End-to-End Entity-Predicate Association Reasoning for Dynamic Scene Graph Generation

Supplementary Material

6. Additional Results

Temporal modeling. The proposed ARN explicitly models temporal dynamics. In the proposed HA module, predicate queries aggregate temporal context $\mathbf{F}_{vit}$ to capture temporal dependencies. The $\mathbf{F}_{vit}$ has **shape** [T,HW,d], this structure supports temporal modeling across frames.

For **temporal input**, we adopt a fixed-length sliding window of 8 frames. Longer videos are divided into equal-length clips; if padding is needed, duplicated frames are marked and excluded from loss computation, ensuring training fairness. This input form avoids reference-frame sampling (as in OED [37]), reduces the training time, and supports full end-to-end optimization. We also evaluated the **sensitivity** to temporal length (T = 6, 8, 12), and observed only marginal differences (rows 5-7 of the Tab. 5 above), suggesting the model is robust to moderate changes in temporal configuration.

To validate the contribution, we conducted a still-image ablation by using spatial context $\mathbf{F}_{vi}$ instead of $\mathbf{F}_{vit}$. Performance drops after removing temporal modeling (see row 3&6 of Tab. 5 on left). Additionally, Some tail predicates exhibit more decrease (such as "covered by" and "have it on the back") in Fig. 7. Meanwhile, Tab. 5 reports the static baseline results of the prevailing methods, RelTR and OED. ARN also outperforms both RelTR and OED under static baseline settings.

Moreover, we report the mR@K results (see Tab. 6) for PredCls task. In this setting, we must reduce the majority of queries to align ground-truth with queries in PredCls setting, ensuring semantic consistency. SGDet is primarily used to validate our contributions, while PredCls results demonstrate that ARN is also applicable to different types of tasks.

Semantic cue modeling. ARN models triplet representations by capturing context-aware relationships among predicate subclasses, avoiding reliance on explicit statistical co-occurrence priors (such as STKET [31]). This context-driven approach reduces reliance on dataset-specific distributions and mitigates. Furthermore, we incorporate CLIP's cross-modal semantic knowledge via textual alignment, which enhances generalization—particularly for tail predicates. We report the ablation study (row 4&6 in the Tab. 5).

7. Visualize

We visualize the predicate decoder attention map for the predicted dynamic scene graph. As shown in Fig.8, The entity decoder employs two distinct query sets, corresponding separately to subject and object detection. The heatmaps generated by the queries with the highest confidence effectively highlight the spatial regions associated with the subject and object, respectively (Columns 2 and 3). This means that our model can accurately infer paired entity from images.

Meanwhile, as illustrated in columns 3 through 6 of Fig. 8, different branches of the predicate decoder focus on distinct image details to infer the corresponding predicate subclasses. Specifically, the heatmap of the 74-th query emphasizes the target object "cup", whereas the 65-th query

Method	R@10	R@20	R@50
RelTR [7]	20.9	24.6	28.2
OED [37]	33.4	41.3	49.0
Full model w/o temporal	36.8	45.8	53.2
Full model w/o CLIP	34.6	42.9	49.9
Full model (batch=6)	37.3	46.3	53.4
Full model (batch=8)	**37.6**	**46.8**	**54.1**
Full model (batch=12)	36.9	45.9	53.1

Table 5. Ablation Studies on temporal modeling and semantic cue modeling under Recall@K metric.

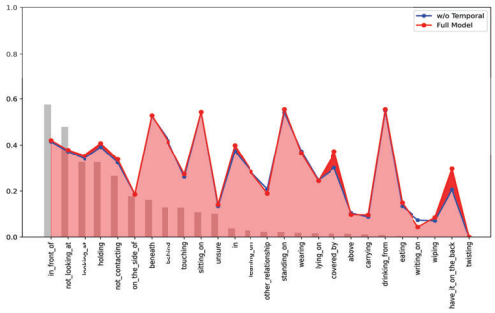

Figure 7. Comparative per predicate class performance for SGDET task. Results are in terms of mR@10 under "No constraint".

