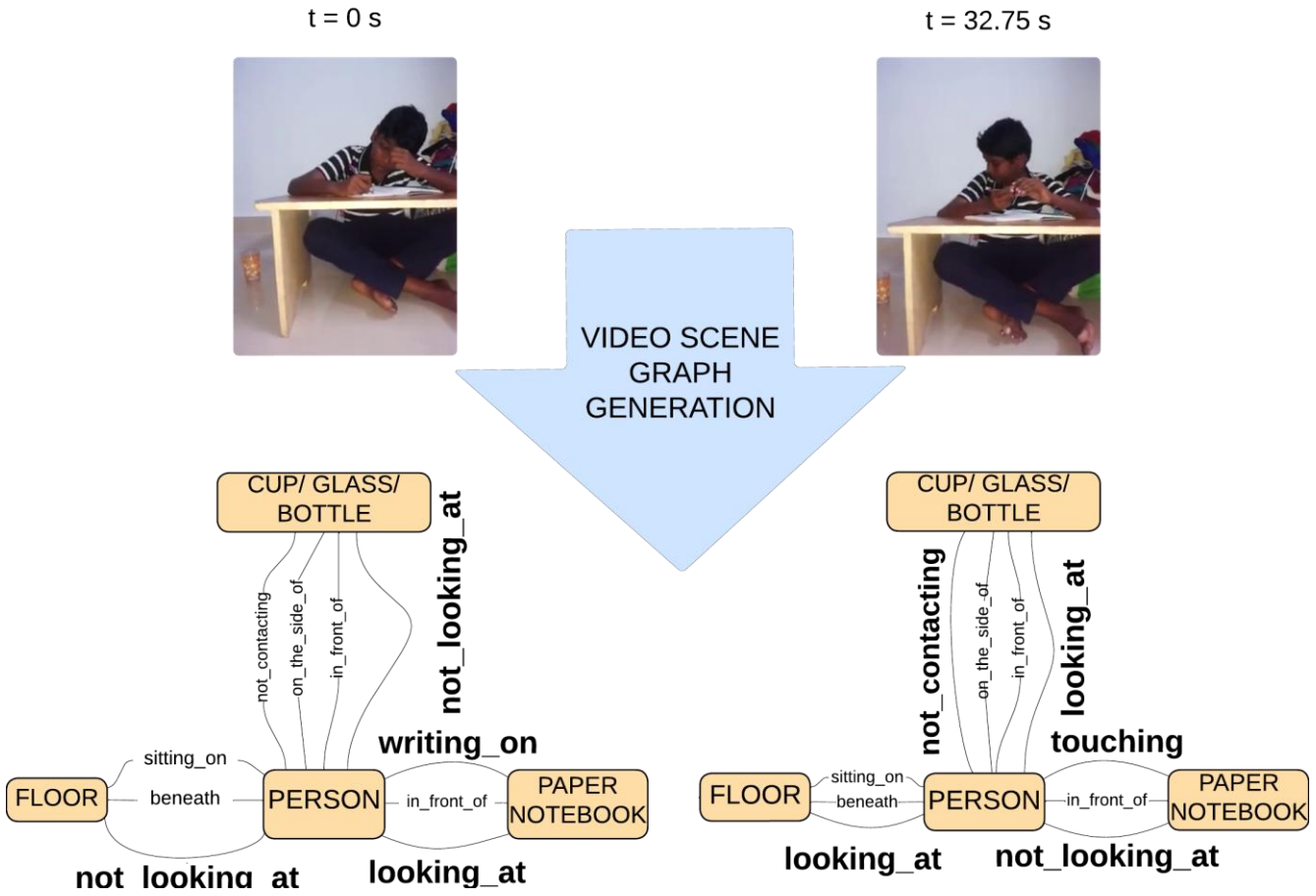
Towards Unbiased and Robust Scene Graph Generation and Anticipation

CVPR 2025

Rohith Peddi[+], Saurabh[*], Ayush Shrivastava[*],
Parag Singla[*], Vibhav Gogate[+]

+ UT Dallas, * IIT Delhi

Task - 1



Video Scene Graph Generation (VidSGG) entails the identification of localized fine-grained relationships between the objects observed in the video, such as (Person, looking at, Paper Notebook) and (Person, not looking at, Paper Notebook) in respective frames

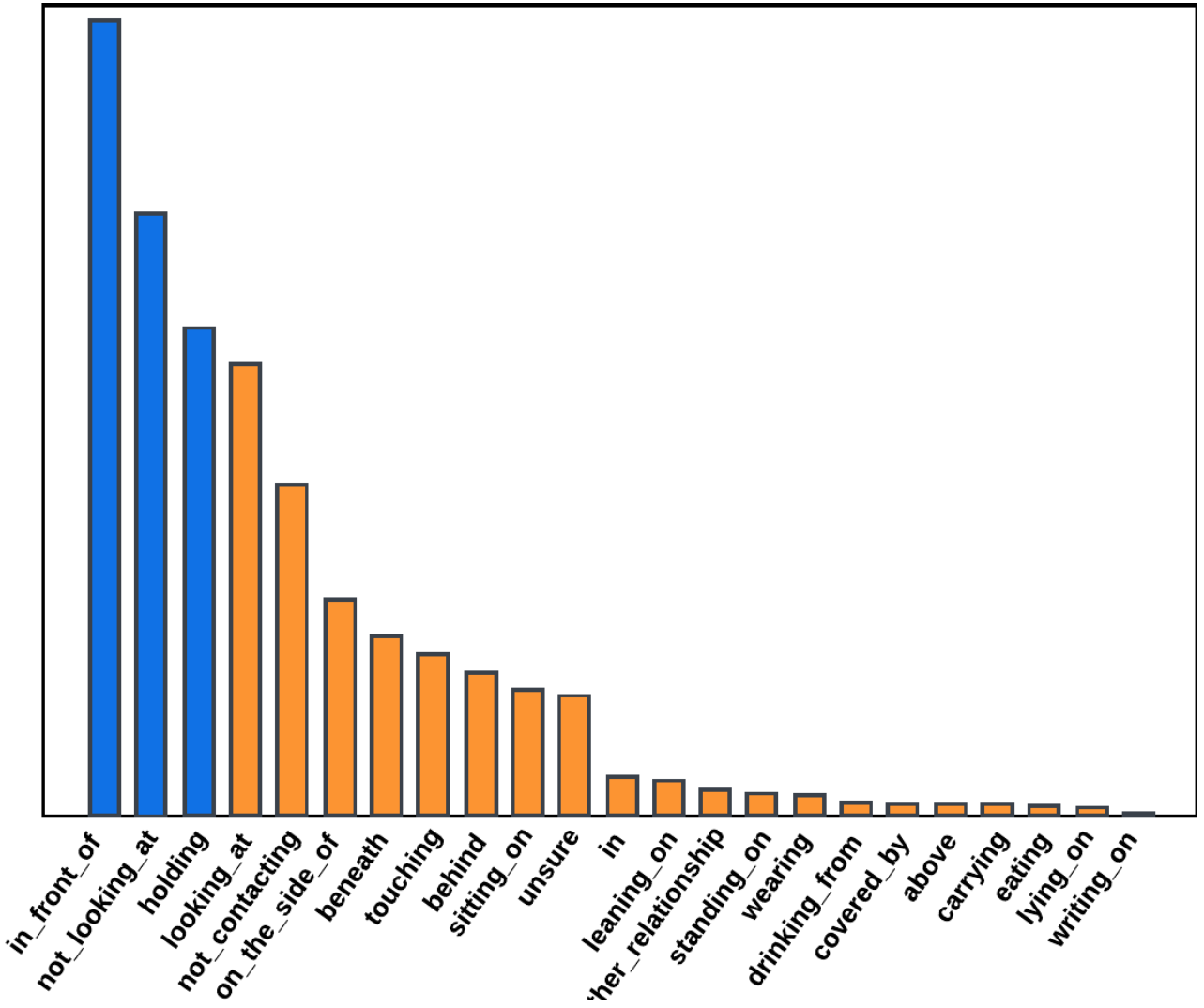
Task - 2

Scene Graph Anticipation aims to anticipate the evolution of relationships
(Person, **looking_at**, Floor) and (Person, **not_contacting**, Cup)
to (Person, **touching**, Cup), and eventually, (Person, **drinking_from**, Cup)

Data Distribution

Datasets

1. Dataset: Action Genome
2. Videos: Charades
3. Objects: 35 classes
4. Relationships: 25 classes
    1. Attention Relationships
    2. Spatial Relationships
    3. Contacting Relationships



Metrics

VidSGG and SGA

Recall @ K

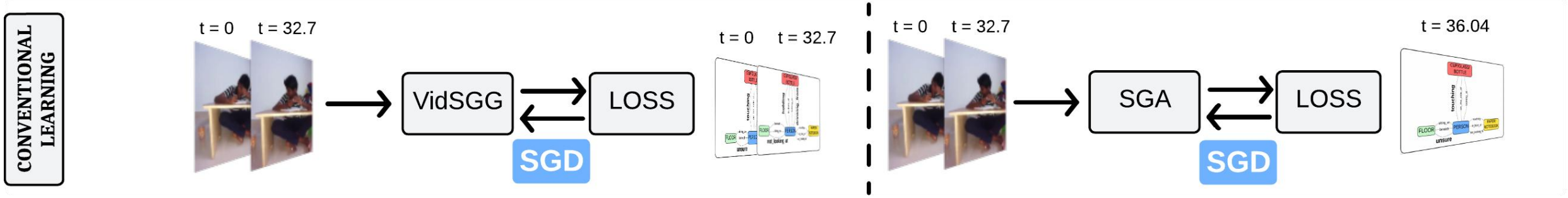
The **fraction** of **ground truth relationships** that are correctly predicted among the **top K predictions**.

Mean Recall @ K

Computed by **first** calculating the **recall for each relationship category** individually and then averaging these recalls over **all** the categories

