ST-HOI: A Spatial-Temporal Baseline for Human-Object Interaction Detection in Videos

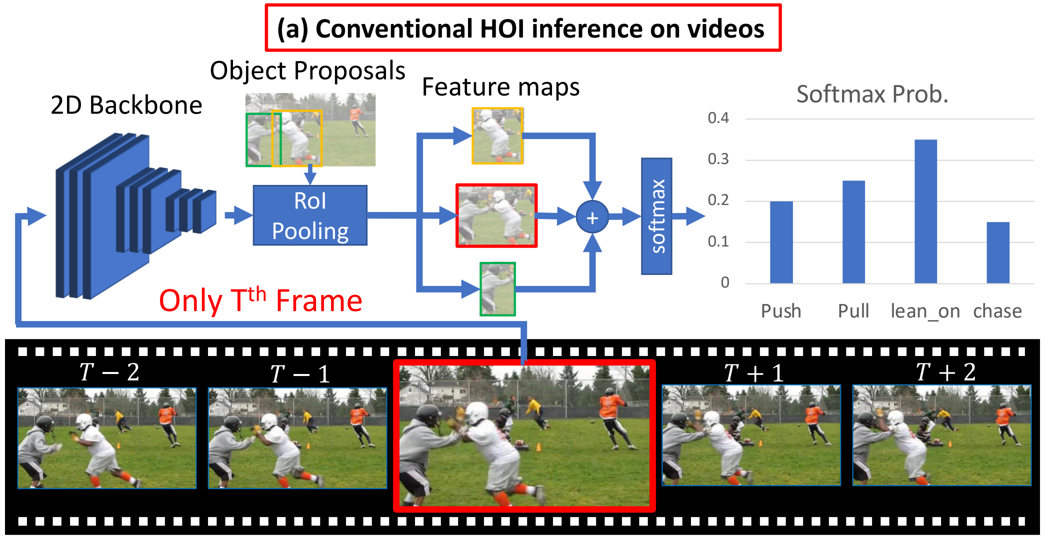
Meng-Jiun Chiou[1], Chun-Yu Liao[2], Li-Wei Wang[2], Roger Zimmermann[1] and Jiashi Feng[1]

[1]National University of Singapore [2]ASUS AICS Department

In ACM ICMR 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval

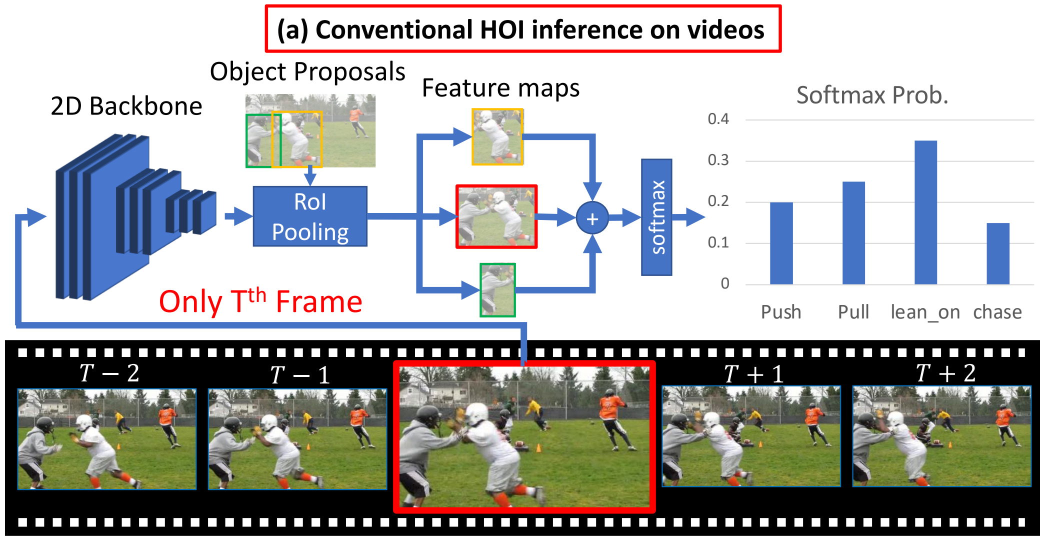
Motivation I- HOI in Videos

- HOI is defined as a relationship between a subject (human) and an object (any class). Can be action or spatial predicate



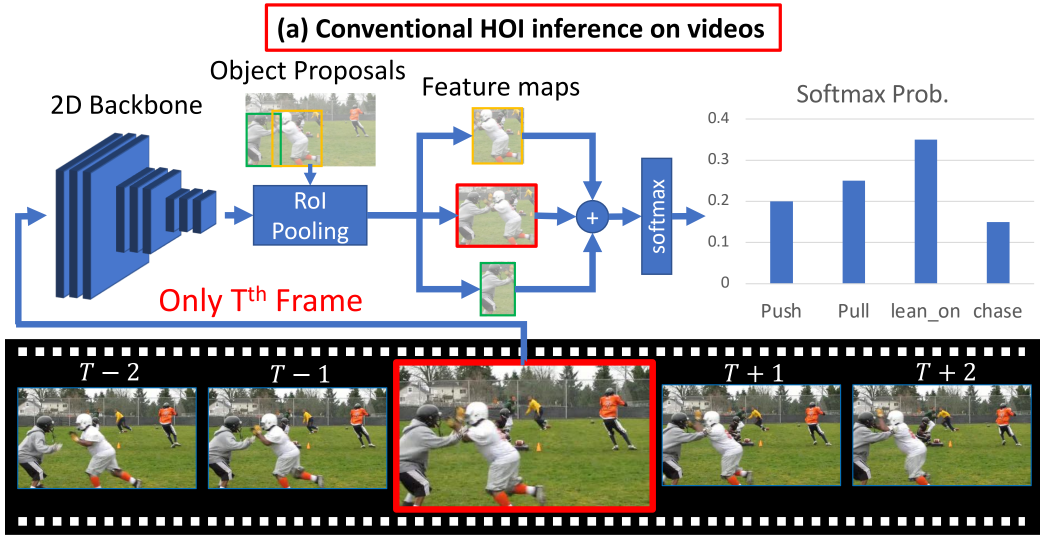
Motivation I - HOI in Videos

- HOI is defined as a relationship between a subject (human) and an object (any class). Can be action or spatial predicate
- Temporal-aware HOIs (*e.g.,* push, pull, open, close) have been predicted **without temporal contexts** in prior work.
  - It is unlikely for both humans and machines to guess from a single video frame that a person is ”opening” or ”closing” a door, where neighboring frames play an essential role.



