Unbiased Scene Graph Generation in Videos

THU-PM-210




Sayak Nag[1] Kyle Min[2] Subarna Tripathi[2] Amit K. Roy-Chowdhury[1]

[1] Video Computing Group
Center for Robotics & Intelligent Systems
UC RIVERSIDE
Marlan and Rosemary Bourns College of Engineering

[2] intel labs

JUNE 18-22, 2023
CVPR
VANCOUVER, CANAD

Dynamic Scene Graph Generation

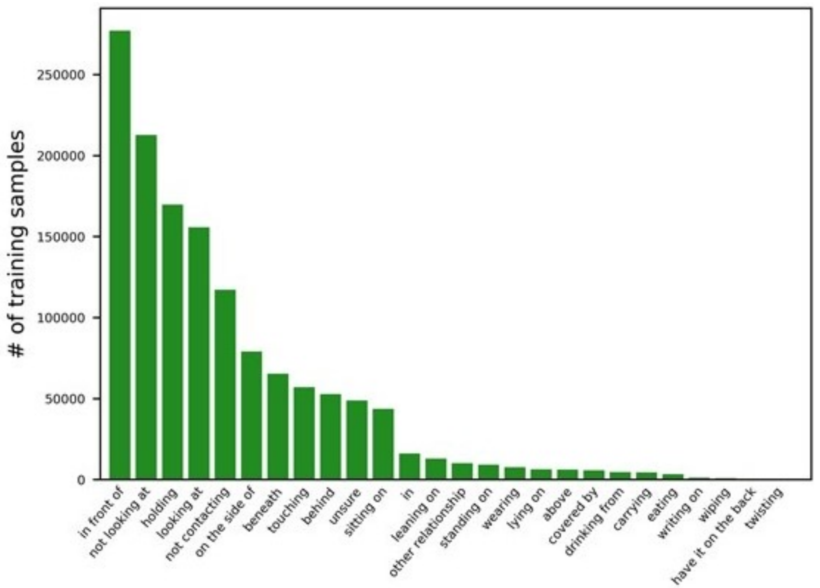
*

Triplet: <subject-predicate-object>
eg: <person-next to- sofa>

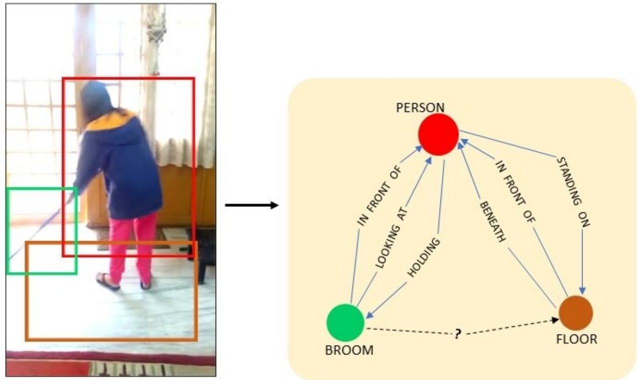
*Action Genome

Inherent Challenges

Long-Tailed Distribution of Predicate Classes

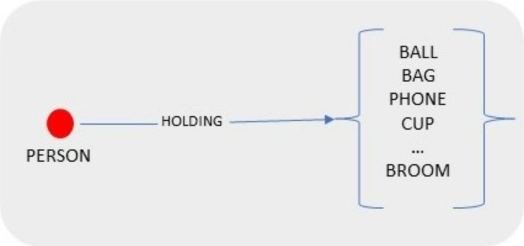


Noisy Labelling



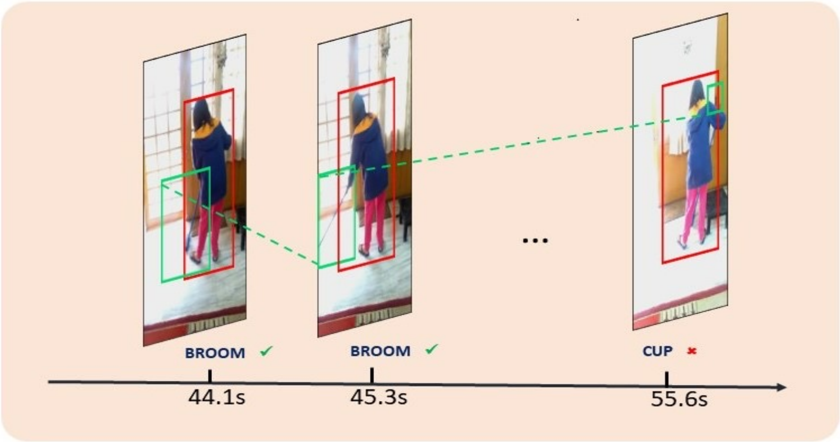
Missing annotations and multiple possible correct predications