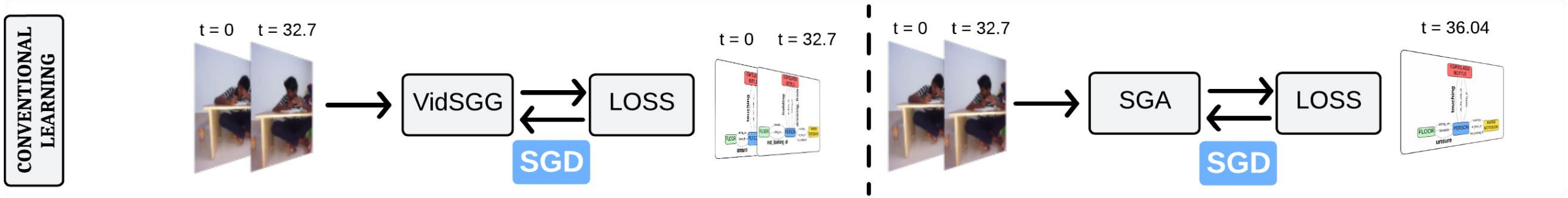
Related Work: Overview

VidSGG and SGA



Related Work: Overview

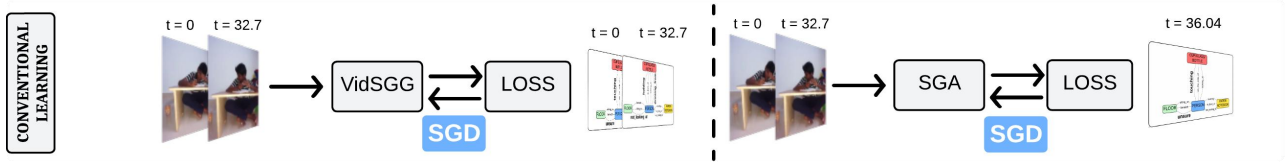
Digression



1. Isn’t this solved?
2. Can’t current large multimodal models get these things out of the box, right?

Related Work: Overview

Digression

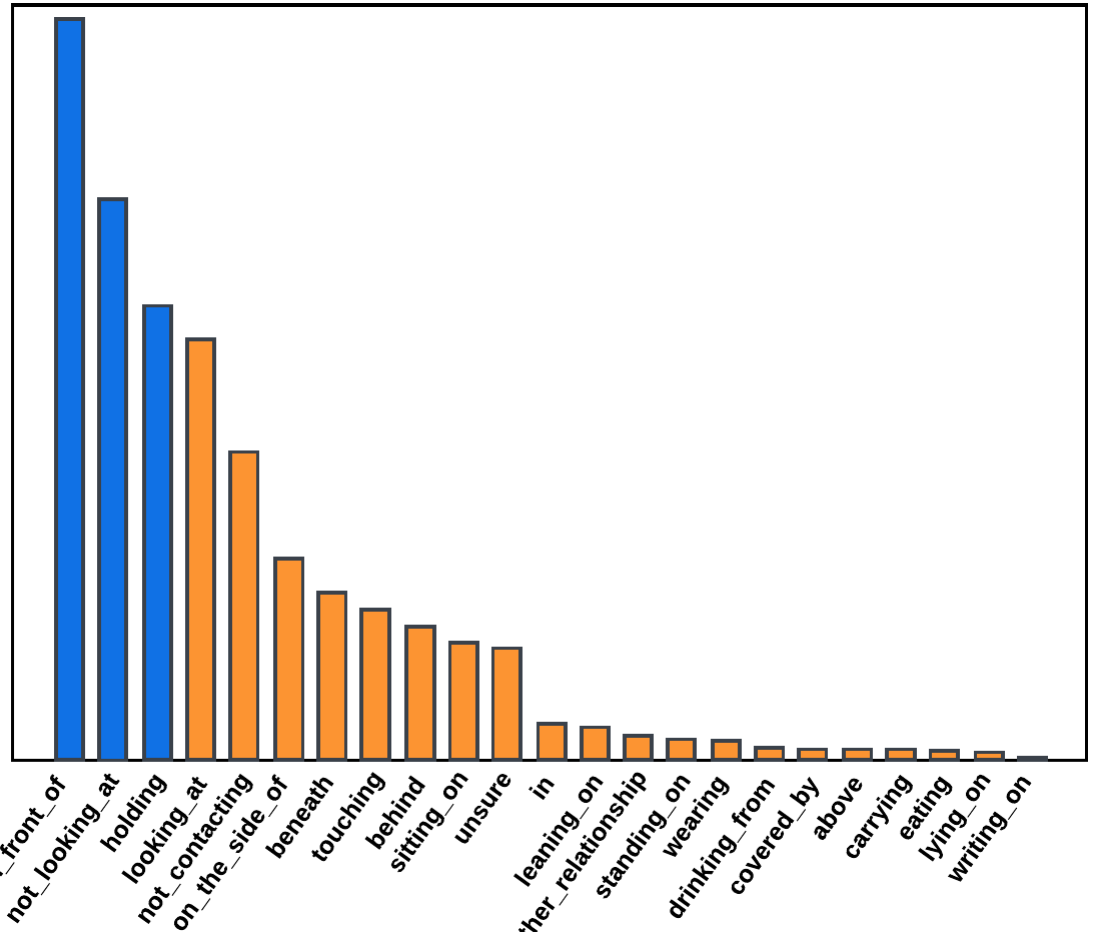


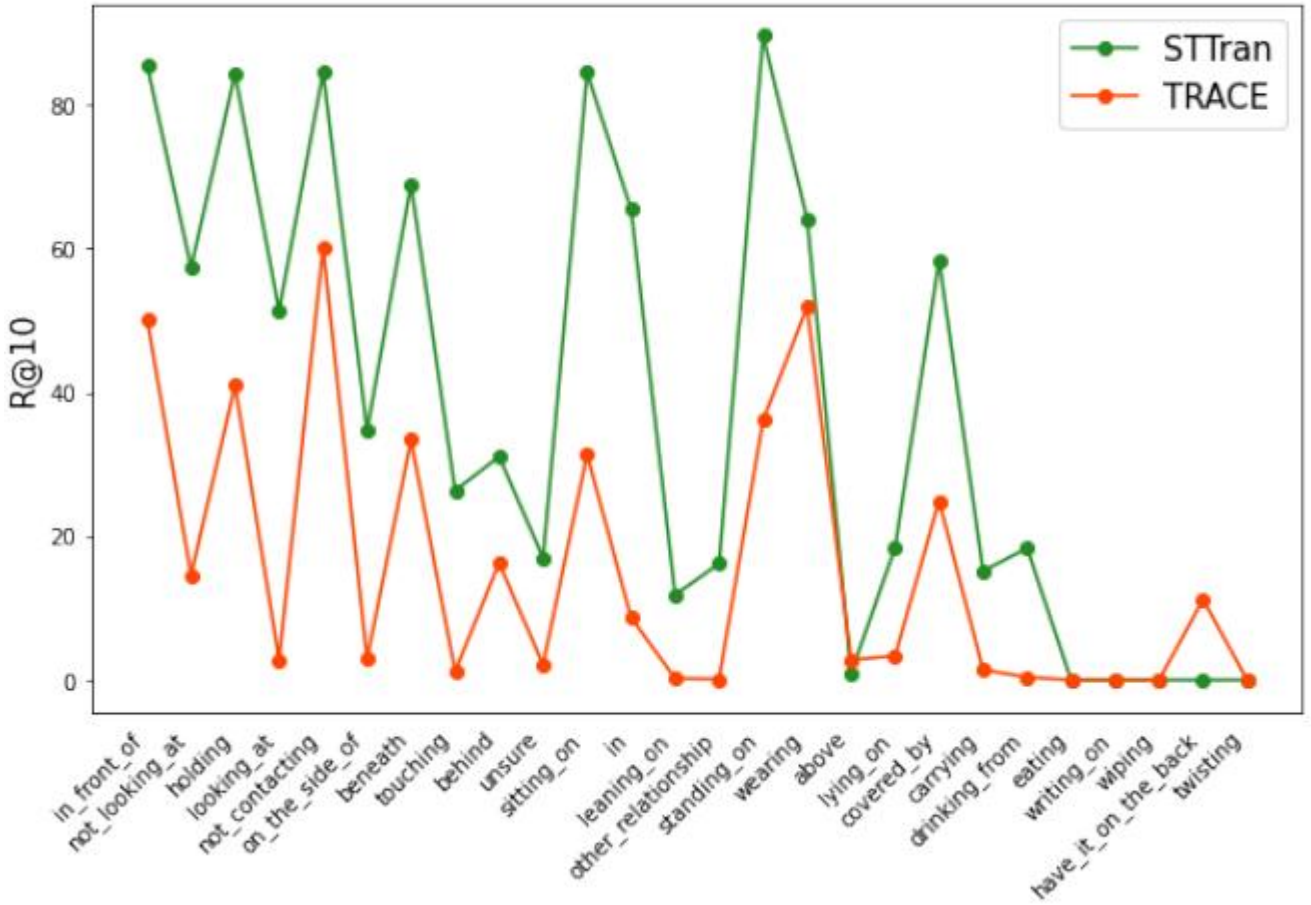
1. Isn't this solved?
2. Can't current large multimodal models get these things out of the box, right?

Video Scene Graph Generation

Dataset	Model	% Data	Recall @1	@10	@20	@50	@100
Action Genome	STTran	-	0.047	0.311	0.457	0.621	0.663
	TEMPURA	-	0.074	0.378	0.485	0.597	0.632
	OED	-	0.074	**0.443**	**0.559**	**0.668**	**0.735**
	Video-LLaVA	Zero-Shot	0.013	0.017	0.017	0.018	0.0178
	LLaVA-OV	Zero-Shot	0.040	0.066	0.067	0.067	0.067
	InternVL2	Zero-Shot	0.049	0.094	0.100	0.111	0.112
	Video-LLaVA	5%	0.045	0.153	0.158	0.158	0.158
	LLaVA-OV	5%	0.123	0.260	0.261	0.261	0.261
	InternVL2	5%	**0.140**	0.274	0.295	0.295	0.295
	Video-LLaVA	100%	0.109	0.289	0.293	0.293	0.293
	LLaVA-OV	100%	0.165	0.386	0.388	0.388	0.388
	InternVL2	100%	**0.189**	0.397	0.445	0.448	0.448

