Towards Flexible Visual Relationship Segmentation

Fangrui Zhu[1]   **Jianwei Yang**[2]   **Huaizu Jiang**[1]
[1]Northeastern University [2]Microsoft Research
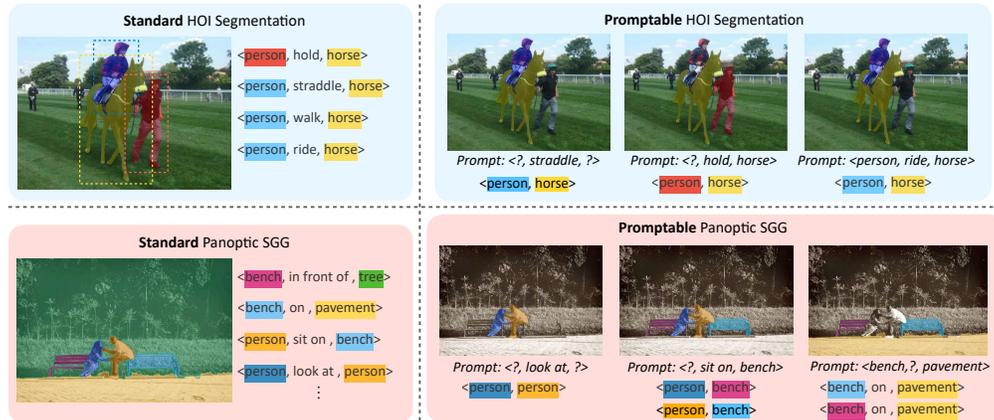`https://neu-vi.github.io/FleVRS`

Figure 1: `FleVRS` is a single model trained to support **standard**, **promptable** and **open-vocabulary** fine-grained visual relationship segmentation (`<subject mask`, relationship categories, `object mask>`). It can take images only or images with structured prompts as inputs, and segment all existing relationships or the ones subject to the text prompts.

Abstract

Visual relationship understanding has been studied separately in human-object interaction (HOI) detection, scene graph generation (SGG), and referring relationships (RR) tasks. Given the complexity and interconnectedness of these tasks, it is crucial to have a flexible framework that can effectively address these tasks in a cohesive manner. In this work, we propose `FleVRS`, a single model that seamlessly integrates the above three aspects in standard and promptable visual relationship segmentation, and further possesses the capability for open-vocabulary segmentation to adapt to novel scenarios. `FleVRS` leverages the synergy between text and image modalities, to ground various types of relationships from images and use textual features from vision-language models to visual conceptual understanding. Empirical validation across various datasets demonstrates that our framework outperforms existing models in standard, promptable, and open-vocabulary tasks, *e.g.*, **+1.9** mAP on HICO-DET, **+11.4** Acc on VRD, **+4.7** mAP on unseen HICO-DET. Our `FleVRS` represents a significant step towards a more intuitive, comprehensive, and scalable understanding of visual relationships.

1   Introduction

An image is not merely a collection of objects. Understanding the visual relationships between different entities at pixel-level through segmentation is a fundamental task in computer vision, which has broad applications in autonomous driving [28, 58], behavior analysis [65, 67], navigation [10, 15, 22, 27], *etc*. Furthermore, segmenting relational objects extends beyond mere detection, playing a

Method	Standard		Promptable	Open-vocabulary	One/Two-stage Model
	HOI	SGG			
RLIPv2 [92]	✓	✓	✗	✓	Two
UniVRD [101]	✓	✓	✗	✗	Two
SSAS [38]	✗	✗	✓	✗	One
GEN-VLKT [47]	✓	✗	✗	✓	One
FleVRS (Ours)	✓	✓	✓	✓	One

Table 1: **Comparisons with previous representative methods in three aspects of model capabilities**. To the best of our knowledge, our `FleVRS` is the first one-stage model capable of performing standard, promptable, and open-vocabulary visual relationship segmentation all at once.

crucial role in improving visual understanding and providing a more comprehensive abstraction on the visual contents and interactions among them.

Ideally, a visual relationship segmentation (VRS) model should demonstrate flexibility across three key dimensions. 1) **Capability of segmenting various types of relationships**, including both human-centric and generic ones. These relationships are defined as triplets in the form of `<subject, predicate, object>`. Human-object interaction (HOI) detection [4, 18], which we adapt into HOI segmentation in our work, exemplifies this capability, such as `<person, ride, horse>` in Fig. 1. Panoptic scene graph generation (SGG) [55, 81, 85], captures generic spatial or semantic relationships among pairs of objects in an image, *e.g.*, `bench on pavement` in Fig. 1. A unified model that can handle these tasks concurrently is essential, as it eliminates the need for separate designs and modifications for each specific task. 2) **Grounding of relational subject-object pairs with different prompts**. Given various textual prompts, the model should output the desired entities and relationships, facilitating a more natural and intuitive user interface. For instance, it should be able to detect just the `person` in an image or all possible interactions between a `person` and a `horse`, as illustrated in Fig. 1. 3) **Open-vocabulary recognition of visual relationships**. In realistic open-world applications, the model should generalize to new scenarios without requiring annotations for new concepts not seen during training. This capability includes detecting novel objects, relationships, and their combinations.

Existing models in visual relationship segmentation (VRS) have targeted aspects of the desired capabilities but fall short of providing a comprehensive solution, as detailed in Tab. 1. Models have typically focused on tasks like human-object interaction (HOI) detection [23, 35, 47, 57, 72, 95, 97, 108] and panoptic SGG [50, 71, 81, 85, 93, 104]. Although models such as [92, 101] have attempted to unify VRS under a single framework, they need additional pretraining on HOI datasets (Tab. 1) and lack features such as promptable segmentation, which allows for dynamic entity and relationship generation based on textual prompts, as well as capabilities for open-vocabulary promptable segmentation. Efforts to detect instances referred to by textual prompts have been made [21, 38, 75, 107], but these models fail to capture all desired entities or relationships comprehensively and struggle with classifying multi-label interactions between the pairs, limiting their effectiveness in complex scenarios. Although recent vision-language grounding models like [29, 52, 83] and multimodal large language models such as [2, 5, 76, 78, 88] exhibit enhanced capabilities in grounding instances specified by free-form text and show strong generalization over novel concepts, they still do not generate the required pairs in the format of segmentation masks. Furthermore, these models require significant computational resources and additional vision models for precise localizations. For open-vocabulary VRS, existing works [47, 91, 92] leverage textual embeddings to transfer knowledge. However, models [91, 92] fall short in grounding diverse prompts, while [47] is exclusively designed for HOI detection, not generic VRS.

To address the limitations in existing models, we introduce `FleVRS`, a flexible one-stage framework capable of performing standard, promptable, and open-vocabulary visual relationship segmentation *simultaneously*. Our approach integrates human-centric (HOI segmentation) and generic VRS (Panoptic SGG) by adopting SAM [36] to unify different types of annotations into segmentation masks and using a query-based Transformer architecture that outputs triplets in the format `<subject, predicate, object>`. The model enhances its interactive capabilities by accepting textual prompts as inputs. These prompts are converted into textual queries that assist the decoder in accurately identifying and localizing objects within the relationships. Additionally, we unify the labels from different datasets into a shared textual space, transforming classification into a process of matching with a set of textual features. Leveraging textual features from the CLIP model [64], we enable the effective matching of visual features with textual knowledge of novel concepts. This design

Figure 2: **Examples of converting HOI detection boxes to masks.** We filter out low-quality masks during training by computing IoU between the mask and box.

inherently supports open-vocabulary and promptable relationship segmentation without pre-defining the number of object or predicate categories, facilitating dynamic and extensive adaptability.

Our `FleVRS` proposes a unified framework that integrates standard, promptable, and open-vocabulary VRS tasks into a single system, as detailed in Tab. 1, providing greater flexibility compared to existing methods. It employs a mask-based approach to effectively manage various VRS tasks, enabling adaptation to different types of annotations, including HOI detection and panoptic SGG. Our architecture incorporates dynamic prompt handling, which supports both prompt-based and open-vocabulary settings, allowing our model to combine promptable queries with open-vocabulary capabilities to ground novel relational objects.

We evaluate our `FleVRS` on standard, promptable, and open-vocabulary VRS tasks, *i.e.*, HOI segmentation [4, 18] and panoptic SGG [85]. Crucially, we demonstrate competitive performance from three perspectives – standard (**40.5** *vs.* **39.1** mAP on HICO-DET [4]), promptable (**56.8** *vs.* **33.5** sIoU on VRD [55]), and open-vocabulary (**31.7** *vs.* **25.6** mAP for "unseen object" on HICO-DET [4]) visual relationship segmentation.

In summary, our main contributions are as follows: 1) We introduce a flexible one-stage framework capable of segmenting both human-centric and generic visual relationships across various datasets. 2) We present a promptable visual relationship learning framework that effectively utilizes diverse textual prompts to ground relationships. 3) We demonstrate competitive performance in both standard close-set and open-vocabulary scenarios, showcasing the model's strong generalization capabilities.

2 Related Work

Visual Relationship Detection (VRD) is split into two lines of works, including human-object interaction (HOI) detection [4, 18] and panoptic scene graph generation (SGG) [37, 85]. They are defined as detecting triplets in the form of `<subject, predicate, object>` triplet, where `subject` or `object` includes object box and category. HOI detection aims to detect human-centric visual relationships, while PSG focuses on generic object pairs' relationships. Previous works [1, 13, 16, 32, 35, 41, 41, 47, 57, 81, 89, 95, 96, 97, 100, 104, 105, 106] usually train specialist models on a single data source and tackle them separately. Departing from this traditional bifurcation, UniVRD [101] initiated the development of a unified model for VRD, with subsequent efforts like [91, 92] advancing relational understanding through large-scale language-image pre-training. Unlike the two-stage approach of [92], which performs object detection before decoding visual relationships, our method employs a one-stage design that decodes objects and their relationships simultaneously. Crucially, our model extends beyond standard VRD capabilities to support promptable and open-vocabulary visual relationship segmentation, enhancing detailed scene comprehension.