Triplet Variability



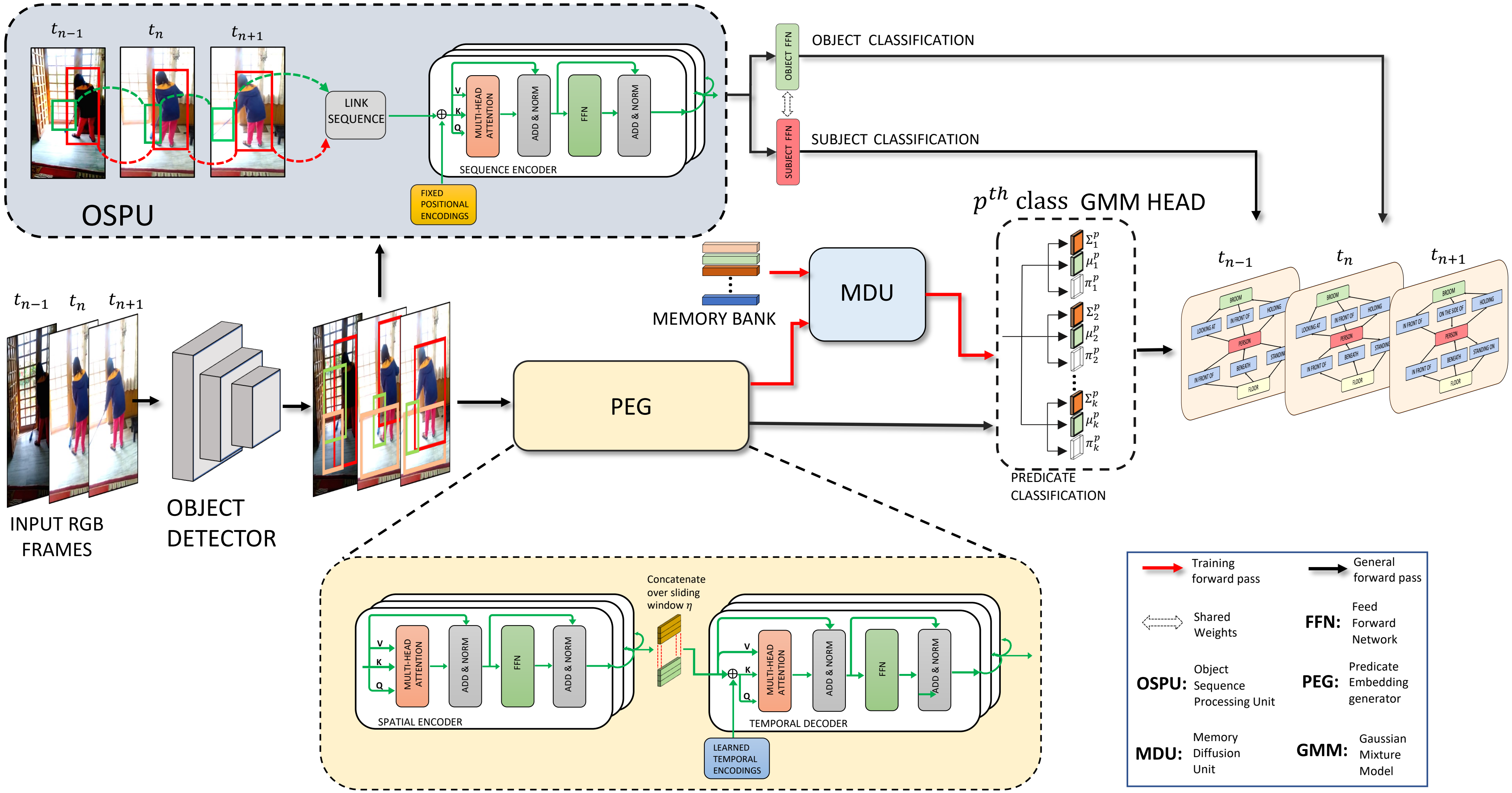
Same relationship with multiple possible subject-object pairs

Temporally Inconsistent Object Classification



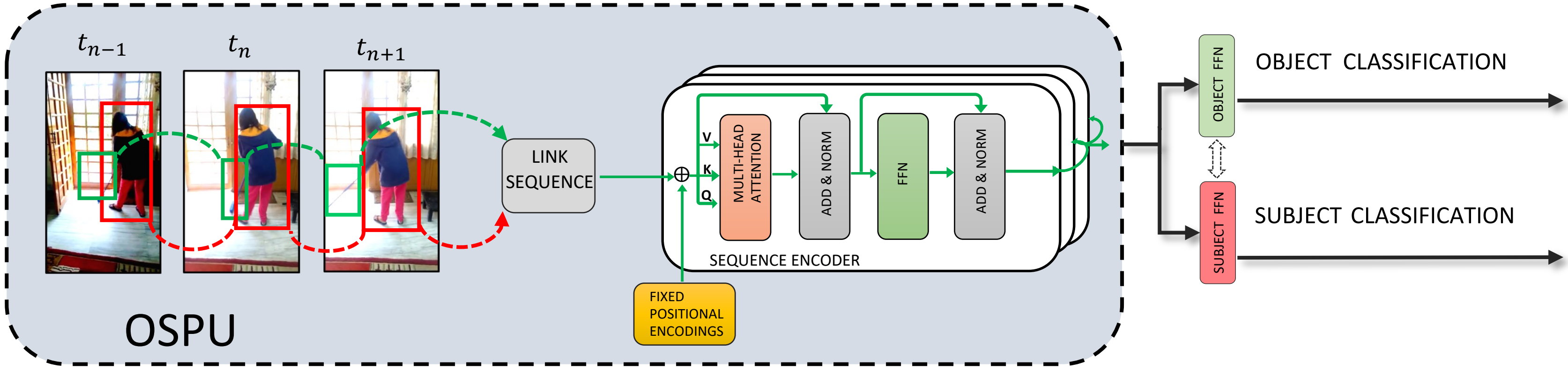
Motion Blur and Occlusion hinders object classification