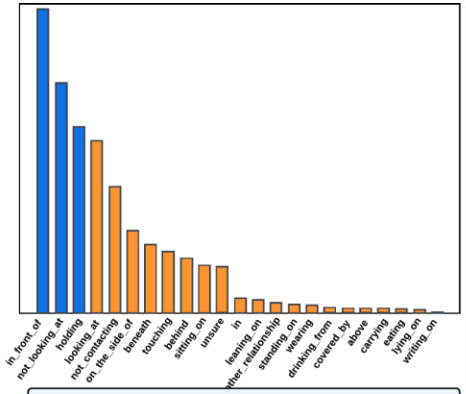
Problem - 1 : Biased Predictions





Problem - 1 : Biased Predictions

VidSGG and SGA



(a) Long-tailed distribution of Action Genome

Head Class

Tail Class

Incorrect prediction

Correct Prediction

Conventional Prediction
Incorrect tail class prediction scores

Unbiased Prediction
Correct tail class prediction scores

(b) Tasks

Observed Frames

t = 0 s

t = 32.75 s

VIDEO SCENE GRAPH GENERATION

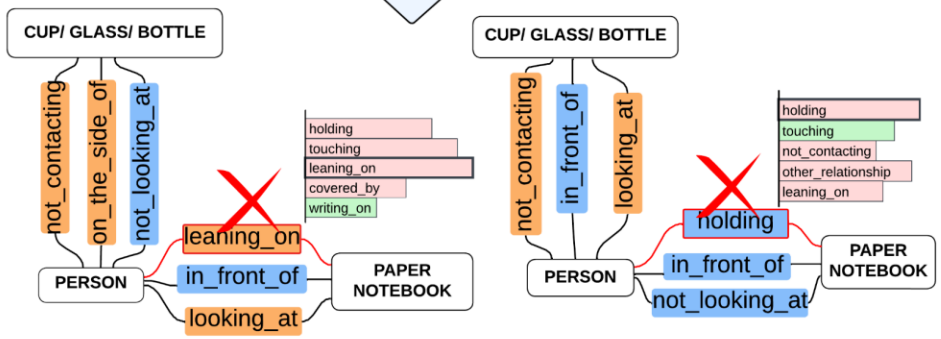
SCENE GRAPH ANTICIPATION

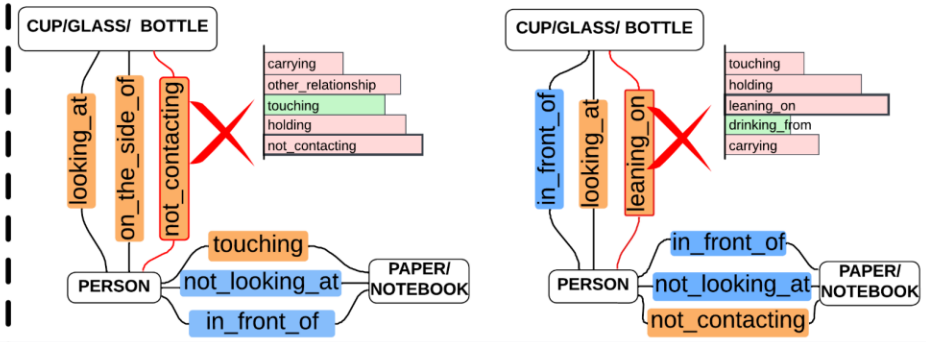
Unobserved Frames

t = 34.42 s

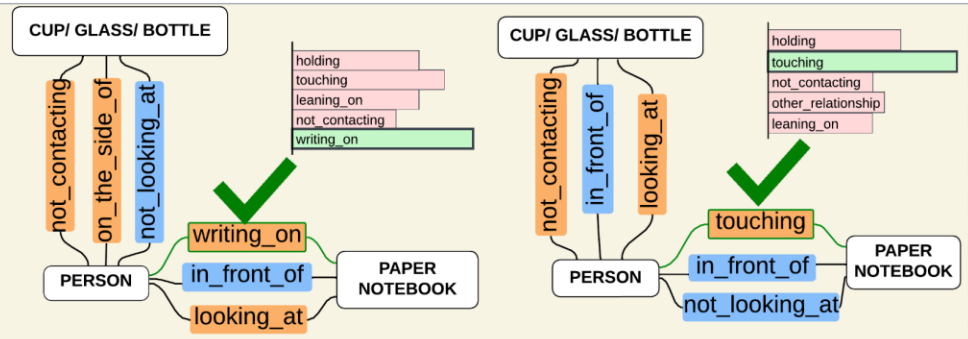
t = 36.04 s

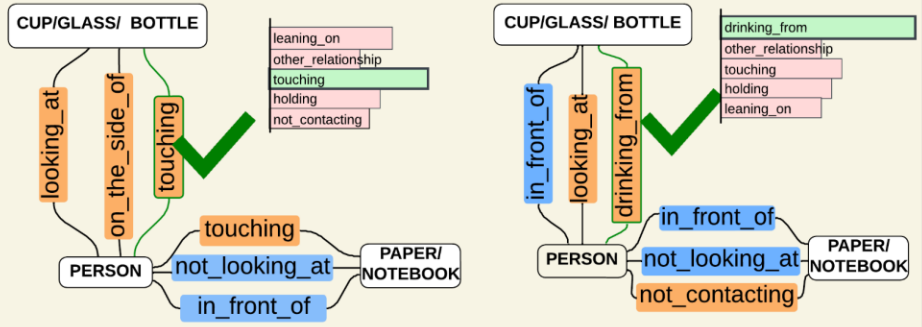
(c) Conventional Learning



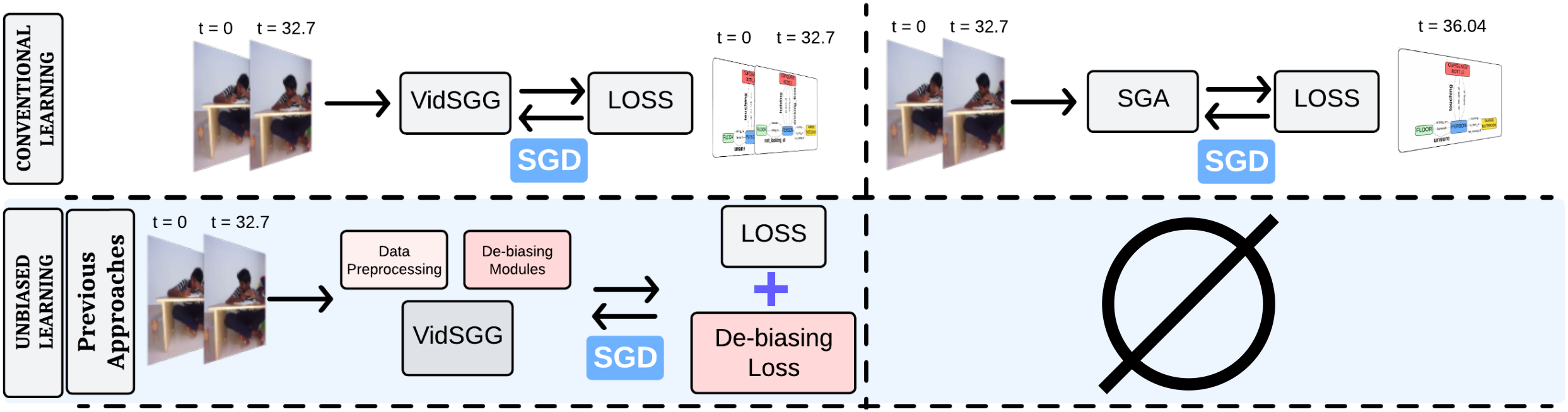


(d) Unbiased Learning

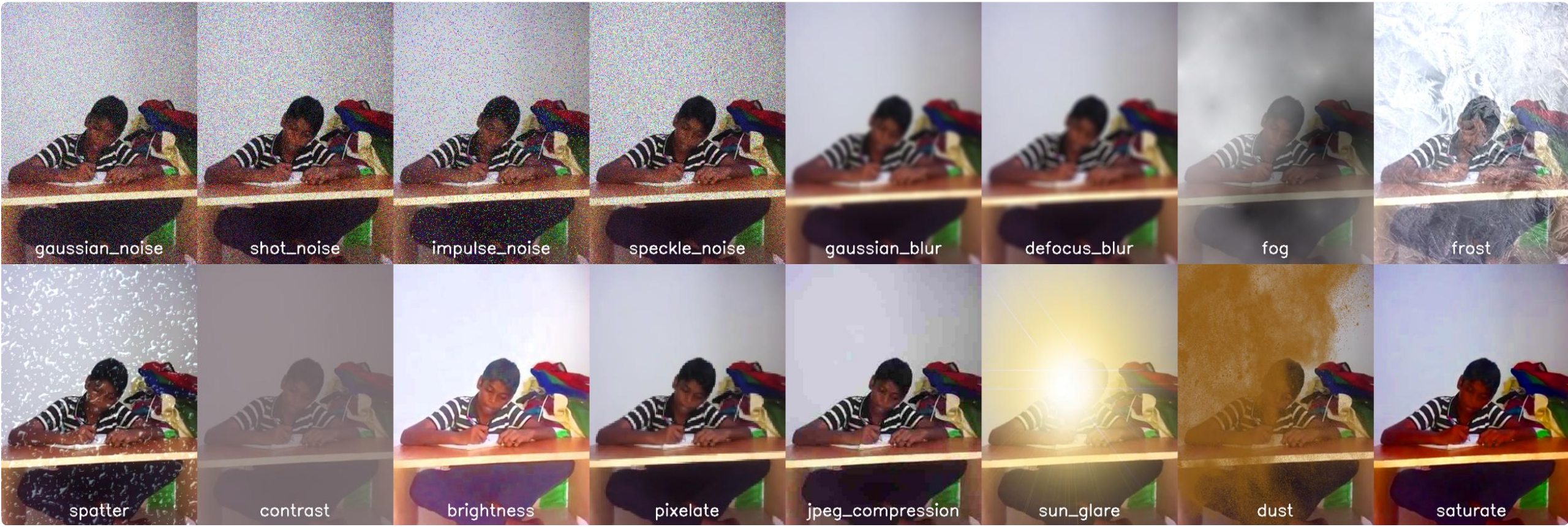




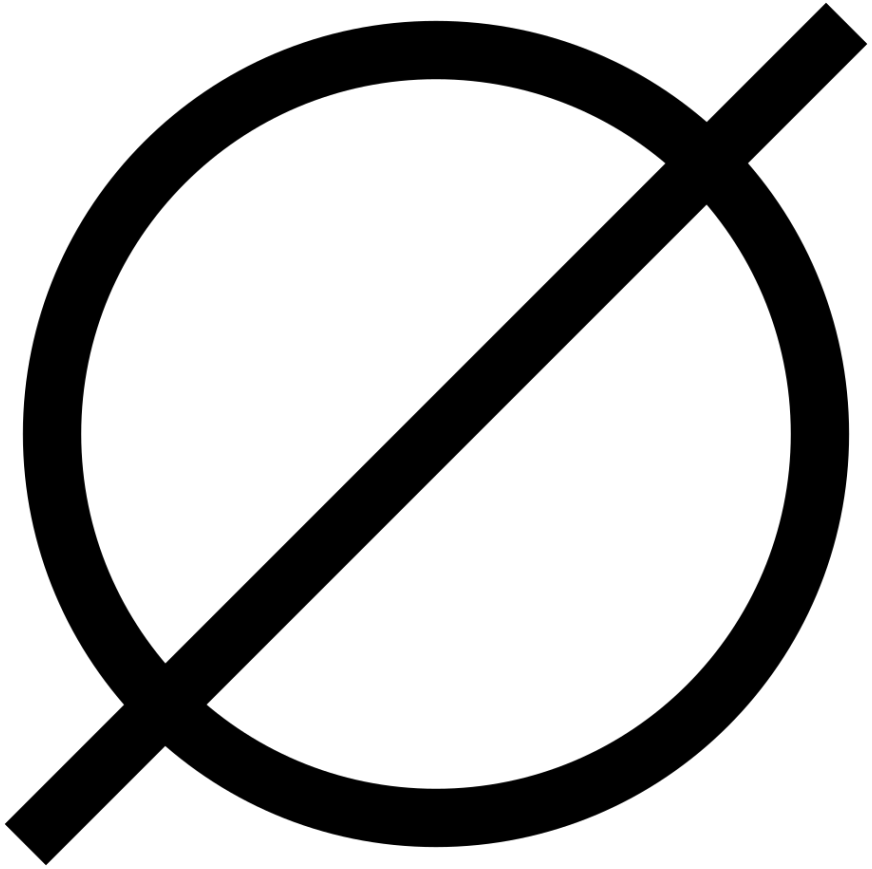
Solutions: In literature



Problem - 2 : Distribution Shifts

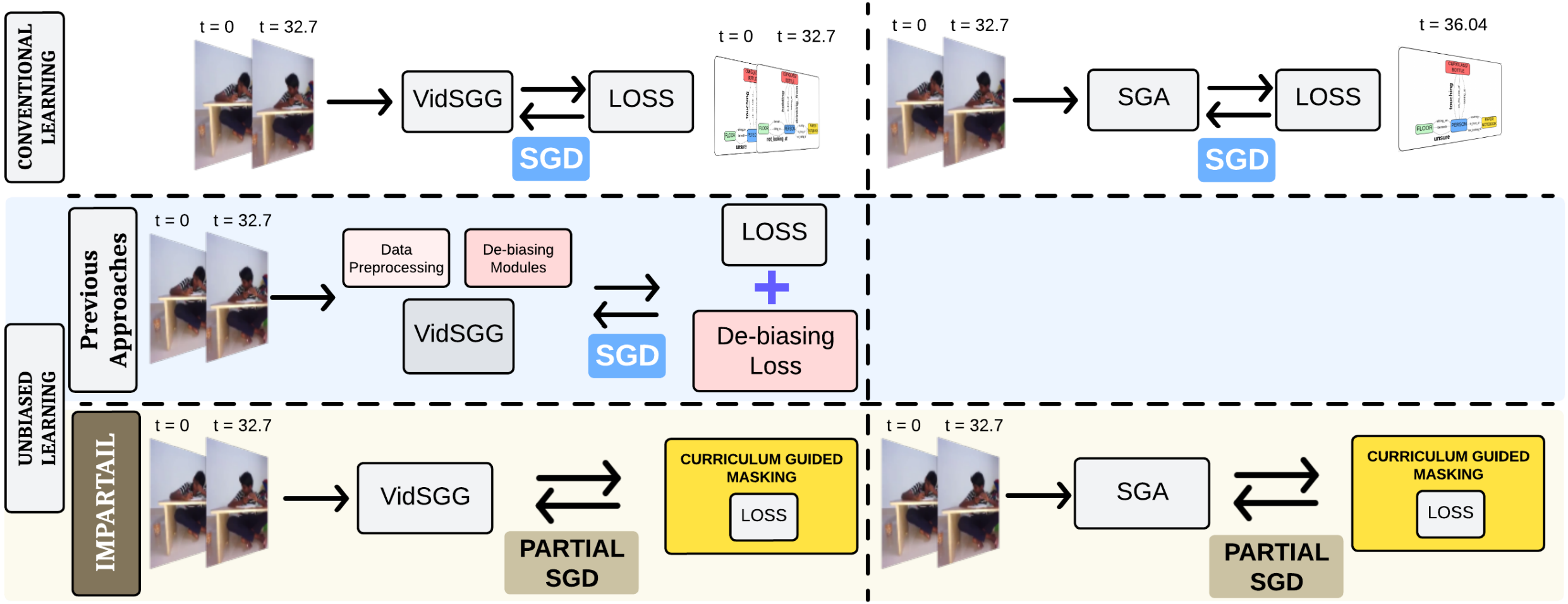


Solutions: In literature



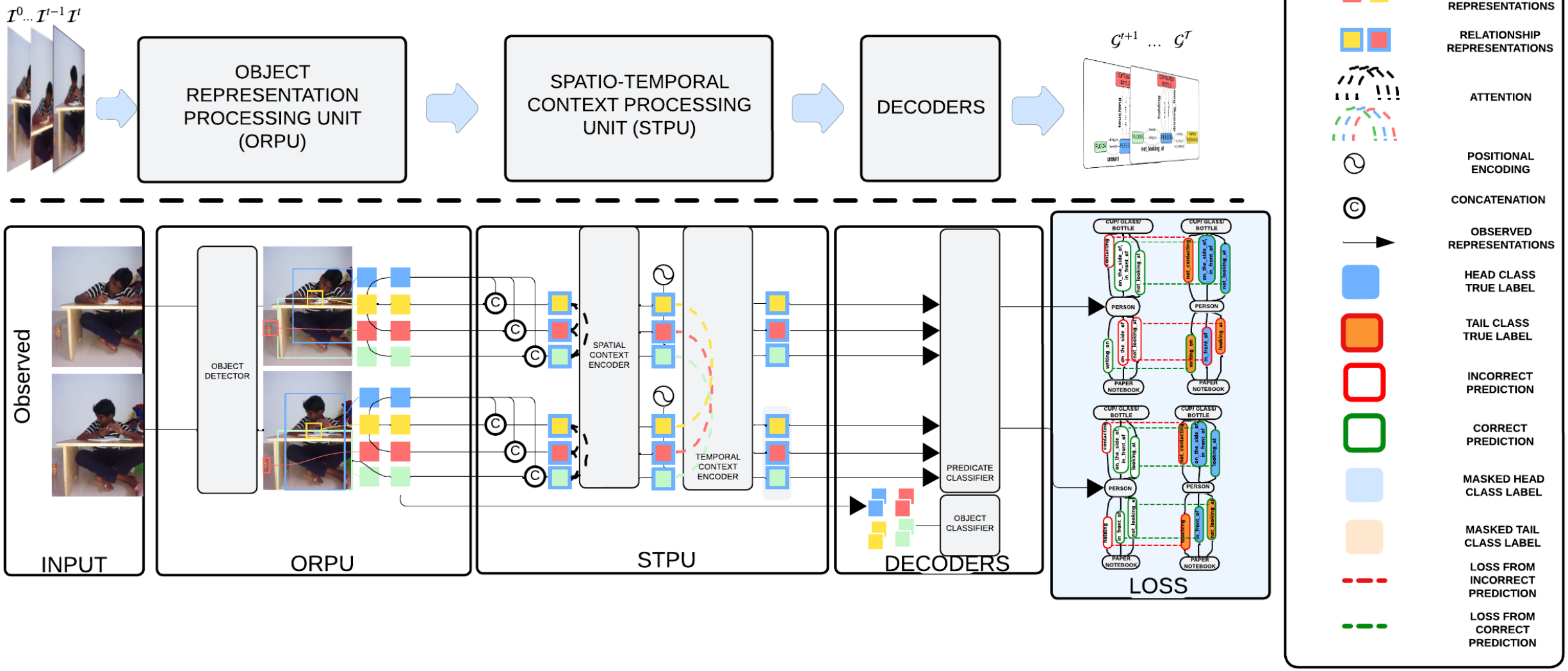
Proposed Solution

Unbiased and Robust VidSGG and SGA

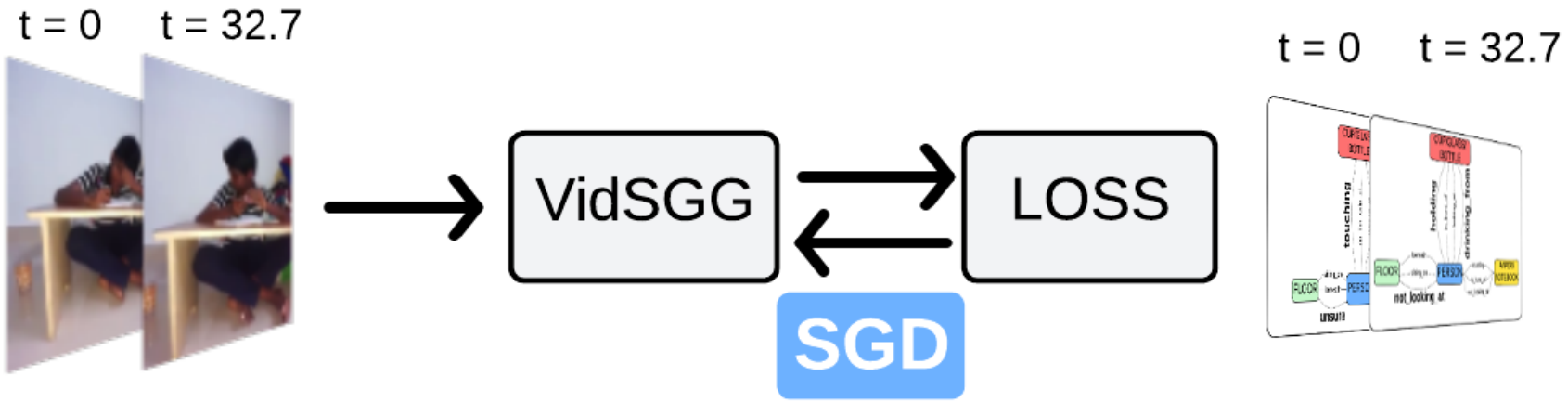


Architectures for Tasks

Video Scene Graph Generation



Architectures for Tasks – Loss Function



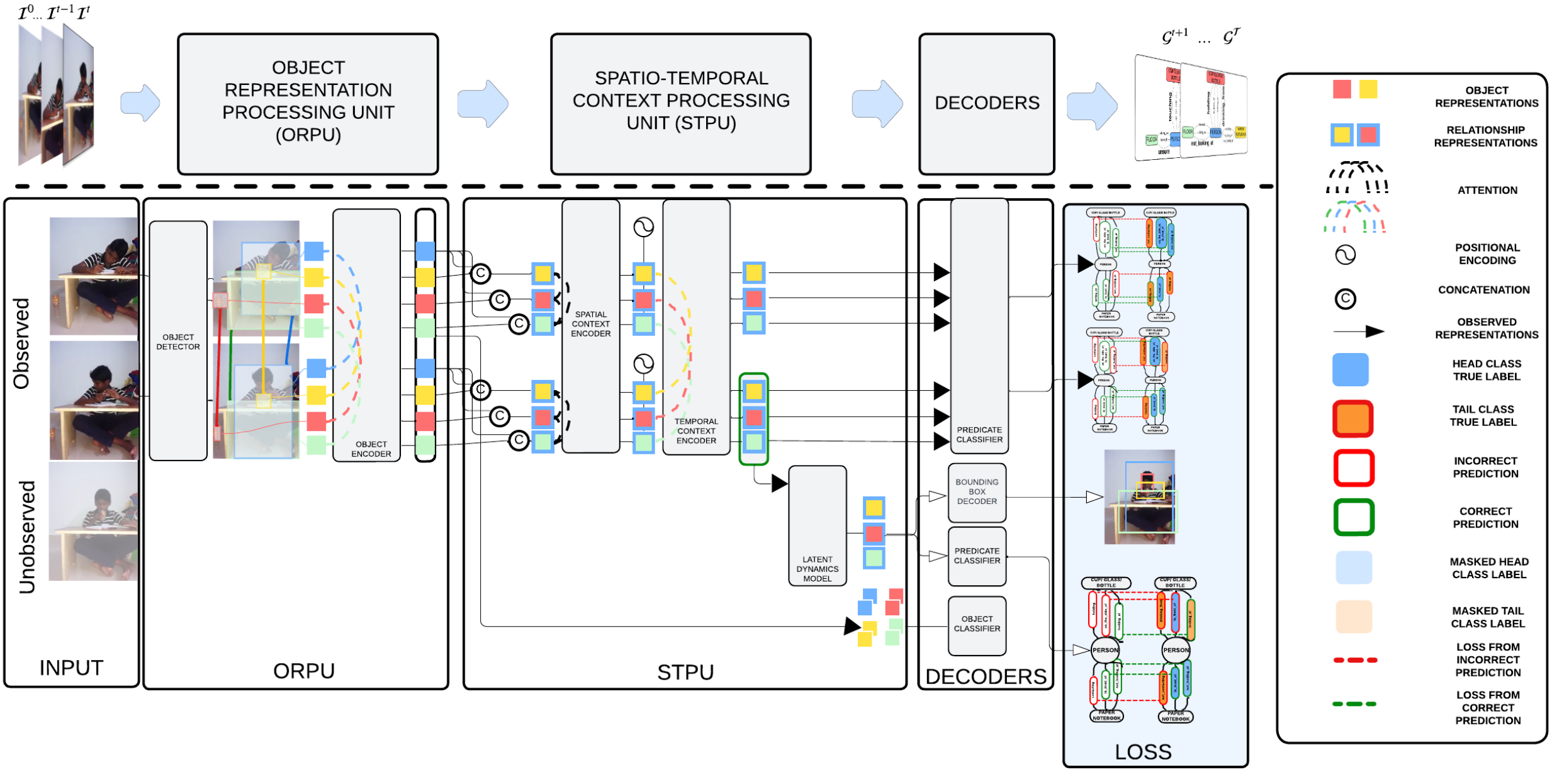