Referring relationship and visual grounding. The most relevant work to ours is referring visual relationship introduced in [38], where the model detects the subject and object depicting the structured relationship `<subject, predicate, object>`. One-stage [38, 75], two-stage [63, 107] and three-stage [21] methods are proposed to localize the two entities' boxes iteratively based on the given structured prompt `<subject, predicate, object>`. Unlike these methods, our approach allows for more *flexible* textual prompts without requiring the complete specification of the triplet. As shown in Fig. 1, our model can handle queries that include a single item (e.g., `predicate`) or a combination of two (e.g., `predicate` and `object`). Additionally, our method is capable of performing standard and open-vocabulary VRS. Visual grounding represents another related area, where models output bounding boxes [7, 8, 20, 29, 40, 52, 83] or object masks [9, 17, 44, 45, 51, 56, 79, 87, 99] in response to textual inputs. This process requires reasoning over the entities mentioned in the text to identify the corresponding objects in the visual space. However, our task fundamentally differs from this.
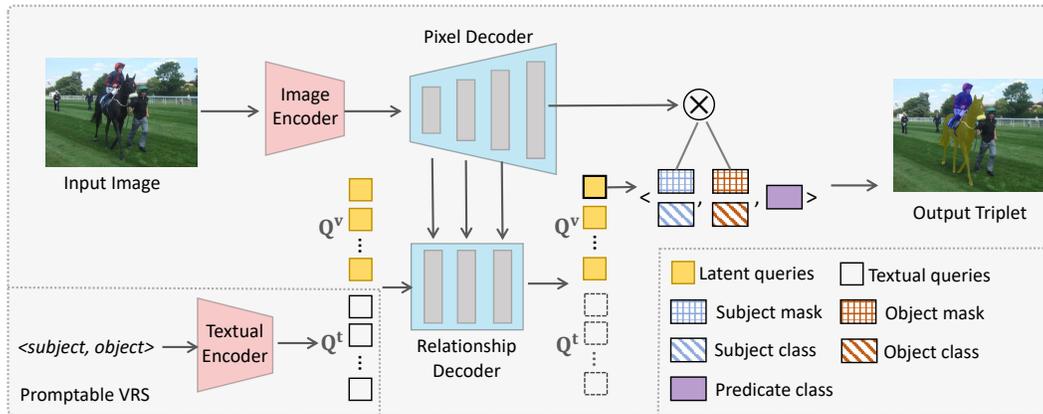
Figure 3: **Overview of** `FleVRS`. In standard VRS, without textual queries, the latent queries perform self- and cross-attention within the relationship decoder to output a triplet for each query. For promptable VRS, the decoder additionally incorporates textual queries $\mathbf{Q^t}$, concatenated with $\mathbf{Q^v}$. This setup similarly predicts triplets, each based on $\mathbf{Q^v}$ outputs aligned with features from the optional textual prompt $\mathbf{Q^t}$.

In our `FleVRS`, the promptable VRS task goes beyond mere identification; it involves outputting segmentation masks for both subject and object pairs along with categorizing their relationships. This capability is essential for understanding and interpreting complex relational dynamics.

Vision and language models Recent advancements in large-scale pre-trained vision-language models (VLM) [39, 64, 68, 69, 82, 94] and multimodal large language models (MLLM) [2, 5, 76, 78, 88] have demonstrated impressive performance and generalization capabilities across a variety of vision and multimodal tasks [36, 109, 110]. However, these models primarily focus on entity-level generalization, with open-vocabulary VRS receiving less attention. While recent efforts in zero-shot HOI detection [47, 61, 80, 92] often utilize CLIP [64] for category classification, their open-vocabulary capabilities lack the flexibility needed for prompt-driven input. Although current VLMs and MLLMs are adept at grounding novel concepts from text, they require significant computational resources and additional visual models, and cannot directly generate comprehensive segmentation masks for subject and object pairs. In contrast, our `FleVRS` provides a lightweight solution that effectively supports various types of open-vocabulary VRS, enabling category classification and the integration of novel concepts from prompts.

3 Method

3.1 Overview

Standard VRS. Given an image $\mathbf{I}$, the goal of *standard* visual relationship segmentation (VRS) is to detect all the visual relationships of interest, either human-centric (*i.e.*, HOI detection) or the generic ones (SGG), in terms of triplets in the form of `<subject, predicate, object>` (masks and object categories of `subject` and `object`, and the `predicate` category). The `subject` is always `human` in HOI detection, whereas it can be any type of object in SGG (may or may not be `human`). We consider the panoptic setting [85] of SGG, where a model needs to generate a more comprehensive scene graph representation based on panoptic segmentation rather than rigid bounding boxes, providing a clear and precise grounding of objects. To produce fine-grained masks, we convert existing bounding box annotations from HOI detection datasets [4, 18] into segmentation masks using the foundation model SAM [36], as illustrated in Fig. 2. We employ a filtering approach based on Intersection over Union (IoU) to filter out inaccurate masks. Details are in Appendix.

Promptable VRS. Our `FleVRS` optionally accepts textual prompts as inputs, enabling users to specify visual relationships for promptable VRS. It accommodates three types of structured text prompts: a single element (e.g., `<?, predicate, ?>`), any two elements (e.g., `<subject, predicate, ?>`, `<subject, ?, object>`), or all three elements. Consequently, the model outputs only the triplets that match the specified elements in the prompt, as depicted in the right column of Fig. 1. Without textual prompts, it functions as a standard VRS model, exhaustively generating all possible triplets, illustrated in the left part of Fig.1.

Open-vocabulary VRS. In practice, it's essential for a VRS model to adapt to new concepts, including new categories of entities (*i.e.*, `subject` and `object`), `predicates`, and their various combinations. Expanding these concept vocabularies to encompass a wider range is particularly challenging due to the vast potential combinations and the long-tail distribution of these categories. Thus, our goal is to equip the model to operate in an open-vocabulary setting, where it can effectively handle these diversities. It it important to note that the above three capabilities are complementary; for example, the text prompts in promptable VRS can include novel object or predicate categories.

To this end, we propose integrating the above three aspects into a *single unified* framework. Since these settings are complementary, a general-purpose model should be capable of performing various combinations of these three functions. Additionally, their inherent similarities make it more intuitive to consolidate them within a flexible, unified approach.

3.2 Model Architecture

Inspired by the success of Transformer-based segmentation models [6, 109], we design a dual-query system for our VRS model, illustrated in Fig. 3. Latent queries, a set of learned embeddings, generate triplets (which may be empty) to formulate output masks and relationship categories. For promptable VRS, textual queries derived from input prompts are incorporated. We employ an image encoder and a pixel decoder to extract visual features, coupled with a relationship decoder that processes `<subject, object>` pairs and their interrelations. For open-vocabulary VRS, our approach shifts from traditional classification to a matching strategy that aligns visual and textual features for both object and predicate categories, enhancing the model's adaptability to new concepts. Each component of this architecture is elaborated further below.

Image Encoder. Specifically, given the image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, it is first fed into the image encoder $\mathbf{Enc_I}$ to obtain multi-scale low-resolution features $\mathbf{F} = \left\{ \mathbf{F}_s \in \mathbb{R}^{C_\mathbf{F} \times \frac{H}{s} \times \frac{W}{s}} \right\}$, where the stride of the feature map $s \in \{4, 8, 16, 32\}$, and $C_\mathbf{F}$ is the number of channels.

Pixel Decoder. A Transformer-based pixel decoder $\mathbf{Dec_P}$ is used to upsample $\mathbf{F}$ and gradually generate high-resolution per-pixel embeddings $\mathbf{P}$. $\mathbf{P}$ is then passed to the relationship decoder $\mathbf{Dec_R}$ to compute cross-attention with query features.

Textual Encoder. When a text prompt is provided for promptable VRS, we use the textual encoder $\mathbf{Enc_T}$ to encode it into a set of textual queries $\mathbf{Q^t} \in \mathbb{R}^{N_t \times C_q}$, where N_t is the number of tokens in the textual queries, and C_q denotes the channel number of query features. In practice, we use the textual encoder from CLIP [64] as $\mathbf{Enc_T}$. The format of the text prompt can be a single item (*e.g.* "*<p>predicate</p>*"), two of them ("*<s>subject</s><p>predicate</p>*"), or all three of them, where "*predicate*" and "*subject*" denote category names of `predicate` and `subject`, respectively. "*<s>*", "*<o>*", "*<p>*" are used as separate tokens between `subject`, `predicate` and `object` in the text prompt. We could use natural language as the textual prompt instead of using a structured format. However, collecting the textual VRD data is not trivial, and we leave it as an extension of our model in future work.

Relationship Decoder. The relationship decoder $\mathbf{Dec_R}$, based on a Transformer decoder design, processes pixel decoder outputs $\mathbf{P}$ and latent queries $\mathbf{Q^v}$ to generate all possible triplets for standard VRS. Inside, masked attention [6] utilizes masks from earlier layers for foreground information. Each $\mathbf{Q^v}$ output feeds into five parallel heads: two mask heads for subject and object masks (M_s, M_o), two class heads for their categories (C_s, C_o), and another class head for relationship prediction (C_p) During training, Hungarian matching aligns predicted triplets with ground truth. For standard VRS inference, triplets above a confidence threshold are considered final predictions. For promptable VRS, $\mathbf{Enc_T}$ transforms text prompts into textual queries $\mathbf{Q^t}$ that are concatenated with $\mathbf{Q^v}$ and input into $\mathbf{Dec_R}$. This process, which uses self- and cross-attention mechanisms, generates `<subject, predicate, object>` triplets, similar to standard VRS. An additional matching loss during training ensures the model predicts triplets as specified by the text prompt. During inference, we calculate similarity scores between the textual query feature (last token's feature of $\mathbf{Q^t}$) and the latent query outputs. We then select entities and relationships specified in the textual prompt from the top k triplets for the final outputs.

Matching with textual features. To enable open-vocabulary VRS, our `FleVRS` uses the CLIP textual encoder [64] to match visual features with candidate textual features for object and predicate

categories. We convert these categories into textual features using prompt templates, such as "*A photo of [predicate-ing]*" for HOI segmentation and "*A photo of something [predicate-ing] (something)*" for panoptic SGG.[1] The model computes matching scores between predicted class embeddings and these textual features, allowing classification beyond the fixed vocabulary of the training set and facilitating open-vocabulary VRS. Textual prompts are similarly encoded, and their features are used to calculate similarity scores for promptable VRS inference.