Motivation I - HOI in Videos

- HOI is defined as a relationship between a subject (human) and an object (any class). Can be action or spatial predicate
- Temporal-aware HOIs (*e.g.*, push, pull, open, close) have been predicted **without temporal contexts** in prior work.
  - It is unlikely for both humans and machines to guess from a single video frame that a person is ”opening” or ”closing” a door, where neighboring frames play an essential role.
- A possible reason for relatively under-explored video HOI is **the lack of dataset and its corresponding setting**



Proposed Method I- VideoHOI

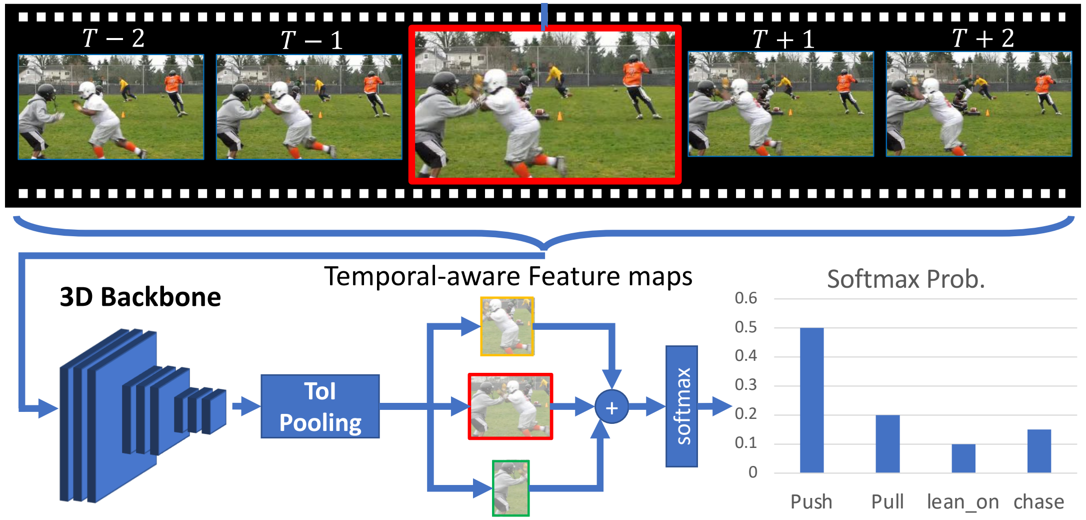
- We establish a benchmark named **VidHOI** (from VidOR), in which we follow the common protocol in video tasks to use a keyframe-centered strategy, where evaluation keyframes are sampled from testing videos with 1-Hz frequency
- With VidHOI we urge the use of video data to predict **VideoHOI**

Motivation II – Preliminary Experiment

- In spatial-temporal action detection (STAD), a popular baseline is to use 3D-CNN to extract person's feature followed by classification. This is similar to HOI methods (*i.e.,*"2D baseline") and differs only in the absence of object features & the 3D backbone.

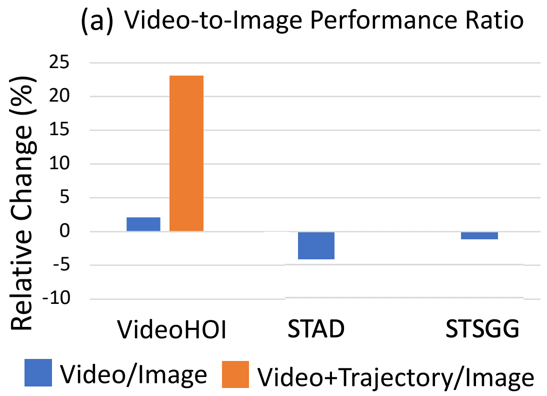
Motivation II – Preliminary Experiment

- In spatial-temporal action detection (STAD), a popular baseline is to use 3D-CNN to extract person's feature followed by classification. This is similar to HOI methods (*i.e.*,"2D baseline") and differs only in the absence of object features & the 3D backbone.
- We thus did a preliminary experiment to make it consider object features as well (*i.e.*, "3D baseline").



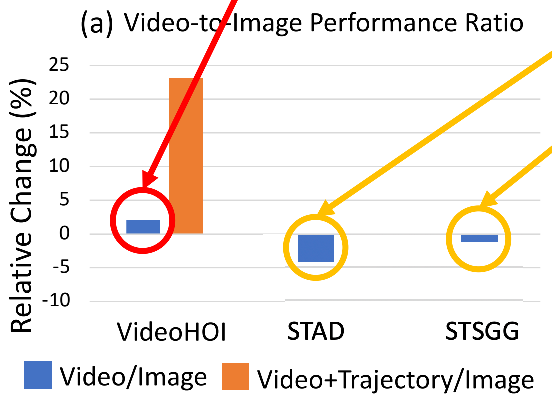
Motivation II – Feature Inconsistency Problem

- However, we found that 3D baseline does not outperform 2D baseline significantly (only ~2%). Worse results have been found in STAD and STSGG literature showing 3D backbones are harmful.



