

Supplementary Materials: Seeing Beyond Classes: Zero-Shot Grounded Situation Recognition via Language Explainer

Anonymous Authors

APPENDIX

This supplementary document is organized as follows:

- The implementation details mentioned in Sec. 4.1.3 are shown in Sec. A.
- The prompts for the verb-centric description (cf., Sec. 3.1.1), complex scene description (cf., Sec. 3.1.2), rephrased template (cf., Sec. 3.2.1), noun filtering (cf., Sec. 3.3.1), and scene-specific noun description (cf., Sec. 3.3.2) are presented in Sec. B.
- Additional experimental and qualitative results (cf., Sec. 4.3 and Sec. 4.4) are reported in Sec. C.

A IMPLEMENTATION DETAILS

Following the prior work of GSR [3], we only predicted the top 500 most frequent noun categories, which significantly reduces the computation time of CLIP by 80%. Specifically, we extracted a subset test dataset comprising 8,551 images from the complete SWiG’s test dataset of 25,200 images, this subset exclusively contains images annotated with nouns falling within the top 500 most frequently occurring categories. For the Grounding DINO model, the

In-Context Examples
Based on the provided verb class, the task is to generate sentences describing useful features for the action depicted in an image.
Let’s take a few examples. Question: What are the useful visual features for the event of ‘stacking’: ‘an AGENT stacks a TOP onto a BOTTOM in a PLACE’.
Answers: -Person’s hands are positioned as if holding something. -The upper object is aligned directly above the lower one. -Muscular tension indicates lifting or holding a weight. -A focused look toward the objects being handled. -Balance is evident in the placement of objects. -The sequence of placement is implied by the positioning. -Items are midway through a transition from separate to stacked. -Shadows confirm the alignment of the items.
Question: What are the useful visual features for the event of ‘twirling’: ‘an AGENT twirls a COAGENT in a PLACE’.
Answers: -Person’s arms extended towards one another. -Clothing and hair of the coagent appear in motion. -Both individuals are positioned as if in a dance step. -Facial expressions show enjoyment or concentration. -The person’s feet are positioned for pivoting or turning. -The flow of coagent’s attire suggests a circular motion. -Other participants watching or waiting their turn. -Floor markings or wear that imply rotational movement. -Balance and poise maintained by both individuals. -Hands clasped or in a hold typical for dancing. -Spectators’ attention focused on the twirling pair.
Instruction
Question: what are the useful visual features for the event of ‘{VERB CLASS}’: ‘{VERB-CENTRIC TEMPLATE}’.
Answers:

Figure 7: Prompt for verb-centric description generation.

In-Context Examples
Based on the verb class, the task is to generate some detailed scene descriptions, as far as possible to list all possible but different scenarios.
Let’s take a few examples. Question: Please describe the scene about ‘biting’, and make each scene as different as possible.
Answers: -An individual is outdoors, pressing their teeth into a wedge of citrus fruit, their facial expression squeezed into a tight grimace. -A small, fuzzy-coated puppy is engaging in a common canine activity, eagerly biting down on a sizable bone it’s secured between its front paws. -Two husky puppies are interacting; one appears to be biting at the other’s face in a playful gesture, typical of young animals practicing social behaviors. -A close-up of an insect, likely a fly, positioned on a surface with its feeding appendage extended, indicative of the insect’s natural feeding behavior. -A young child outdoors, taking a bite from a cone of ice cream with sprinkles, under the watchful eye of an adult in the background. -Two large felines are engaging with each other amidst a rocky backdrop, capturing a typical display of their interaction within their habitat.
Question: Please describe the scene about ‘marching’, and make each scene as different as possible.
Answers: -A group of individuals is assembled indoors, likely a gymnasium, with some holding drums and others with brass instruments, preparing for a marching band performance. -Rows of uniformed individuals march in sync, clad in military attire and helmets, bearing rifles, which conveys a disciplined, organized military parade or ceremony. -A group is marching in formation in what appears to be a festive parade, accompanied by onlookers and decorated with bunting. -A group is seen marching in a stadium with a stage set up and spectators in the stands, with a backdrop of an urban skyline at what appears to be dusk. -Individuals in traditional dress with bagpipes are marching down a road, with flags and onlookers in the background.
Instruction
Question: Please describe the scene about ‘{VERB CLASS}’, and make each scene as different as possible.
Answers:

