

GSRFormer: Grounded Situation Recognition Transformer with Alternate Semantic Attention Refinement

Zhi-Qi Cheng
Carnegie Mellon University
zhiqic@cs.cmu.edu

Qi Dai
Microsoft Research
qid@microsoft.com

Siyao Li
Carnegie Mellon University
siyaol@andrew.cmu.edu

Teruko Mitamura
Carnegie Mellon University
teruko@cs.cmu.edu

Alexander G. Hauptmann
Carnegie Mellon University
Alex@cs.cmu.edu

ABSTRACT

Grounded Situation Recognition (GSR) aims to generate structured semantic summaries of images for “human-like” event understanding. Specifically, GSR task not only detects the salient activity verb (e.g. *buying*), but also predicts all corresponding semantic roles (e.g. *agent* and *goods*). Inspired by object detection and image captioning tasks, existing methods typically employ a two-stage framework: 1) detect the activity verb, and then 2) predict semantic roles based on the detected verb. Obviously, this illogical framework constitutes a huge obstacle to semantic understanding. First, pre-detecting verbs solely without semantic roles inevitably fail to distinguish many similar daily activities (e.g., *offering* and *giving*, *buying* and *selling*). Second, predicting semantic roles in a closed auto-regressive manner can hardly exploit the semantic relations among the verb and roles. To this end, in this paper we propose a novel two-stage framework that focuses on utilizing such bidirectional relations within verbs and roles. In the first stage, instead of pre-detecting the verb, we postpone the detection step and assume a pseudo label, where an intermediate representation for each corresponding semantic role is learned from images. In the second stage, we exploit transformer layers to unearth the potential semantic relations within both verbs and semantic roles. With the help of a set of support images, an alternate learning scheme is designed to simultaneously optimize the results: update the verb using nouns corresponding to the image, and update nouns using verbs from support images. Extensive experimental results on challenging SWiG benchmarks show that our renovated framework outperforms other state-of-the-art methods under various metrics¹.

CCS CONCEPTS

• Computing methodologies → Vision and Language.

KEYWORDS

Grounded Situation Recognition; Transformer Framework

¹Code is available at <https://github.com/zhiqic/GSRFormer>



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs International 4.0 License.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9203-7/22/10.
<https://doi.org/10.1145/3503161.3547943>



Figure 1: Given an input image, Grounded Situation Recognition (GSR) not only detects the salient verb category (e.g., *sprinkling*), but also predicts all corresponding semantic roles for sprinkling, such as *agent: man*, *item: spice*, and *source: cup*, etc.


ACM Reference Format:

Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander G. Hauptmann. 2022. GSRFormer: Grounded Situation Recognition Transformer with Alternate Semantic Attention Refinement. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, Oct. 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3503161.3547943>


1 INTRODUCTION

Understanding complex events in a way that obeys human cognitive habits is one of the core tasks of computer vision and multimedia. As shown in Figure 1, “human-like” event understanding goes beyond traditional object- and action-centric detection and recognition tasks [11, 13, 24, 30, 31, 39]. Different from image captioning [9, 32, 43, 78] and scene graph generation [20, 67, 68, 71], which use natural language or object-relational graphs to describe scenes, human-friendly event understanding must be event-centric, that is, not only to identify what activities happened, but also to recognize how objects participate in activities, i.e., answer questions like “*who is doing what with some tools at someplace*.”

To meet the demands of “human-like” event understanding, inspired by previous research on semantic role labeling [21, 45, 53, 59] in text, Grounded Situation Recognition (GSR) [47, 73] is proposed for event understanding in the multimedia domain. As illustrated in Figure 1, GSR not only detects the salient activity (verb) in the image (e.g., *sprinkling*), but also recognizes all semantic roles (e.g., *agent is man*, *source is cup*). To further determine semantic roles in images, GSR also provides visually grounded information (i.e., bounding boxes) for noun entities. From the perspective of the theory of frame semantics [26], GSR task can be considered as a multimedia extension to earlier lexical databases such as FrameNet [3] and PropBank [33]. By describing activities with verbs and grounded semantic roles, GSR can provide a visually grounded



VERB: BUYING		VERB: GIVING	
AGENT	woman	AGENT	man
GOODS	flower	ITEM	flower
PAYMENT	credit card	RECIPIENT	woman
SELLER	man	PLACE	outdoor
PLACE	shop		



VERB: REPAIRING		VERB: FIXING	
AGENT	man	AGENT	person
ITEM	laptop	OBJECT	water faucet
PROBLEM	∅	OBJECTPART	handle
TOOL	screwdriver	TOOL	screwdriver
PLACE	garage	PLACE	kitchen

Figure 2: Some examples illustrate the importance of semantic relations in GSR tasks. Semantic relations are bidirectionally constrained, i.e., (left) noun entities can distinguish similar activities (verbs), and (right) similar verbs can control the occurrence of semantic roles (nouns).

verb-frame, which benefits many downstream scene understanding tasks, such as information retrieval [15, 16, 44, 56], question answering [1, 7, 35, 67], recommended system [10, 12, 14, 55], and multimedia understanding [29, 51, 77, 78].

To sum up, the essence of GSR task is using semantic relations to generate verb-frames for event understanding. However, inspired by object detection and image captioning, almost all current GSR methods [19, 36, 42, 47, 54] adopt a two-stage framework. As shown in Figure 3 (a-b), two-stage framework 1) first blindly pre-detects verbs to reduce the search space, and then 2) predicts the semantic roles in an auto-regressive (RNN) or parallel (Transformer) manner depending on the detected verbs. Such two-stage frameworks obviously neglect the semantic relations among verbs and semantic roles. On the one hand, pre-detecting the verb without noun entities inevitably fails to distinguish some similar daily activities. For example, as shown in Figure 2 (left), it is hard to distinguish similar verbs (e.g., *buying* and *giving*) without the help of any noun entities. On the other hand, based on pre-identified verbs, applying auto-regression in a closed space would accumulate errors and thus miss the semantic relationships. As shown in Figure 2, once verb *buying* is wrongly predicted as *giving* in the first stage, the semantic roles *payment: credit* and *place: shop* could be neglected in the second stage.

Similar to our starting point, previous work CoFormer [18] and SituFormer [65] also argue that the existing two-stage framework is problematic. As shown in Figure 3 (c), a three-stage framework is thus proposed to further optimize verbs with predicted nouns. Following traditional two-stage works, the first two stages perform the verb and noun entities detection, respectively. Then, in the third stage, the predicted noun entities are used to refine the verb. However, such disentangled framework still has the following flaws. First, the bidirectional semantic relations between verbs and noun entities cannot be fully exploited. They either use the noun roles to refine the verbs (SituFormer), or only use verbs to refine the noun roles (CoFormer) while ignoring each other. Second, the framework is redundant. It has two parallel transformer verb and noun detectors, but does not learn semantic relations between them during the encoding phase. Third, the refinement process is not scalable. It can be treated as a one-time noun-to-verb optimization, which apparently cannot be expanded.

To address these issues, we focus on how to exploit such bidirectional semantic relations within verbs and noun entities, which can constrain each other. Rather than explicitly pre-detecting the verbs at the first stage, we postpone decision verbs, thus simply assuming a pseudo-category and learning an intermediate representation for each noun entity. We then devise an iterative framework to capture the semantic relations among verbs and nouns and alternately learn

their features. As such, we streamline the redundant structures and make them flexibly handle semantic roles in parallel.