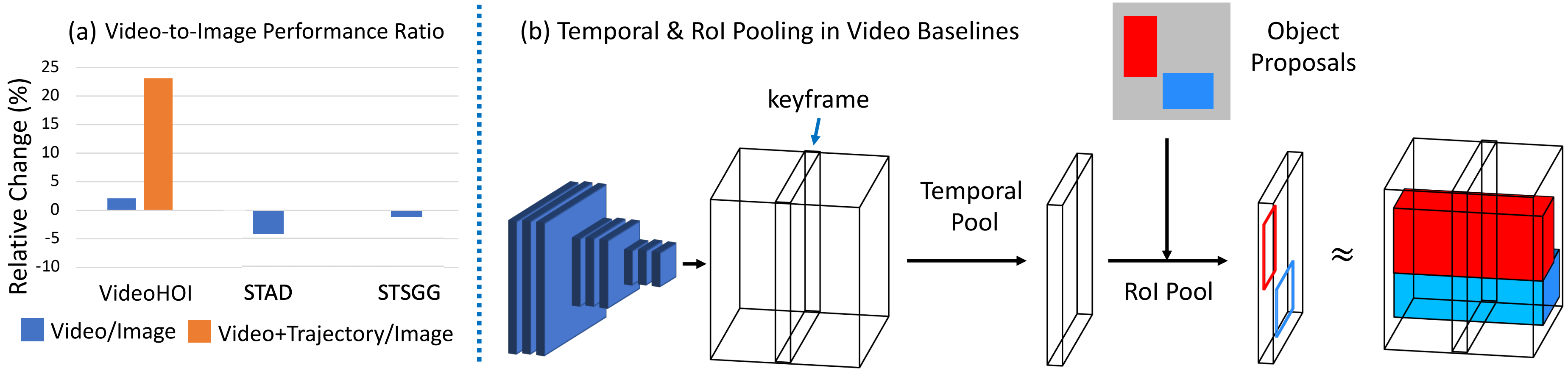
Motivation II – Feature Inconsistency Problem

- However, we found that 3D baseline does not outperform 2D baseline significantly (only ~2%). Worse results have been found in STAD and STSGG literature showing 3D backbones are harmful.



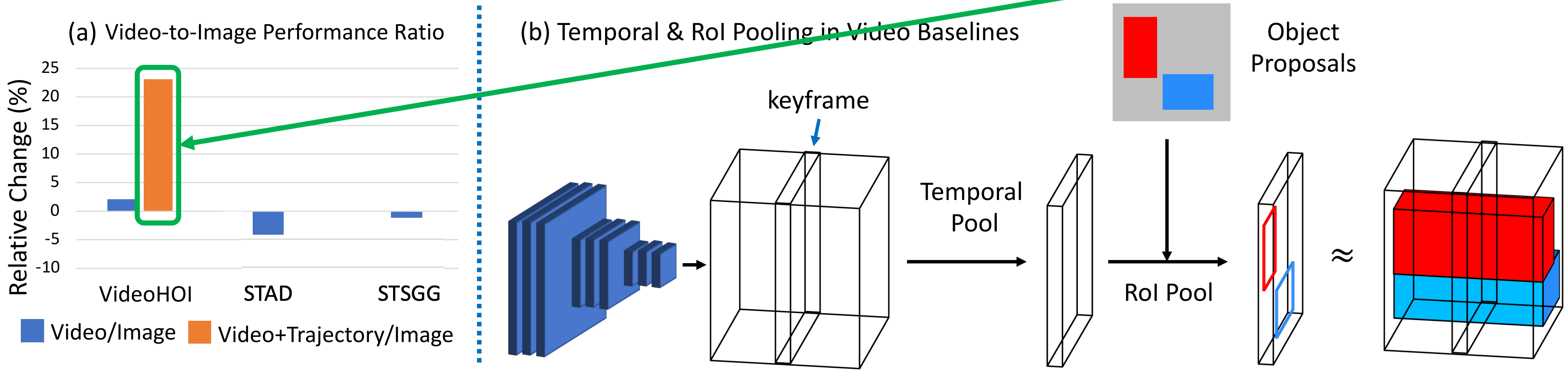
Motivation II – Feature Inconsistency Problem

- However, we found that 3D baseline does not outperform 2D baseline significantly (only ~2%). Worse results have been found in STAD and STSGG literature showing 3D backbones are harmful.
- We probed the reason and found that Temporal-RoI pooling does not work correctly by cropping feature of the same region through the video segment (cuboid). This does not consider the way objects move



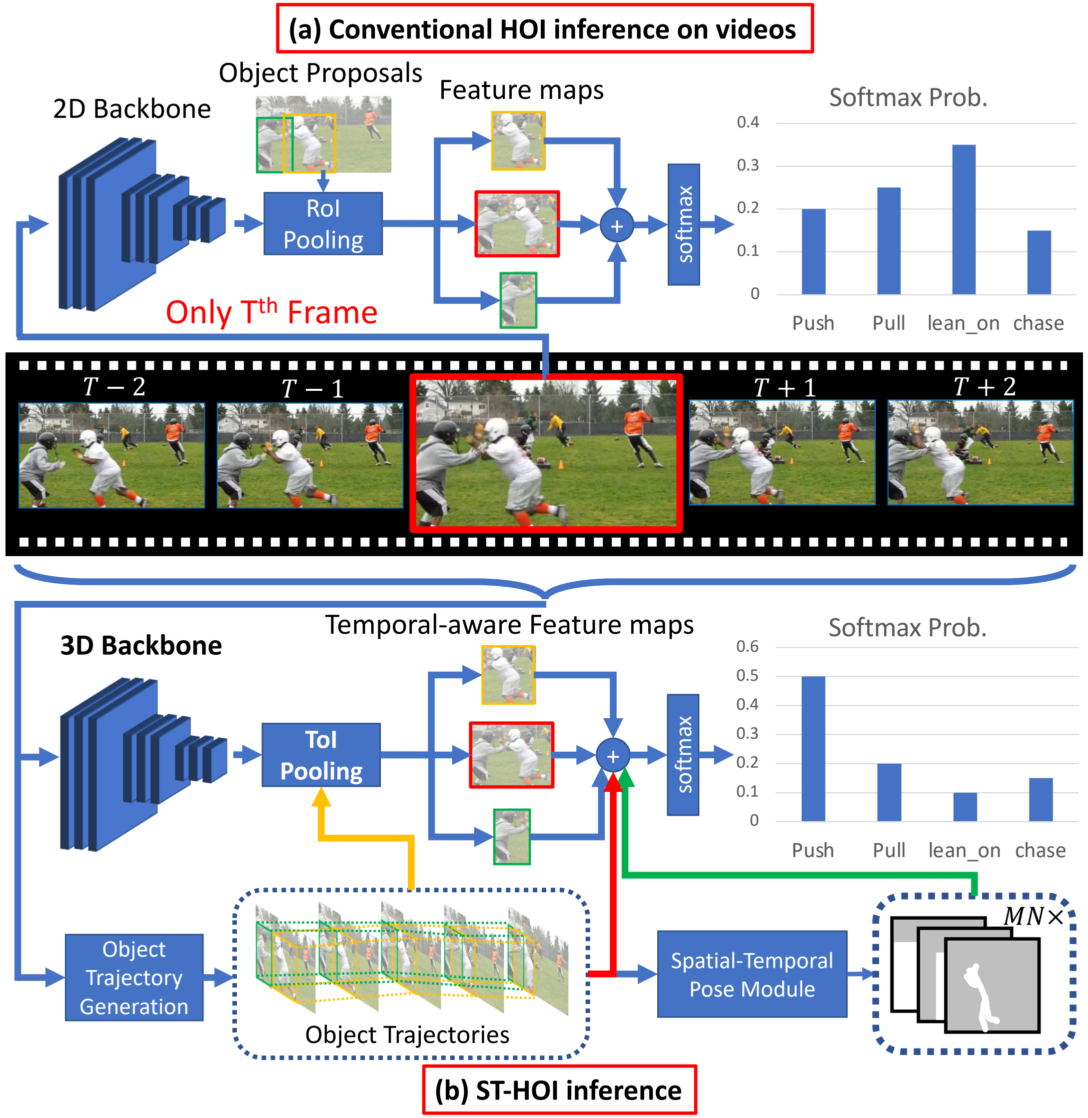
Motivation II – Feature Inconsistency Problem