3.3 Loss functions

We use Hungarian matching during training to find the matched triplets with ground truth ones. For standard VRS, we compute focal losses $\mathcal{L}_b^s$, $\mathcal{L}_b^o$ and dice losses $\mathcal{L}_d^s$, $\mathcal{L}_d^o$ on `subject` and `object` mask predictions, cross-entropy losses $\mathcal{L}_c^s$, $\mathcal{L}_c^o$, $\mathcal{L}_c^p$ on `subject`, `object`, and `predicate` category classifications, which can be written as

$$\mathcal{L} = \lambda_b \sum_{i \in \{s,o\}} \mathcal{L}_b^i + \lambda_d \sum_{j \in \{s,o\}} \mathcal{L}_d^j + \sum_{k \in \{s,o,p\}} \lambda_c^k \mathcal{L}_c^k, \tag{1}$$

where λ_b, λ_d, and λ_c are hyper-parameters for adjusting the weights of each loss. λ_c^s, λ_c^o, λ_c^p are different classification loss weights for `subject`, `object`, and `predicate`. For promptable VRS, we adopt an additional matching loss $\mathcal{L}_g$ between the matched triplet class embedding and the textual query feature (the last token feature of $\mathbf{Q^t}$), which is in the form of cross-entropy loss. The final training loss is written as

$$\mathcal{L} = \lambda_b \sum_{i \in \{s,o\}} \mathcal{L}_b^i + \lambda_d \sum_{j \in \{s,o\}} \mathcal{L}_d^i + \sum_{k \in \{s,o,p\}} \lambda_c^k \mathcal{L}_c^k + \lambda_g \mathcal{L}_g, \tag{2}$$

where λ_g controls the weight of $\mathcal{L}_g$. $\mathcal{L}_c$ depends on the text prompt. For example, given <`subject`, `predicate`>, there will not have $\mathcal{L}_c^s$ and $\mathcal{L}_c^p$ terms in Eq. (2), with `subject` and `predicate` categories being given. See the appendix for the concrete values of loss weights.

4 Experiments

4.1 Experimental Settings

Datasets For HOI segmentation, we utilize two public benchmarks: HICO-DET [4] and V-COCO [18]. To fit Our `FleVRS`, we use SAM [36] to transform box annotations into masks and apply Non-Maximum Suppression (NMS) to remove overlapping masks with an IoU threshold greater than 0.1. We omit `no_interaction` annotations from HICO-DET due to incomplete annotation, leaving 44,329 images (35,801 training, 8,528 testing) with 520 HOI classes from 80 objects and 116 actions.[2] V-COCO is built from COCO [49], comprising 10,396 images (5,400 training, 4,964 testing), featuring 80 objects and 29 actions, and includes 263 HOI classes. Both datasets align with COCO's object categories. For panoptic SGG, we use the PSG dataset [85], sourced from COCO and VG [37] intersections, containing 48,749 images (46,572 training, 2,177 testing) with 133 objects and 56 predicates.

Data Structure for open-vocabulary HOI segmentation Following prior studies [3, 23], we evaluate HICO-DET under three scenarios: (1) Unseen Composition (UC), where some HOI classes are absent despite all object and verb categories being present; (2) Unseen Object (UO), where certain object classes and their corresponding HOI triplets are excluded from training; and (3) Unseen Verb (UV), where specific verb classes and their associated triplets are similarly omitted. In UC, the Rare First (RF-UC) approach targets tail HOI classes, while Non-rare First (NF-UC) focuses on head categories. Originally, UC included 120/480/600 categories for unseen/seen/full sets, which reduces to 115/405/520 after removing `no_interaction` annotations. For UO, we select 12 unseen objects from 80, resulting in 88/432 unseen/seen HOI categories.

Evaluation Metric For standard HOI segmentation, we convert the predicted masks to bounding boxes to compare with current methods, and follow the setting in [4] to use the mean Average Precision (mAP) for evaluation. We also turn the outputs of other methods into masks and report

---

[1] Omit "*something*" for spatial relationships.
[2] interchangeable with "verb", "predicate".

(a) Text prompt: <?, object, predicate>

<?, flip, skateboard>

<?, wash, bus>

<?, ride, boat>

<?, hold, baseball glove>

(b) Text prompt: <subject, predicate, ?>

<person, hold, ?>

<person, pull, ?>

<person, hold, ?>

<person, straddle, ?>

(c) Text prompt: <subject, ?, object>

<person, ?, motorcycle>
[ride, sit on]

<person, ?, laptop>
[hold, read]

<person, ?, kite>
[hold, carry]
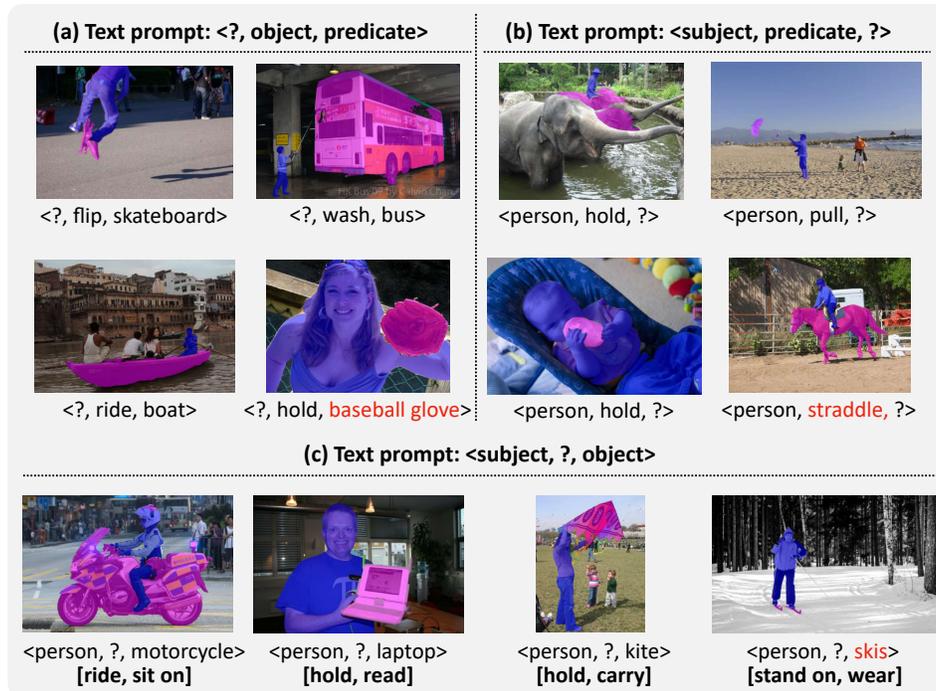
<person, ?, skis>
[stand on, wear]

Figure 4: **Qualitative results of promptable VRS on HICO-DET [4] test set.** We show visualizations of `subject` and `object` masks and relationship category outputs, given three types of text prompts. In (c), we show the predicted predicates in bold characters. Unseen objects and predicates are denoted in red characters.

mask mAP for thorough comparison. An HOI triplet prediction is a true positive if (1) both predicted human and object bounding boxes/masks have IoU larger than 0.5 *w.r.t.* GT boxes/masks; (2) Both the predicted object and verb categories are correct. For HICO-DET, we evaluate the three different category sets: all 520 HOI categories (Full), 112 HOI categories (less than 10 training instances) (Rare), and the other 408 HOI categories (Non-Rare). For VCOCO, we report the role mAPs in two scenarios: (1) S1: 29 actions including 4 body motions; (2) S2: 25 actions without the no-object HOI categories. For standard panoptic SGG, following [85], we use $R@K$ and $mR@K$ metrics, which calculate the triplet recall and mean recall for every predicate category, given the top K triplets from the model. A successful recall requires both subject and object to have mask-based IoU larger than 0.5 compared to their GT masks, with the correct predicate classification in the triplet.

Implementation Details Following [6, 109], we use 100 latent queries and 9 decoder layers in the relationship decoder. We adopt Focal-T/L [84] for the Image Encoder and DaViT-B/L for the pixel decoder. We use the textual encoder from CLIP to encode input text prompt and subject, object, and predicate categories. During training, we set the input image to be 640×640, with batch size of 64. We optimize our network with AdamW [54] with a weight decay of 10^{-4}. We train all models for 30 epochs with an initial learning rate of 10^{-4} decreased by 10 times at the 20th epoch. To improve training efficiency, we initialize Our `FleVRS` using the pre-trained weights from [109]. For all experiments, the parameters of the textual encoder are frozen except its logit scales. The loss weights λ_b, λ_d, λ_c and λ_{grd} (superscript omitted) are set to 1,1,2, and 2. More details are in the appendix.

4.2 Standard VRS

We evaluate our method on three benchmarks, *i.e.* HICO-DET [4], VCOCO [18] for HOI segmentation, and PSG [85] for the panoptic SGG.

HOI segmentation Since Our `FleVRS` leverages mask supervision, either converting mask results into bounding boxes or transforming bounding boxes from previous methods' output into masks does not facilitate a completely equitable comparison. For the utmost fairness in comparison, we report both box mAP and mask mAP from the above ways. As shown in Table 2, Our `FleVRS` shows superior

Model	Backbone	Default (%)		
		box/mask mAP$_F$	box/mask mAP$_R$	box/mask mAP$_N$
Bottom-up methods				
SCG [96]	ResNet-50	31.3 / 31.3	24.7 / 25.0	33.3 / 35.5
UPT [97]	ResNet-101	32.6 / 34.9	28.6 / 29.4	33.8 / 36.1
STIP [100]	ResNet-50	32.2 / 30.8	28.2 /28.6	33.4 / 32.5
ViPLO [60]	ViT-B	<u>37.2</u> / <u>39.1</u>	<u>35.5</u> / <u>37.8</u>	<u>37.8</u> / <u>39.7</u>
Additional training with object detection data				
UniVRD [101]	ViT-L	37.4 / -	28.9 / -	39.9 / -
PViC [98]	Swin-L	44.3 / -	44.6 / -	44.2 / -
RLIPv2 [92]	Swin-L	45.1 / 48.6	45.6 / 44.3	43.2 / 49.8
Single-stage methods				
HOTR [33]	ResNet-50	25.1 / 26.5	17.3 / 18.5	27.4 / 29.0
QPIC [70]	ResNet-101	29.9 / 30.5	23.0 / 23.1	31.7 / 33.1
CDN [95]	ResNet-101	32.1 / 33.9	27.2 / 28.9	33.5 / 36.0
RLIP [91](VG+COCO)	ResNet-50	32.8 / 34.4	26.9 / 27.7	34.6 / 36.5
GEN-VLKT [47]	ResNet-101	35.0 / 35.6	31.2 / 32.6	36.1 / 37.8
ERNet [48]	EfficientNetV2-XL	35.9 / -	30.1 / -	38.3 / -
MUREN [35]	ResNet-50	32.9 / 35.4	28.7 / 30.1	34.1 / 37.6
Ours	Focal-L	**38.1 / 40.5**	**33.0 / 34.9**	**39.5 / 42.4**

Table 2: **Quantitative results on the HICO-DET test set.** We report both box and mask mAP under the *Default* setting [4] containing the *Full* (F), *Rare* (R), and *Non-Rare* (N) sets. `no_interaction` class is removed in mask mAP. The best score is highlighted in bold, and the second-best score is underscored. '-' means the model did not release weights and we cannot get the mask mAP. Due to space limit, we show the complete table with more models in the appendix.