Technically, our proposed method, called GSRFormer, is built based on the transformer structure [62] due to its parallel processing capability. As shown in Figure 4, GSRFormer adopts a two-stage architecture that consists of an encoder and a decoder. In the encoder part, we first utilize stacked Multi-Head Attention (MHA) layers to learn the feature of the verb. By assuming a pseudo category of the verb, we further learn the intermediate representations of the corresponding noun entities from the image. In the decoder part, MHA layers are employed to mine the implicit relations among both verbs and nouns. By leveraging a set of support images, the model alternately optimizes the verbs and nouns in a loop: update the features of verbs using nouns, and vice versa. Our framework successfully learns the semantically rich representations for both verbs and nouns and thus performs accurate recognition. To conclude, our contributions are mainly three folds:

- We reveal the problems of existing frameworks and point out that learning the bidirectional semantic relations is the core for accurate role recognition.
- We propose a two-stage framework with transformer structures to iteratively refine activity verbs and noun entities. It flexibly mines the potentially open semantic relations within verbs and nouns and alternately updates their features.
- Extensive experiments on challenging SWiG benchmarks fully demonstrate that our proposed framework outperforms other state-of-the-art methods under various metrics.

2 RELATED WORK

History of SR to GSR tasks. Although deep learning achieves satisfactory performance in basic vision tasks such as action recognition [6, 30, 39, 61, 64, 70, 79] and object detection [5, 11, 24, 40, 48, 58], they still cannot fully understand events in natural scenes. To address this problem, image captioning [1, 23, 49, 52, 69, 75] and scene graph generation [8, 20, 37, 66, 68, 71, 76] attempt to reason and describe scene content through natural language or relational graphs of objects. However, these efforts have still failed to understand events consistent with human cognition, i.e., identifying what happened and who was involved in what roles.

In this context, Yatskar *et al.* [73] first proposed Situation Recognition (SR) task and annotated imSitu dataset as the benchmark. However, the original SR task cannot point to where the involved noun entities are located in the image. To further address the visual grounding of the entities, Pratt *et al.* [47] redefines Grounded Situation Recognition (GSR) task and proposes SWiG dataset by providing bounding box annotations on imSitu dataset. GSR task can be seen as a further extension of SR task.

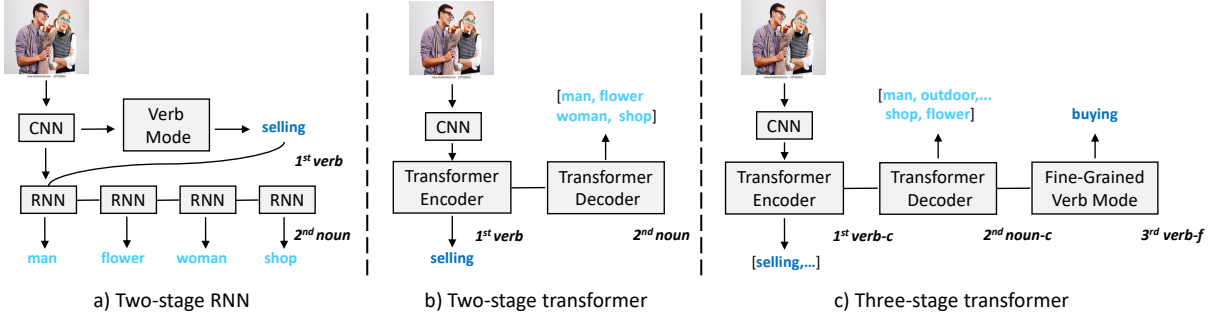


Figure 3: GSR task learning framework. (a) Two-stage RNN pre-detects verbs in the first stage and predicts noun entities in an auto-regressive manner in the second stage. (b) Two-stage transformer first uses the encoder to detect verbs and then uses the decoder to predict noun entities. (c) Three-stage transformer [18, 65] adopts a coarse-to-fine refinement process. It still obeys the two-stage idea, where the first stage detects a set of similar verbs and the second stage recognizes noun entities. The third stage utilizes noun entities to refine verbs.

Challenges of GSR tasks. The challenges of GSR task are to handle the semantic relations in the scene. Yatskar *et al.* [73] proposed a conditional random field (CRF) model in the initial phase. However, follow-up work [72] pointed out that CRF method cannot effectively utilize semantic relations. Since then, a lot of works have started to investigate how to model the relations among semantic roles. The previous technical routes mainly include Recurrent Neural Network (RNN) [42, 47, 63, 63], Graph Neural Network (GNN) [36, 54] and Relational Reasoning [4, 22], etc.

Transformers in GSR task. After great success in NLP tasks, transformer structure [62] was introduced to solve various computer vision problems, including image generation [17, 46], image recognition [25, 41, 60], object detection [5, 80], object segmentation [74], image captioning [23] and scene graph generation [20], etc. To further exploit the strength of the transformer in GSR task, Cho *et al.* [19] proposed the first transformer framework (GSRTR) by replacing the object queries in transformer object detector (DETR [80]) with semantic role queries. Wei *et al.* [65] recently proposed SituFormer, which uses two transformer-based verb and noun detectors to improve the performance. In addition, Cho *et al.* [18] proposed CoFormer that tends to leverage the semantic relations between the verb and noun roles to refine the prediction.

Problem of framework. To elaborate, as shown in Figure 3, almost all existing GSR methods [4, 19, 22, 42, 47, 54, 63] adopt a two-stage framework based on RNN or transformer. In RNN structure, Pratt *et al.* [47] proposes a Joint Situation Localizer (JSL) model, which consists of a verb classifier in the first stage and an RNN-based object detector in the second stage. In transformer structure, GSRTR [19] uses a transformer encoder to detect verbs in the first stage and a transformer decoder to predict semantic roles in the second stage. These two-stage frameworks evidently cannot exploit the semantic relations. First, recklessly detecting verbs in the first stage will inevitably misidentify similar activities. Second, the closed auto-regressive strategy of the second stage not only fails to correct misrecognized verbs, but leads to more mispredicted semantic roles. Although recent works (SituFormer [65] and CoFormer [18]) are also aware of framework issues and adopt a three-stage framework to refine prediction results from coarse to fine, they are still unable to optimize the results with bidirectional semantic relationships (i.e., from both verbs and nouns).

3 PROPOSED METHOD

3.1 Problem Formulation

Definition of GSR task. Given an image I , GSR aims to generate a structured verb frame $\mathcal{F}_v = \{v, \mathcal{R}_v\}^2$. As shown in Figure 1, GSR not only recognizes the salient verb $v \in \mathcal{V}$, but also detects all corresponding semantic roles $\mathcal{R}_v = \{(r, n_r, b_r) \mid \text{for } r \in \mathcal{R}_v\}$, where $\mathcal{R}_v = \{r_1, \dots, r_m\}$ is the set of role types for verb v . For instance, the verb-frame in Figure 1 can be instantiated as $\mathcal{F}_v = (\text{Sprinkling}, \{(Agent, Man, \square), (Item, Spice, \square), (Source, Cup, \square), (Destination, Pan, \square), (Place, Kitchen, \emptyset_b)\})$. Semantic roles \mathcal{R}_v is a collection of triples, where each role contains the role type r , the noun entity $n_r \in \mathcal{N}$ and the corresponding bounding box $b_r \in \mathbb{R}^4$. Note that not all semantic roles have corresponding nouns and bounding boxes, i.e., n_r or b_r can be equal to $\{\emptyset\}$.

Problems of existing frameworks. Inspired by object detection and image captioning, as shown in Figure 3, the existing GSR methods [19, 36, 42, 47, 54] widely adopt a two-stage framework, i.e., 1) identifying salient verbs v in the first stage, and 2) detecting the corresponding semantic roles \mathcal{R}_v in the second stage:

$$\mathcal{P}(\mathcal{F}_v|I) = \underbrace{\mathcal{P}(v|I)}_{\text{verb}} \underbrace{\mathcal{P}(\mathcal{R}_v|v, I)}_{\text{noun}}. \quad (1)$$

There are multiple highly similar activities in GSR tasks, such as *buying* and *giving*, as shown in Figure 2. This two-stage framework is inherently unable to utilize the noun entities in the first stage to distinguish similar verbs, let alone misidentify semantic roles due to the verb prediction errors accumulated in the second stage.