FRAMEWORK



TEMPURA: TEmporal consistency and Memory Prototype guided Uncertainty Attenuation for Unbiased dynamic SGG

Temporal Consistency in Object Classification



INPUT: $\mathcal{T}^{j}_{t_j k_j} = \{\boldsymbol{v}_i^t, \boldsymbol{v}_i^{t+1}, \ldots, \boldsymbol{v}_i^{t+k}\}$, $1 \leq t_j, k_j \leq T$,

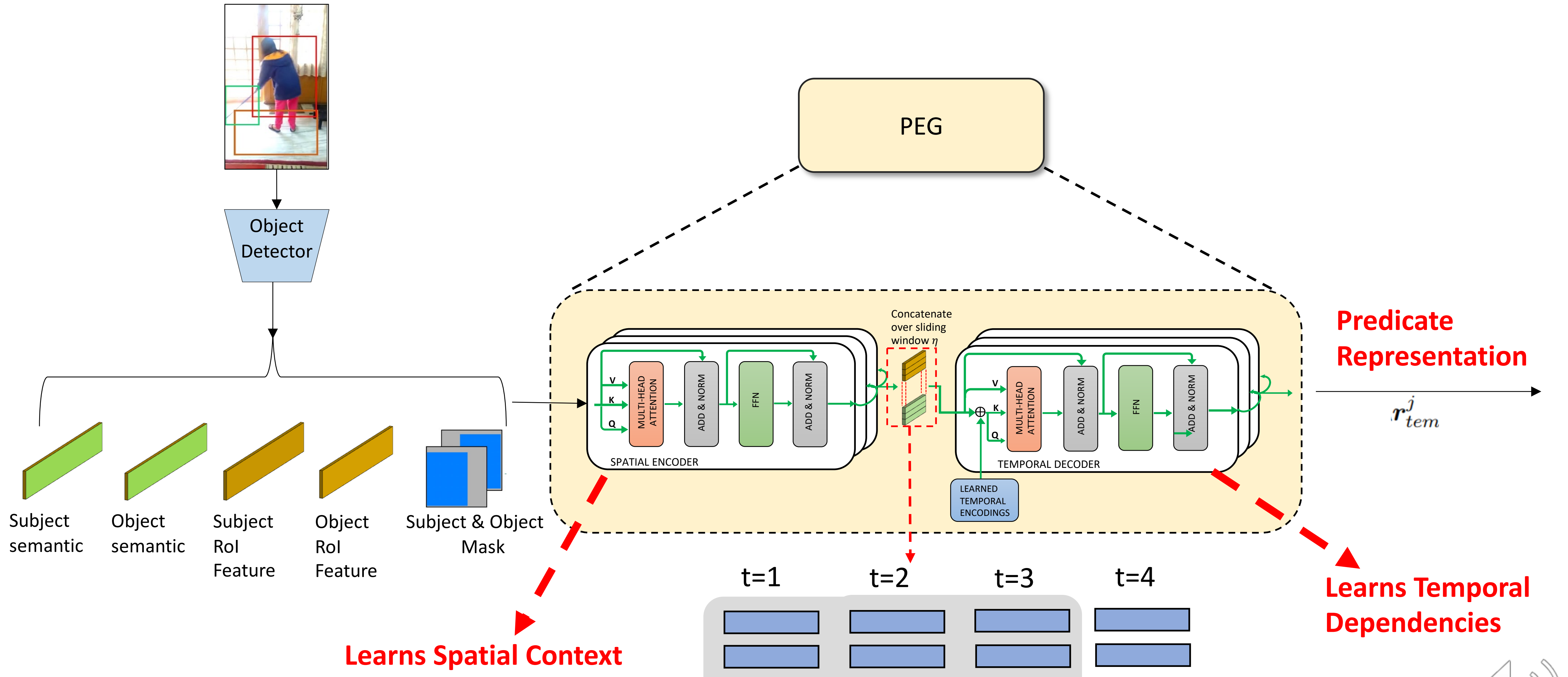
each $\boldsymbol{v}_i^t$ has same detected class.