- However, we found that 3D baseline does not outperform 2D baseline significantly (only ~2%). Worse results have been found in STAD and STSGG literature showing 3D backbones are harmful.
- We probed the reason and found that Temporal-RoI pooling does not work correctly by cropping feature of the same region through the video segment (cuboid). This does not consider the way objects move
- We try to recover this missing information by appending trajectory to the subject/object visual feature and achieve a ~23% improvement

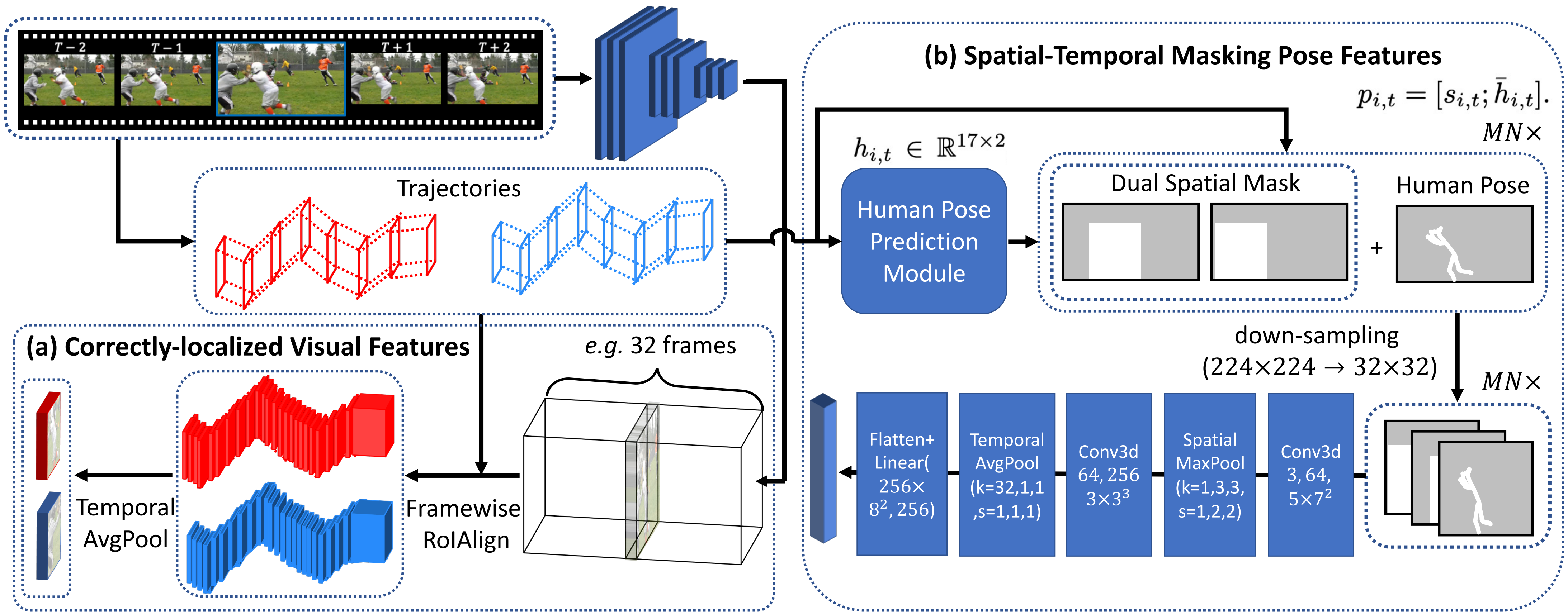


Proposed Method II – Trajectory-based Feature

- We propose **ST-HOI** with three trajectory-based spatial-temporal features:
  - Correctly-localized Visual Feature
  - Spatial-Temporal Masking Pose Feature
  - Trajectory Feature