To address these issues, Wei *et al.* [65] recently proposed a three-stage framework (SituFormer) as,

$$\mathcal{P}(\mathcal{F}_v|I) = \underbrace{\mathcal{P}(\mathcal{V}_c|I)}_{\text{verb-c}} \underbrace{\mathcal{P}(\{\mathcal{R}_v\}_c|\mathcal{V}_c, I)}_{\text{noun-c}} \underbrace{\mathcal{P}(v, \mathcal{R}_v|\mathcal{V}_c, \{\mathcal{R}_v\}_c, I)}_{\text{verb-f}}, \quad (2)$$

where the main idea is to refine verb predictions in a coarse-to-fine manner. As shown in Figure 3 (c), it uses the first two stages to identify a set of candidate verbs \mathcal{V}_c and corresponding noun entities $\{\mathcal{R}_v\}_c$. Then the third stage uses a ranking loss to refine the verb v with a set of support images through similarity retrieval.

²These predefined verb frames are filtered from PropBank [33] or FrameNet [3, 27].

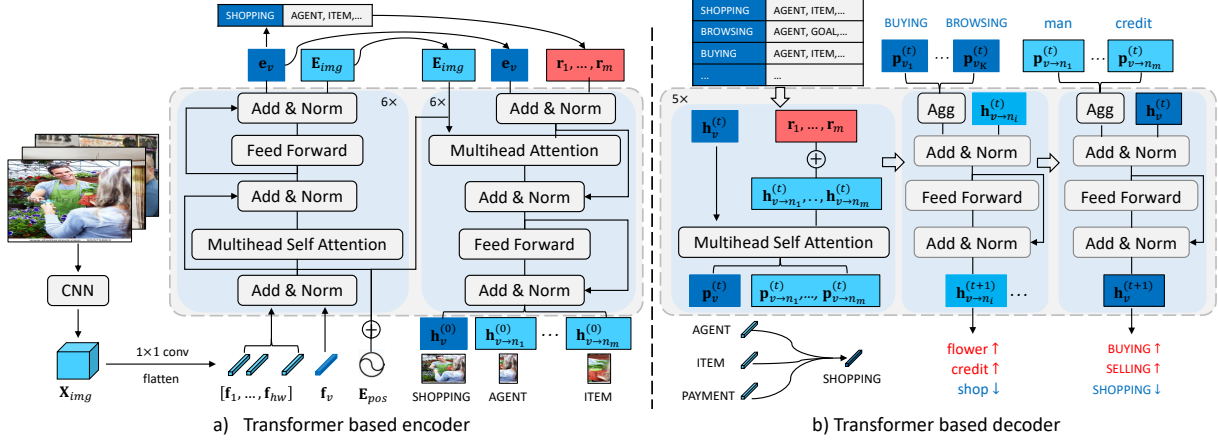


Figure 4: GSRFormer follows the classic encoding and decoding framework. Transformer encoder utilizes two multi-head attention modules to learn the intermediate features for verbs and semantic roles. Transformer decoder consists of four parts. 1) A set of similar activities (verbs) are retrieved using features from the encoder. 2) A multi-head attention layer is used to compute the messages p , thus capturing semantic relations among verbs and nouns. 3-4) The features of nouns and verbs are updated alternately with the computed messages. Note that we stack multiple decoder layers for iterative optimization (refinement).

Similarly, Cho *et al.* [18] also proposed a three-stage framework (CoFormer). From the framework perspective, CoFormer is similar to SituFormer. 1) The first stage coarsely predicts noun roles (Glance). 2) The second stage uses the predicted noun roles to help predict the verb (Gaze-Step1). 3) The third stage refines the candidate noun roles in Gaze-Step1 using the predicted verb (Gaze-Step2).

Obviously, these revised frameworks are still unreasonable due to the following issues. 1) It uses only noun entities to refine the verb, but totally ignores semantic constraints in the opposite direction (i.e., from verbs to noun entities). 2) Although it has two redundant transformer-based noun and verb detectors, it does not learn the semantic relations between them at the encoding stage. 3) The coarse-to-fine refinement process can only be done once, and it is impossible to iteratively optimize both verbs and nouns.

3.2 Overview of GSRFormer

To solve these problems, we reconstruct a framework (GSRFormer) with transformers to fully exploit the semantic relations of GSR tasks. We reformulate GSR task as,

$$\mathcal{P}(\mathcal{F}_v|I) = \underbrace{\mathcal{P}(\tilde{v}, \mathcal{H}_v|I)}_{\text{step-1: encoding}} \underbrace{\mathcal{P}(v, \mathcal{R}_v|\mathcal{H}_v, I)}_{\text{step-2: decoding}}, \quad (3)$$

where \tilde{v} is the assumed pseudo verb category, and \mathcal{H}_v is the intermediate feature for the verb and associated roles. The godsend is that it yields the essence of the GSR task, which is to extract comprehensive semantic relations and then iteratively refine the results. As shown in Figure 4, GSRFormer consists of a transformer-based encoder and decoder, respectively. In the first step (Sec. 3.3), instead of pre-detecting verbs, the transformer encoder learns intermediate representations \mathcal{H}_v of verbs and corresponding semantic roles from images, respectively. In the second step (Sec. 3.4), we first use the features obtained in the encoder to retrieve the representations of Top- K similar verbs as the support verb set $\{\mathcal{H}_v\}_s$. To mine the various semantic relations, the neural message passing mechanism [28] is then used to flexibly associate relevant relations to each verb or role, thus effectively updating their representations. Finally, we

take full advantage of the transformer structure to perform multiple iterations to refine verbs v and semantic roles \mathcal{R}_v .

3.3 Transformer-based Encoder

The transformer encoder is devised to learn intermediate representations of verbs and semantic roles from images using two multi-head attention modules, respectively.

Representation of verbs. As shown in Figure 4, given an image I , CNN backbone first extracts the feature map $X_{img} \in \mathbb{R}^{c \times h \times w}$. Since the input to the transformer encoder is a sequence of tokens, a 1×1 convolutional layer and a flatten operator are used to convert the X_{img} into a sequence of “visual” tokens $[f_1, \dots, f_{hw}]$, where each token $f_i \in \mathbb{R}^d$ is compressed as a d -dim visual feature. Inspired by the classification token used in ViT [25], we initialize a learnable verb token $f_v \in \mathbb{R}^d$ to stands for verb v . Then all visual and verb token sequences are fed into the first encoding module as,

$$[e_v, e_1, \dots, e_{hw}] = \text{MHA}_{img}^{\text{verb}}([f_v, f_1, \dots, f_{hw}] \oplus E_{pos}), \quad (4)$$

where \oplus is element-wise addition and E_{pos} is positional embedding used to distinguish relative positions in the sequence. $\text{MHA}_{img}^{\text{verb}}$ is a set of stacked multi-head self-attention blocks. As shown in Figure 4, each block consists of a multi-head self-attention layer and a feed-forward network, and the layer normalization [2] (Add & Norm) is used before both of them³. The role of $\text{MHA}_{img}^{\text{verb}}$ module is to use multi-head self-attention to learn the intermediate representation of the verb from image features. The output can be divided into optimized 1) image features $E_{img} = [e_1, \dots, e_{hw}] \in \mathbb{R}^{d \times hw}$ and 2) verb feature $e_v \in \mathbb{R}^d$. Here verb feature e_v is fed to a classifier to determine a pseudo verb category \tilde{v} . After obtaining \tilde{v} , we can fetch the corresponding semantic roles \mathcal{R}_v and initialize them to a sequence $[r_1, \dots, r_m]$, where each initialized role $r_i \in \mathbb{R}^d$ is a d -dim visual feature.

³ E_{pos} has the same dimension as sequences of visual and verb tokens. For more implementation details, see original papers [19, 62].