1. Object Classification Loss
2. Predicate Classification Loss

$$\underbrace{\mathcal{L}_i = \sum_{t=1}^{\bar{T}} \mathcal{L}_i^t;}_{(1)} \quad \underbrace{\mathcal{L}_{\text{gen}} = \sum_{t=1}^{\bar{T}} \mathcal{L}_{\text{gen}}^t, \mathcal{L}_{\text{gen}}^t = \sum_{ij} \mathcal{L}_{p_{ij}^t}}_{\text{Predicate Classification Loss (2)}}$$

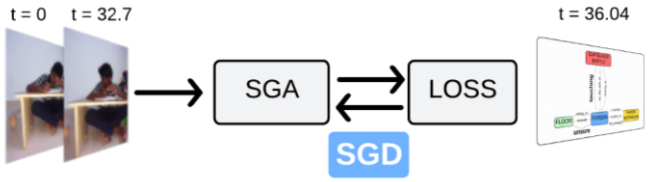
$$\mathcal{L} = \sum_{t=1}^{\bar{T}} \left(\lambda_1 \mathcal{L}_{\text{gen}}^t + \lambda_2 \sum_i \mathcal{L}_i^t \right)$$

Architectures for Tasks

Scene Graph Anticipation



Architectures for Tasks – Loss Function



1. Loss Over Observed Representations
    1. Object Classification Loss
    2. Observed Predicate Classification Loss
2. Loss Over Anticipated Representations
    1. Anticipated Predicate Classification Loss
    2. Bounding Box Regression Loss
    3. Representation Reconstruction Loss

$$\underbrace{\mathcal{L}_i = \sum_{t=1}^{\bar{T}} \mathcal{L}_i^t, \quad \mathcal{L}_i^t = -\sum_{n=1}^{|\mathcal{C}|} y_{i,n}^t \log(\hat{\mathbf{c}}_{i,n}^t);}_{\text{Object Classification Loss (I)}} \quad \underbrace{\mathcal{L}_{\text{gen}} = \sum_{t=1}^{\bar{T}} \mathcal{L}_{\text{gen}}^t, \quad \mathcal{L}_{\text{gen}}^t = \sum_{ij} \mathcal{L}_{p_{ij}^t}}_{\text{Predicate Classification Loss (II)}}$$