Method	mR@20	mR@50
TEMPURA[29]	85.1	98.0
OED[37]	**89.6**	97.2
TD^2[24]	-	98.2
Ours	85.6	**99.0**

Table 6. Comparison results for PredCls task, in terms of Mean Recall@K metric.

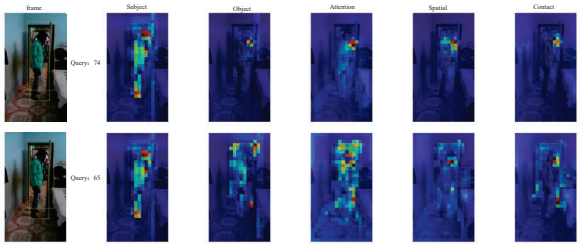

Figure 8. Visualization of attention maps of pair-wise entity feature and and different predicate subcategories.

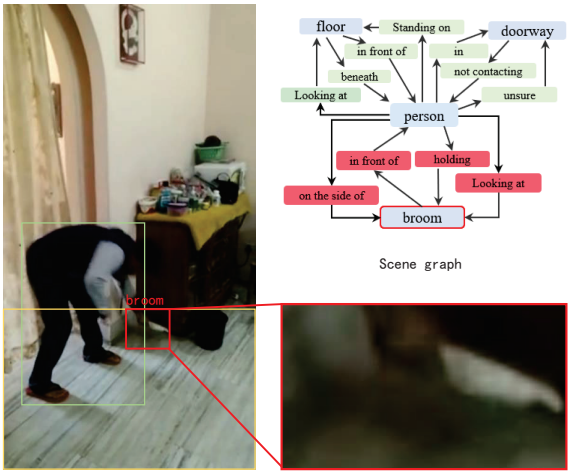

Figure 9. Instance of failure case.

predominantly attends to the ”doorway”. This indicate subtle variations in attention across different semantic regions facilitate fine-grained semantic understanding.

8. Failure Cases

We analyzed the results and reported failure cases to summarize the limitations of our model. A failure case is illustrated in Fig. 9. Although the generated scene graph correctly detects the object “floor” and accurately predicts all its associated predicates, the object “broom” is misclassified. Despite the relatively clear visibility of the “floor”, the “broom” is incorrectly recognized, possibly due to the image’s low resolution. We conjecture that the model’s misclassification arises from insufficient visual clarity, causing incorrect semantic categorization despite accurately attending to the object’s spatial region.

As shown in Fig. 10, to further analyze this failure case, we visualize attention heatmaps of the two most confident object queries from the failure case. It can be observed that the model first successfully localizes and recognizes the object “floor”, followed by identifying the object “shelf”, despite it not being annotated in the current image. We conjecture that such failure cases could potentially be improved by the higher-resolution training data.

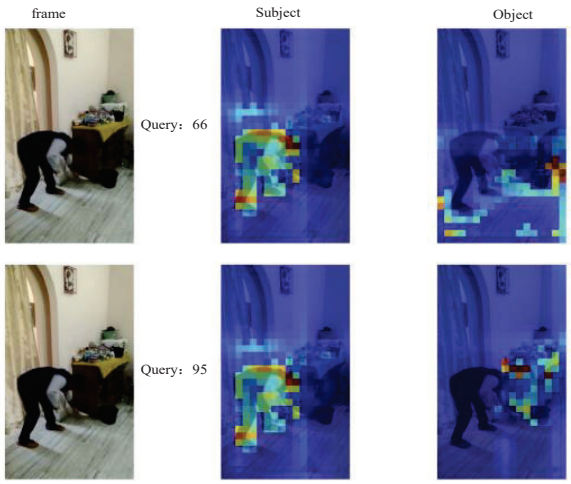

Figure 10. Visualization of attention maps of the failure case.