Representation of semantic roles. We further learn intermediate representations of corresponding semantic roles plus verb $\mathcal{H}_v = [\mathbf{h}_v^{(0)}, \mathbf{h}_{v \rightarrow n_1}^{(0)}, \dots, \mathbf{h}_{v \rightarrow n_m}^{(0)}]$ from images. Specifically, \mathbf{E}_{img} and \mathbf{e}_v are used as the input to the second encoding module as,

$$\begin{aligned} & \left[\mathbf{h}_v^{(0)}, \mathbf{h}_{v \rightarrow n_1}^{(0)}, \dots, \mathbf{h}_{v \rightarrow n_m}^{(0)} \right] = \\ & \text{MHA}_{\text{img}}^{\text{roles}} \left(\underbrace{\mathbf{E}_{\text{pos}} \oplus \mathbf{E}_{\text{img}}}_{\text{key/value}}, \underbrace{[\mathbf{e}_v, \mathbf{r}_1, \dots, \mathbf{r}_m]}_{\text{query}} \right), \end{aligned} \quad (5)$$

where we concatenate verb feature \mathbf{e}_v and corresponding semantic role embeddings $[\mathbf{r}_1, \dots, \mathbf{r}_m]$ as query. As shown in Figure 4, similar to the previous encoding module, $\text{MHA}_{\text{img}}^{\text{roles}}$ is also a set of stacked multi-head attention blocks. The output representation \mathcal{H}_v is then utilized in the decoder part. After GSRFormer encoder, we assume that the feature \mathcal{H}_v has captured the semantics from the image.

3.4 Transformer-based Decoder

The transformer decoder aims to utilize the semantic relations among a set of similar verbs and corresponding semantic roles to simultaneously refine their features. While transformer itself is a powerful model that can mine the implicit relations of all pairwise interactions, it still struggles to leverage additional domain-specific knowledge. In our framework, we expect to exploit a set of support verbs (excluding their corresponding nouns) to profit the noun refinement, while only using the nouns in one single image to refine its verb. This *a priori* knowledge can hardly be included when directly applying the transformer since it considers ALL pairwise relations. To this end, we borrow ideas from the neural message passing [28], where the “messages” are first computed by transformers, and then each verb or noun entity is updated using “messages” arising from the relations to other appropriate entities. As shown on the right of Figure 4, the transformer decoder is mainly composed of the following four parts.

Support verbs set. There are many very similar verbs in GSR tasks, as shown in Figure 2. Inspired by previous work [65], we use the features $\mathcal{H}_v = [\mathbf{h}_v, \mathbf{h}_{v \rightarrow n_1}, \dots, \mathbf{h}_{v \rightarrow n_m}]$ learned from the encoder to retrieve the features of top- K similar verbs $\{\mathcal{H}_v\}_s = \{\mathcal{H}_{v_1}, \dots, \mathcal{H}_{v_K}\}$ as support verbs set,

$$\{\mathcal{H}_v\}_s = \arg \max_{\mathcal{H}_{v_j} \mid v_j \in \mathcal{D}_{\mathcal{T}}} \text{top-}K \ S(\mathcal{H}_v, \mathcal{H}_{v_j}), \quad (6)$$

where $\mathcal{D}_{\mathcal{T}}$ is the set of all training images. The similarity score $S(\cdot, \cdot)$ is the average cosine similarity of semantic roles as,

$$S(\mathcal{H}_v, \mathcal{H}_{v_j}) = \frac{1}{m} \sum_{i=1}^m \text{sim}(\mathbf{h}_{v \rightarrow n_i}^{(0)}, \mathbf{h}_{v_j \rightarrow n_i}^{(0)}), \quad (7)$$

where $\text{sim}(\cdot)$ is cosine similarity. $\mathbf{h}_{v \rightarrow n_i}$ and $\mathbf{h}_{v_j \rightarrow n_i}$ are the noun entity features for the verbs v and v_j , respectively.

Semantic relation message computation. Given retrieved support set $\{\mathcal{H}_v\}_s$, we aim to update the feature representation of each element in \mathcal{H}_v by leveraging the relations between or within \mathcal{H}_v and $\{\mathcal{H}_v\}_s$. Following the standard message passing paradigm, we compute one message \mathbf{p} for each involved verb and noun in one

image by a multi-head attention (MHA) layer as,

$$\begin{aligned} & \left[\mathbf{p}_v^{(t)}, \mathbf{p}_{v \rightarrow n_1}^{(t)}, \dots, \mathbf{p}_{v \rightarrow n_m}^{(t)} \right] = \\ & \text{MHA}_{\text{verb}}^{\text{roles}} \left(\left[\mathbf{h}_v^{(t)}, \mathbf{r}_1 \oplus \mathbf{h}_{v \rightarrow n_1}^{(t)}, \dots, \mathbf{r}_m \oplus \mathbf{h}_{v \rightarrow n_m}^{(t)} \right] \right), \end{aligned} \quad (8)$$

where \oplus is element-wise addition, and t implies the t -th iteration. $\mathbf{h}_v^{(t)}$ and $\mathbf{h}_{v \rightarrow n_i}^{(t)}$ are the learned verb and noun entity features, respectively. Here we replace the positional encoding in the original MHA layer with the semantic role embedding $\mathbf{r}(\cdot)$. Thus $\text{MHA}_{\text{verb}}^{\text{roles}}$ is actually a multi-head self-attention model. We perform the relation message computation for the entire support set. The obtained verb message $\mathbf{p}_v^{(t)}$ and the noun entity message $\mathbf{p}_{v \rightarrow n_i}^{(t)}$ contain all the semantic information within a single image. Below we will consider semantic relations in multiple verbs (i.e., support verbs set).

Refine noun entity with verbs. We utilize the semantic relations (messages) from the verbs of support set to refine the noun entities.

To update the representation of a noun entity $\mathbf{h}_{v \rightarrow n_i}^{(t)}$, a single update message $\mathbf{p}_{v_{\text{all}}}^{(t)}$ is computed by aggregating the messages of support set verbs $\{\mathbf{p}_{v_1}^{(t)}, \dots, \mathbf{p}_{v_K}^{(t)}\}$ as Eq. 9. The aggregation function $\text{Agg}(\cdot)$ can be any permutation-invariant function (e.g., element-wise sum and max), and here we employ a gated update function [38]. After the messages are fused, a transformer sublayer (FFN and LN) are used to updated representation with residual connection:

$$\mathbf{p}_{v_{\text{all}}}^{(t)} = \text{Agg}(\{\mathbf{p}_{v_k}^{(t)} \mid \text{for } 1 \leq k \leq K\}), \quad (9)$$

$$\mathbf{q}_{v \rightarrow n_i}^{(t)} = \text{LN} \left(\mathbf{h}_{v \rightarrow n_i}^{(t)} + \mathbf{p}_{v_{\text{all}}}^{(t)} \right), \quad (10)$$

$$\mathbf{h}_{v \rightarrow n_i}^{(t+1)} = \text{LN} \left(\mathbf{q}_{v \rightarrow n_i}^{(t)} + \text{FFN} \left(\mathbf{q}_{v \rightarrow n_i}^{(t)} \right) \right), \quad (11)$$

where $\text{LN}(\cdot)$ is layer normalization [2] and $\text{FFN}(\cdot)$ is a feed-forward neural network (commonly with one large intermediate layer).

Refine verb with noun entities. Similarly, when updating the verb feature, we utilize the messages of nouns only from the single associated image, as shown in Eq. 12-14,

$$\mathbf{p}_{n_{\text{all}}}^{(t)} = \text{Agg}(\{\mathbf{p}_{v \rightarrow n_i}^{(t)} \mid \text{for } 1 \leq i \leq m\}), \quad (12)$$

$$\mathbf{q}_v^{(t)} = \text{LN} \left(\mathbf{h}_v^{(t)} + \mathbf{p}_{n_{\text{all}}}^{(t)} \right), \quad (13)$$

$$\mathbf{h}_v^{(t+1)} = \text{LN} \left(\mathbf{q}_v^{(t)} + \text{FFN} \left(\mathbf{q}_v^{(t)} \right) \right). \quad (14)$$

Note that the above two refining processes can be accomplished alternately. Unlike the previous work [65], which only optimized from rough noun entities to verbs, our framework takes full advantage of the flexibility of the transformer structure to efficiently perform multiple refinement iterations.