Trajectory-based Spatial-Temporal Features



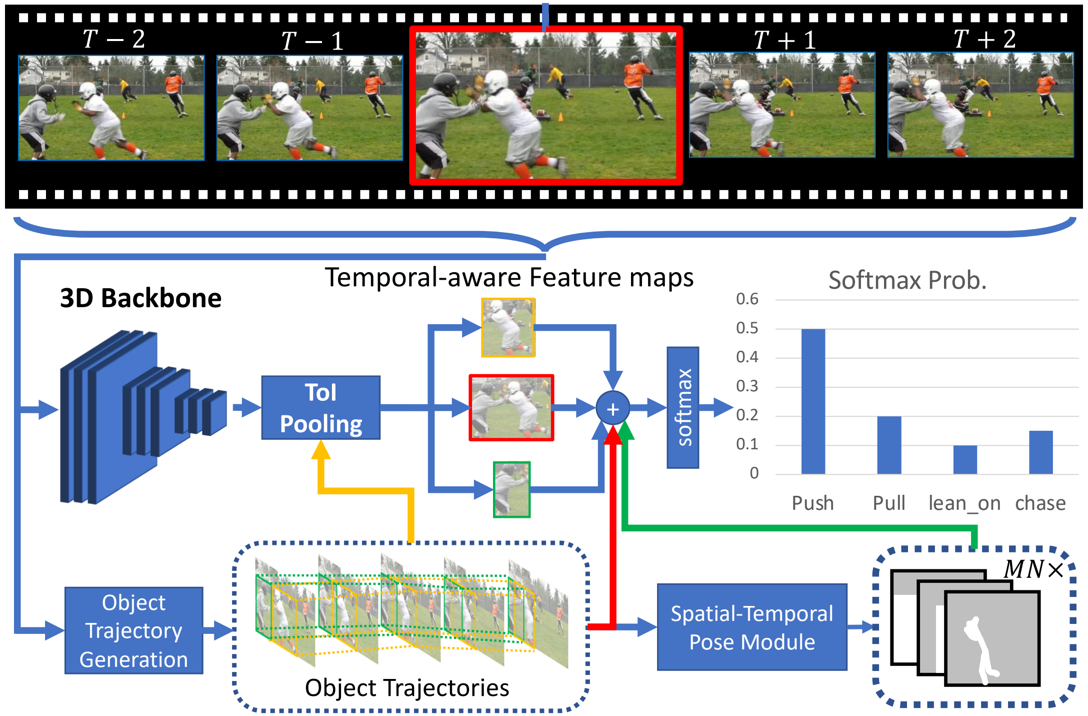
$$\bar{v}_i = \frac{1}{T}\sum_{t=1}^{T} \mathrm{RoIAlign}(v_t, j_{i,t}),$$

Prediction and Training

- We simply concatenate all features

$$v_{\mathrm{so}} = [\bar{v}_s; \bar{v}_u; \bar{v}_o; j_s; j_o; \bar{p}_{so}],$$

- A multilabel problem -> train with binary cross entropy loss
- Two modes during testing:
  - *Oracle* uses GT boxes for test set
  - *Detection* uses predicted boxes
- We use pretrained pose estimation model (FastPose)



(b) ST-HOI inference

Dataset

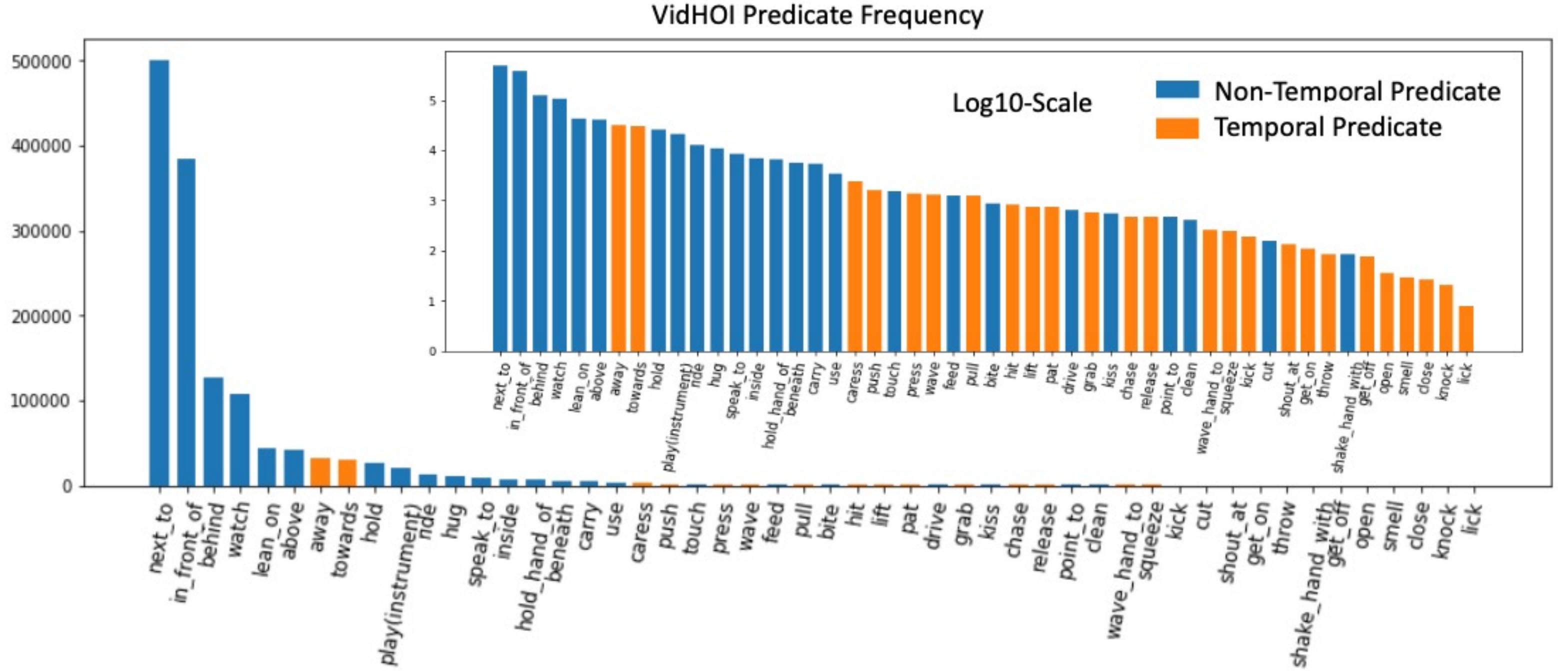
- Keyframe-centered evaluation strategy: test frames sampled in 1 fps
- 78 object classes and 50 predicates
- 557 (Full) HOI classes including 315 (Rare) or 242 (Non-rare)

Table 1: A comparison of our benchmark VidHOI with existing STAD (AVA [11]), image-based (HICO-DET [3] and V-COCO [12]) and video-based (CAD-120 [21] and Action Genome [20]) HOI datasets. VidHOI is the only dataset that provides temporal information from video clips and complete multi-person and interacting-object annotations. VidHOI also provides the most annotated keyframes and defines the most HOI categories in the existing video datasets. †Two less categories as we combine `adult`, `child` and `baby` into a single category, `person`.