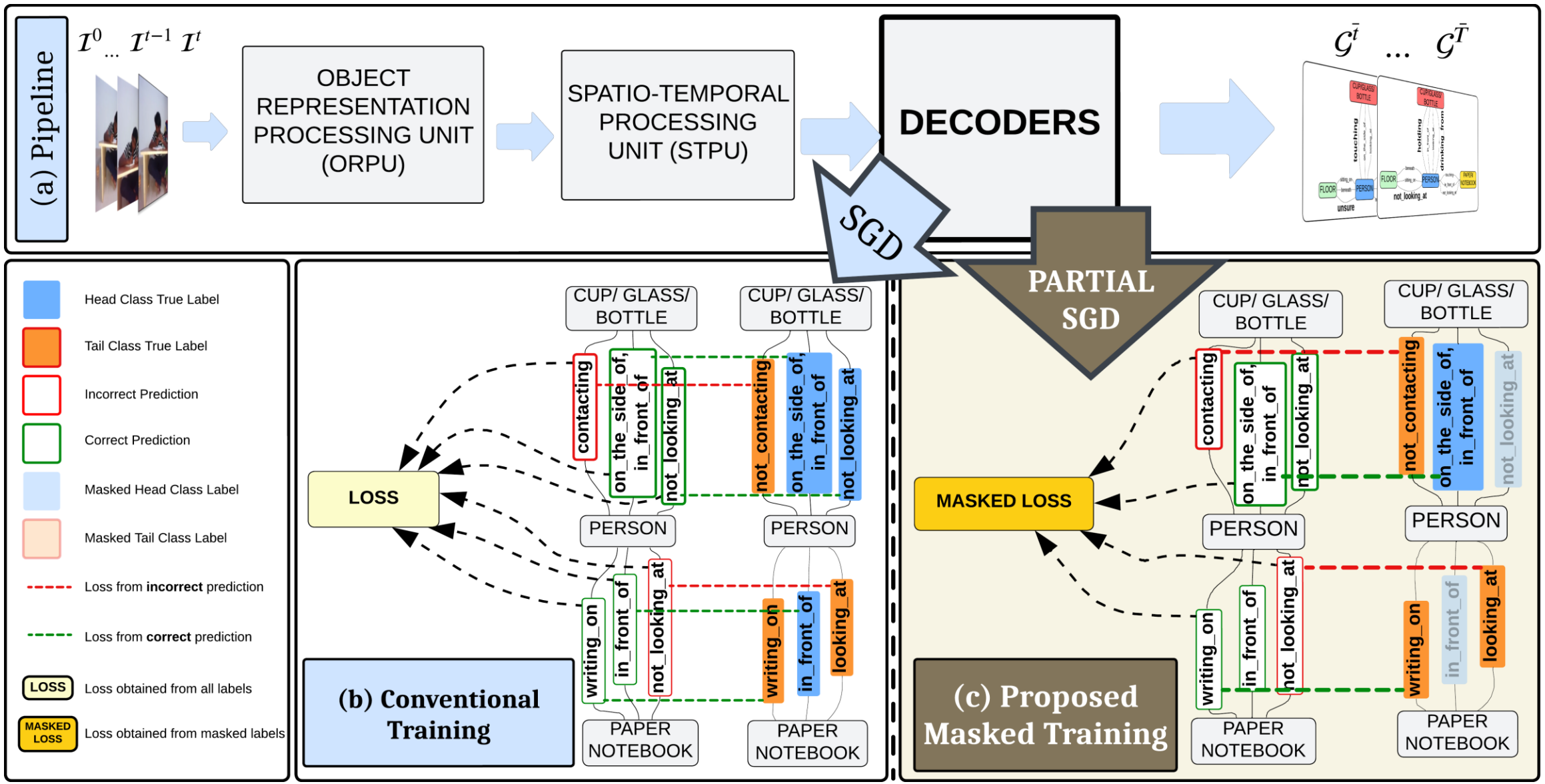
$$\mathcal{L}_{\text{ant}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathcal{L}_{\text{ant}}^t, \quad \mathcal{L}_{\text{ant}}^t = \sum_{ij} \mathcal{L}_{p_{ij}^t}$$

$$\mathcal{L}_{\text{boxes}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathcal{L}_{\text{boxes}}^t, \quad \mathcal{L}_{\text{boxes}}^t = \sum_{k \in \text{boxes}} \mathrm{L}_{\text{smooth}}(b_k^t - \hat{b}_k^t)$$

$$\mathcal{L}_{\text{recon}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathcal{L}_{\text{recon}}^t, \quad \mathcal{L}_{\text{recon}}^t = \frac{1}{N(t) \times N(t)} \sum_{ij}^{(N(t) \times N(t))} \mathrm{L}_{\text{smooth}}(\mathbf{z}_{ij}^t - \hat{\mathbf{z}}_{ij}^t)$$

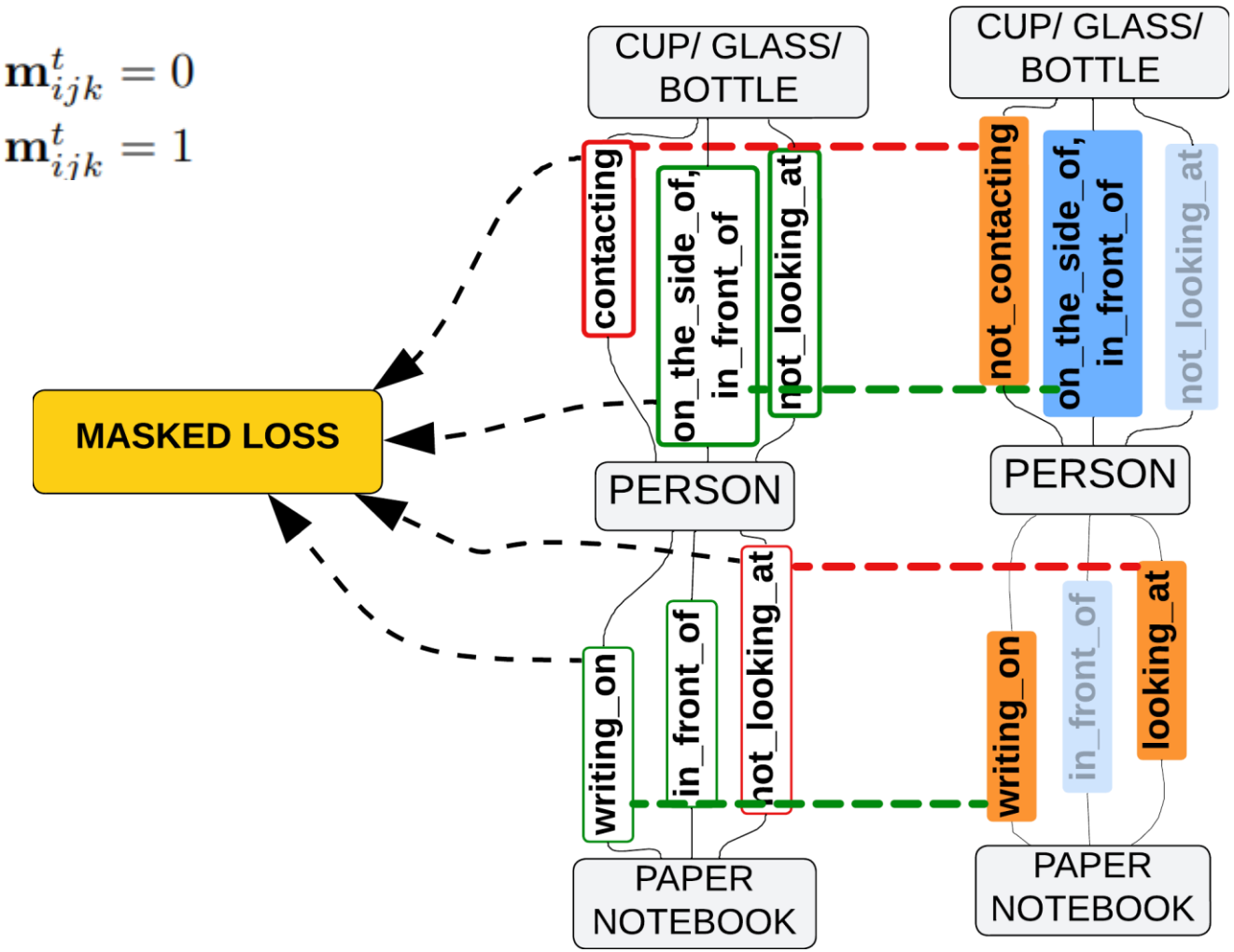
$$\mathcal{L} = \underbrace{\sum_{t=1}^{\bar{T}} \left(\lambda_1 \mathcal{L}_{\text{gen}}^t + \lambda_2 \sum_i \mathcal{L}_i^t \right)}_{\text{Loss Over Observed Representations}} + \underbrace{\sum_{T=3}^{\bar{T}-1} \left(\lambda_3 \mathcal{L}_{\text{ant}}^{(1:T)} + \lambda_4 \mathcal{L}_{\text{boxes}}^{(1:T)} + \lambda_5 \mathcal{L}_{\text{recon}}^{(1:T)} \right)}_{\text{Loss Over Anticipated Representations}}$$

Proposed Solution: Masked Training

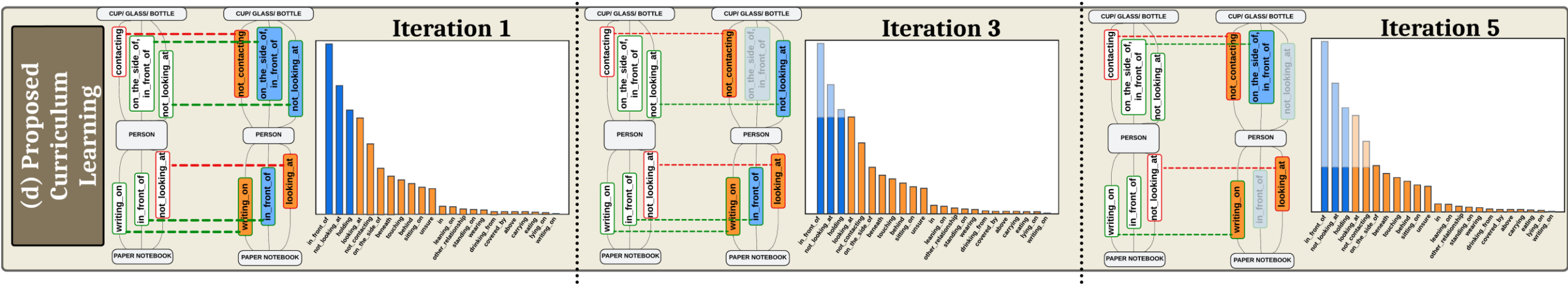


Proposed Solution: Masked Loss

$$\mathscr{L}_{p_{ijk}^t} = (1 - \mathbf{m}_{ijk}^t) * \mathcal{L}_{p_{ijk}^t} = \begin{cases} \mathcal{L}_{p_{ijk}^t} & \text{if } \mathbf{m}_{ijk}^t = 0 \\ 0 & \text{if } \mathbf{m}_{ijk}^t = 1 \end{cases}$$