Model	Backbone	AP$_{role}^{S\#1}$	AP$_{role}^{S\#2}$
Bottom-up methods			
VSGNet [72]	ResNet-152	51.8 / -	57.0 / -
ACP [34]	ResNet-152	53.2 / -	- / -
IDN [43]	ResNet-50	53.3 / -	60.3 / -
STIP [100]	ResNet-50	**66.0** / 66.2	**70.7 / 70.5**
Additional training with object detection data			
UniVRD [101]	ViT-L	65.1 / -	66.3 / -
PViC [98]	Swin-L	64.1 / -	70.2 / -
RLIPv2 [92]	Swin-L	72.1 / 71.7	74.1 / 73.5
Single-stage methods			
HOTR [33]	ResNet-50	55.2 / 55.0	64.4 / 64.1
DIRV [11]	EfficientDet-d3	56.1 / -	- / -
CDN [95]	ResNet-101	63.9 / 61.3	65.8 / 63.2
RLIP [91]	ResNet-50	61.9 / 61.3	64.2 / 64.0
GEN-VLKT [47]	ResNet-101	63.6 / 61.8	65.9 / 64.0
ERNet [48]	EfficientNetV2-XL	64.2 / -	- / -
Ours	Focal-L	<u>65.2</u> / **66.5**	<u>66.5</u> / <u>67.9</u>

Table 3: **Quantitative results on V-COCO.** We report both box and mask mAP. The best score is highlighted in bold, and the second-best score is underscored. '-' means the model did not release weights and we cannot get the mask mAP. Due to space limit, we show the complete table with more models in the appendix.

performance over current single-stage methods in terms of box and mask mAP on HICO-DET. We also achieve competitive performance on VCOCO [18], as shown in Table 3. The advantages of Our `FleVRS` come from: (1) one-stage Transformer-based design with fine-grained training supervision for VRS. With subject and object masks, the model has more accurate supervision, compared with box annotations that contain redundancy [85]. (2) good language-visual alignment with the large-scale pretrained model [64]. Our `FleVRS` achieves competitive results without additional training on large-scale detection datasets [101]. Among one-stage HOI methods, our approach is simpler and able to tackle different datasets without modifications to the structure.

Panoptic SGG From Table 4, Our `FleVRS` can achieve competitive results in terms of $R@50$ and $R@100$ without elaborated designs for PSG, compared with most of previous work. Our `FleVRS` is not superior to HiLo [104], which is mainly due to the long-tail distribution of the dataset and the

Method	Backbone	R/mR@20	R/mR@50	R/mR@100
Adapted from SGG methods				
IMP [81]	VGG-16	17.9 / 7.35	19.5 / 7.88	20.1 / 8.02
MOTIFS [93]	VGG-16	20.9 / 9.60	22.5 / 10.1	23.1 / 10.3
VCTree [71]	VGG-16	21.7 / 9.68	23.3 / 10.2	23.7 / 10.3
GPSNet [50]	VGG-16	18.4 / 6.52	20.0 / 6.97	20.6 / 7.17
One-stage PSG methods				
PSGTR [85]	ResNet-101	**28.2 / 15.4**	**32.1 / 20.3**	**35.3 / 21.5**
PSGFormer [85]	ResNet-101	18.0 / 14.2	20.1 / 18.3	21.0 / 19.8
Training with additional data				
HiLo [104]	Swin-L	40.6 / 29.7	48.7 / 37.6	51.4 / 40.9
Ours	Focal-L	<u>27.0 / 15.4</u>	<u>31.0 / 18.3</u>	<u>31.7 / 18.8</u>

Table 4: **Quantitative results on PSG.** The best score is highlighted in bold, and the second-best score is underscored.

limitation of using CLIP to encode abstract relationships (*e.g.*, entering, exiting). The model tends to predict high-frequency relationships and is hard to understand and predict low-frequency ones.

Ablation study. We ablate Our `FleVRS` by testing different encoding strategies for relationships via the textual encoder in 7. Specifically, we compare encoding `object` and `predicate` categories as `<person, predicate, object>` triplets or separately, associating the results with either triplet cross-entropy (CE) loss or disentangled CE loss. Results reveal that while HICO-DET benefits from the disentangled CE loss, allowing better generalization to novel concepts, VCOCO performs better with triplet CE loss due to the challenge of distinguishing verbs without corresponding objects in various contexts (*e.g.*, differentiating "eat" in "a person eating an apple" *vs* "a person eating"). Further experiments with various backbones demonstrate performance enhancements with larger models. Additionally, incorporating a box head for supervision alongside mask supervision enhances performance, which is attributed to the masked attention mechanism inspired by [6]. Exploring the potential synergies of training across multiple datasets, we find that while unified training improves VCOCO's performance due to its smaller size, HICO-DET and PSG show limited gains. This disparity is likely due to the different predicate categories used in PSG compared to HICO-DET and VCOCO.

Comparison with previous works. UniVRD uses a two-stage approach, where the model first detects independent objects and then decodes relationships between them, retrieving boxes from the initial detection stage. In contrast, our method employs a one-stage approach, where each query directly corresponds to a <subject, object, predicate> triplet. This transition improves time efficiency from O(MxN) to O(K), where M is the number of subject boxes, N is the number of object boxes, and K is the number of interactive pairs. Our approach also provides greater flexibility by learning a unified representation that encompasses object detection, subject-object association, and relationship classification in a single model.

In terms of training data, we use much fewer training data (x50 less, without using VG [37] and Objects365 [66]) and Our `FleVRS` with the Focal-L [84] backbone is much smaller than UniVRD [101] (164M vs 640M) with LiT(ViT-H/14), we achieve comparable results(37.4 vs 38.1 on HICO-DET). While our method does not match RLIPv2 [92] in performance, this is due to different design philosophies and goals. RLIPv2 is a two-stage approach optimized for large-scale pre-training and relies on separately trained detectors. Our `FleVRS`, however, is not designed for pretraining and does not include a separately trained detector.

	No subject	No object	Only predicate	
	S-IoU	O-IoU	S-IoU	O-IoU
Conv-based methods				
VRD [55]	0.208	0.008	0.024	0.026
SSAS [38]	0.335	0.363	0.334	0.365
Ours	**0.568**	**0.364**	**0.556**	**0.366**

Table 5: Comparison of promptable VRD results with the baseline on VRD dataset [55].

Our focus is on enhancing the flexibility of the VRS model without *directly* training on extensive curated data(x50 more, VG and Objects365). Thus, the differences in performance are attributed to the scale and design objectives. We further discuss the FLOPs and the number parameters of the backbone compared to previous works in the Appendix.

4.3 Promptable VRS

We evaluate the ability of promptable VRS on the VRD dataset [55], to compare with [38]. As in Table 5, Our `FleVRS` can locate entities given flexible text query inputs and performs better localizing

subjects and objects. Our `FleVRS` gets particular better results on localizing subjects (0.568 *vs* 0.335, 0.556 *vs* 0.334), which is mainly because there are fewer categories in subjects compared with objects and lots of subjects are humans, making it easier to segment subjects. We further evaluate our promptable VRS approach on HICO-DET and PSG, as they contain rich relationship labels. Since there are no previous baselines, we show qualitative results in Fig. 4. We visualize the subject and object masks with the highest matching score for each example. We can see that the model is able to localize `subject` and `object` masks and predict their relationships given the structured textual prompt. We further perform postprocessing way to search triplets from standard VRS output, which serves as another baseline to show the effectiveness of our method. Please refer to section E in the appendix for results and discussions of fair comparison.

Difference with standard REC tasks The referring expression comprehension (REC) tasks on benchmarks like RefCOCO [30], RefCOCO+ [59], and RefCOCOg [90] are designed to detect objects based on free-form textual phrases, such as "a ball and a cat" or "Two pandas lie on a climber." In contrast, the promptable VRS task in our work focuses on detecting subject-object pairs within a structured prompt format, such as `<?, sit_on, bench>` or `<person, ?, horse>`, as illustrated in Fig. 1 of the main paper. Our `FleVRS` is designed to encode and compute similarity scores for each of these elements separately. Our primary focus is on relational object segmentation based on a single structured query, which differs significantly from the objectives of REC benchmarks.

4.4 Open-vocabulary VRS

We conduct open-vocabulary experiments following the defined zero-shot HOI detection setting [23, 25, 26, 47] on HICO-DET. As shown in Table 6, Our `FleVRS` surpasses previous single-dataset methods across all settings, with its open-vocabulary capabilities stemming from the knowledge transferred from CLIP [64]. GEN-VLKT [47] also leverages CLIP to facilitate open-vocabulary capabilities by encoding `<person, predicate, object>` as a triplet and using it for HOI category classification. In contrast, our approach separates the encoding of `predicate` and `object`, enhancing the model's generalization ability over novel concepts.

Method	Unseen	Seen	Full
Rare First Unseen Composition			
VCL [24]	10.06	24.28	21.43
ATL [25]	9.18	24.67	21.57
FCL [26]	13.16	24.23	22.01
GEN-VLKT [47]	21.36	32.91	30.56
Ours	**26.06**	**39.61**	**36.60**
Non-rare First Unseen Composition			
VCL [24]	16.22	18.52	18.06
ATL [25]	18.25	18.78	18.67
FCL [26]	18.66	19.55	19.37
GEN-VLKT [47]	25.05	23.38	23.71
Ours	**26.62**	**31.17**	**30.17**
Unseen Object			
FCL [26]	0.00	13.71	11.43
ATL [25]	5.05	14.69	13.08
GEN-VLKT [47]	10.51	28.92	25.63
Ours	**14.48**	**35.28**	**31.71**
Unseen Verb			
GEN-VLKT [47]	20.96	30.23	28.74
Ours	**21.50**	**35.63**	**33.09**

Table 6: Results of open-vocabulary HOI detection on HICO-DET.