After completing the refinement of noun entities and verb features for T iterations, we employ a lightweight MLP over the verb and noun entity features, respectively, to achieve the classification of verbs and the regression of noun entities with bounding boxes,

$$v = \text{MLP}(\mathbf{h}_v^T), \quad (15)$$

$$\{n_i, \mathbf{b}_i\} = \text{MLP}(\mathbf{h}_{v \rightarrow n_i}^T). \quad (16)$$

We discuss the training process in detail in the following section. In the ablation studies, we discuss the effect of the number of loops and refinement order on the results in detail.

Table 1: Performance (%) comparisons of GSRFormer (ours) and baseline methods on SWiG dataset development set.

Models	Top-1-Verb					Top-5-Verb					Ground-Truth-Verb			
	verb	value	val-all	grnd	grnd-all	verb	value	val-all	grnd	grnd-all	value	val-all	grnd	grnd-all
Methods for Situation Recognition														
CRF [73]	32.25	24.56	14.28	–	–	58.64	42.68	22.75	–	–	65.90	29.50	–	–
CRF+DataAug [72]	34.20	25.39	15.61	–	–	62.21	46.72	25.66	–	–	70.80	34.82	–	–
VGG+RNN [42]	36.11	27.74	16.60	–	–	63.11	47.09	26.48	–	–	70.48	35.56	–	–
FC-Graph [36]	36.93	27.52	19.15	–	–	61.80	45.23	29.98	–	–	68.89	41.07	–	–
CAQ [22]	37.96	30.15	18.58	–	–	64.99	50.30	29.17	–	–	73.62	38.71	–	–
Kernel-Graph [54]	43.21	35.18	19.46	–	–	68.5	56.32	30.56	–	–	73.14	41.68	–	–
Methods for Grounded Situation Recognition														
ISL [47]	38.83	30.47	18.23	22.47	7.64	65.74	50.29	28.59	36.90	11.66	72.77	37.49	52.92	15.00
JSL [47]	39.60	31.18	18.85	25.03	10.16	67.71	52.06	29.73	41.25	15.07	73.53	38.32	57.50	19.29
GSRTTR [19]	41.06	32.52	19.63	26.04	10.44	69.46	53.69	30.66	42.61	15.98	74.27	39.24	58.33	20.19
SituFormer [65]	44.32	35.35	22.10	29.17	13.33	71.01	55.85	33.38	45.78	19.77	76.08	42.15	61.82	24.65
CoFormer [18]	44.41	35.87	22.47	29.37	12.94	72.98	57.58	34.09	46.70	19.06	76.17	42.11	61.15	23.09
GSRFormer (ours)	46.64	37.69	23.58	31.61	14.42	73.43	58.75	35.82	48.42	21.67	78.76	44.71	63.95	25.85

3.5 Training Objectives

We use the same data augmentation and batch training strategy as previous work [19]. The training details for the encoder and decoder are as follows.

Training of encoder. We use the cross-entropy loss function to train the encoder to obtain the pseudo-verb category as,

$$\mathcal{L}_{\text{verb-e}} = \mathcal{L}_{\text{CE}}(v^{gt}, \tilde{v}), \quad (17)$$

where the ground-truth verb category is denoted as v^{gt} and the predicted pseudo-verb category is \tilde{v} . Note that the first multi-head attention module of the encoder performs gradient back-propagation only when training the encoder, and does not participate in parameter updates when training the decoder.

Training of decoder. When training the decoder, we need to optimize the categories of verbs and nouns as well as the bounding boxes of nouns. Following previous work [19, 65], the losses of the decoder are calculated as,

$$\mathcal{L}_{\text{verb-d}} = \mathcal{L}_{\text{CE}}(v^{gt}, v), \quad (18)$$

$$\mathcal{L}_{\text{nouns}} = \sum_{i=1}^m \left[\mathcal{L}_{\text{CE}}(n_i^{gt}, n_i) + \mathcal{L}_{\text{box}}(\mathbf{b}_i^{gt}, \mathbf{b}_i) \right], \quad (19)$$

where we use the cross-entropy loss function to train the decoder to get the true verb categories v . For the noun loss function, n_i^{gt} and \mathbf{b}_i^{gt} denote the ground-truth noun category and bounding box, while n_i and \mathbf{b}_i are the predicted noun category and bounding box. \mathcal{L}_{CE} is the cross-entropy loss for noun classification. \mathcal{L}_{box} consists of the generalized IoU loss [50] and the $L1$ regression loss.

Process of inference. Similar to previous methods [19, 42, 47, 65], GSRFormer also requires inference in the encoder and decoder separately. Compared to the three-stage SituFormer [65], our inference process is more straight. At inference time, GSRFormer first predicts a pseudo-verb category and then constructs the corresponding semantic roles to learn the representations using the encoder. Based on the output features from the encoder, a set of similar support verbs is then retrieved in the training set as the input to the decoder. Finally, the decoder produces the verb and noun predictions.

4 EXPERIMENT

4.1 Experimental Settings

Datasets. Our experiments are carried out on the challenging SWiG benchmark [47]. SWiG dataset builds on the original Situation Recognition (SR) imSitu dataset [73] by adding bounding box (bbox) annotations for all visible semantic roles (63.9% of roles have bbox annotations). Since each image in imSitu is annotated with three verb frames by three annotators, SWiG contains 126,102 images with 504 verbs and 190 semantic role types, and each verb is followed by 1 to 6 semantic roles (3.55 on average). We followed the official splits to construct the training/validation/testing set with sizes of 75K/25K/25K, respectively.

Evaluation metrics. We use the same five evaluation metrics as Pratt *et al.* [47], including 1) **verb**: the accuracy of verb prediction, 2) **value**: the accuracy of noun prediction for each semantic role, 3) **val-all**: the accuracy of noun prediction for the whole semantic role set, 4) **grnd**: the accuracy of noun prediction with correct grounding (bbox) for each semantic role, 5) **grnd-all**: the accuracy of noun prediction with grounding (bbox) for the whole semantic role set. Note that we consider a grounding is correct if the IoU between the predicted and ground-truth bbox is above 0.5. Meanwhile, we report the above metrics in three evaluation settings: 1) **Top-1-verb**, 2) **Top-5-verb** and 3) **Ground-Truth-Verb**, which select verbs based on top-1 prediction, top-5 predictions, and corresponding ground truth, respectively. If verb predictions are incorrect in the Top-1-verb and Top-5-verb settings, the corresponding semantic role predictions are also considered false.

4.2 Comparisons with State-of-the-Arts

Baseline models. Existing SR models can be classified into: 1) **CRF** [73]: CRF-based model, 2) **CRF+DataAug** [72]: CRF-based model with data augmentation, 3) **VGG+RNN** [42]: RNN-based prediction model with VGG backbone, 4) **FC-Graph** [36]: GNN-based model with fully connected semantic roles, 5) **CAQ** [22]: Query-based model with top-down attention, 6) **Kernel-Graph** [54]: GNN-based model with mixture kernel attention. Correspondingly, existing GSR models can be divided as: 1) **ISL** [47]: RNN-based method has independent semantic role values and grounding predictions. 2) **JSL** [47]: RNN-based methods jointly predict semantic role values and their basis. 3) **GSRTTR** [19]: Transformer two-stage model has both

Table 2: Performance (%) comparisons of GSRFormer (ours) and baseline methods on SWiG dataset test set.

