

Supplementary Material

S1. Prompt Template and Structured Description Example

This section provides the complete prompt template used for relation-predicate semantic description generation and an illustrative input-output example for the predicate *hanging from*. The placeholder $\mathcal{V}_o$ denotes the VG-150 object vocabulary, and $[p_j]$ denotes the target predicate.

Input Prompt

Instruction. You are a Scene Graph Generation Knowledge Expert. The reference object vocabulary is $\mathcal{V}_o$. All subject and object examples must be selected from this vocabulary, and the response must be written in English.

For the target predicate $[p_j]$, provide three examples: (1) **Strong Visual Example**, where interaction or posture is central; (2) **Strong Spatial Example**, where relative location is central; and (3) **Rare/Uncommon Example**, which is valid but less frequent. For each example, provide one triplet (*Subject, Predicate, Object*), followed by **[Visual Cues]**, **[Spatial Logic]**, and **[Physical Commonsense]**.

To illustrate the required format, the prompt provides the following example for the predicate *riding*.

1. Strong Visual Example

Triplet: (person, riding, horse)

[Visual Cues]: Person straddling the animal; legs on either side; hands holding reins or neck.

[Spatial Logic]: Subject is vertically above and centrally aligned with the object.

[Physical Commonsense]: The horse supports the subject's weight, requiring dynamic balance.

2. Strong Spatial Example

Triplet: (person, riding, bike)

[Visual Cues]: Person seated on the saddle; feet on pedals; hands on handlebars.

[Spatial Logic]: The subject contacts the seat, pedals, and handlebars.

[Physical Commonsense]: The bicycle provides mechanical support and propulsion.

3. Rare/Uncommon Example

Triplet: (dog, riding, surfboard)

[Visual Cues]: Dog standing on all four legs on the board while adjusting balance.

[Spatial Logic]: The subject remains within the top surface area of the object.

[Physical Commonsense]: The surfboard supports the dog while floating on water.

Generated Structured Description

Target Predicate: *hanging from*

1. Strong Visual Example

Triplet: (person, hanging from, branch)

[Visual Cues]: Arms extended to grasp the branch; body suspended below it.

[Spatial Logic]: The subject is below the object and aligned with the support point.

[Physical Commonsense]: Grip and friction support the hanging state.

2. Strong Spatial Example

Triplet: (lamp, hanging from, ceiling)

[Visual Cues]: The lamp is suspended by a cable or chain with the base facing downward.

[**Spatial Logic**]: The object is vertically attached to an overhead support.

[**Physical Commonsense**]: Suspension is maintained by tensile support from above.

3. Rare/Uncommon Example

Triplet: (bag, hanging from, handle)

[**Visual Cues**]: The bag hangs from the handle and may tilt slightly.

[**Spatial Logic**]: The handle serves as the support point and the bag remains below it.

[**Physical Commonsense**]: The handle bears the load and maintains balance.

S2. Complete Hyperparameter Settings

Table S2 summarizes the hyperparameter settings used in the experiments. The remaining detector and relation-encoder configurations follow the PE-Net setting. The pre-trained Faster R-CNN with ResNeXt-101-FPN is kept fixed during relation training.

Table S2 Complete hyperparameter settings used in the experiments.

Hyperparameter	Value
Optimizer	SGD, momentum = 0.9, weight decay = 1×10^{-4}
Batch size	8
Gradient-norm clipping	5.0
Stage-1 training iterations	60k
Stage-1 initial learning rate	1×10^{-3}
Initial temperature τ	0.07
Number of nearest negative prototypes k_1	10
Number of nearest prototype neighbors k_2	1
Feature-prototype margin γ_1	1.0
Prototype-separation margin γ_2	7.0
Stage-2 training iterations	60k
Stage-2 initial learning rate	1×10^{-4}
Base sample count Q_{base}	40
Tolerance radius R	0.3
Boundary-loss coefficient $\lambda^{(2)}$	0.1
Maximum frequency-compensation coefficient λ_t	0.05
Compensated log-scale clipping range	$[-3.5, 0.5]$
Scale adjustment coefficient ρ	0.55
Distance-calibration weight λ	0.3
Predicate-semantic vector dimension	384
Visual-semantic fusion hidden dimension	512
Number of attention heads	8
Dropout rate	0.1

S3. Virtual-Sample Allocation Statistics

Table S3 reports the minimum and maximum virtual-sample counts obtained from Eq. (22) under different Q_{base} settings for the 50 relation categories. Across all settings, each relation category receives more than one virtual sample. Under the default setting $Q_{base} = 40$, the virtual-sample count ranges from 34 to 61.

Table S3 Virtual-sample allocation statistics under different base sample counts.

Q_{base}	Minimum Q_j	Maximum Q_j
2	2	3
10	8	15
20	17	31
30	25	46
40	34	61
60	51	92
80	68	122