Variant	HICO-DET mask mAP$_F$	VCOCO mask AP$_{role}^{S\#1}$	PSG R/mR@20
Different losses			
Disentangled CE loss	**40.5**	62.1	**27.0 / 15.4**
Triplet CE loss	36.8	**66.5**	25.5 / 14.6
Disentangled CE loss + Triplet CE loss	39.0	64.5	26.5 / 14.8
Different visual backbones			
Focal Tiny	34.2	59.8	25.8 / 15.0
Focal Large	**40.5**	**66.5**	**27.0 / 15.4**
Different design choices			
Box head only	33.0	62.0	-
Mask head only	40.5	66.5	27.0 / 15.4
Mask and box head	**41.2**	**67.0**	-
Different training datasets			
Single source	**40.5**	66.5	27.0
HICO-DET+VCOCO	40.3	**66.9**	-
HICO-DET+VCOCO+PSG	40.0	66.4	**27.6**

Table 7: Ablations of different loss types, backbones, design choices and training sets. We adopt the Focal-L backbone by default.

5 Conclusion

In this work, we present a novel approach for visual relationship segmentation that integrates the three critical aspects of a flexible VRS model: standard VRS, promptable querying, and open-vocabulary capabilities. Our `FleVRS` demonstrates the ability to not only support HOI segmentation and panoptic SGG but also to do so in response to various textual prompts and across a spectrum of previously unseen objects and interactions. By harnessing the synergistic potential of textual and visual features, our model delivers promising experimental results on existing benchmark datasets. We hope our work can serve as a solid stepping stone for pursuing more flexible visual relationship segmentation models.

References

[1] Abdelkarim, S., Agarwal, A., Achlioptas, P., Chen, J., Huang, J., Li, B., Church, K., and Elhoseiny, M. (2021). Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[2] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. (2023). Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv*.

[3] Bansal, A., Rambhatla, S. S., Shrivastava, A., and Chellappa, R. (2020). Detecting human-object interactions via functional generalization. In *AAAI*.

[4] Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., and Deng, J. (2018). Learning to detect human-object interactions. In *WACV*.

[5] Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. (2023). Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

[6] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *CVPR*.

[7] Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., and Zhang, L. (2021). Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382.

[8] Deng, J., Yang, Z., Chen, T., Zhou, W., and Li, H. (2021). Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779.

[9] Ding, Z., Wang, J., and Tu, Z. (2022). Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*.

[10] Du, H., Yu, X., and Zheng, L. (2020). Learning object relation graph and tentative policy for visual navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 19–34. Springer.

[11] Fang, H.-S., Xie, Y., Shao, D., and Lu, C. (2021). Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[12] Gao, C., Zou, Y., and Huang, J.-B. (2018). ican: Instance-centric attention network for human-object interaction detection. In *BMVC*.

[13] Gao, C., Xu, J., Zou, Y., and Huang, J.-B. (2020). DRG: Dual relation graph for human-object interaction detection. In *ECCV*.

[14] Gkioxari, G., Girshick, R., Dollár, P., and He, K. (2018). Detecting and recognizing human-object interactions. In *CVPR*.

[15] Gu, J., Stefani, E., Wu, Q., Thomason, J., and Wang, X. (2022). Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623.

[16] Gu, J., Wang, Y., Zhao, N., Xiong, W., Liu, Q., Zhang, Z., Zhang, H., Zhang, J., Jung, H., and Wang, X. E. (2024a). Swapanything: Enabling arbitrary object swapping in personalized visual editing. *arXiv preprint arXiv:2404.05717*.

[17] Gu, J., Fang, Y., Skorokhodov, I., Wonka, P., Du, X., Tulyakov, S., and Wang, X. E. (2024b). Via: A spatiotemporal video adaptation framework for global and local video editing. *arXiv preprint arXiv:2406.12831*.

[18] Gupta, S. and Malik, J. (2015). Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.

[19] Gupta, T., Schwing, A., and Hoiem, D. (2019). No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*.

[20] Han, Z., Zhu, F., Lao, Q., and Jiang, H. (2024). Zero-shot referring expression comprehension via structural similarity between images and captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14364–14374.

[21] He, C., Zhu, H., Gao, J., Chen, K., and Nevatia, R. (2020). Cparr: Category-based proposal analysis for referring relationships. In *CVPR workshops*.

[22] Hong, Y., Rodriguez, C., Qi, Y., Wu, Q., and Gould, S. (2020). Language and visual entity relationship graph for agent navigation. *Advances in Neural Information Processing Systems*, **33**, 7685–7696.

[23] Hou, Z., Peng, X., Qiao, Y., and Tao, D. (2020a). Visual compositional learning for human-object interaction detection. In *ECCV*.

[24] Hou, Z., Peng, X., Qiao, Y., and Tao, D. (2020b). Visual compositional learning for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 584–600. Springer.

[25] Hou, Z., Yu, B., Qiao, Y., Peng, X., and Tao, D. (2021a). Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504.

[26] Hou, Z., Yu, B., Qiao, Y., Peng, X., and Tao, D. (2021b). Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655.

[27] Hu, X., Lin, Y., Wang, S., Wu, Z., and Lv, K. (2023). Agent-centric relation graph for object visual navigation. *IEEE Transactions on Circuits and Systems for Video Technology*.

[28] Huang, Z., Mo, X., and Lv, C. (2022). Multi-modal motion prediction with transformer-based neural network for autonomous driving. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2605–2611. IEEE.

[29] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. (2021). Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

[30] Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. (2014). Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

[31] Kim, B., Choi, T., Kang, J., and Kim, H. J. (2020a). Uniondet: Union-level detector towards real-time human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 498–514. Springer.

[32] Kim, B., Lee, J., Kang, J., Kim, E.-S., and Kim, H. J. (2021a). HOTR: End-to-end human-object interaction detection with transformers. In *CVPR*.

[33] Kim, B., Lee, J., Kang, J., Kim, E.-S., and Kim, H. J. (2021b). Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83.

[34] Kim, D.-J., Sun, X., Choi, J., Lin, S., and Kweon, I. S. (2020b). Detecting human-object interactions with action co-occurrence priors. In *ECCV*.

[35] Kim, S., Jung, D., and Cho, M. (2023). Relational context learning for human-object interaction detection. In *CVPR*.

[36] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., *et al.* (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

[37] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., *et al.* (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, **123**, 32–73.

[38] Krishna, R., Chami, I., Bernstein, M., and Fei-Fei, L. (2018). Referring relationships. In *CVPR*.

[39] Li, J., Li, D., Xiong, C., and Hoi, S. (2022a). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

[40] Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., *et al.* (2022b). Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

[41] Li, R., Zhang, S., Wan, B., and He, X. (2021). Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *CVPR*.

[42] Li, Y.-L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.-S., Wang, Y., and Lu, C. (2019). Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594.

[43] Li, Y.-L., Liu, X., Wu, X., Li, Y., and Lu, C. (2020). Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, **33**, 5011–5022.

[44] Liang, C., Wu, Y., Zhou, T., Wang, W., Yang, Z., Wei, Y., and Yang, Y. (2021). Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*.

[45] Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., and Marculescu, D. (2023). Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*.

[46] Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., and Feng, J. (2020). Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490.

[47] Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., and Liu, S. (2022). Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*.

[48] Lim, J., Baskaran, V. M., Lim, J. M.-Y., Wong, K., See, J., and Tistarelli, M. (2023). Ernet: An efficient and reliable human-object interaction detection network. *IEEE Transactions on Image Processing*.

[49] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.

[50] Lin, X., Ding, C., Zeng, J., and Tao, D. (2020). Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753.

[51] Liu, C., Ding, H., and Jiang, X. (2023a). Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601.

[52] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., *et al.* (2023b). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

[53] Liu, Y., Chen, Q., and Zisserman, A. (2020). Amplifying key cues for human-object-interaction detection. In *ECCV*.

[54] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[55] Lu, C., Krishna, R., Bernstein, M., and Fei-Fei, L. (2016). Visual relationship detection with language priors. In *ECCV*.

[56] Lüddecke, T. and Ecker, A. (2022). Image segmentation using text and image prompts. In *CVPR*.

[57] Ma, S., Wang, Y., Wang, S., and Wei, Y. (2023). Fgahoi: Fine-grained anchors for human-object interaction detection. *arXiv preprint arXiv:2301.04019*.

[58] Ma, X., Li, J., Kochenderfer, M. J., Isele, D., and Fujimura, K. (2021). Reinforcement learning for autonomous driving with latent state inference and spatial-temporal relationships. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6064–6071. IEEE.

[59] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

[60] Park, J., Park, J.-W., and Lee, J.-S. (2023). Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162.

[61] Peyre, J., Laptev, I., Schmid, C., and Sivic, J. (2019). Detecting unseen visual relations using analogies. In *ICCV*.

[62] Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C. (2018). Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 401–417.

[63] Raboh, M., Herzig, R., Berant, J., Chechik, G., and Globerson, A. (2020). Differentiable scene graphs. In *WACV*.

[64] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.* (2021). Learning transferable visual models from natural language supervision. In *ICML*.

[65] Shadlen, M. N., Britten, K. H., Newsome, W. T., and Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, **16**(4), 1486–1510.

[66] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., and Sun, J. (2019). Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*.

[67] Shic, F. and Scassellati, B. (2007). A behavioral analysis of computational models of visual attention. *International journal of computer vision*, **73**, 159–177.

[68] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. (2022). Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

[69] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

[70] Tamura, M., Ohashi, H., and Yoshinaga, T. (2021). Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*.

[71] Tang, K., Zhang, H., Wu, B., Luo, W., and Liu, W. (2019). Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628.

[72] Ulutan, O., Iftekhar, A., and Manjunath, B. S. (2020). VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*.

[73] Wan, B., Zhou, D., Liu, Y., Li, R., and He, X. (2019). Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478.

[74] Wang, H., Zheng, W.-s., and Yingbiao, L. (2020a). Contextual heterogeneous graph network for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 248–264. Springer.

[75] Wang, H., Du, Y., Zhang, Y., Li, S., and Zhang, L. (2022a). One-stage visual relationship referring with transformers and adaptive message passing. *IEEE Transactions on Image Processing*.

[76] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022b). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

[77] Wang, T., Yang, T., Danelljan, M., Khan, F. S., Zhang, X., and Sun, J. (2020b). Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125.

[78] Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., *et al.* (2024). Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, **36**.

[79] Wu, J., Jiang, Y., Sun, P., Yuan, Z., and Luo, P. (2022). Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984.

[80] Xu, B., Wong, Y., Li, J., Zhao, Q., and Kankanhalli, M. S. (2019). Learning to detect human-object interactions with knowledge. In *CVPR*.

[81] Xu, D., Zhu, Y., Choy, C. B., and Fei-Fei, L. (2017). Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419.

[82] Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. (2022). Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144.

[83] Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., and Lu, H. (2023). Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336.

[84] Yang, J., Li, C., Dai, X., and Gao, J. (2022a). Focal modulation networks. *NeurIPS*.