Figure 8: Prompt for complex scene description generation.

box threshold, text threshold, and iou threshold are set to 0.2, 0.23, and 0.5 respectively. In both verb classification step (cf., Sec. 3.1.3) and noun classification step (cf., 3.3.3), we set the balance factor λ (Eq. 6 and Eq. 10) to 0.3 as default. In the description weighting step (cf., Sec. 3.1.2), we utilized complex scene descriptions containing fifteen items. Our implementation of LEX was based on PyTorch and one NVIDIA GTX 3090 GPU. For reproducibility, we will provide the complete codebase.

B PROMPTS

In this section, we present prompts for verb-centric description (cf., Sec. 3.1.1), complex scene description (cf., Sec. 3.1.2), rephrased template (cf., Sec. 3.2.1), noun filtering (cf., Sec. 3.3.1), and scene-specific noun description (cf., Sec. 3.3.2). These prompts utilize examples to guide LLMs to produce analogous outputs through an in-context learning manner [1].

Verb-Centric Description Prompt. This prompt contributes to providing more detailed visual feature descriptions regarding scene

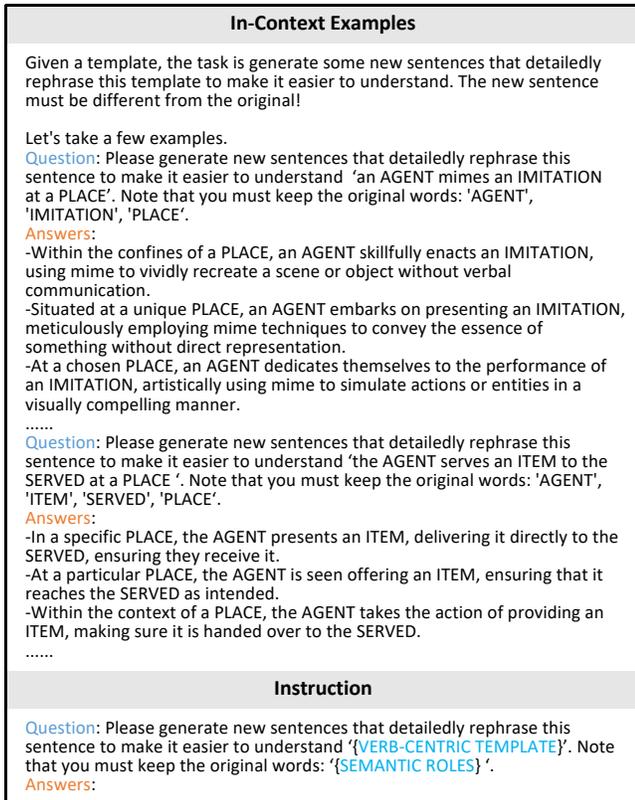


Figure 9: Prompt for rephrased template generation.

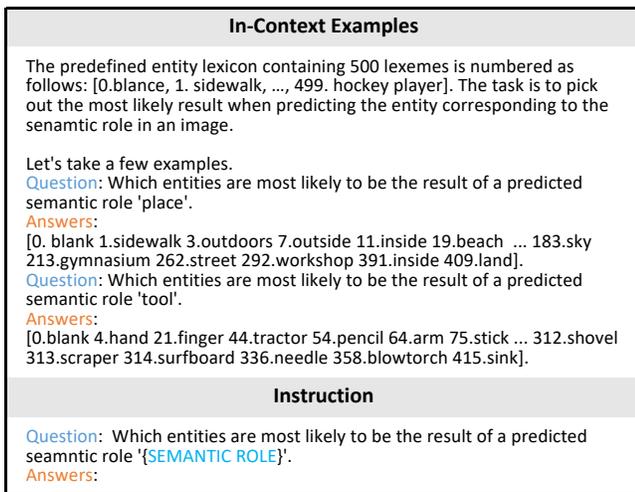


Figure 10: Prompt for noun filtering.

information to enhance verb recognition. As presented in Figure 7, by specifying verb classes and providing instructions (i.e., "What are the useful..."), the prompt guides LLMs to generate descriptions tailored to the respective verb classes.