Dataset	Video dataset?	Localized object?	Video hours	# Videos	# Annotated images/frames	# Objects categories	# Predicate categories	# HOI categories	# HOI Instances
HICO-DET [3]	✗	✓	-	-	47K	80	117	600	150K
V-COCO [12]	✗	✓	-	-	10K	80	25	259	16K
AVA [11]	✓	✗	108	437	3.7M	-	49	80	1.6M
CAD-120 [21]	✓	✓	0.57	0.5K	61K	13	6	10	32K
Action Genome [20]	✓	△	82	10K	234K	35	25	157	1.7M
VidHOI	✓	✓	70	7122	**7.3M**	78†	50	**557**	755K

Evaluation Metrics

- Mean Average Precision w.r.t. class frequencies: (a) Full, (b) Non-rare and (c) rare
- Mean Average Precision w.r.t. modalities: (a) Temporal and (b) Spatial



Quantitative Results I

Table 2: Results of the baselines and our ST-HOI on Vid-HOI validation set (numbers in mAP). There are two evaluation modes: `Detection` and `Oracle`, which differ only in the use of predicted or ground truth trajectories during inference. T: Trajectory features. V: Correctly-localized visual features. P: Spatial-temporal masking pose features. "%" means the full mAP change compared to the 2D model.

	Model	Full	Non-rare	Rare	%
Oracle	2D model [39]	14.1	22.9	11.3	-
	3D model	14.4	23.0	12.6	2.1
	Ours-T	17.3	26.9	16.8	22.7
	Ours-T+V	17.3	26.9	16.3	22.7
	Ours-T+P	17.4	27.1	16.4	23.4
	Ours-T+V+P	**17.6**	**27.2**	**17.3**	**24.8**
Detection	2D model [39]	2.6	4.7	1.7	-
	3D model	2.6	4.9	1.9	0.0
	Ours-T	3.0	5.5	2.0	15.4
	Ours-T+V	3.1	5.8	2.0	19.2
	Ours-T+P	**3.2**	**6.1**	2.0	**23.1**
	Ours-T+V+P	3.1	5.9	**2.1**	19.2

Quantitative Results I

Table 2: Results of the baselines and our ST-HOI on VidHOI validation set (numbers in mAP). There are two evaluation modes: `Detection` and `Oracle`, which differ only in the use of predicted or ground truth trajectories during inference. T: Trajectory features. V: Correctly-localized visual features. P: Spatial-temporal masking pose features. "%" means the full mAP change compared to the 2D model.

	Model	Full	Non-rare	Rare	%
Oracle	2D model [39]	14.1	22.9	11.3	-
	3D model	14.4	23.0	12.6	2.1
	Ours-T	17.3	26.9	16.8	22.7
	Ours-T+V	17.3	26.9	16.3	22.7
	Ours-T+P	17.4	27.1	16.4	23.4
	Ours-T+V+P	**17.6**	**27.2**	**17.3**	**24.8**
Detection	2D model [39]	2.6	4.7	1.7	-
	3D model	2.6	4.9	1.9	0.0
	Ours-T	3.0	5.5	2.0	15.4
	Ours-T+V	3.1	5.8	2.0	19.2
	Ours-T+P	**3.2**	**6.1**	2.0	**23.1**
	Ours-T+V+P	3.1	5.9	**2.1**	19.2

Trajectory is very useful

Quantitative Results I

Table 2: Results of the baselines and our ST-HOI on VidHOI validation set (numbers in mAP). There are two evaluation modes: `Detection` and `Oracle`, which differ only in the use of predicted or ground truth trajectories during inference. T: Trajectory features. V: Correctly-localized visual features. P: Spatial-temporal masking pose features. "%" means the full mAP change compared to the 2D model.

	Model	Full	Non-rare	Rare	%
Oracle	2D model [39]	14.1	22.9	11.3	-
	3D model	14.4	23.0	12.6	2.1
	Ours-T	17.3	26.9	16.8	22.7
	Ours-T+V	17.3	26.9	16.3	22.7
	Ours-T+P	17.4	27.1	16.4	23.4
	Ours-T+V+P	**17.6**	**27.2**	**17.3**	**24.8**
Detection	2D model [39]	2.6	4.7	1.7	-
	3D model	2.6	4.9	1.9	0.0
	Ours-T	3.0	5.5	2.0	15.4
	Ours-T+V	3.1	5.8	2.0	19.2
	Ours-T+P	**3.2**	**6.1**	2.0	**23.1**
	Ours-T+V+P	3.1	5.9	**2.1**	19.2

Full model gets the highest performance in Oracle mode

Performance improvement saturates when adding V/P feats

Quantitative Results I

Table 2: Results of the baselines and our ST-HOI on VidHOI validation set (numbers in mAP). There are two evaluation modes: `Detection` and `Oracle`, which differ only in the use of predicted or ground truth trajectories during inference. T: Trajectory features. V: Correctly-localized visual features. P: Spatial-temporal masking pose features. "%" means the full mAP change compared to the 2D model.

	Model	Full	Non-rare	Rare	%
Oracle	2D model [39]	14.1	22.9	11.3	-
	3D model	14.4	23.0	12.6	2.1
	Ours-T	17.3	26.9	16.8	22.7
	Ours-T+V	17.3	26.9	16.3	22.7
	Ours-T+P	17.4	27.1	16.4	23.4
	Ours-T+V+P	**17.6**	**27.2**	**17.3**	**24.8**
Detection	2D model [39]	2.6	4.7	1.7	-
	3D model	2.6	4.9	1.9	0.0
	Ours-T	3.0	5.5	2.0	15.4
	Ours-T+V	3.1	5.8	2.0	19.2
	Ours-T+P	**3.2**	**6.1**	2.0	**23.1**
	Ours-T+V+P	3.1	5.9	**2.1**	19.2

The ground truth trajectories (T) may have provided enough ”correctly-localized” spatial-temporal information.