[85] Yang, J., Ang, Y. Z., Guo, Z., Zhou, K., Zhang, W., and Liu, Z. (2022b). Panoptic scene graph generation. In *ECCV*.

[86] Yang, J., Peng, W., Li, X., Guo, Z., Chen, L., Li, B., Ma, Z., Zhou, K., Zhang, W., Loy, C. C., *et al.* (2023). Panoptic video scene graph generation. In *CVPR*.

[87] Yi, M., Cui, Q., Wu, H., Yang, C., Yoshie, O., and Lu, H. (2023). A simple framework for text-supervised semantic segmentation. In *CVPR*.

[88] You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.-F., and Yang, Y. (2023). Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

[89] Yu, J., Chai, Y., Wang, Y., Hu, Y., and Wu, Q. (2021). Cogtree: Cognition tree loss for unbiased scene graph generation. In *IJCAI*.

[90] Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. (2016). Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

[91] Yuan, H., Jiang, J., Albanie, S., Feng, T., Huang, Z., Ni, D., and Tang, M. (2022). Rlip: Relational language-image pre-training for human-object interaction detection. In *Advances in Neural Information Processing Systems*.

[92] Yuan, H., Zhang, S., Wang, X., Albanie, S., Pan, Y., Feng, T., Jiang, J., Ni, D., Zhang, Y., and Zhao, D. (2023). Rlipv2: Fast scaling of relational language-image pre-training. In *ICCV*.

[93] Zellers, R., Yatskar, M., Thomson, S., and Choi, Y. (2018). Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840.

[94] Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. (2022). Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

[95] Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., and Li, X. (2021a). Mining the benefits of two-stage and one-stage hoi detection. In *NeurIPS*.

[96] Zhang, F. Z., Campbell, D., and Gould, S. (2021b). Spatially conditioned graphs for detecting human-object interactions. In *ICCV*.

[97] Zhang, F. Z., Campbell, D., and Gould, S. (2022a). Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*.

[98] Zhang, F. Z., Yuan, Y., Campbell, D., Zhong, Z., and Gould, S. (2023a). Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[99] Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., and Zhang, L. (2023b). A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031.

[100] Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., and Chen, C.-W. (2022b). Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*.

[101] Zhao, L., Yuan, L., Gong, B., Cui, Y., Schroff, F., Yang, M.-H., Adam, H., and Liu, T. (2023). Unified visual relationship detection with vision and language models. *arXiv preprint arXiv:2303.08998*.

[102] Zhong, X., Ding, C., Qu, X., and Tao, D. (2020). Polysemy deciphering network for human-object interaction detection. In *ECCV*.

[103] Zhong, X., Qu, X., Ding, C., and Tao, D. (2021). Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243.

[104] Zhou, Z., Shi, M., and Caesar, H. (2023a). Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. In *ICCV*.

[105] Zhou, Z., Shi, M., and Caesar, H. (2023b). Vlprompt: Vision-language prompting for panoptic scene graph generation. *arXiv:2311.16492*.

[106] Zhu, F., Xie, Y., Xie, W., and Jiang, H. (2023). Diagnosing human-object interaction detectors. *arXiv preprint arXiv:2308.08529*.

[107] Zhu, J. and Wang, H. (2021). Multiscale conditional relationship graph network for referring relationships in images. *IEEE Transactions on Cognitive and Developmental Systems*.

[108] Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., *et al.* (2021). End-to-end human object interaction detection with hoi transformer. In *CVPR*.

[109] Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., *et al.* (2023a). Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127.

[110] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., and Lee, Y. J. (2023b). Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.

A Limitations and Future Work

Deploying our model in real-world scenarios requires specialized pretraining data for relationship understanding, which is notably scarce. The lack of automated annotation pipelines and dependence on the CLIP model pose scalability challenges due to specific resource requirements. Ideally, we aim for a single general-purpose framework that can be trained on multiple datasets and enhance performance across various tasks and benchmarks. However, achieving this remains a challenge with our current model. We leave the exploration of how to synergize different datasets and develop effective training strategies to future work. While integrating free-form text inputs is more natural as large language models evolve, it necessitates additional preprocessing to align with our framework. Furthermore, the absence of comparable methods for promptable VRS makes complete fair benchmarking difficult.

B Model Structure Details

We use Focal T/L [84] network as the image encoder $\mathbf{Enc_I}$. Given the image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we pass it to $\mathbf{Enc_I}$ and obtain multi-scale features of different strides and channals $\mathbf{F} = \{\mathbf{F}_s | s = 4, 8, 16, 32\}$, where s is the stride.

Then, the pixel decoder $\mathbf{Dec_p}$ gradually upsample $\mathbf{F}$ to generate high-resolution per-pixel embeddings $\mathbf{P} = \{\mathbf{P}_i | i = 1, 2, 3, 4\}$, where i is the layer number and different $\mathbf{P}_i$s have the same channel number but different resolutions. $\mathbf{P}$ will then input to $\mathbf{Dec_R}$.

Under standard VRS, $\mathbf{Dec_R}$ takes latent queries $\mathbf{Q^v}$ and $\mathbf{P}$ as inputs. Under promptable VRS, we use the textual encoder $\mathbf{Enc_T}$ to encode the textual prompt into a set of textual queries $\mathbf{Q^t}$ and concatenate $\mathbf{Q^t}$ with the latent queries $\mathbf{Q^v}$, and input them to $\mathbf{Dec_R}$. Inside $\mathbf{Dec_R}$, cross- and self-attention are computed among queries and per-pixel embeddings, where masked attention [6] is adopted to enhance the foreground regions of predicted masks.

On top of latent queries output $\mathbf{Q}_o^{\mathbf{v}} \in \mathbb{R}^{N_v \times C_q}$ (N_v is the number of latent queries, C_q is the channel number), there are five heads, producing predictions in parallel. They are two mask heads $f_{M_s}(\cdot)$, $f_{M_o}(\cdot)$ for predicting subject and object masks (M_s, M_o), and two class heads $g_{C_s}(\cdot)$, $g_{C_o}(\cdot)$ for predicting their object categories (C_s, C_o). Another class head g_{C_p} is used to predict relationships C_p for this `<subject, object>` pair. Detailed operations can be written as

$$M_s = \text{Up}\left[\mathbf{P}_4 \cdot f_{M_s}(\mathbf{Q}_o^{\mathbf{v}})\right], \tag{3}$$

$$M_o = \text{Up}\left[\mathbf{P}_4 \cdot f_{M_o}(\mathbf{Q}_o^{\mathbf{v}})\right], \tag{4}$$

$$C_s = \mathbf{T}_s \cdot g_{C_s}(\mathbf{Q}_o^{\mathbf{v}}), \tag{5}$$

$$C_o = \mathbf{T}_o \cdot g_{C_o}(\mathbf{Q}_o^{\mathbf{v}}), \tag{6}$$

$$C_p = \mathbf{T}_p \cdot g_{C_p}(\mathbf{Q}_o^{\mathbf{v}}), \tag{7}$$

where $\text{Up}[\cdot]$ denotes the upsampling operation, $\mathbf{T}_s$, $\mathbf{T}_o$ and $\mathbf{T}_p$ denote candidate textual features of subject, object and predicate categories that are encoded by CLIP [64]. The mask embeddings $f_{M_s}(\mathbf{Q}_o^{\mathbf{v}})$ and $f_{M_o}(\mathbf{Q}_o^{\mathbf{v}})$ compute the dot products with the last layer's per-pixel embedding $\mathbf{P}_4$, respectively, and upsample to the original resolution as final mask predictions.

Training. We employ Hungarian matching to align predicted triplets with ground truth, calculating mask and category classification losses on these matches. For promptable VRS, we introduce a matching loss to assess the similarity between the matched triplet embedding and the textual prompt's feature, formulated as a cross-entropy loss. The triplet embedding combines class embeddings from class heads. For the textual prompt's feature, we use the last token's feature from textual queries $\mathbf{Q^t}$. For instance, with a textual prompt like `<subject, predicate, ?>`, where the subject and predicate are specified, the similarity measurement utilizes the summation of their class embeddings $g_{C_s}(\mathbf{Q}_o^{\mathbf{v}}) + g_{C_p}(\mathbf{Q}_o^{\mathbf{v}})$.

Inference. Under standard VRS, we compute the confidence score of each triplet, which comes from the product of subject, object, and predicate classification scores. We take top k_s ($k_s = 100$) triplets to compute mean average precision for HOI segmentation and mean recall for panoptic SGG. Under promptable VRS, we compute similarities between the textual prompt's feature and triplet embeddings and choose k_f ($k_f = 10$) as the final predicted triplets.

(a) Text prompt: <?, object, predicate>

<?, walk, bicycle>   <?, load, bus>

<?, sit on, coach>   <?, jump, skateboard>

(b) Text prompt: <subject, predicate, ?>

<person, load, ?>   <person, sit on, ?>

<person, lie on, ?>   <person, straddle, ?>

(c) Text prompt: <subject, ?, object>

<person, ?, dining table>   <person, ?, skateboard>   <person, ?, hotdog>   <person, ?, knife>
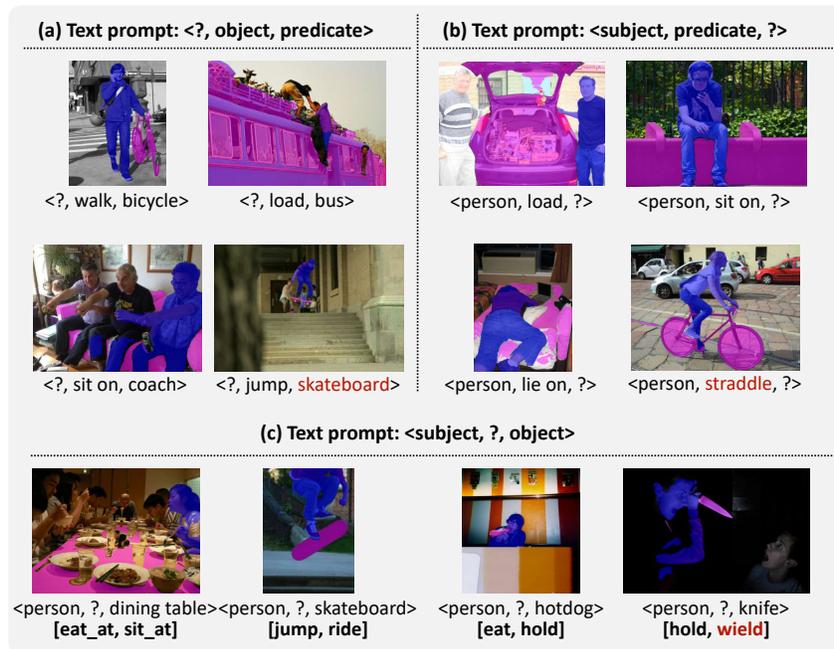**[eat_at, sit_at]**   **[jump, ride]**   **[eat, hold]**   **[hold, wield]**

Figure 5: **Qualitative results of promptable and open-vocabulary VRS on HICO-DET [4] test set.** We show visualizations of the predicted triplet with the highest matching score, including subject, object masks, and predicted predicate categories. There are three types of textual prompts shown in (a), (b), and (c), with unseen concepts in the rightmost columns. In (c), we show the predicted predicates in bold characters. Unseen objects and predicates are denoted in red characters. Note that the subject is always "person" in HICO-DET.