Complex Scene Description Prompt. This prompt intends to offer comprehensive descriptions of the visual scene to replace

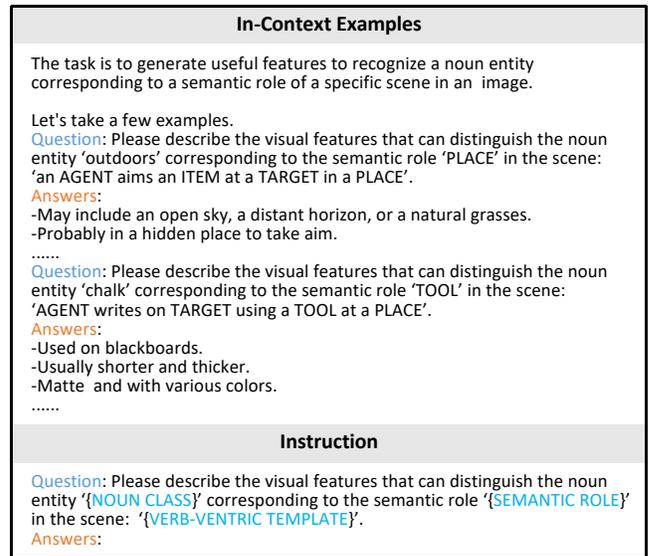


Figure 11: Prompt for scene-specific noun description generation.

annotated images. As shown in Figure 8, we provided verb classes and instructions (i.e., "Please describe the scene...") to prompt LLMs to generate distinct and detailed scene descriptions about each specific verb category.

Rephrased Template Prompt. This prompt aims to enhance the comprehensibility of generated descriptions while ensuring consistency with semantic roles to enable Grounding DINO to produce more precise candidate bounding boxes. As shown in Figure 9, we utilized verb-centric templates, instructions (i.e., "Please generate new..."), and constraints (i.e., "Note that...") to prompt LLMs to produce sentences. These sentences maintain the original semantic roles outlined in the provided templates while being easier to understand.

Noun filtering Prompt. This prompt is utilized to filter unreasonable noun categories for specific scene contexts. As depicted in Figure 10, we provided the semantic role, and instructions (i.e., "Which entities...") to make LLMs select corresponding plausible noun categories.

Scene-specific Noun Description. This prompt endeavors to facilitate the generation of specific visual feature descriptions conditioned on distinctive scene information (i.e., semantic roles and noun categories) to enhance noun recognition. As displayed in Figure 11, we utilized verb-centric templates, related semantic roles, and instructions (i.e., "Please describe the visual...") to guide LLMs to generate descriptions tailored to elucidate the visual characteristics associated with distinct roles within a given scene context.

C FURTHER ANALYSIS

C.1 Additional Ablation Studies

C.1.1 Ablation on different balance factor λ . In both the verb classification (cf., Eq. 6 in Sec. 3.1.3) and noun classification (cf., Eq. 10 in Sec. 3.3.3) steps, we employed a hyperparameter of balance factor λ to weight the contributions of class-based and description-based

Table 6: Top-1 and Top-5 verb accuracy for different λ in verb classification.

λ	Top-1-Verb verb \uparrow	Top-5-Verb verb \uparrow
0.0	30.18	55.49
0.3	32.41	58.34
0.5	31.79	57.75
0.7	29.29	54.97
1.0	22.41	46.84

Table 7: Ground-Truth-Verb setting result for different λ in noun classification.

λ	Ground-Truth-Verb			
	Value \uparrow	val-all \uparrow	grnd \uparrow	grnd-all \uparrow
0.0	29.39	4.51	23.22	3.01
0.3	29.92	4.68	23.57	3.08
0.5	28.65	4.05	22.46	2.53
0.7	27.25	3.71	21.30	2.27
1.0	25.31	3.16	19.67	1.93

prompts. The larger value of λ indicates less dependence on class-based prompts, and vice versa. We varied λ from 0.0 to 1.0, where $\lambda = 0.0$ represents completely dependent on class-based prompts.