OSPU LOSS: $\mathcal{L}_o + \mathcal{L}_{intra}$

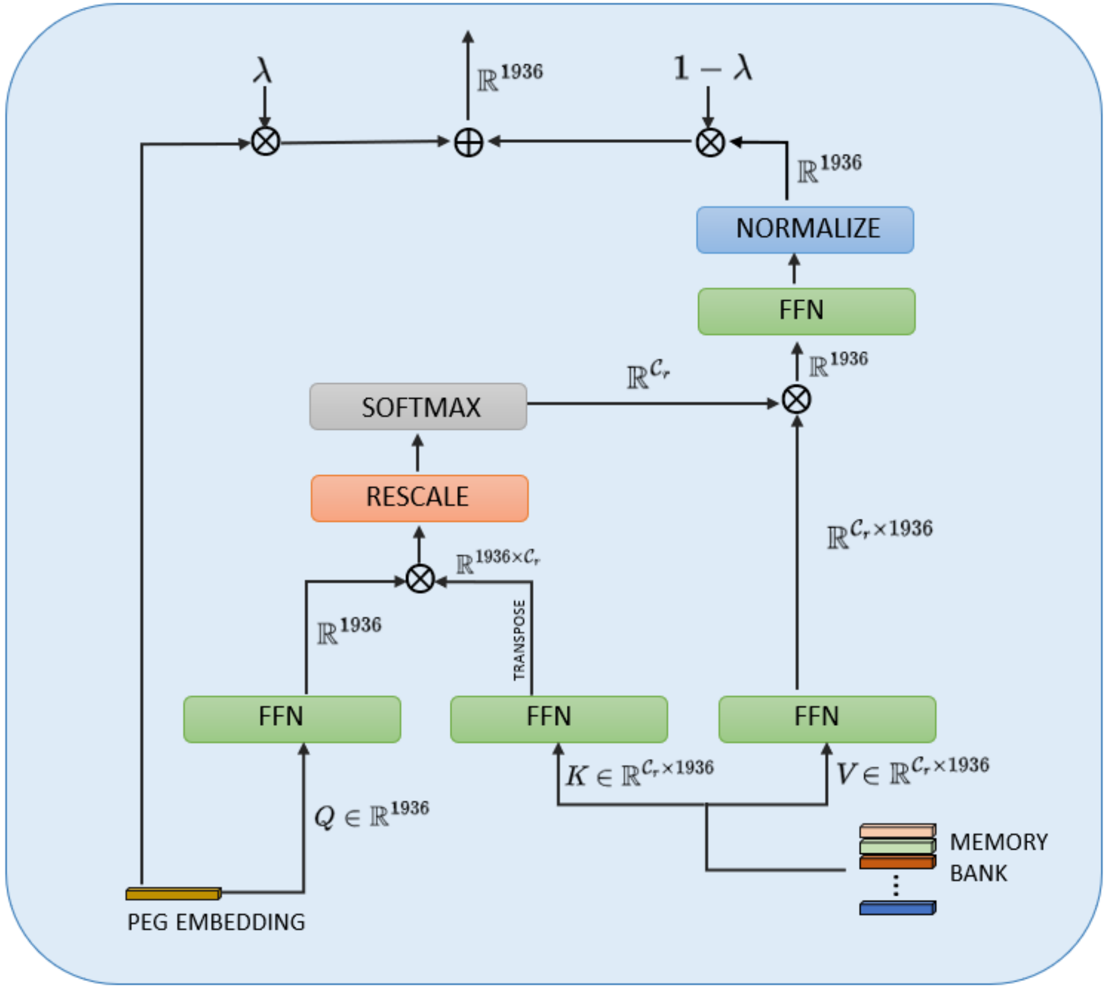
$$\mathcal{L}_{intra} = \sum_i \sum_j ||\hat{\boldsymbol{x}}_{o_i} - \hat{\boldsymbol{x}}^{+}_{o_j}||_2^2 + \sum_k max(0, 1 - ||\hat{\boldsymbol{x}}_{o_i} - \hat{\boldsymbol{x}}^{-}_{o_k}||_2^2)$$

Intra-Video supervised contrastive loss -> Enhances the temporal consistency of positive pairs

Predicate Representations



Memory Guided Training



Memory Diffusion Unit

- **Memory bank:** $\boldsymbol{\Omega}_R = \{\boldsymbol{\omega}_p\}_{p=1}^{\mathcal{C}_r}$

  Memory -> Prototype $\boldsymbol{\omega}_p = \frac{1}{N_{y_{r_p}}} \sum_{j=1}^{N_{y_{r_p}}} \boldsymbol{r}_{tem}^j \ \forall p \in \mathcal{Y}_r$

- **OUTPUT:** $\hat{\boldsymbol{r}}_{tem}^j = \lambda \boldsymbol{r}_{tem}^j + (1-\lambda)\boldsymbol{r}_{mem}^j$; $0 < \lambda \leq 1$

  $\boldsymbol{r}_{mem}^j = \mathbb{A}(\boldsymbol{r}_{tem}^j, \boldsymbol{\Omega}_r)$ is cross-attention b/w. $\boldsymbol{r}_{tem}^j$ and $\boldsymbol{\Omega}_r$.

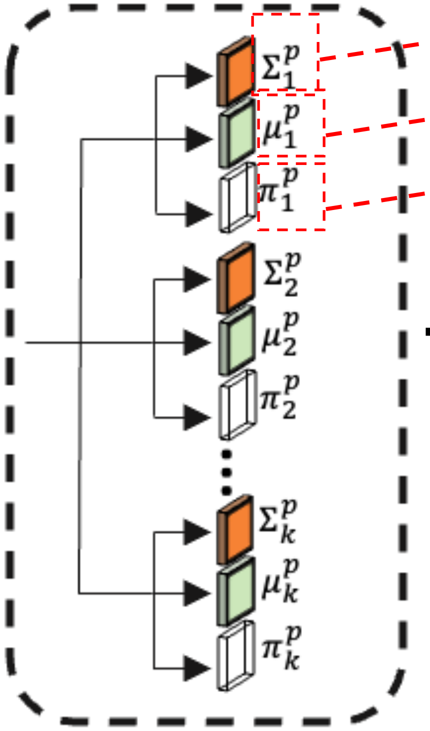
- **If λ is too high, debiasing fails, and if too low model becomes biased towards data-poor classes.**