C Implementation details

We set the input image size to be 640×640, with batch size as 64. The model is optimized with AdamW [54] with a weight decay of 10^{-4}. We set $N_v = 100$ and $N_v = 200$ for Focal-T and Focal-L backbones, respectively, and $C_q = 512$. The structure of pixel decoder $\mathbf{Dec_p}$ is a Transformer encoder with 6 encoder layers and 8 heads. The structure of relationship decoder $\mathbf{Dec_R}$ is a Transformer decoder with 9 decoder layers.

Standard HOI segmentation The model only takes the image as input without textual prompts. Since the subject class is always "person" in HOI segmentation, we omit the subject class head. The model is trained with 30 epochs, with an initial learning rate of 10^{-4} (10^{-5} for the image encoder) decreased by 10 times at the 20th epoch. The loss weights λ_b, λ_d, λ_c^o and λ_c^p are set to be 2, 1, 1, 2.

Promptable HOI segmentation To enable promptable HOI segmentation, we build three types of textual prompts: 1) "*<s>person</s><p>predicate</p>*"; 2) "*<p>predicate</p><o>object</o>*"; 3) "*<s>person</s><o>object</o>*". We evaluate the model on HICO-DET [4] since it contains richer human-object interactions than VCOCO [18]. During training, we randomly sample various types of text prompts and simultaneously train different objectives using distinct loss terms. To prevent the model from learning shortcuts, we select one ground truth triplet per training image and pair it with a randomly chosen textual prompt type. This approach ensures a balanced distribution of labeled training data for promptable VRS across different prompt types.

Standard panoptic SGG We use the subject class head to predict the subject category and the model does not have textual prompts as inputs. The model is trained with 60 epochs, with an initial learning rate of 10^{-4} (10^{-5} for the image encoder) decreased by 10 times at the 40th epoch. The loss weights λ_b, λ_d, λ_c^s, λ_c^o and λ_c^p are set to be 2, 1, 1, 1, 2.

Promptable panoptic SGG Similar to promptable HOI segmentation, there are three types of textual prompts: 1) "*<s>subject</s><p>predicate</p>*"; 2) "*<p>predicate</p><o>object</o>*"; 3) "*<s>subject</s><o>object</o>*". Similarly, during training, we randomly

(a) Text prompt: <?, object, predicate>

<?, standing on, road> <?, on, floor>

<?, driving on, road> <?, over, pavement>

(b) Text prompt: <subject, predicate, ?>

<truck, on, ?> <person, hold, ?>

<person, ride, ?> <car, driving on, ?>

(c) Text prompt: <subject, ?, object>

<bottle, ?, cabinet>
[on]

<coach, ?, floor>
[over]

<person, ?, surfboard>
[in, lie on, carry, touch]

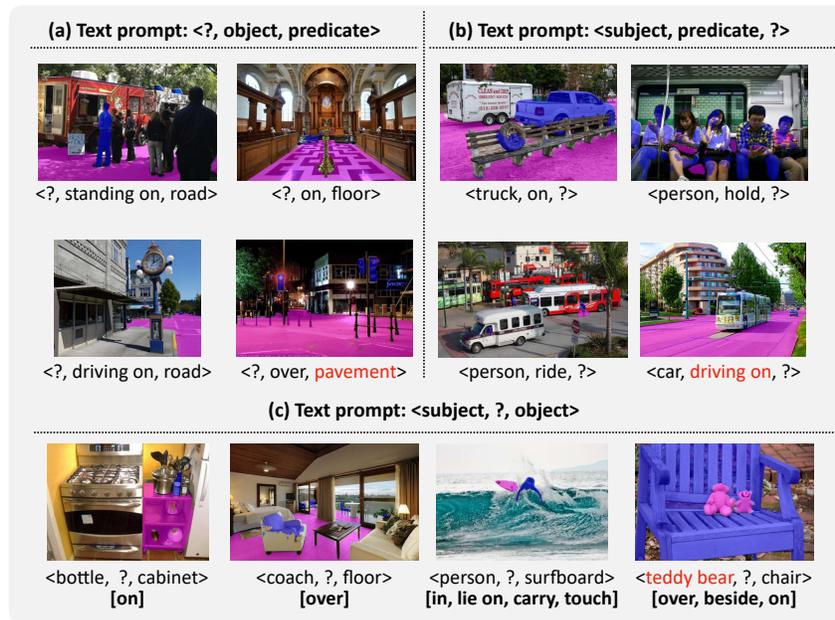<teddy bear, ?, chair>
[over, beside, on]

Figure 6: **Qualitative results of promptable and open-vocabulary VRS on PSG [85] test set.** We show visualizations of the predicted triplet with the highest matching score, including subject, object masks, and predicted predicate categories. There are three types of textual prompts shown in (a), (b), and (c), with unseen concepts in the rightmost columns. In (c), we show the predicted predicates in bold characters. Unseen objects and predicates are denoted in red characters.

sample different types of text prompts, and different objectives are trained simultaneously with different loss terms. We also keep a balanced distribution of labeled training data for panoptic SGG. We set the weight of grounding loss λ_g to be 2, while other weights are the same as standard panoptic SGG.

Open-vocabulary promptable VRS We adopt the zero-shot setting from [47] for open-vocabulary HOI segmentation. To streamline our approach, we integrate the open-vocabulary promptable setting within the broader context of open-vocabulary VRS. In this setting, 'open-vocabulary' refers to handling both seen and unseen categories in the input textual prompts. We randomly exclude object and predicate categories during training and assess our model on these as well as on seen categories. Given the absence of existing benchmarks for this specific challenge, we present qualitative results demonstrating our model's proficiency in open-vocabulary promptable VRS.

D Qualitative results of promptable and open-vocabulary VRS

We show qualitative results of promptable and open-vocabulary VRS on HOI segmentation and panoptic SGG by giving the model different types of structured textual prompts. For simplicity, we show examples of omitting only one component of the triplet. We can see that our model is able to localize the correct subject and object and complement the missing element corresponding to the given textual prompt, *e.g.* <person, ?, dining_table> in Fig. 5(c) and <truck, on, ?> in Fig. 6(b). The model can also predict multiple interactions for the same subject-object pair, as shown in Fig. 5(c) and Fig. 6(c). We further trained two versions by removing unseen objects and unseen predicates, respectively. We show that our model can detect novel objects and predicates by feeding unseen concepts in textual prompts, as in the rightmost columns of Fig. 5 and Fig. 6. Fig. 6(b) shows the model outputs multiple instances in one subject mask due to similar patterns occurring in the training set. Note that the flexible VRD task is more difficult on the PSG [85] dataset due to its complexity of scenes, while we make the first attempt and our model is still able to show promising grounding results.

Model	Backbone	Default (%)		
		box/mask mAP$_F$	box/mask mAP$_R$	box/mask mAP$_N$
Bottom-up methods				
InteractNet [14]	ResNet-50	9.9 / -	7.2 / -	10.8 / -
iCAN [12]	ResNet-50	14.8 / -	10.5 / -	16.2 / -
No-Frills [19]	ResNet-152	17.2 / -	12.2 / -	18.7 / -
DRG [13]	ResNet-50	24.5 / -	19.5 / -	26.0 / -
VSGNet [72]	ResNet-152	19.8 / -	16.1 / -	20.9 / -
FCMNet [53]	ResNet-50	20.4 / -	17.3 / -	21.6 / -
IDN [43]	ResNet-50	23.4 / -	22.5 / -	23.6 / -
ATL [25]	ResNet-101	23.8 / -	17.4 / -	25.7 / -
SCG [96]	ResNet-50	31.3 / 31.3	24.7 / 25.0	33.3 / 35.5
UPT [97]	ResNet-101	32.6 / 34.9	28.6 / 29.4	33.8 / 36.1
STIP [100]	ResNet-50	32.2 / 30.8	28.2 /28.6	33.4 / 32.5
ViPLO [60]	ViT-B	<u>37.2</u> / <u>39.1</u>	<u>35.5</u> / <u>37.8</u>	<u>37.8</u> / <u>39.7</u>
Additional training with object detection data				
UniVRD [101]	ViT-L	37.4 / -	28.9 / -	39.9 / -
PViC [98]	Swin-L	44.3 / -	44.6 / -	44.2 / -
RLIPv2 [92]	Swin-L	45.1 / 48.6	45.6 / 44.3	43.2 / 49.8
Single-stage methods				
DIRV [11]	EfficientDet-d3	21.8 / -	16.4 / -	23.4 / -
PPDM-Hourglass [46]	DLA-34	21.9 / -	13.9 / -	24.3 / -
HOI-Transformer [108]	ResNet-101	26.6 / -	19.2 / -	28.8 / -
GGNet [103]	Hourglass-104	29.2	22.1 / -	30.8 / -
HOTR [33]	ResNet-50	25.1 / 26.5	17.3 / 18.5	27.4 / 29.0
QPIC [70]	ResNet-101	29.9 / 30.5	23.0 / 23.1	31.7 / 33.1
CDN [95]	ResNet-101	32.1 / 33.9	27.2 / 28.9	33.5 / 36.0
RLIP [91](VG+COCO)	ResNet-50	32.8 / 34.4	26.9 / 27.7	34.6 / 36.5
GEN-VLKT [47]	ResNet-101	35.0 / 35.6	31.2 / 32.6	36.1 / 37.8
ERNet [48]	EfficientNetV2-XL	35.9 / -	30.1 / -	38.3 / -
MUREN [35]	ResNet-50	32.9 / 35.4	28.7 / 30.1	34.1 / 37.6
Ours	Focal-L	**38.1 / 40.5**	**33.0 / 34.9**	**39.5 / 42.4**

Table 8: **Quantitative results on the HICO-DET test set.** We report both box and mask mAP under the *Default* setting [4] containing the *Full* (F), *Rare* (R), and *Non-Rare* (N) sets. no_interaction class is removed in mask mAP. The best score is highlighted in bold, and the second-best score is underscored. '-' means the model did not release weights and we cannot get the mask mAP.

