and a weight decay of $10e-4$. For the object detection loss weights, we set the $\lambda_{ce} = 5.0$ and $\lambda_{dice} = 5.0$, and $\lambda_{cls} = 2.0$ for predictions matched with ground truth and 0.1 for the "no object". In the final loss, we set the loss weights $\alpha = 1.0, \beta = 1.0$ and $\beta = 0.8, \eta = 0.6$.

F. Extended Experimental Results

We exhibit more experimental results here.

The Impact of the Overlap Ratio. We delve into the analysis of the overlap ratio and its crucial role in influencing the performance of USG-Par. To this end, the $\mathcal{S}-\mathcal{I}$, $\mathcal{S}-\mathcal{I}$, and $\mathcal{S}-\mathcal{I}$ datasets are divided into five groups with overlap ratios ranging from 0.0 to 1.0. The results, presented in Fig. 11, indicate that as the overlap ratio increases, the model achieves more accurate SG generation. This improvement can be attributed to the increased similarity between the two modalities, where complementary information enhances

$\mathcal{S}$	$\mathcal{I}$	$\mathcal{V}$	$\mathcal{D}$	$\mathcal{S}-\mathcal{I}$	$\mathcal{S}-\mathcal{V}$	$\mathcal{S}-\mathcal{D}$	$\mathcal{I}-\mathcal{V}$	$\mathcal{I}-\mathcal{D}$
28.4	36.9	5.3	38.7	21.6	14.5	7.9	10.7	18.7
25.1	×	×	×	-	-	-	-	-
27.6	34.0	×	×	16.0	-	-	-	-
28.1	34.9	4.9	×	19.2	14.0	-	-	-

Table 1. The ablation of modality unification. mR@20 scores are reported.

$\mathcal{I}-\mathcal{V}$	$\mathcal{I}-\mathcal{D}$
• *Only Corresponding Supervision.*	
25.1	18.4
• *No Supervision.*	
23.4	16.8

Table 2. Emergent zero-shot on the $\mathcal{I}-\mathcal{V}/\mathcal{D}$ dataset, where the model is only trained on $\mathcal{S}-\mathcal{I}/\mathcal{V}/\mathcal{D}$ data.

SG recognition performance. Furthermore, a higher overlap ratio allows the model to establish more precise associations between objects across modalities.

The Effect of Modality Unification. Tab. 1 compares the performance of our method across different numbers of unified modalities. The results demonstrate that incorporating additional modalities consistently improves performance across all datasets. This observation highlights the complementary nature of information across modalities when representing a scene. By unifying multiple modalities, our model effectively leverages this complementary information, resulting in enhanced performance compared to using a single modality alone.

The Emergent Capability of USG-Par. Tab. 2 presents the emergent zero-shot SG generation performance on $\mathcal{I}-\mathcal{V}$ and $\mathcal{I}-\mathcal{D}$, where USG-Par is trained solely on $\mathcal{S}-\mathcal{I}/\mathcal{V}/\mathcal{D}$. The results indicate that our model achieves performance comparable to supervised learning approaches. This demonstrates the strong capability of USG-Par to align image, video, and 3D modalities effectively, leveraging text as a unifying bridge.

More Visualizations. Here, we provide more visualizations of generated USG from various input modality combinations, including 1) text + image in Fig. 12, 2) text + video in Fig. 13, and 2) image + 3D in Fig. 14.

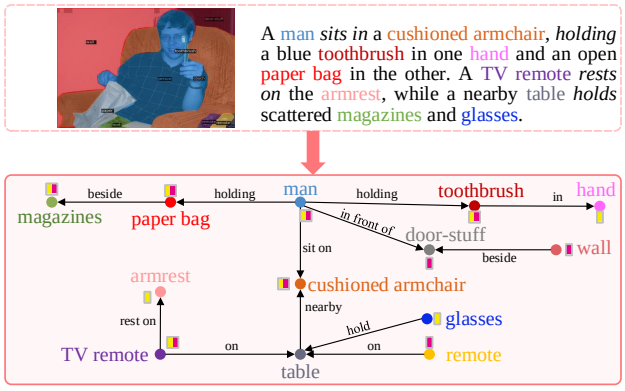

Figure 12. Visualization of USG generated from text and image

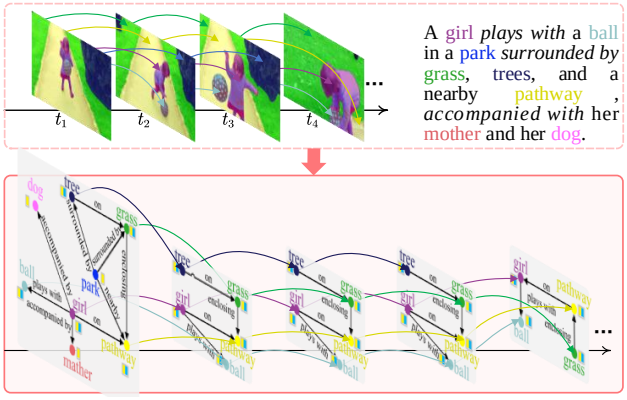

Figure 13. Visualization of USG generated from text and video

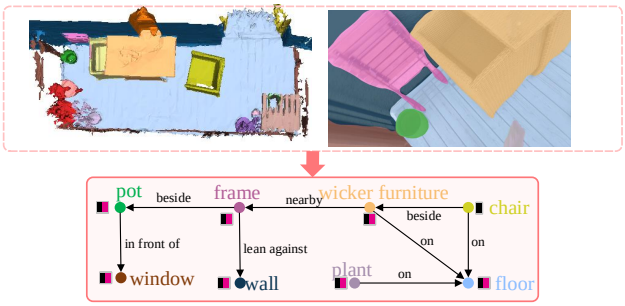

Figure 14. Visualization of USG generated from 3D and the corresponding view image.