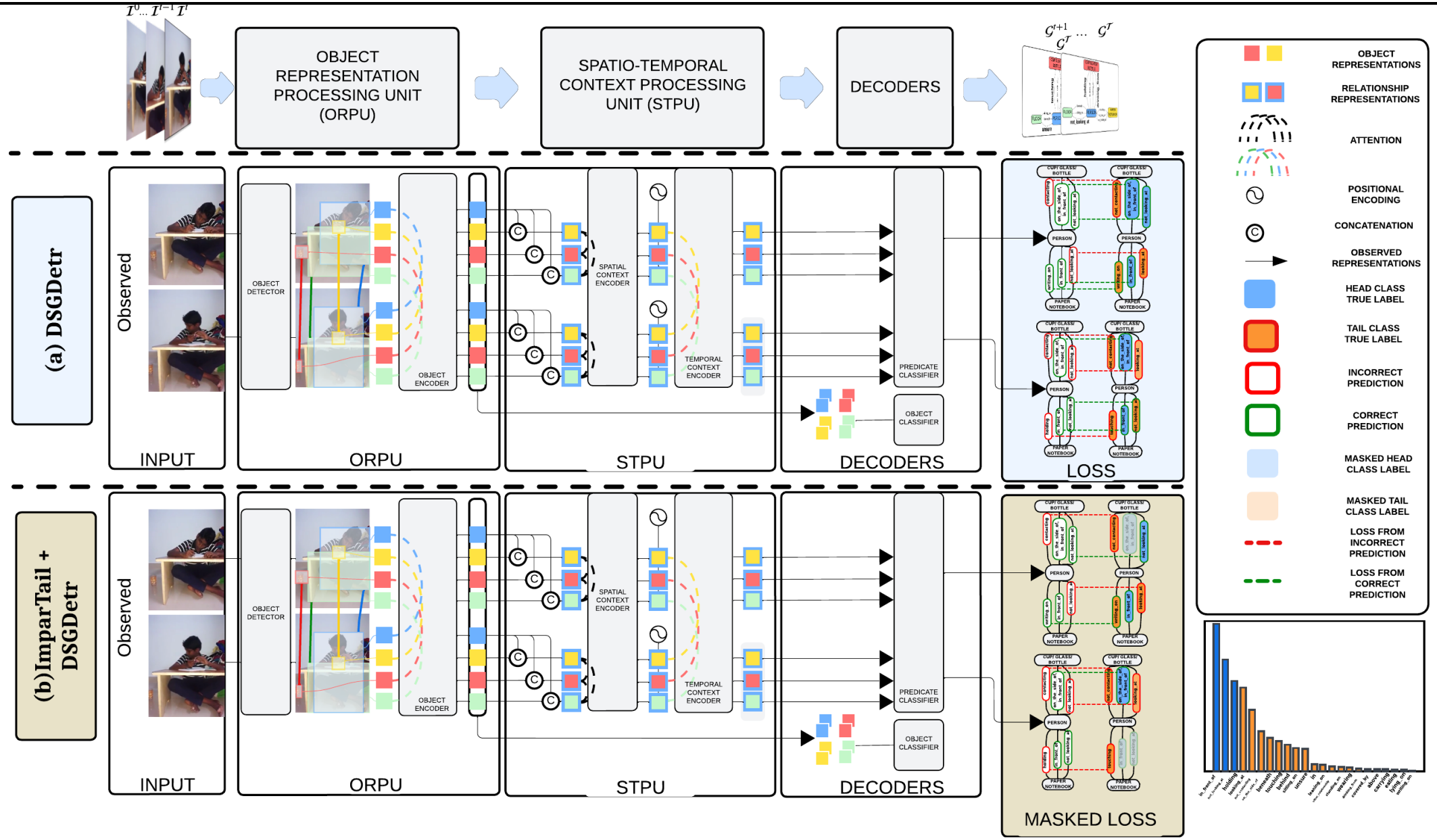


Proposed Solution: Mask Generation

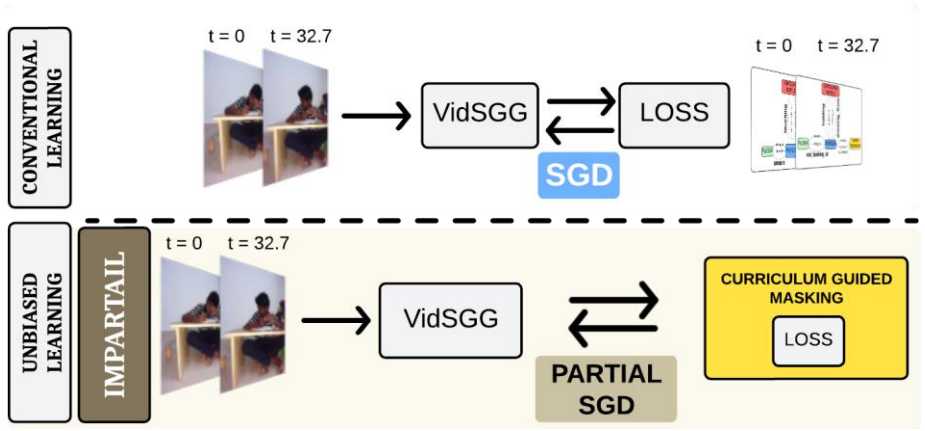


Proposed Solution: In Model

Video Scene Graph Generation



Proposed Solution: In Model



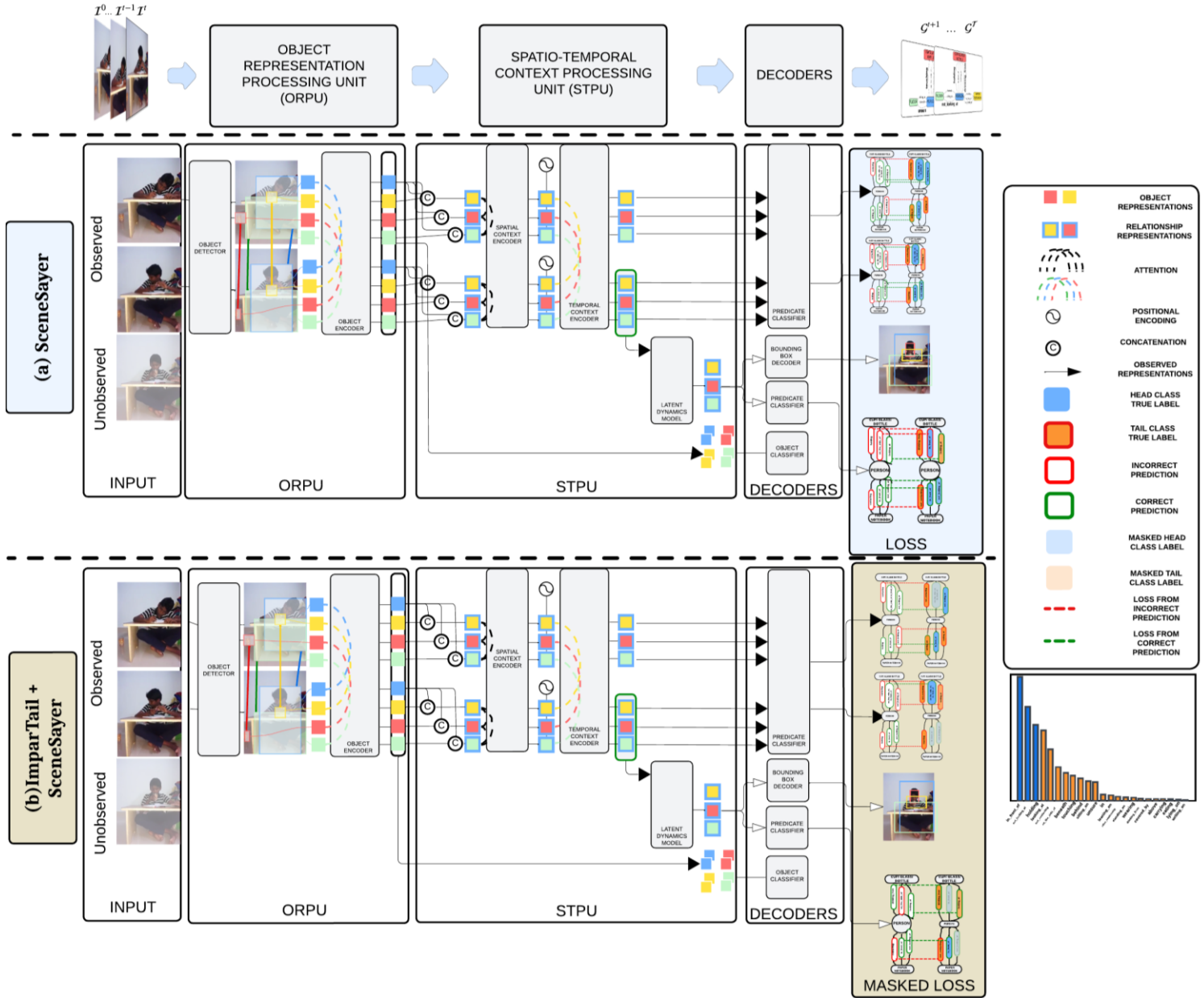
$$\underbrace{\mathcal{L}_i = \sum_{t=1}^{\bar{T}} \mathcal{L}_i^t;}_{(1)} \quad \underbrace{\mathcal{L}_{\text{gen}} = \sum_{t=1}^{\bar{T}} \mathcal{L}_{\text{gen}}^t, \mathcal{L}_{\text{gen}}^t = \sum_{ij} \mathcal{L}_{p_{ij}^t}}_{\text{Predicate Classification Loss (2)}}$$

$$\mathcal{L} = \sum_{t=1}^{\bar{T}} \left(\lambda_1 \mathcal{L}_{\text{gen}}^t + \lambda_2 \sum_i \mathcal{L}_i^t \right)$$

$$\underbrace{\mathcal{L}_i = \sum_{t=1}^{\bar{T}} \mathcal{L}_i^t;}_{(1)} \quad \underbrace{\mathscr{L}_{\text{gen}} = \sum_{t=1}^{\bar{T}} \mathscr{L}_{\text{gen}}^t, \mathscr{L}_{\text{gen}}^t = \sum_{ij} \mathbf{m}_{ij}^t * \mathcal{L}_{p_{ij}^t}}_{\textbf{Masked} \text{ Predicate Classification Loss (2)}}$$

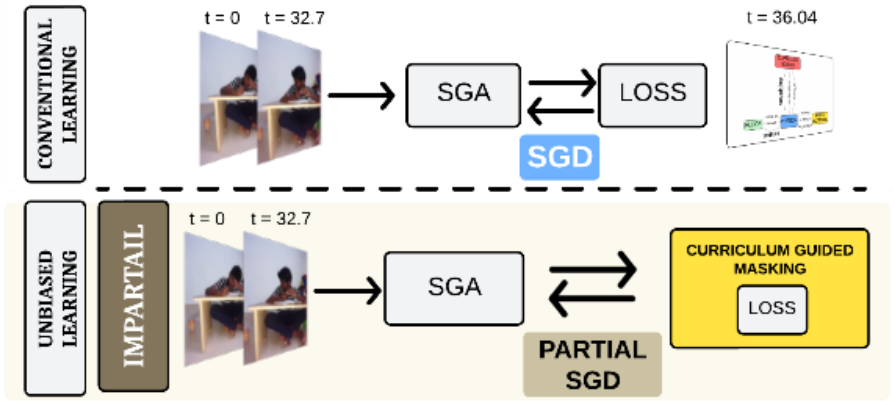
$$\mathcal{L} = \sum_{t=1}^{\bar{T}} \left(\lambda_1 \mathscr{L}_{\text{gen}}^t + \lambda_2 \sum_i \mathcal{L}_i^t \right)$$

Proposed Solution: In Model



Proposed Solution: In Model

Scene Graph Anticipation



$$\underbrace{\mathcal{L}_i = \sum_{t=1}^{\bar{T}} \mathcal{L}_i^t, \quad \mathcal{L}_i^t = -\sum_{n=1}^{|\mathcal{C}|} y_{i,n}^t \log(\hat{\mathbf{c}}_{i,n}^t);}_{\text{Object Classification Loss (I)}} \quad \underbrace{\mathcal{L}_{\text{gen}} = \sum_{t=1}^{\bar{T}} \mathcal{L}_{\text{gen}}^t, \quad \mathcal{L}_{\text{gen}}^t = \sum_{ij} \mathcal{L}_{p_{ij}^t}}_{\text{Predicate Classification Loss (II)}}$$

$$\mathcal{L}_{\text{ant}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathcal{L}_{\text{ant}}^t, \quad \mathcal{L}_{\text{ant}}^t = \sum_{ij} \mathcal{L}_{p_{ij}^t}$$

$$\mathcal{L}_{\text{boxes}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathcal{L}_{\text{boxes}}^t, \quad \mathcal{L}_{\text{boxes}}^t = \sum_{k \in \text{boxes}} \text{L}_{\text{smooth}}(b_k^t - \hat{b}_k^t)$$

$$\mathcal{L}_{\text{recon}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathcal{L}_{\text{recon}}^t, \quad \mathcal{L}_{\text{recon}}^t = \frac{1}{N(t) \times N(t)} \sum_{ij}^{(N(t)\times N(t))} \text{L}_{\text{smooth}}(\mathbf{z}_{ij}^t - \hat{\mathbf{z}}_{ij}^t)$$

$$\mathcal{L} = \underbrace{\sum_{t=1}^{\bar{T}} \left(\lambda_1 \mathcal{L}_{\text{gen}}^t + \lambda_2 \sum_i \mathcal{L}_i^t \right)}_{\text{Loss Over Observed Representations}} + \underbrace{\sum_{T=3}^{\bar{T}-1} \left(\lambda_3 \mathcal{L}_{\text{ant}}^{(1:T)} + \lambda_4 \mathcal{L}_{\text{boxes}}^{(1:T)} + \lambda_5 \mathcal{L}_{\text{recon}}^{(1:T)} \right)}_{\text{Loss Over Anticipated Representations}}$$