Quantitative Results I

Table 2: Results of the baselines and our ST-HOI on VidHOI validation set (numbers in mAP). There are two evaluation modes: `Detection` and `Oracle`, which differ only in the use of predicted or ground truth trajectories during inference. T: Trajectory features. V: Correctly-localized visual features. P: Spatial-temporal masking pose features. "%" means the full mAP change compared to the 2D model.

	Model	Full	Non-rare	Rare	%
Oracle	2D model [39]	14.1	22.9	11.3	-
	3D model	14.4	23.0	12.6	2.1
	Ours-T	17.3	26.9	16.8	22.7
	Ours-T+V	17.3	26.9	16.3	22.7
	Ours-T+P	17.4	27.1	16.4	23.4
	Ours-T+V+P	**17.6**	**27.2**	**17.3**	**24.8**
Detection	2D model [39]	2.6	4.7	1.7	-
	3D model	2.6	4.9	1.9	0.0
	Ours-T	3.0	5.5	2.0	15.4
	Ours-T+V	3.1	5.8	2.0	19.2
	Ours-T+P	**3.2**	**6.1**	2.0	**23.1**
	Ours-T+V+P	3.1	5.9	**2.1**	19.2

Strong long-tail effect (but natural)

Quantitative Results II

- Under most of circumstances naively replacing 2D backbones with 3D ones doesn't help VideoHOI detection
- Again, both temporal predicates (*e.g.* towards, away, pull) and spatial (next to, behind, beneath) predicates benefit from the additional temporal-aware features

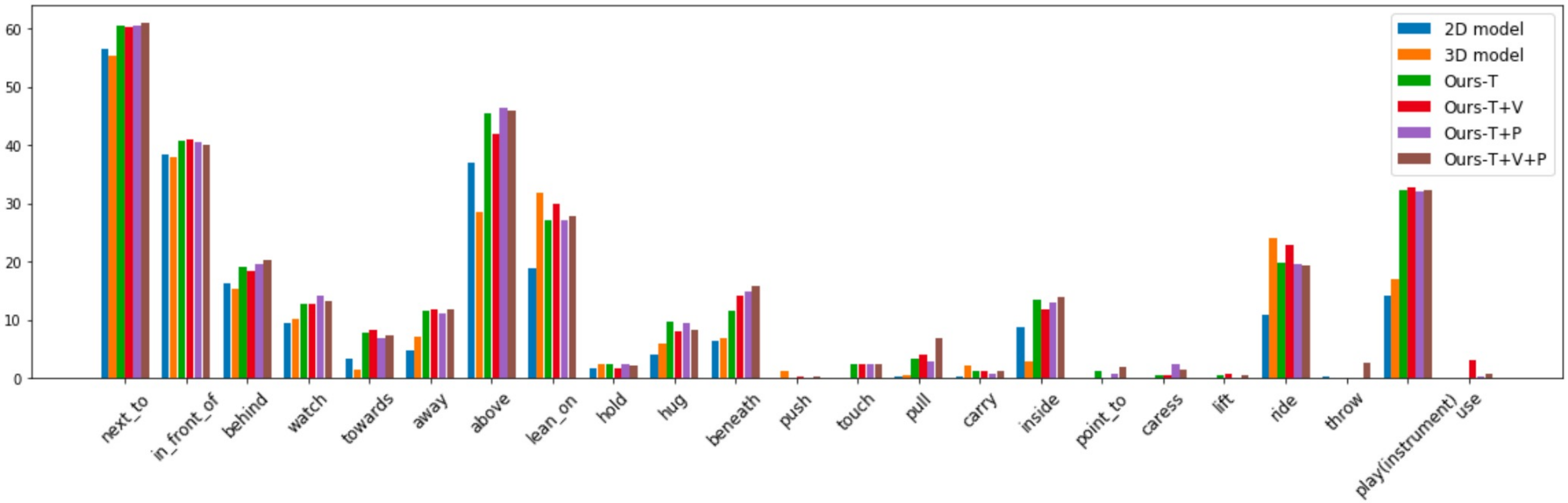


Figure 4. Performance comparison in predicate-wise mAP (pmAP). The performance boost after adding trajectory features is observed for most of the predicates. Interestingly, both spatial (*e.g.* `next to`, `behind`, `beneath`) and temporal (*e.g.* `towards`, `away`, `pull`) predicates benefit from the temporal-aware features. Predicates are sorted by the number of occurrence. Models are in `Oracle` mode.

Quantitative Results III

Temporal-predicates are helped a lot with our proposed model, in sharp contrast to 2D/3D baselines

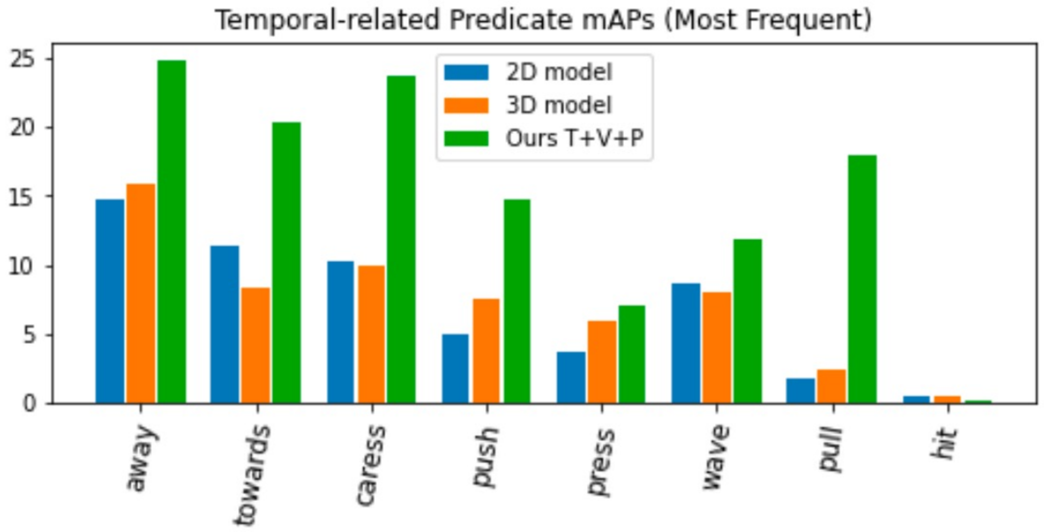


Table 3: Results of temporal-related and spatial (non-temporal) related triplet mAP. T%/S% means relative temporal/spatial mAP change compared to 2D model [39].