- **Memory bank of current epoch is computed from the predicate embeddings of previous epoch.**

- **MDU used during training only and acts as a structural meta-regularizer.**

Uncertainty Attenuated Predicate Classification

p^{th} class GMM HEAD



Variance of 1st component

Mean of 1st component

Mixing Coefficient of 1st component

$$\hat{y}^i_{r_p} = \sum_{k=1}^{\mathcal{K}} \pi^k_{i,p} \mathcal{N}(\mu^k_{i,p}, \Sigma^k_{i,p})$$

Reparameterization trick

$$\hat{y}^i_{r_p} = \sum_{k=1}^{\mathcal{K}} \pi^k_{i,p} \sigma(\hat{c}^i_{p,k}); \quad \hat{c}^i_{p,k} = \mu^k_{i,p} + \varepsilon \sqrt{\Sigma^k_{i,p}},$$

$$\varepsilon \sim \mathcal{N}(0,1)$$

Predictive Uncertainty

Aleatoric

$$U^p_{al}(\mathbf{z}_i) = \sum_{k=1}^{\mathcal{K}} \pi^k_{i,p} \Sigma^k_{i,p}$$

Epistemic

$$U^p_{ep}(\mathbf{z}_i) = \sum_{k=1}^{\mathcal{K}} \pi^k_{i,p} || \mu^k_{i,p} - \sum_{j=1}^{\mathcal{K}} \pi^j_{i,p} \mu^j_{i,p} ||_2^2$$

- **Aleatoric: Data-specific Uncertainty**
- **Epistemic: Model-specific Uncertainty**

GMM LOSS: $\mathcal{L}_p = -\sum_{i=1}^{N_{r,p}} \sum_{p=1}^{\mathcal{C}_r} y^i_{r_p} \log \sum_{k=1}^{\mathcal{K}} \pi^k_p \sigma(\hat{c}^i_{p,k}) \longrightarrow$ **Penalizes the model if aleatoric uncertainty**

TOTAL LOSS: $\mathcal{L}_{total} = \mathcal{L}_p + \mathcal{L}_o + \mathcal{L}_{intra}$

Experimental Setup

Dataset

Action Genome
- 35 object classes
- 26 HOI Predicates

SGG Tasks

- PREDCLS: object bounding box and labels given
- SGCLS: object bounding box given
- SGDET: end to end SGG

Evaluation Setups

- **With Constraint**: At most one edge of allowed b/w subject-object pairs
- **No Constraint**: Multiple edges allowed b/w object pairs

Performance Metrics

- **Recall@K**: Recall computed over entire dataset making it biased towards data rich classes
- **mean-Recall@K**: Recall computed over each predicate class and then averaged

Comparative Results

Table 1. Comparative results for SGDET task, on AG [24], in terms of *m-Recall@K* and *Recall@K*.

Method	With Constraint						No Constraints					
	mR@10	mR@20	mR@50	R@10	R@20	R@50	mR@10	mR@20	mR@50	R@10	R@20	R@50
RelDN [64]	3.3	3.3	3.3	9.1	9.1	9.1	7.5	18.8	33.7	13.6	23.0	36.6
HCRD supervised [15]	-	8.3	9.1	-	27.9	30.4	-	-	-	-	-	-
TRACE [56]	8.2	8.2	8.2	13.9	14.5	14.5	22.8	31.3	41.8	26.5	35.6	45.3
ISGG [28]	-	19.7	22.9	-	29.2	35.3	-	-	-	-	-	-
STTran [9]	16.6	20.8	22.2	25.2	34.1	37.0	20.9	29.7	39.2	24.6	36.2	48.8
STTran-TPI [58]	15.6	20.2	21.8	26.2	34.6	37.4	-	-	-	-	-	-
APT [37]	-	-	-	26.3	**36.1**	**38.3**	-	-	-	25.7	37.9	**50.1**
TEMPURA	**18.5**	**22.6**	**23.7**	**28.1**	33.4	34.9	**24.7**	**33.9**	**43.7**	**29.8**	**38.1**	46.4

Table 2. Comparative results for SGG tasks: PREDCLS and SGCLS, on AG [24], in terms of *m-Recall@K*.

Method	With Constraint						No Constraints					
	PredCLS			SGCLS			PredCLS			SGCLS		
	mR@10	mR@20	mR@50	mR@10	mR@20	mR@50	mR@10	mR@20	mR@50	mR@10	mR@20	mR@50
RelDN [64]	6.2	6.2	6.2	3.4	3.4	3.4	31.2	63.1	75.5	18.6	36.9	42.6
TRACE [56]	15.2	15.2	15.2	8.9	8.9	8.9	50.9	73.6	82.7	31.9	42.7	46.3
STTran [9]	37.8	40.1	40.2	27.2	28.0	28.0	51.4	67.7	82.7	40.7	50.1	58.8
STTran-TPI [58]	37.3	40.6	40.6	28.3	29.3	29.3	-	-	-	-	-	-
TEMPURA	**42.9**	**46.3**	**46.3**	**34.0**	**35.2**	**35.2**	**61.5**	**85.1**	**98.0**	**48.3**	**61.1**	**66.4**

Table 3. Comparative results for SGG tasks: PREDCLS and SGCLS, on AG [24], in terms of *Recall@K*.