$$\underbrace{\mathcal{L}_i = \sum_{t=1}^{\bar{T}} \mathcal{L}_i^t, \quad \mathcal{L}_i^t = -\sum_{n=1}^{|\mathcal{C}|} y_{i,n}^t \log(\hat{\mathbf{c}}_{i,n}^t);}_{\text{Object Classification Loss (I)}} \quad \underbrace{\mathscr{L}_{\text{gen}} = \sum_{t=1}^{\bar{T}} \mathscr{L}_{\text{gen}}^t, \quad \mathscr{L}_{\text{gen}}^t = \sum_{ij} \mathbf{m}_{ij}^t * \mathcal{L}_{p_{ij}^t}}_{\text{Predicate Classification Loss (II)}}$$

$$\mathscr{L}_{\text{ant}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathscr{L}_{\text{ant}}^t, \quad \mathscr{L}_{\text{ant}}^t = \sum_{ij} \mathbf{m}_{ij}^t * \mathcal{L}_{p_{ij}^t}$$

$$\mathcal{L}_{\text{boxes}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathcal{L}_{\text{boxes}}^t, \quad \mathcal{L}_{\text{boxes}}^t = \sum_{k \in \text{boxes}} \text{L}_{\text{smooth}}(b_k^t - \hat{b}_k^t)$$

$$\mathcal{L}_{\text{recon}}^{(1:T)} = \sum_{t=T+1}^{\min(T+H,\bar{T})} \mathcal{L}_{\text{recon}}^t, \quad \mathcal{L}_{\text{recon}}^t = \frac{1}{N(t) \times N(t)} \sum_{ij}^{(N(t)\times N(t))} \text{L}_{\text{smooth}}(\mathbf{z}_{ij}^t - \hat{\mathbf{z}}_{ij}^t)$$

$$\mathcal{L} = \underbrace{\sum_{t=1}^{\bar{T}} \left(\lambda_1 \mathscr{L}_{\text{gen}}^t + \lambda_2 \sum_i \mathcal{L}_i^t \right)}_{\text{Loss Over Observed Representations}} + \underbrace{\sum_{T=3}^{\bar{T}-1} \left(\lambda_3 \mathscr{L}_{\text{ant}}^{(1:T)} + \lambda_4 \mathcal{L}_{\text{boxes}}^{(1:T)} + \lambda_5 \mathcal{L}_{\text{recon}}^{(1:T)} \right)}_{\text{Loss Over Anticipated Representations}}$$

Results

Settings: VidSGG

Scene Graph Detection

- **Input:**
  - Frames in a video
- **Output:**
  - Localized relationship predicates

Scene Graph Classification

- **Input:**
  - Frames in a video
  - Object bounding boxes
- **Output:**
  - Localized relationship predicates

Predicate Classification

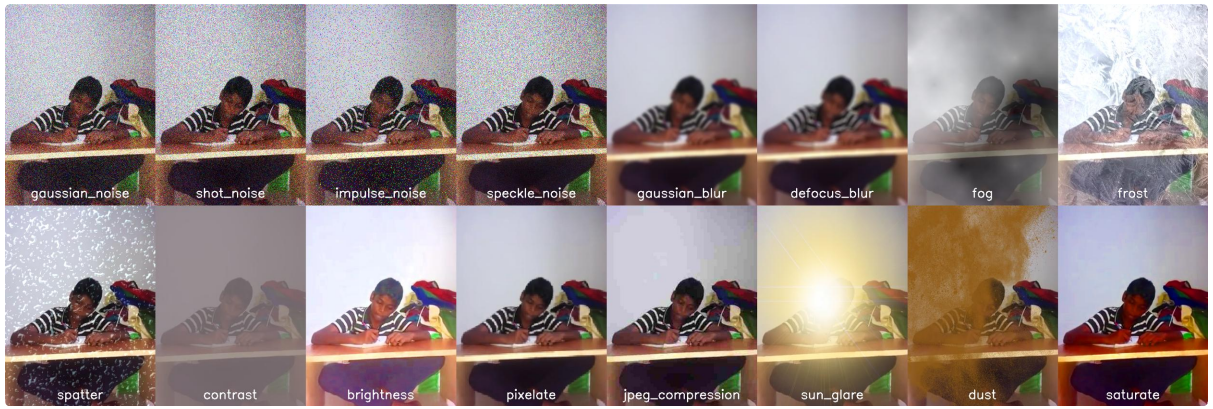
- **Input:**
  - Frames in a video
  - Object bounding boxes
  - Object labels
- **Output:**
  - Localized relationship predicates

Results - 1: Video Scene Graph Generation

Table 1. Mean Recall Results for VidSGG.

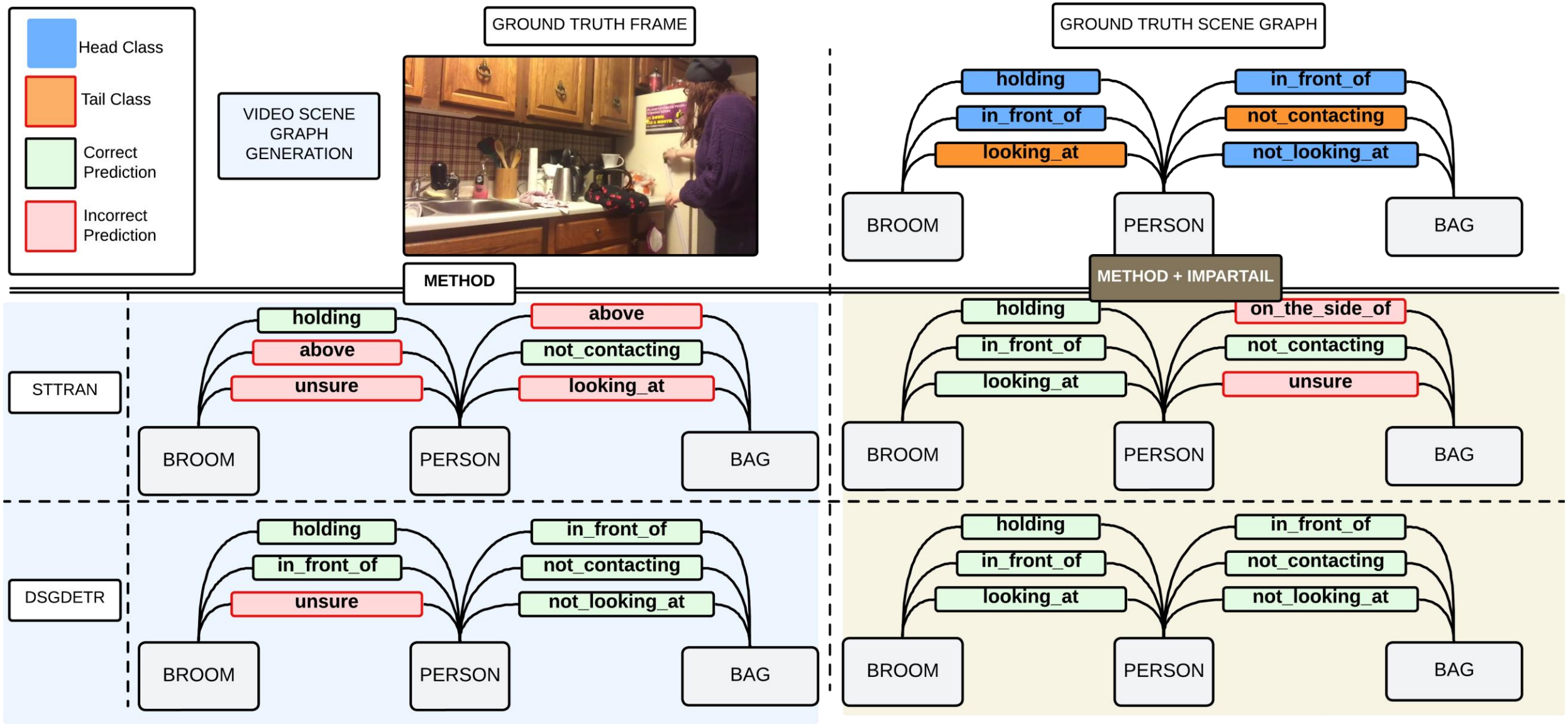
Mode	Method	With Constraint			No Constraint			Semi Constraint		
		mR@10	mR@20	mR@50	mR@10	mR@20	mR@50	mR@10	mR@20	mR@50
SGDET	STTran [7]	8.0	16.6	19.3	19.3	26.9	35.6	7.7	18.2	30.4
	+ImpartTail (Ours)	9.4 (+17.5%)	21.5 (+29.5%)	25.9 (+34.2%)	23.5 (+21.8%)	33.6 (+24.9%)	43.8 (+23.0%)	8.6 (+11.7%)	21.8 (+19.8%)	38.3 (+26.0%)
	DSGDetr [11]	6.7	14.7	19.1	23.3	29.8	36.0	6.5	16.0	30.4
	+ImpartTail (Ours)	7.5 (+11.9%)	17.8 (+21.1%)	23.7 (+24.1%)	27.5 (+18.0%)	35.2 (+18.1%)	43.3 (+20.3%)	7.3 (+12.3%)	18.4 (+15.0%)	36.6 (+20.4%)
SGCLS	STTran [7]	25.0	27.5	27.6	38.8	47.1	59.9	29.5	39.9	40.9
	+ImpartTail (Ours)	32.3 (+29.2%)	36.2 (+31.5%)	36.2 (+31.2%)	47.4 (+22.2%)	57.5 (+22.1%)	66.6 (+11.2%)	36.2 (+22.7%)	50.5 (+26.6%)	52.2 (+27.6%)
	DSGDetr [11]	25.6	28.1	28.1	39.9	49.4	64.6	30.1	40.6	41.6
	+ImpartTail (Ours)	32.2 (+25.8%)	36.0 (+28.1%)	36.0 (+28.1%)	48.8 (+22.3%)	59.6 (+20.6%)	70.1 (+8.5%)	36.8 (+22.3%)	52.4 (+29.1%)	54.9 (+32.0%)
PREDCLS	STTran [7]	30.5	34.7	34.8	45.7	63.4	80.5	36.6	51.8	53.8
	+ImpartTail (Ours)	44.0 (+44.3%)	52.7 (+51.9%)	52.9 (+52.0%)	65.5 (+43.3%)	82.0 (+29.3%)	93.0 (+15.5%)	47.7 (+30.3%)	69.7 (+34.6%)	73.4 (+36.4%)
	DSGDetr [11]	31.5	36.1	36.2	45.6	64.4	80.5	36.5	52.5	55.2
	+ImpartTail (Ours)	41.0 (+30.2%)	48.1 (+33.2%)	48.2 (+33.1%)	59.4 (+30.3%)	76.2 (+18.3%)	89.8 (+11.6%)	43.9 (+20.3%)	65.4 (+24.6%)	69.8 (+26.4%)