Table 6 illustrates the impact of varying balance factor λ in verb classification. As λ increases from 0.0 to 0.3, the performance in Top1/5 verbs improved due to incorporating a certain proportion of description-based prompts with class-based prompts providing a richer visual feature description to make the verbs more distinguishable. Then, reaching a peak at $\lambda = 0.3$ (e.g., 32.41% in Top-1 verb and 58.34% in Top-5 verb), which suggests a best balanced dependence between class-based and description-based prompts. However, as continues to increase λ towards 1.0, the performance begins to decline. This decline could indicate that an over-reliance on description-based prompts may influenced by noises or irrelevant information, leading to less accuracy.

Table 7 reports the effects of altering balance factor λ in noun classification. Similar to verb classification, the best results are obtained at $\lambda = 0.3$ (e.g., 29.92% under value metrics and 4.68% under val-all metrics). This indicates that the collaboration between class-based prompts and description-based prompts yields better results for noun classification accuracy.

C.1.2 Ablation on the number of complex scene descriptions. In the description weighting step (cf., Sec. 3.1.2), we replace an annotated image with complex scene descriptions generated from LLMs. Here we evaluated the impact of different the number N_s of complex scene descriptions on verb recognition. Table 8 shows that with an increase in N_s , there is a gradual improvement under all verb metrics. It indicated that augmenting the pool of complex scene descriptions enhanced the discrimination of the scene. However, once the description quantity surpasses a certain threshold (e.g., $N_s = 20$), further increments appear to yield negligible enhancements in performance. The reason may be that when the number reaches a certain number, similar scene descriptions will appear and the scene diversity is saturated.

Table 8: Top-1 and Top-5 verb accuracy for different number of scene descriptions N_s .

N_s	Top-1-Verb verb \uparrow	Top-5-Verb verb \uparrow
5	32.28	57.92
10	32.31	58.18
15	32.41	58.34
20	32.46	58.38
25	32.45	58.36
30	32.44	58.37

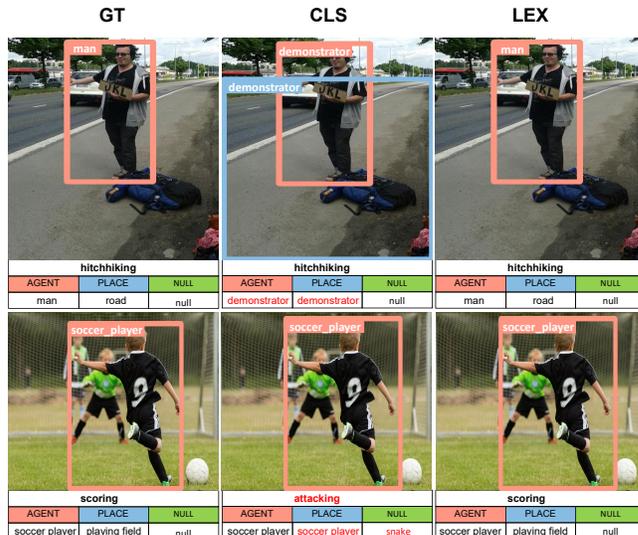


Figure 12: Examples of GT, CLS, and LEX under Top-1 Verb settings on the test dataset of SWiG [2]. The red font denotes the incorrect prediction, and “null” denotes the corresponding semantic role does not exist.

C.2 Additional Qualitative Analysis

C.2.1 Visualization of Role Grounding via Different Templates. To prove the effectiveness of the rephrased templates, we compare the localization effect of the templates generated by the grounding explainer (LEX) and the original verb-centric template (CLS) in Figure 13. It can be concluded that the CLS may overlook or incorrectly ground semantic roles through its verb-centric template design. For example, in the first row, AGENT was erroneously co-localized with TOOL grounding. In the second row, VICTIM was erroneously grounded at the location of AGENT, while AGENT remained not grounded. In contrast, the LEX utilized contextual information and a better-understood rephrased template can improve the precision of grounding semantic roles (e.g., the right side in Figure 13).