Method	With Constraint						No Constraints					
	PredCLS			SGCLS			PredCLS			SGCLS		
	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50
RelDN [64]	20.3	20.3	20.3	11.0	11.0	11.0	44.2	75.4	89.2	25.0	41.9	47.9
TRACE [56]	27.5	27.5	27.5	14.8	14.8	14.8	72.6	91.6	96.4	37.1	46.7	50.5
STTran [9]	68.6	71.8	71.8	46.4	47.5	47.5	77.9	94.2	99.1	54.0	63.7	66.4
STTran-TPI [58]	**69.7**	72.6	72.6	**47.2**	48.3	48.3	-	-	-	-	-	-
APT [37]	69.4	**73.8**	**73.8**	**47.2**	**48.9**	**48.9**	78.5	**95.1**	99.2	55.1	**65.1**	**68.7**
TEMPURA	68.8	71.5	71.5	**47.2**	48.3	48.3	**80.4**	94.2	**99.4**	**56.3**	64.7	67.9

Optimal λ

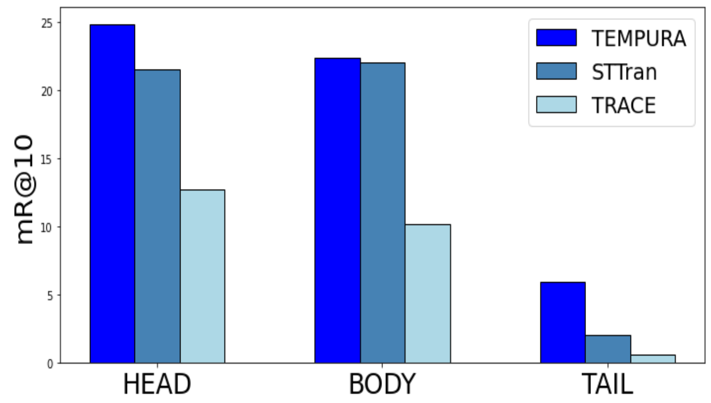
- PREDCLS: λ=0.5
- SGCLS: λ=0.3
- SGDET: λ=0.5

- Outperforms prior methods in terms of mR@K
- Does not compromise on R@K

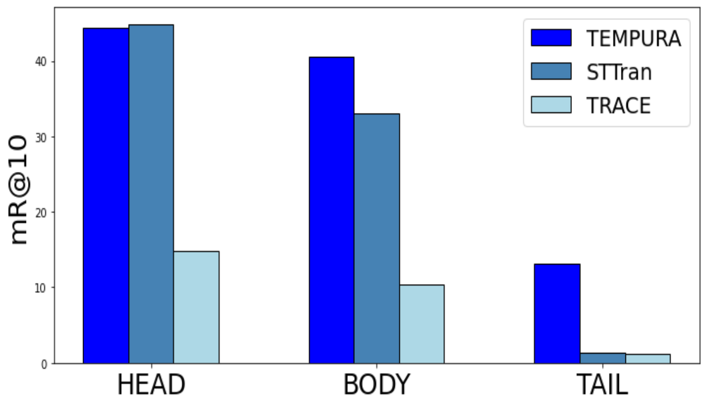
Comparative Results

- *HEAD* $\geq$ 100000 training samples
- 8000 training samples $\leq$ *BODY* $<$ 100000 training samples
- *TAIL* $<$ 8000 training samples