Models	Top-1-Verb					Top-5-Verb					Ground-Truth-Verb			
	verb	value	val-all	grnd	grnd-all	verb	value	val-all	grnd	grnd-all	value	val-all	grnd	grnd-all
Methods for Grounded Situation Recognition														
ISL [47]	39.36	30.09	18.62	22.73	7.72	65.51	50.16	28.47	36.60	11.56	72.42	37.10	52.19	14.58
JSL [47]	39.94	31.44	18.87	24.86	9.66	67.60	51.88	29.39	40.60	14.72	73.21	37.82	56.57	18.45
GSRTTR [19]	40.63	32.15	19.28	25.49	10.10	69.81	54.13	31.01	42.50	15.88	74.11	39.00	57.45	19.67
SituFormer [65]	44.20	35.24	21.86	29.22	13.41	71.21	55.75	33.27	46.00	20.10	75.85	42.13	61.89	24.89
CoFormer [18]	44.66	35.98	22.22	29.05	12.21	73.31	57.76	33.98	46.25	18.37	75.95	41.87	60.11	22.12
GSRFormer (ours)	46.53	37.48	23.32	31.53	14.23	73.44	58.84	35.82	48.43	21.41	78.81	44.68	63.87	25.35

Table 3: Effectiveness of each component of GSRFormer.

Components	verb	value	val-al	grnd	grnd-all
w/o Encoder-1st	35.30	25.44	15.69	21.27	7.60
Gains (Δ)	-11.23	-12.04	-7.63	-10.26	-6.63
w/o Encoder-2nd	32.84	24.21	14.60	20.91	7.66
Gains (Δ)	-13.69	-13.27	-8.72	-10.62	-6.57
w/o Decoder	34.94	25.08	14.28	20.99	7.79
Gains (Δ)	-11.59	-12.40	-9.04	-10.54	-6.44
w/o Iteration	39.10	31.30	19.11	26.18	11.92
Gains (Δ)	-7.43	-6.18	-4.21	-5.35	-2.31
w/o Alternate	35.81	27.87	16.39	22.09	9.17
Gains (Δ)	-10.72	-9.61	-6.93	-9.44	-5.06
w/o Message	38.06	29.87	17.69	21.83	7.77
Gains (Δ)	-8.47	-7.61	-5.63	-9.70	-6.46
GSR-Former (ours)	46.53	37.48	23.32	31.53	14.23

a verb predictor and semantic role detector. 4) **SituFormer** [65]: Transformer three-stage model consists of a coarse-to-fine verb predictor and a semantic role detector. 4) **CoFormer** [18]: Transformer three-stage model exploits the semantic relations between the verb and noun roles to improve results.

Results under Ground-Truth-Verb setting. The Ground-Truth-Verb setting evaluates whether the system can understand events in a human-cognitive manner. The numerical value of this setting describes how well the machine predictions match the human annotations. The experimental results are shown in Table 1 and Table 2. In general, our GSRFormer outperforms other state-of-the-art methods. Compared with SituFormer[65], which also adopts the transformer structure, GSRFormer improves the accuracy of noun prediction in single (value) and all semantic roles (val-all) by 2.96% and 2.55%, respectively. Furthermore, GSRFormer achieves similar improvements under the vision grounding setting (i.e., grnd and grnd-all), which shows that GSRFormer can effectively learn visual information from natural scenes.

Results under Top-N-Verb settings. We use the Top-N-Verb setting to evaluate the accuracy of predicting verb categories. The results in Table 1 and Table 2 show that our GSRFormer outperforms other state-of-the-art methods. It is well known that SituFormer [65] adopts a three-stage framework to improve the accuracy of verb prediction. Compared to SituFormer, GSRFormer can further push the verb prediction accuracy by 2.33% (absolute) under the Top-1-Verb setting. In addition, GSRFormer also achieves a more splendid improvement on SituFormer in single noun prediction (2.24% on value) under the Top-1-Verb setting. Unlike SituFormer, which only uses nouns to improve verbs, our GSRFormer adopts a bidirectional refinement strategy to iterative optimize the results. The 2.24% increase in noun prediction accuracy fully reveals the effectiveness of our proposed alternative semantic refinement.

Table 4: Effects of adopting two opposite refinement orders.

Order	Ground-Truth-Verb				
	verb	value	val-all	grnd	grnd-all
Refine-Verb-First	-	75.45	42.29	61.61	24.72
Refine-Noun-First	-	78.81	44.68	63.87	25.35
Top-1-Verb					
Refine-Verb-First	45.23	36.26	22.57	31.12	13.65
Refine-Noun-First	46.53	37.48	23.32	31.53	14.23

Table 5: Effects of utilizing different aggregate functions.

Aggregate Functions	value	val-all	grnd	grnd-all
Element-wise Sum	75.64	42.79	61.31	24.38
Max-Pooling	76.30	42.94	61.85	24.23
Aggregated Message [34]	77.79	43.58	62.30	24.76
Gated Function [38]	78.81	44.68	63.87	25.35

4.3 Ablation Studies

Effectiveness of encoder. We perform ablation studies on two multi-head attention modules of the encoder (denoted as Encoder-1st and Encoder-2nd), as shown in Table 3. Experimental results show that both structures significantly improve performance (over 10% absolute improvement). This fully demonstrates the effectiveness of the encoder module. Encoder-1st is valuable in understanding visual grounding details (comparable on grnd-all metrics), while Encoder-2nd is more conducive to mining verbs and semantic roles. **Effectiveness of decoder.** We validate the entire decoder module. By removing the decoder, the modified model resembles a simple two-stage approach [19]. The results in Table 3 show that the absolute improvement is also over 10%, which fully demonstrates the effectiveness of the decoder module. Below we will analyze each part of the decoder in detail.

Effectiveness of iterative refinement. We verify iterative refinement. Without iterative refinement (i.e., only 1 transformer layer in the decoder instead of 5), the performance (2%-8% drop) will be similar to traditional RNN-based methods (e.g., JSL[47]). It hints that the real improvement to the transformer structure is the ability to make iterative improvements.

Effectiveness of alternate optimization. We verify alternate optimization. By removing this, we mean to update the verb with support verbs, and update nouns with other nouns. We are surprised to find that alternate optimization resulted in more performance drop than iterative refinement, except for the visual relevant grnd-all metrics. This fully illustrates that the core of iterative refinement is to exploit the semantic relationship between verbs and nouns (roles) in both directions.

Effectiveness of message computation. We compare with that removes the message passing mechanism, i.e., directly using features of transformer encoder to learn the semantic relations. As



Figure 5: (a) Comparison of GSRFormer and GSRTTR [19], where GSRFormer predicts more correct verbs and nouns under Top-5-Verb setting. (b) An example demonstrates the application of GSRFormer, which can serve cross-modal semantic question answering.

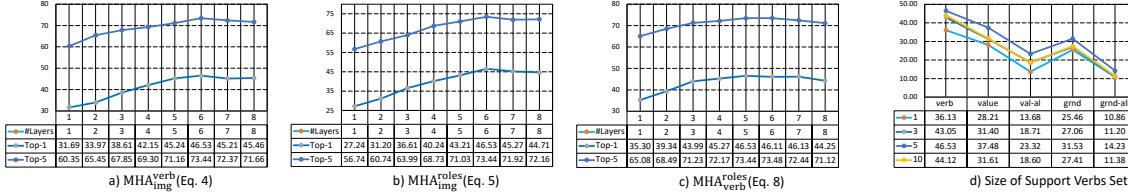


Figure 6: (a)-(c) show the verb prediction accuracy under the Top-1-Verb and Top-5-Verb settings. Here we verify the effect of different numbers of stacked layers on three multi-head attention structures. (d) reveal the effect of the size of the support verbs under the Top-1-Verb setting.

shown in Table 3, message passing not only effectively uses semantic relations to improve verb and noun prediction, but also learns the visual grounding information (grnd and grnd-all metrics).