C.2.2 Visualization of Zero-Shot GSR Results. In this section, we presented the qualitative comparisons of LES and the CLS baseline in Figure 12. It shows that LEX can rectify errors in verb classification (e.g., the second row in Figure 12, incorrect prediction of “attacking” is corrected to the ground truth label “scoring”) made by CLS, as well as errors in role grounding (e.g., the error localization

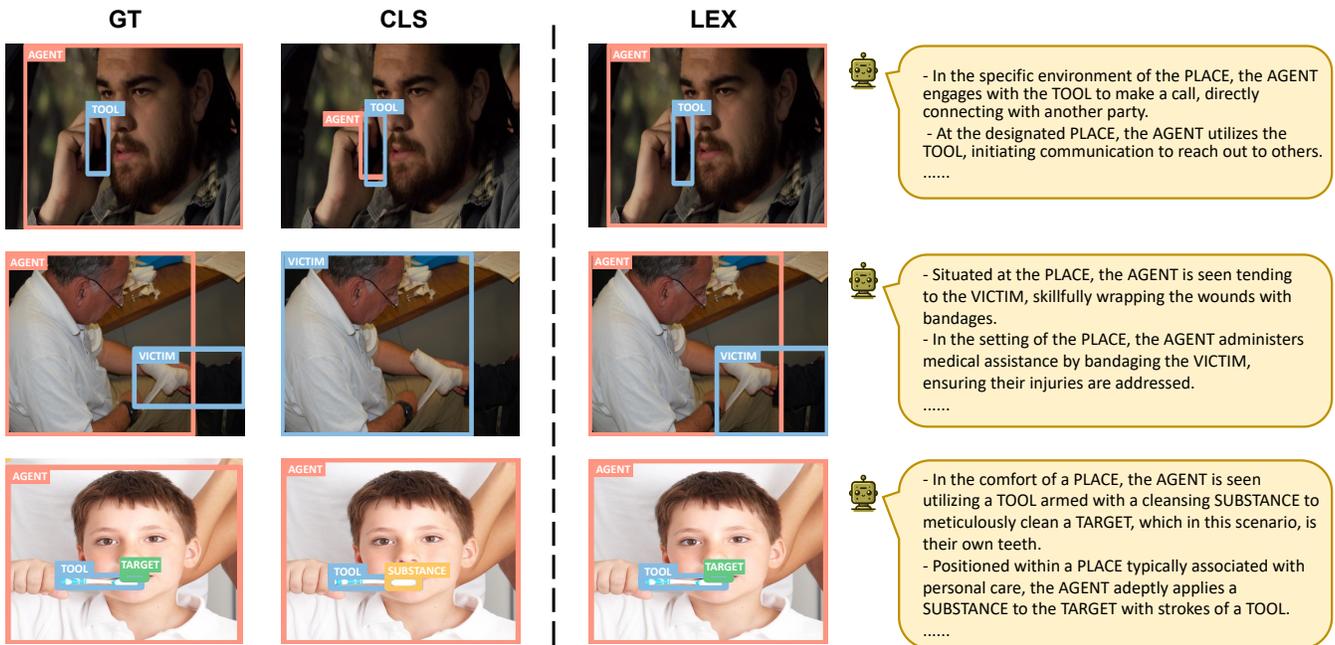


Figure 13: Examples of the semantic role grounding results in GT, CLS, and LEX on the test dataset of SWiG [2]. The right side shows the rephrased template description prompts, which are used to generate the candidate bounding boxes.

of the semantic role of *PLACE* in the first row) and noun classification (e.g., the noun “man” corresponding to *AGENT* was wrongly predicted as “demonstrator” in the first row). This demonstrates the effectiveness of the various modules of LEX.

REFERENCES

- [1] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).
- [2] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *ECCV*. Springer, 314–332.
- [3] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. 2022. Rethinking the two-stage framework for grounded situation recognition. In *AAAI*, Vol. 36. 2651–2658.