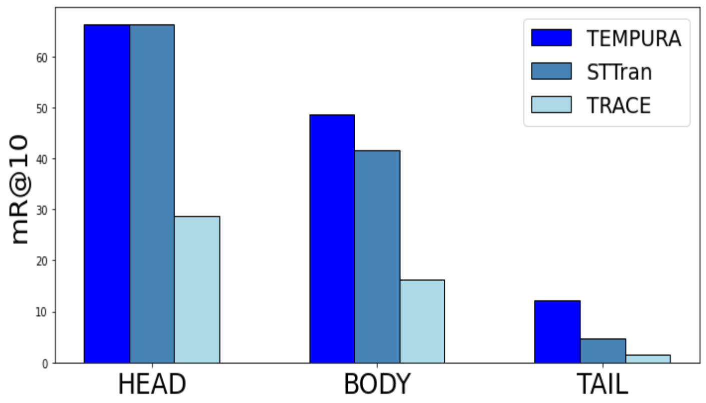
With Constraint



SGDET

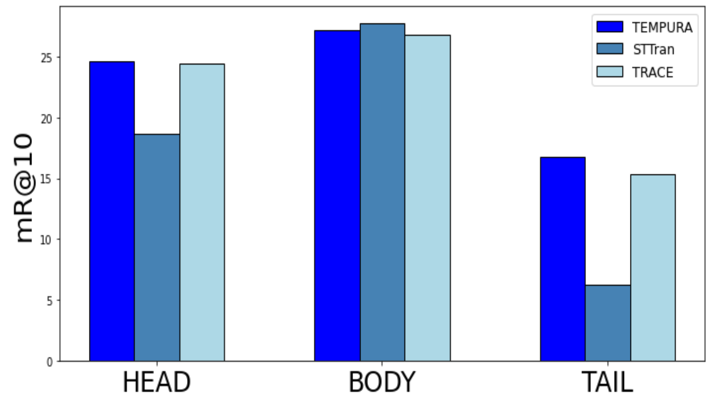


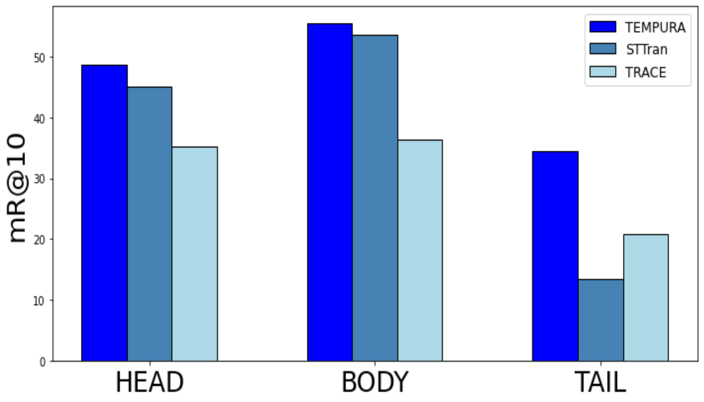
SGCLS

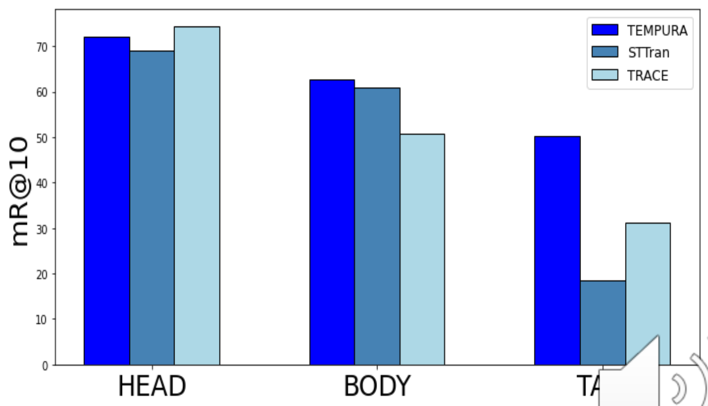


PREDCLS

No Constraint







Qualitative Visualization

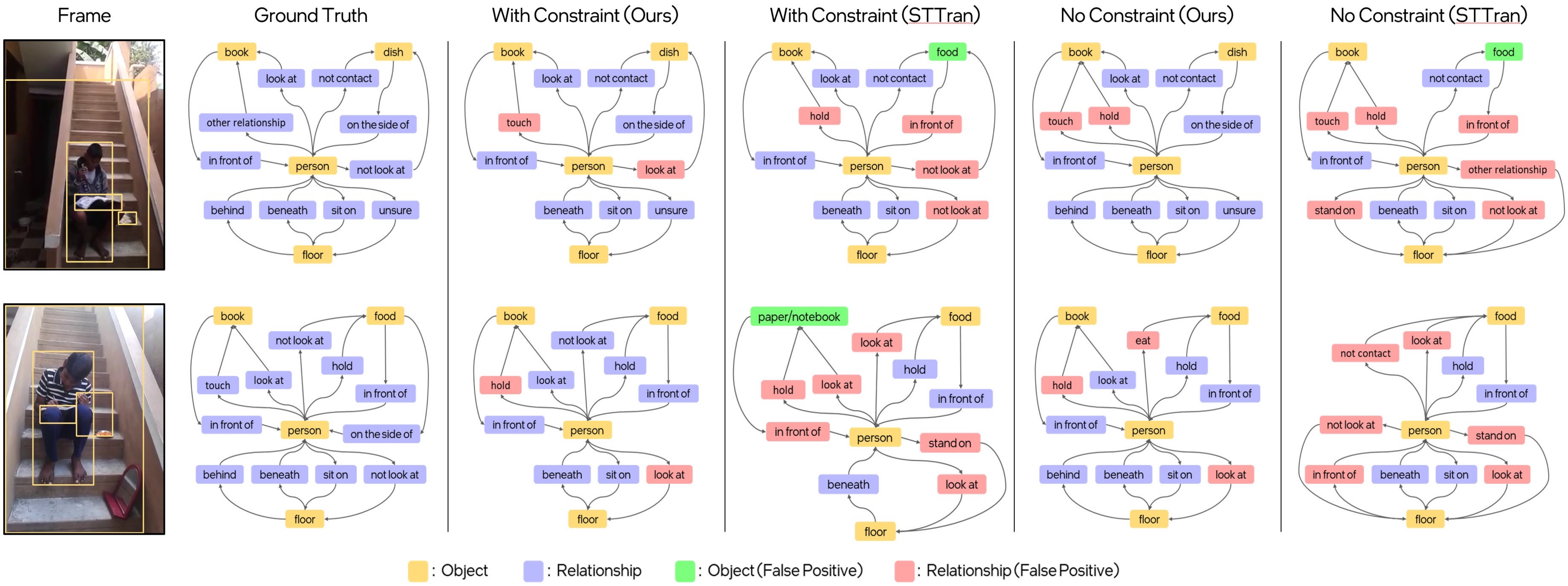
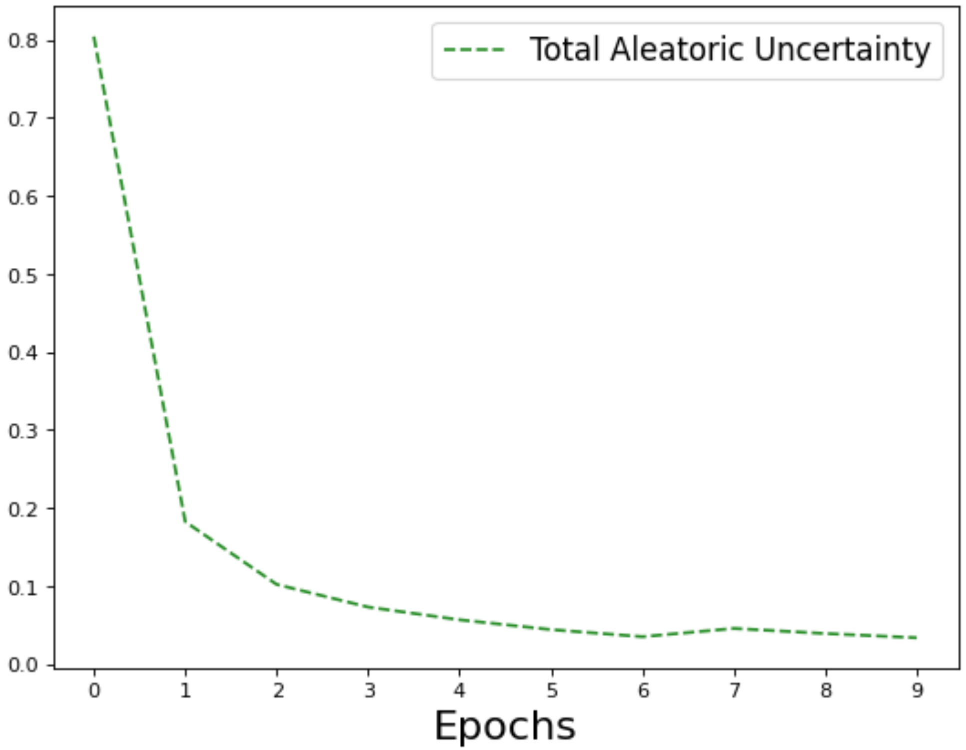
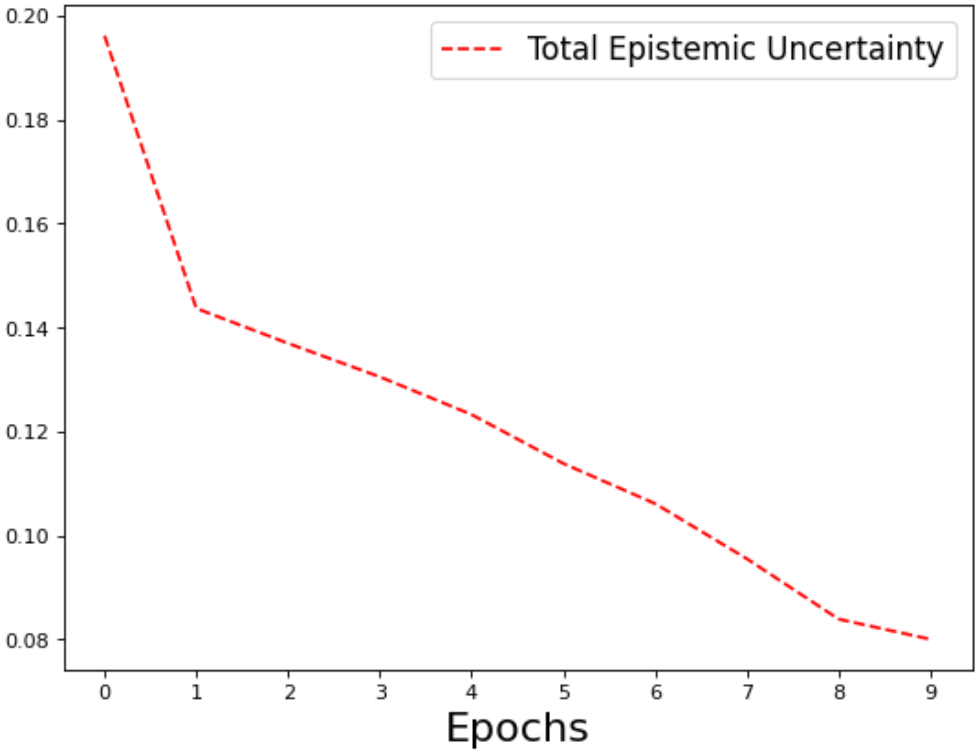


Figure 8. **Comparative qualitative results**. From left to right: input video frames, ground truth scene graphs, scene graphs generated by TEMPURA, and the scene graphs generated by the baseline STTran [10]. Incorrect object and predicate predictions are shown in green and pink, respectively

Uncertainty Analysis

Is the predictive uncertainty being attenuated?





Poster Session: June 22
Poster ID: 210

Paper

https://arxiv.org/abs/2304.00733

Codebase

https://github.com/sayaknag/unbiasedSGG.git