Effects of refinement order. Table 4 shows that refining the noun first gives better results. This is the exact opposite of the previous work [19, 42, 47]. We speculate that this is because the set of noun entities is more relevant to the visual groundings, so predicting nouns first leads to fewer errors than verbs. For example, humans naturally recognize noun entities as evidence for judgments when stating verbs. This inspires us to prioritize noun entities.

Effects of aggregate functions. We test four aggregate functions. Table 5 shows that Gated Function [38] achieves the best results and Element-Wise Sum does not reach satisfactory results. This contrast points out that aggregation functions should strive to highlight semantically relevant features while ignoring unnecessary noise. In future work, we will explore more fusion strategies.

Effects of the number of stacked layers. We validate the effect of stacking layers in three multi-head attention structures. As shown in Figure 6, the best performance is achieved when six layers are stacked in the encoder and five layers in the decoder. As the number of stacked layers increases, the performance first increases and then decreases. We attribute this performance change to the noise of stacking too many layers.

Effects of the size of support verb set. As shown in Figure 6 (d), the best performance is when the support set size is 5. The performance drops when the size is 1, indicating that similar verbs can support semantic understanding. Moreover, when the size is 10, the performance does not further improve, which means that expanding support verbs also raises noise.

4.4 Visualization and Application

We visualize the results of GSRFormer in Figure 5 (a). With the help of alternate semantic refinement, GSRFormer predicts some

almost indistinguishable action verbs and semantic roles (see red font). We also show an example of the semantic question-answering application in Figure 5 (b). For example, when we ask questions like "How are women's eyebrows made?", GSRFormer can not only utilize the generated verb-frame to understand the question, but also provide structured answers with rich image facts.

5 CONCLUSION

In this paper, we first reveal the problems of existing frameworks and point out that the use of semantic relations is the root of the GSR task. To this end, we propose GSRFormer, a two-stage transformer framework that utilizes bidirectional semantic relations to iteratively refine predictions of verb and noun entities. Experimental results on SWiG dataset show that it outperforms other state-of-the-art methods. In the future, we will further explore to explain the semantic structure of GSRFormer and extend it to other semantic analysis tasks.

ACKNOWLEDGMENTS

This work was supported by the Air Force Research Laboratory under agreement number FA8750-19-2-0200; the Defense Advanced Research Projects Agency (DARPA) grants funded under the GAILA program (award HR00111990063); the Defense Advanced Research Projects Agency (DARPA) grants funded under the AIDA program (FA8750-18-20018).

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

REFERENCES

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [3] Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.
- [4] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1989–1998.
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European conference on computer vision*. Springer, 213–229.
- [6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [7] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10800–10809.
- [8] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. 2019. Counterfactual Critic Multi-Agent Training for Scene Graph Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [9] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5659–5667.
- [10] Tao Chen, Xiangnan He, and Min-Yen Kan. 2016. Context-aware image tweet modelling and recommendation. In *Proceedings of the ACM international conference on Multimedia*. 1018–1027.
- [11] Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. 2021. Robust and accurate object detection via adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16622–16631.
- [12] Zhi-Qi Cheng, Yang Liu, Xiao Wu, and Xian-Sheng Hua. 2016. Video ecommerce: Towards online video advertising. In *Proceedings of the 24th ACM international conference on Multimedia*. 1365–1374.
- [13] Zhi-Qi Cheng, Xiao Wu, Siyu Huang, Jun-Xiu Li, Alexander G Hauptmann, and Qiang Peng. 2018. Learning to transfer: Generalizable attribute learning with multitask neural model search. In *Proceedings of the ACM international conference on Multimedia*. 90–98.
- [14] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. 2017. Video eCommerce++: Toward large scale online video advertising. *IEEE transactions on multimedia* 19, 6 (2017), 1170–1183.
- [15] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. 2017. Video2shop: Exact matching clothes in videos to online shopping images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4048–4056.
- [16] Zhi-Qi Cheng, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. 2017. On the selection of anchors and targets for video hyperlinking. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 287–293.
- [17] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509* (2019).
- [18] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. 2022. Collaborative Transformers for Grounded Situation Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [19] Junhyeong Cho, Youngseok Yoon, Hyeonjun Lee, and Suha Kwak. 2021. Grounded Situation Recognition with Transformers. *Proceedings of the British Machine Vision Conference* (2021).
- [20] Yuren Cong, Wentong Liao, Hanno Ackermann, Michael Ying Yang, and Bodo Rosenhahn. 2021. Spatial-Temporal Transformer for Dynamic Scene Graph Generation. *arXiv preprint arXiv:2107.12309* (2021).
- [21] Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. Unifying cross-lingual Semantic Role Labeling with heterogeneous linguistic resources. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. 338–351.
- [22] Thilini Cooray, Ngai-Man Cheung, and Wei Lu. 2020. Attention-based context aware reasoning for situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4736–4745.
- [23] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10578–10587.
- [24] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. 2021. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7373–7382.
- [25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [26] Charles J Fillmore and Collin F Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, Vol. 6.
- [27] Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. 2003. Back-ground to framenet. *International journal of lexicography* 16, 3 (2003), 235–250.
- [28] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the International conference on machine learning*. PMLR, 1263–1272.
- [29] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. 2022. Group Contextualization for Video Recognition. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 928–938.
- [30] Jun-Yan He, Xiao Wu, Zhi-Qi Cheng, Zhaoquan Yuan, and Yu-Gang Jiang. 2021. DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition. *Neurocomputing* 444 (2021), 319–331.
- [31] Siyu Huang, Xi Li, Zhi-Qi Cheng, Zhongfei Zhang, and Alexander Hauptmann. 2018. Gnas: A greedy neural architecture search method for multi-attribute learning. In *Proceedings of the ACM international conference on Multimedia*. 2049–2057.
- [32] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1655–1663.
- [33] Paul R Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank.. In *LREC. Citeseer*, 1989–1993.
- [34] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [35] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10313–10322.
- [36] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. 2017. Situation recognition with graph neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4173–4182.
- [37] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*. 1261–1270.
- [38] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. 2016. Gated Graph Sequence Neural Networks. In *Proceedings of the International Conference on Learning Representations*.
- [39] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. 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*.
- [40] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.
- [41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.
- [42] Arun Mallya and Svetlana Lazebnik. 2017. Recurrent models for situation recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 455–463.
- [43] Phuong Anh Nguyen, Qing Li, Zhi-Qi Cheng, Yi-Jie Lu, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. 2017. Vireo@ trecvid 2017: Video-to-text, ad-hoc video search and video hyperlinking. (2017).
- [44] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*. 3456–3465.
- [45] Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies* 3, 1 (2010), 1–103.
- [46] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4055–4064.
- [47] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Proceedings of the European Conference on Computer Vision*. Springer, 314–332.
- [48] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
- [49] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7008–7024.

- [50] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 658–666.
- [51] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual Semantic Role Labeling for Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5589–5600.
- [52] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [53] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199* (2018).
- [54] Mohammed Suhail and Leonid Sigal. 2019. Mixture-kernel graph attention network for situation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10363–10372.
- [55] Guang-Lu Sun, Zhi-Qi Cheng, Xiao Wu, and Qiang Peng. 2018. Personalized clothing recommendation combining user social circle and fashion style consistency. *Multimedia Tools and Applications* 77, 14 (2018), 17731–17754.
- [56] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. 982–997.
- [57] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
- [58] Mingxing Tan, Ruoming Pang, and Quoc V Le. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10781–10790.
- [59] Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [60] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.
- [61] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.
- [62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [63] Paul Vicol, Makarand Tapaswi, Lluís Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8581–8590.
- [64] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*. 3551–3558.
- [65] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. 2022. Rethinking the Two-Stage Framework for Grounded Situation Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [66] Meng Wei, Chun Yuan, Xiaoyu Yue, and Kuo Zhong. 2020. Hose-net: Higher order structure embedded network for scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1846–1854.
- [67] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2021. Video as Conditional Graph Hierarchy for Multi-Granular Question Answering. *arXiv preprint arXiv:2112.06197* (2021).
- [68] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5410–5419.
- [69] Guanghui Xu, Shuaicheng Niu, Minghui Tan, Yucheng Luo, Qing Du, and Qi Wu. 2021. Towards accurate text-based image captioning with content diversity exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12637–12646.
- [70] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. 2020. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 591–600.
- [71] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision*. 670–685.
- [72] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. 2017. Commonly uncommon: Semantic sparsity in situation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7196–7205.
- [73] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5534–5542.
- [74] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10502–10511.
- [75] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4651–4659.
- [76] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5831–5840.
- [77] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. 2021. Token shift transformer for video classification. In *Proc. of the ACM International Conference on Multimedia*. 917–925.
- [78] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14830–14840.
- [79] Zhichen Zhao, Huimin Ma, and Shaodi You. 2017. Single image action recognition using semantic body part actions. In *Proceedings of the IEEE international conference on computer vision*. 3391–3399.
- [80] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of the International Conference on Learning Representations*.