E Quantitative results of standard VRS

Due to the large number of works on HOI detection, we show the complete comparison with previous methods in Fig. 8 and Fig. 9. Our model achieves competitive results on both datasets, especially compared with other single-stage methods. From Table 9, MUREN [35] gets the best result on VCOCO (68.8 *vs.* 65.2, 68.2 *vs.* 66.5), but cannot achieve a similarly strong result on HICO-DET (32.9 *vs.* 38.1, 35.4 *vs.* 40.5), where the verb categories are more complicated.

Fair Comparison. Since existing models use bounding box annotations to train and evaluate mAP, we ensure fair comparisons by converting our model's output masks into bounding boxes to compute box mAP. Additionally, we apply released weights from previous methods, transform their output boxes into segmentation masks using SAM [36], and report mask mAP. In both metrics, our model demonstrates superior performance.

We further train the existing HOI detectors CDN [95], STIP, GEN-VLKT the same SAM generated data used in our paper, which leads to worse accuracy on HICO-DET, as in Tab. 10. Thus, the major performance improvements of our work are due to both the SAM-labeled data and our architectural design.

At the same time, we also train our model with bounding boxes only, where we get decreased accuracy (mAP of 30.7 *vs* 36.3). We attribute it to the network architecture derived from Mask2Former [6], which is mainly designed for pixel-wise segmentation tasks.

FLOPs and the number parameters of the backbone compared to previous works. As in Table 2 and 3 of the main paper, we have done extensive comparisons with previous methods, including backbones on ResNet-50/101/152, EfficientNet, Hourglass, Swin Transformers, and LiT architectures. For previous methods that utilize ResNet backbone for HOI detection and PSG, our comparison

Model	Backbone	$AP_{role}^{S\#1}$	$AP_{role}^{S\#2}$
Bottom-up methods			
InteractNet [14]	ResNet-50	40.0 / -	- / -
GPNN [62]	ResNet-50	44.0 / -	- / -
iCAN [12]	ResNet-50	45.3 / -	52.4 / -
TIN [42]	ResNet-50	47.8 / -	54.2 / -
DRG [13]	ResNet-50	51.0 / -	- / -
IP-Net [77]	ResNet-50	51.0 / -	- / -
VSGNet [72]	ResNet-152	51.8 / -	57.0 / -
PMFNet [73]	ResNet-50	52.0 / -	- / -
PD-Net [102]	ResNet-50	52.6 / -	- / -
CHGNet [74]	ResNet-50	52.7 / -	- / -
FCMNet [53]	ResNet-50	53.1 / -	- / -
ACP [34]	ResNet-152	53.2 / -	- / -
IDN [43]	ResNet-50	53.3 / -	60.3 / -
STIP [100]	ResNet-50	66.0 / 66.2	70.7 / 70.5
Additional training with object detection data			
VCL [24]	ResNet-101	48.3 / -	- / -
SCG [96]	ResNet-50	54.2 / 49.2	60.9 / 53.4
UPT [97]	ResNet-101	61.3 / 60.3	67.1 / 65.6
UniVRD [101]	ViT-L	65.1 / -	66.3 / -
PViC [98]	Swin-L	64.1 / -	70.2 / -
RLIPv2 [92]	Swin-L	72.1 / 71.7	74.1 / 73.5
Single-stage methods			
UnionDet [31]	ResNet-50	47.5 / -	56.2 / -
HOI-Transformer [108]	ResNet-101	52.9 / -	- / -
GGNet [103]	Hourglass-104	54.7 / -	- / -
HOTR [33]	ResNet-50	55.2 / 55.0	64.4 / 64.1
DIRV [11]	EfficientDet-d3	56.1 / -	- / -
QPIC [70]	ResNet-101	58.3 / -	60.7 / -
CDN [95]	ResNet-101	63.9 / 61.3	65.8 / 63.2
RLIP [91]	ResNet-50	61.9 / 61.3	64.2 / 64.0
GEN-VLKT [47]	ResNet-101	63.6 / 61.8	65.9 / 64.0
ERNet [48]	EfficientNetV2-XL	64.2 / -	- / -
MUREN [35]	ResNet-50	**68.8 / 68.2**	**71.0 / 70.2**
Ours	Focal-L	65.2 / 66.5	66.5 / 67.9

Table 9: **Quantitative results on V-COCO.** We report both box and mask mAP. The best score is highlighted in bold, and the second-best score is underscored. '-' means the model did not release weights and we cannot get the mask mAP.

Model	Trained with original boxes	Trained with SAM masks
CDN	31.4	28.5
STIP	32.2	29.7
GEN-VLKT	35.6	32.1

Table 10: **Results of box mAP on HICO-DET test set.** We train existing HOI detectors with a mask head, by using the masks we generated through SAM.

includes VSGNet, ACP, No-Frills, which use ResNet-152. To the best of our knowledge, larger ResNet, such as ResNet-200, and ResNet-269, are not used in previous methods on related tasks. ResNet, we have included the largest model ResNet-152, which has 65M parameters and 15 GFLOPs. Other baselines are not using the ResNet backbone, for example, UniVRD is using LiT(ViT-H/14) backbone. It has 632M parameters and 162 GFLOPs, a lot more than our 198M parameters and 15.6 GFLOPs, but still performs worse than our model.

F Fair comparison of promptable VRS

Postprocessing of standard VRS outputs. Since no existing models share the same settings as promptable VRS, we create a baseline for fair comparison. Typically, promptable VRS can be addressed by filtering standard VRS outputs. We post-process outputs from our standard VRS model to extract the desired triplets and compared their mAP with those from promptable VRS. The post-processed results yield a lower mAP (15.7 vs. 26.8), primarily because the selected triplets often have lower confidence scores. Additionally, the post-processing approach is slower, taking 8 seconds compared to 5 seconds for directly prompting the model to retrieve the desired triplet.

Grounding ability compared with prompt-based vision-language models. Although promptable VRS is similar to vision-language models like GLIP [40] and MDETR [29] in grounding capabilities, it has distinct objectives. Unlike these models, which focus on entities, promptable VRS outputs triplets, making direct comparisons infeasible. Previous models are not equipped to handle the promptable relationship understanding task directly. To explore this, we modify our structural design to incorporate multiple text prompts as inputs, which are individually processed with their matching scores aggregated for classification. This experimental setup, however, results in reduced performance, increased inference time (26s vs. 5s), and higher GPU memory usage (5G vs. 3G). Thus, we argue that the proposed structure is suitable for tackling promptable VRS.

G   Masks generated by SAM

Clarifications of choosing segmentation masks We firstly illustrate the importance of choosing segmentation masks over boxes in Fig. 7. Traditional bounding boxes often include overlapping and ambiguous information, leading to redundancy. Segmentation masks, by accurately delineating object boundaries, provide a more precise and clear representation, reducing such redundancy, which is also illustrated in [85] and [86]. Besides, segmentation masks provide enhanced visual understanding and comprehensive contextual analysis. Additionally, object detection models often struggle to precisely extract foreground objects, which is why they are typically combined with segmentation models like SAM for fine-grained image tasks. Our model, however, presents a unified model that can localize both subjects and objects, along with their corresponding segmentation masks.

Figure 7: **Illustration of the importance of using masks instead of bounding boxes.** We show examples where one object is occluded by other objects. We show both bounding box annotations and masks generated with SAM, where only the masks can correctly locate the pure object.

Noise handling in using masks generated by SAM. To address potential noise and inaccuracies in masks generated by SAM, we employ a filtering approach based on Intersection over Union (IoU). We compute the IoU between the generated masks and the original box annotations. Masks with an IoU score below a threshold of 0.2 are considered to have significant deviations from the ground truth and are filtered out. This threshold is chosen to balance the trade-off between including sufficient mask data and excluding those with substantial inaccuracies. The chosen IoU threshold helps ensure that only masks with a reasonable overlap with the ground truth annotations are retained. This threshold is set based on empirical evaluation and aims to minimize the impact of masks that are too noisy or incorrect, while still retaining as much useful data as possible. After using this strategy, we conduct analysis on 200 samples. We tested various thresholds and found this gets the best balance between denoising and data retaining(95% valid data retraining).

More visualizations of generated masks. We have included additional visualizations in Fig. 8 to illustrate the fine-grained masks generated from the bounding box annotations of existing HOI detection datasets. These visualizations indicate that converting to masks significantly reduces the redundancy in the box annotations. Additionally, as shown in Fig. 8 (d), filtering with IoU helps eliminate low-quality masks.
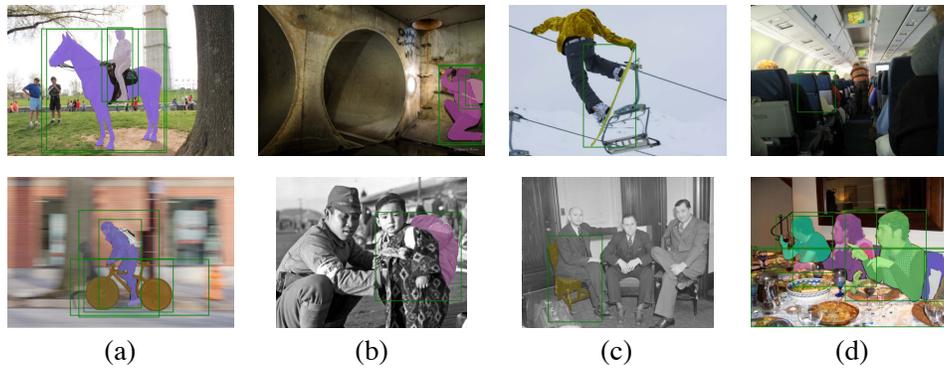
Figure 8: **Samples of fine-grained masks generated by converting existing bounding box annotations with SAM.** Samples are chosen from the HICO-DET dataset. Green boxes are original box annotations. Duplicated boxes are suppressed after converting to the mask, as shown in (a). There are also failure cases where no masks are generated with the given box annotations, as in (d).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and section 1 introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not have theoretical result in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide implementation details in section 4 and supplementary section B.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In section 4 and supplementary section B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In section 4 and supplementary.

Guidelines:
- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have made sure.

Guidelines:
- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Do not have societal impact.

Guidelines:
- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Do not have such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Citations are complete.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets released.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.