Results - 2: Robust Video Scene Graph Generation



$\mathcal{S}$	Mode	Corruption	Method	With Constraint			No Constraint						Semi Constraint		
				mR@10	mR@20	mR@50	R@10	R@20	R@50	mR@10	mR@20	mR@50	mR@10	mR@20	mR@50
3	SGCLS	Gaussian Noise	DSGDetr [11]	9.6	10.3	10.3	20.9	25.4	26.8	15.7	19.4	23.4	11.4	15.3	15.7
			+ImparTail (Ours)	**13.7** (+42.7%)	**14.9** (+44.7%)	**15.0** (+45.6%)	**21.0** (+0.5%)	**27.3** (+7.5%)	**30.1** (+12.3%)	**20.8** (+32.5%)	**25.4** (+30.9%)	**29.1** (+24.4%)	**15.6** (+36.8%)	**21.5** (+40.5%)	**22.2** (+41.4%)
		Fog	DSGDetr [11]	22.6	24.9	24.9	**47.8**	**58.4**	62.0	35.5	43.4	54.6	26.6	36.1	37.2
			+ImparTail (Ours)	**28.3** (+25.2%)	**31.8** (+27.7%)	**31.9** (+28.1%)	42.6 (-10.9%)	56.0 (-4.1%)	**62.1** (+0.2%)	**43.8** (+23.4%)	**53.0** (+22.1%)	**61.7** (+13.0%)	**31.8** (+19.5%)	**45.7** (+26.6%)	**48.2** (+29.6%)
		Frost	DSGDetr [11]	16.7	18.5	18.5	**34.5**	**42.3**	45.1	26.8	33.0	40.1	19.6	26.8	27.7
			+ImparTail (Ours)	**22.4** (+34.1%)	**25.0** (+35.1%)	**25.1** (+35.7%)	31.9 (-7.5%)	42.0 (-0.7%)	**47.1** (+4.4%)	**34.9** (+30.2%)	**42.3** (+28.2%)	**48.2** (+20.2%)	**25.9** (+32.1%)	**36.5** (+36.2%)	**38.4** (+38.6%)
		Brightness	DSGDetr [11]	23.6	25.7	25.7	**50.8**	**61.9**	65.5	36.8	45.4	57.5	27.6	37.5	38.6
			+ImparTail (Ours)	**29.8** (+26.3%)	**33.2** (+29.2%)	**33.2** (+29.2%)	45.5 (-10.4%)	59.9 (-3.2%)	**66.1** (+0.9%)	**45.0** (+22.3%)	**55.4** (+22.0%)	**65.3** (+13.6%)	**33.7** (+22.1%)	**48.1** (+28.3%)	**50.6** (+31.1%)
		Sun Glare	DSGDetr [11]	12.1	13.2	13.2	**26.3**	32.5	34.7	19.3	24.4	30.2	14.2	19.2	19.6
			+ImparTail (Ours)	**17.3** (+43.0%)	**19.4** (+47.0%)	**19.4** (+47.0%)	25.8 (-1.9%)	**34.3** (+5.5%)	**38.5** (+11.0%)	**26.6** (+37.8%)	**32.2** (+32.0%)	**37.3** (+23.5%)	**19.4** (+36.6%)	**27.6** (+43.7%)	**29.0** (+48.0%)
5	PREDCLS	Gaussian Noise	STTran [7]	20.0	22.3	22.4	**64.2**	**87.6**	99.0	31.4	52.5	79.7	26.0	36.6	38.5
			+ImparTail (Ours)	**37.6** (+88.0%)	**43.8** (+96.4%)	**43.9** (+96.0%)	62.5 (-2.6%)	84.6 (-3.4%)	**99.0** (0.0%)	**57.5** (+83.1%)	**77.7** (+48.0%)	**92.7** (+16.3%)	**42.2** (+62.3%)	**60.0** (+63.9%)	**62.9** (+63.4%)
		Fog	STTran [7]	26.5	30.2	30.3	**70.2**	**91.1**	**99.1**	41.6	61.0	80.5	33.2	46.8	48.7
			+ImparTail (Ours)	**42.6** (+60.8%)	**50.9** (+68.5%)	**51.1** (+68.6%)	64.8 (-7.7%)	86.3 (-5.3%)	98.8 (-0.3%)	**63.8** (+53.4%)	**80.2** (+31.5%)	**92.7** (+15.2%)	**46.2** (+39.2%)	**65.5** (+40.0%)	**68.2** (+40.0%)
		Frost	STTran [7]	25.6	29.2	29.2	**69.4**	**90.7**	**99.1**	41.0	60.9	80.5	32.7	46.1	48.0
			+ImparTail (Ours)	**41.0** (+60.2%)	**49.0** (+67.8%)	**49.2** (+68.5%)	62.2 (-10.4%)	84.3 (-7.1%)	98.5 (-0.6%)	**62.5** (+52.4%)	**78.6** (+29.1%)	**92.7** (+15.2%)	**45.1** (+37.9%)	**62.9** (+36.4%)	**65.1** (+35.6%)
		Brightness	STTran [7]	28.2	32.0	32.1	**71.3**	**91.6**	**99.2**	42.8	62.0	80.4	34.5	49.0	51.2
			+ImparTail (Ours)	**42.3** (+50.0%)	**50.4** (+57.5%)	**50.5** (+57.3%)	65.9 (-7.6%)	87.2 (-4.8%)	98.9 (-0.3%)	**64.0** (+49.5%)	**80.8** (+30.3%)	**92.8** (+15.4%)	**46.9** (+35.9%)	**67.8** (+38.4%)	**71.0** (+38.7%)
		Sun Glare	STTran [7]	22.5	25.1	25.2	**66.7**	**89.9**	**99.1**	36.7	56.7	80.0	28.9	40.5	42.3
			+ImparTail (Ours)	**40.2** (+78.7%)	**47.5** (+89.2%)	**47.7** (+89.3%)	57.9 (-13.2%)	81.7 (-9.1%)	98.0 (-1.1%)	**60.3** (+64.3%)	**77.5** (+36.7%)	**92.7** (+15.9%)	**43.3** (+49.8%)	**59.3** (+46.4%)	**61.0** (+44.2%)

Qualitative Results: Video Scene Graph Generation



Results

Settings: SGA

Action Genome Scenes (AGS)

- **Input:**
  - Frames in a video
- **Output (No Localization):**
  - Observed relationship predicates
  - Anticipated relationship predicates

Partially Grounded Action Genome Scenes (PGAGS)

- **Input:**
  - Frames in a video
  - Object bounding boxes
- **Output (No Localization):**
  - Observed relationship predicates
  - Anticipated relationship predicates

Grounded Action Genome Scenes (GAGS)

- **Input:**
  - Frames in a video
  - Object bounding boxes
  - Object labels
- **Output (No Localization):**
  - Observed relationship predicates
  - Anticipated relationship predicates

Results - 3: Scene Graph Anticipation

Table 2. Mean Recall Results for SGA.

$\mathcal{F}$	Method	AGS						PGAGS						GAGS					
		With Constraint			No Constraint			With Constraint			No Constraint			With Constraint			No Constraint		
		@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
0.5	STTran++ [37]	7.9	16.4	18.4	**13.9**	**21.3**	38.5	14.3	15.8	15.8	20.9	32.5	50.1	17.8	20.9	21.0	**25.2**	**39.4**	63.5
	+IMPARTAIL (Ours)	**9.3**	**18.7**	**20.9**	12.1	20.0	**39.6**	**20.2**	**22.0**	**22.0**	**24.0**	**34.9**	**50.5**	**19.9**	**22.7**	**22.8**	23.7	39.2	**64.1**
	DSGDetr++ [37]	7.4	13.4	14.6	11.8	18.2	36.1	15.0	16.3	16.3	19.9	32.3	50.6	17.1	20.0	20.0	23.2	37.3	62.9
	+IMPARTAIL (Ours)	**8.9**	**17.0**	**18.6**	**13.1**	**21.6**	**39.6**	**18.6**	**20.1**	**20.1**	**22.6**	**35.2**	**52.5**	**21.2**	**24.5**	**24.6**	**28.2**	**42.2**	**64.9**
	SceneSayerODE [37]	5.8	12.6	16.9	14.0	22.3	36.5	11.2	12.8	12.8	16.9	26.3	45.7	17.5	20.7	20.9	24.9	38.0	61.8
	+IMPARTAIL (Ours)	**6.8**	**16.1**	**22.0**	**15.6**	**24.8**	**39.7**	**14.5**	**16.4**	**16.4**	**22.7**	**33.6**	**49.7**	**19.3**	**23.2**	**23.5**	**26.5**	**40.9**	**63.2**
	SceneSayerSDE [37]	6.4	13.7	18.3	15.4	23.7	38.7	15.2	17.5	17.5	22.9	34.3	**51.0**	**18.2**	**21.7**	**21.8**	25.0	39.0	62.7
	+IMPARTAIL (Ours)	**7.4**	**19.1**	**27.7**	**21.8**	**31.4**	**45.4**	**15.7**	**17.9**	**17.9**	**23.6**	**34.3**	50.6	17.8	21.2	21.4	**27.0**	**40.7**	**63.6**
0.7	STTran++ [37]	9.1	18.2	20.2	**15.7**	**23.7**	41.9	17.2	18.6	18.6	25.3	38.3	56.1	21.9	25.0	25.0	**31.2**	47.0	75.4
	+IMPARTAIL (Ours)	**10.9**	**21.9**	**24.1**	14.0	23.2	**43.7**	**21.0**	**22.7**	**22.7**	**28.0**	**41.7**	**57.1**	**25.8**	**29.1**	**29.1**	31.1	**49.2**	**76.5**
	DSGDetr++ [37]	8.4	14.8	16.0	13.2	20.0	38.8	18.1	19.4	19.4	24.8	39.5	57.3	20.8	23.8	23.8	28.6	46.1	73.8
	+IMPARTAIL (Ours)	**10.5**	**19.5**	**21.2**	**14.9**	**24.8**	**43.9**	**20.6**	**21.8**	**21.8**	**26.3**	**41.0**	**58.1**	**28.3**	**32.5**	**32.5**	**31.4**	**49.7**	**75.7**
	SceneSayerODE [37]	6.7	**14.0**	**18.5**	16.4	24.9	40.5	13.6	15.1	15.1	20.5	32.4	52.8	20.7	24.0	24.0	29.8	45.2	72.0
	+IMPARTAIL (Ours)	**6.8**	13.9	18.2	**17.5**	**25.8**	**41.1**	**22.2**	**25.6**	**25.7**	**30.7**	**43.9**	**55.9**	**23.2**	**27.5**	**27.5**	**31.7**	**49.9**	**73.8**
	SceneSayerSDE [37]	7.1	14.6	19.3	17.3	26.1	42.5	17.9	19.9	19.9	27.0	40.2	57.2	**21.0**	**24.6**	**24.6**	30.2	45.4	72.8
	+IMPARTAIL (Ours)	**8.6**	**21.3**	**29.3**	**25.6**	**35.1**	**50.0**	**25.9**	**30.0**	**30.1**	**35.5**	**48.2**	**58.5**	20.9	24.4	24.4	**31.6**	**47.9**	**73.4**

Results - 4: Robust Scene Graph Anticipation

Table 4. Robustness Evaluation Results for SGA.

$\mathcal{F}$	Mode	Corruption	Method	With Constraint		
				mR@10	**mR@20**	**mR@50**
0.5	PGAGS	Gaussian Noise	STTran+ [37]	**7.9**	**8.4**	**8.4**
			+ImparTail (Ours)	5.1 (-35.4%)	5.4 (-35.7%)	5.4 (-35.7%)
			DSGDetr+ [37]	5.5	5.8	5.8
			+ImparTail (Ours)	**7.5** (+36.4%)	**8.0** (+37.9%)	**8.0** (+37.9%)
			STTran++ [37]	5.9	6.4	6.4
			+ImparTail (Ours)	**9.4** (+59.3%)	**10.2** (+59.4%)	**10.2** (+59.4%)
			DSGDetr++ [37]	5.7	6.1	6.1
			+ImparTail (Ours)	**8.3** (+45.6%)	**8.8** (+44.3%)	**8.8** (+44.3%)
		Frost	STTran+ [37]	8.2	**9.0**	**9.0**
			+ImparTail (Ours)	**8.3** (+1.2%)	8.7 (-3.3%)	8.7 (-3.3%)
			DSGDetr+ [37]	9.0	9.9	9.9
			+ImparTail (Ours)	**12.4** (+37.8%)	**13.3** (+34.3%)	**13.3** (+34.3%)
			STTran++ [37]	9.6	10.5	10.5
			+ImparTail (Ours)	**13.8** (+43.7%)	**15.3** (+45.7%)	**15.3** (+45.7%)
			DSGDetr++ [37]	9.9	10.8	10.8
			+ImparTail (Ours)	**13.2** (+33.3%)	**14.4** (+33.3%)	**14.4** (+33.3%)
		Brightness	STTran+ [37]	11.0	12.1	12.1
			+ImparTail (Ours)	**11.7** (+6.4%)	**12.5** (+3.3%)	**12.5** (+3.3%)
			DSGDetr+ [37]	12.2	13.5	13.5
			+ImparTail (Ours)	**15.9** (+30.3%)	**17.2** (+27.4%)	**17.2** (+27.4%)
			STTran++ [37]	12.8	14.1	14.1
			+ImparTail (Ours)	**17.7** (+38.3%)	**19.3** (+36.9%)	**19.3** (+36.9%)
			DSGDetr++ [37]	13.6	14.7	14.7
			+ImparTail (Ours)	**16.6** (+22.1%)	**18.1** (+23.1%)	**18.1** (+23.1%)