A IMPLEMENTATION DETAILS

Network structure. We use ImageNet pre-trained ResNet-50 as the backbone. To facilitate computation, we use 1×1 convolution to compress the dimension of image features to 512, which is consistent with the hidden dimension of each semantic role query and verb token. To further speed up inference, the dimension of verb embedding and role embedding is 256. Correspondingly, we use learnable 2D embeddings for positional encodings, which have the same dimensions as sequences of visual and verb tokens. The number of heads for all multi-head attention blocks is 8. The sizes of the hidden dimensions of the four structures are 2048, 1024, 1024, and 1024, and the dropout rates are 0.15, 0.3, 0.3, and 0.2, respectively. The bounding box regressor is three fully connected layers with ReLU activation function and 1024 hidden dimensions, using a dropout rate of 0.2. Label smoothing regularization [57] is used for target verb and noun labels with label smoothing factors of 0.3 and 0.2, respectively. We set the support verb set size to 5. The number of stacked layers of the encoder is set to 6, and the number of stacked layers of the decoder is set to 5.

Training details. Although using the data augmentation strategies similar to DETR [5] can improve the experimental results, for a fair comparison, we employ the same data augmentation procedure as previous work [19]. Specifically, Random Color Jittering, Random Grayscale Scaling, Random Scaling, and Random Horizontal Flipping are employed. The Hue, Saturation, and Lightness Scales in random Color Jittering are set to 0.1. Random Grayscale Scaling is set to 0.3. Random Scaling is set to 0.5, 0.75, and 1.0. The probability of Random Horizontal Flip is set to 0.5. The number of semantic roles varies from 1 to 6, depending on the verb category. To speed up training, inspired by previous work [19], we utilize zero padding for each output of the noun prediction branch to ensure batch training. Because there are as many semantic role queries as semantic roles, we directly ignore the padding output in the loss computation. When training the decoder, we need to alternately compute verb and noun losses separately. Since there are three nouns per semantic role, the final noun loss is the sum of the three noun losses. In addition, we also illustrate more visualization comparisons and application examples, as shown in Figure 7-8.

	VERB: GIVING → BUYING <ol style="list-style-type: none"> Giving Buying ↑ AGENT woman Selling Selling - GOODS flower Providing Giving ↓ PAYMENT credit card Offering Providing ↓ SELLER man Applying Offering ↓ PLACE shop 		VERB: SELLING → GIVING <ol style="list-style-type: none"> Selling Giving ↑ AGENT man Buying Offering ↑ ITEM flower Providing Buying ↓ RECIPIENT woman Offering Providing ↓ PLACE outdoors Applying Selling ↓
	VERB: READING → SIGNING <ol style="list-style-type: none"> Reading Signing ↑ AGENT woman Writing Writing - OBJECTIVE document Signing Sketching ↑ TOOL pen Sketching Helping ↑ PLACE room Helping Reading ↓ 		VERB: STARTING → COACHING <ol style="list-style-type: none"> Starting Coaching ↑ AGENT woman Resting Resting - STUDENT people Coaching Starting ↓ SKILL basketball Aiming Aiming - PLACE stadium Giving Giving -
	VERB: EXTERMINATING → VACUUMING <ol style="list-style-type: none"> Exterminating Vacuuming ↑ AGENT man Dusting Clearing ↑ SURFACE rug Mopping Mopping - TOOL vacuum Scrubbing Scrubbing - PLACE house Applying Dusting ↓ 		VERB: MAKING → BAKING <ol style="list-style-type: none"> Making Baking ↑ AGENT woman Frying Cooking ↑ FOOD cake Cooking Making ↓ FOODCONTAINER pan Microwaving Pouring ↑ HEATSOURCE ∅ Stirring Stirring - PLACE kitchen

Figure 7: Comparison between GSRFormer and GSRT [19]. The first two columns compare the verbs detected by GSRT and GSRFormer, respectively. The last two columns are the semantic roles predicted by GSRFormer. With the help of an iterative refinement mechanism, GSRFormer predicts more correct verbs and nouns (marked in red font) under the Top-5-Verb setting. These examples also clearly illustrate the importance of semantic relations for GSR tasks, i.e., bidirectional semantic ties can mutually refine the predictions of verbs and nouns.

What will the teacher do? What are the responsibilities of the teacher?	VERB: LECTURING <ol style="list-style-type: none"> AGENT teacher AUDIENCE student PLACE classroom 	VERB: TEACHING <ol style="list-style-type: none"> AGENT teacher AUDIENCE student PLACE classroom 	VERB: TRAINING <ol style="list-style-type: none"> AGENT teacher AUDIENCE student PLACE classroom 	VERB: ERASING <ol style="list-style-type: none"> AGENT teacher ERASED writing SOURCE blackboard PLACE classroom
How do people maintain their house? What people use to maintain their houses?	VERB: EXTERMINATING <ol style="list-style-type: none"> AGENT man PLACE house INSTRUMENT nozzle 	VERB: CLEANING <ol style="list-style-type: none"> AGENT man SOURCE brick TOOL sponge PLACE house 	VERB: MOPPING <ol style="list-style-type: none"> AGENT man SURFACE floor PLACE house 	VERB: SEALING <ol style="list-style-type: none"> AGENT man ITEM plastic wrap SEALANT tape PLACE house
How do people cook? What do people cook with and where do people cook?	VERB: FRYING <ol style="list-style-type: none"> AGENT man FOOD meat CONTAINER frying pan PLACE Kitchen 	VERB: BAKING <ol style="list-style-type: none"> AGENT man FOOD cake CONTAINER ∅ HEAT SOURCE ∅ PLACE kitchen 	VERB: CHOPPING <ol style="list-style-type: none"> AGENT man ITEM meat TOOL knife PLACE kitchen 	VERB: MASHING <ol style="list-style-type: none"> AGENT man ITEM food TOOL masher PLACE kitchen
How do people get treated? What treatment do people use and where do they treat it?	VERB: CHECKING <ol style="list-style-type: none"> AGENT nurse PATIENT woman ASPECT health TOOLS stethoscope PLACE hospital 	VERB: OPERATING <ol style="list-style-type: none"> AGENT man ITEM ∅ TOOL hand PLACE hospital 	VERB: BANDAGING <ol style="list-style-type: none"> AGENT nurse VICTIM man PLACE hospital 	VERB: INJECTING <ol style="list-style-type: none"> AGENT doctor SUBSTANCE ∅ SOURCE syringe DESTINATION arm PLACE hospital

Figure 8: Application example. GSFormer can serve cross-modal semantic question answering and reasoning on the basis of human event understanding. For example, when questioned "How do people cook? What do people cook with, and where do people cook?", GSFormer can list cooking procedures and steps. Compared to image captioning and scene graphs, GSFormer can not only utilize the generated structured verb-frame to apprehend the questions, but also provide answers with intuitive image facts to help users understand.