References

[1] Anurag Arnab, Chen Sun, and Cordelia Schmid. Unified graph structured models for video understanding. In *ICCV*, pages 8097–8106, 2021. 3

[2] Anurag Arnab, Chen Sun, and Cordelia Schmid. Unified graph structured models for video understanding. In *ICCV*, pages 8097–8106, 2021. 4

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2

[4] Shivam Chandhok. Scenegpt: A language model for 3d scene understanding. *CoRR*, abs/2408.06926, 2024. 2

[5] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in RGB-D scans using natural language. In *ECCV*, pages 202–221, 2020. 8

[6] Zhanwen Chen, Saed Rezayi, and Sheng Li. More knowledge, less bias: Unbiasing scene graph generation with explicit ontological adjustment. In *WACV*, pages 4012–4021, 2023. 4

[7] 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. 5, 8

[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, 3

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 2432–2443, 2017. 8

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. *CoRR*, abs/1804.02748, 2018. 6

[11] Mingtao Feng, Haoran Hou, Liang Zhang, Yulan Guo, Hongshan Yu, Yaonan Wang, and Ajmal Mian. Exploring hierarchical spatial layout cues for 3d point cloud based scene graph prediction. *IEEE Transactions on Multimedia*, 2023. 4

[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. 4

[13] Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. In *WACV*, pages 5119–5128, 2023. 2

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *CVPR*, pages 18973–18990, 2022. 6

[15] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *ICRA*, pages 5021–5028, 2024. 4

[16] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 8

[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. 2, 6

[19] Bowen Jiang, Zhijun Zhuang, and Camillo Jose Taylor. Enhancing scene graph generation with hierarchical relationships and commonsense knowledge. *CoRR*, abs/2311.12889, 2023. 4

[20] 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. 2

[21] Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. LLM4SGG: large language models for weakly supervised scene graph generation. In *CVPR*, pages 28306–28316, 2024. 7

[22] 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. 2, 6, 7

[23] 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. 2

[24] Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. Compositional feature augmentation for unbiased scene graph generation. In *ICCV*, pages 21628–21638, 2023. 2

[25] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *CVPR*, pages 11109–11119, 2021. 2

[26] Rongjie Li, Songyang Zhang, and Xuming He. SGTR: end-to-end scene graph generation with transformer. In *CVPR*, pages 19464–19474, 2022. 3

[27] 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. 8

[28] 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

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 6, 7

[30] 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. 2

[31] Hengyue Liu and Bir Bhanu. Repsgg: Novel representations of entities and relationships for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2024. 2

[32] Hengyue Liu, Ning Yan, Masood S. Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *CVPR*, pages 11546–11556, 2021. 2

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[34] Yichao Lu, Himanshu Rai, Jason Chang, Boris Knyazev, Guang Wei Yu, Shashank Shekhar, Graham W. Taylor, and Maksims Volkovs. Context-aware scene graph generation with seq2seq transformers. In *ICCV*, pages 15911–15921, 2021. 2

[35] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K. Roy-Chowdhury. Unbiased scene graph generation in videos. In *CVPR*, pages 22803–22813, 2023. 3

[36] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *NeurIPS*, pages 2171–2180, 2017. 2

[37] OpenAI. Gpt-4 technical report. `https://openai.com/research/gpt-4`, 2023. `https://openai.com/research/gpt-4`. 7

[38] 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. 6, 8

[39] Sonia Raychaudhuri, Tommaso Campari, Unnat Jain, Manolis Savva, and Angel X. Chang. Reduce, reuse, recycle: Modular multi-object navigation. *CoRR*, abs/2304.03696, 2023. 4

[40] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 2

[41] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: assembling open-world models for diverse visual tasks. *CoRR*, abs/2401.14159, 2024. 7, 8

[42] 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. 2

[43] 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. 2

[44] 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. 6

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 7

[46] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 6

[47] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451, 2017. 4

[48] Gopika Sudhakaran, Devendra Singh Dhami, Kristian Kersting, and Stefan Roth. Vision relation transformer for unbiased scene graph generation. In *ICCV*, pages 21825–21836, 2023. 2

[49] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. In *AAAI*, pages 8968–8975, 2020. 4

[50] 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

[51] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *ICCV*, pages 13668–13677, 2021. 3, 4

[52] 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, 6, 8

[53] 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, 3

[54] 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. 2, 3

[55] 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, 4

[56] 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

[57] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. In *ICML*, 2024. 2

[58] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 3097–3106, 2017. 2, 6

[59] 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, 3, 6

[60] 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. 2, 6, 8

[61] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. In *IJCAI*, pages 1274–1280, 2021. 4

[62] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, pages 1068–1076, 2017. 4

[63] 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. 8

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

[65] Ce Zhang, Simon Stepputtis, Joseph Campbell, Katia P. Sycara, and Yaqi Xie. Hiker-sgg: Hierarchical knowledge enhanced robust scene graph generation. In *CVPR*, pages 28233–28243, 2024. 4

[66] 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

[67] Zhuosheng Zhang and Aston Zhang. You only look at screens: Multimodal chain-of-action agents. In *ACL*, pages 3132–3149, 2024. 2

[68] Chaofan Zheng, Xinyu Lyu, Yuyu Guo, Pengpeng Zeng, Jingkuan Song, and Lianli Gao. Learning to generate scene graph from head to tail. In *ICME*, pages 1–6, 2022. 2