		Temporal	T%	Spatial	S%
Oracle	2D model [39]	8.3	-	18.6	-
	3D model	7.7	-7.2	20.9	12.3
	Ours-T	**14.4**	**73.5**	24.7	32.8
	Ours-T+V	13.6	63.9	24.6	32.3
	Ours-T+P	12.9	55.4	**25.0**	**34.4**
	Ours-T+V+P	**14.4**	**73.5**	**25.0**	**34.4**
Detection	2D model [39]	1.5	-	2.7	-
	3D model	1.6	6.7	2.9	7.4
	Ours-T	1.8	20.0	**3.3**	**23.6**
	Ours-T+V	1.8	20.0	**3.3**	**23.6**
	Ours-T+P	1.8	20.0	**3.3**	**23.6**
	Ours-T+V+P	**1.9**	**26.7**	**3.3**	**23.6**

Quantitative Results III

Trajectories are especially helpful for temporal-related predicates

Temporal-predicates are helped a lot with our proposed model, in sharp contrast to 2D/3D baselines

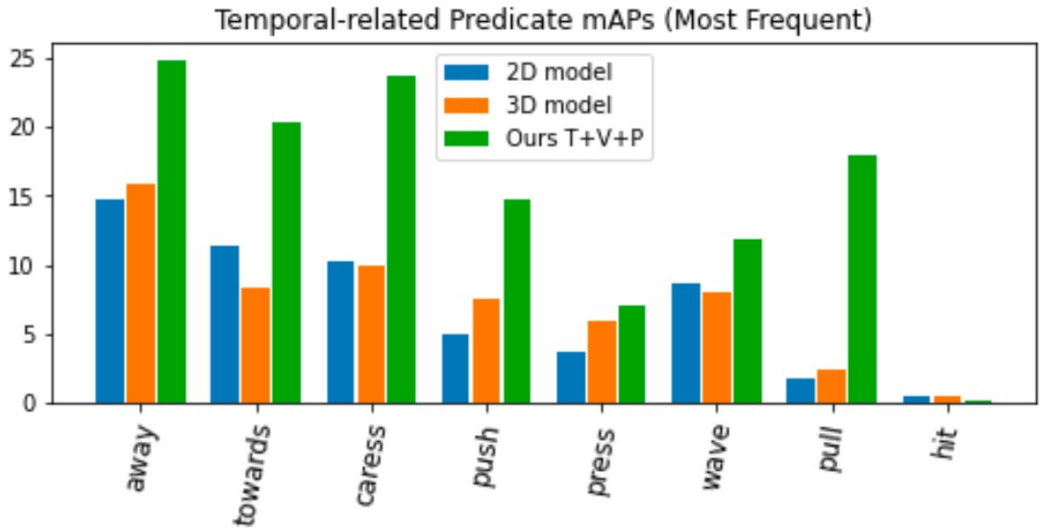
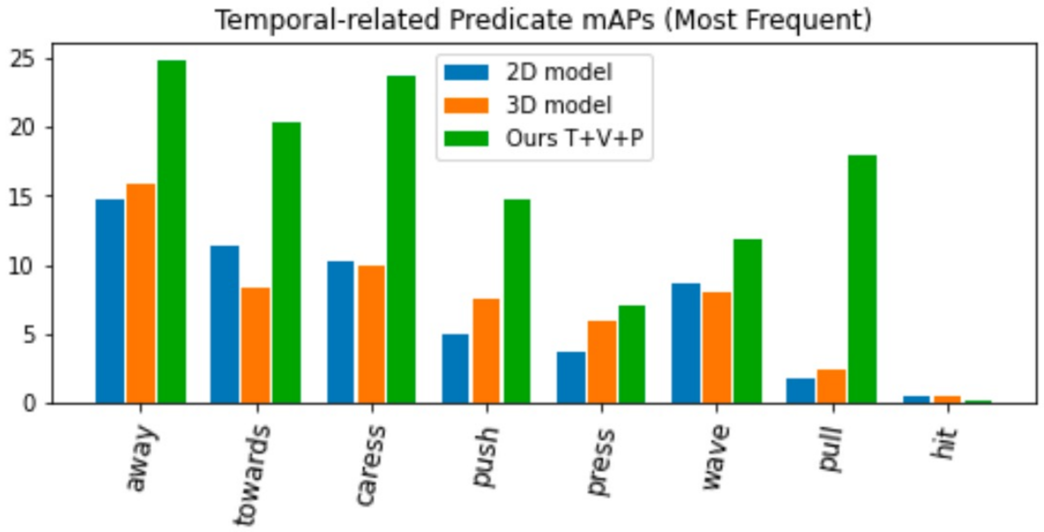


Table 3: Results of temporal-related and spatial (non-temporal) related triplet mAP. T%/S% means relative temporal/spatial mAP change compared to 2D model [39].

		Temporal	T%	Spatial	S%
Oracle	2D model [39]	8.3	-	18.6	-
	3D model	7.7	-7.2	20.9	12.3
	Ours-T	**14.4**	**73.5**	24.7	32.8
	Ours-T+V	13.6	63.9	24.6	32.3
	Ours-T+P	12.9	55.4	**25.0**	**34.4**
	Ours-T+V+P	**14.4**	**73.5**	**25.0**	**34.4**
Detection	2D model [39]	1.5	-	2.7	-
	3D model	1.6	6.7	2.9	7.4
	Ours-T	1.8	20.0	**3.3**	**23.6**
	Ours-T+V	1.8	20.0	**3.3**	**23.6**
	Ours-T+P	1.8	20.0	**3.3**	**23.6**
	Ours-T+V+P	**1.9**	**26.7**	**3.3**	**23.6**

Quantitative Results III

Temporal-predicates are helped a lot with our proposed model, in sharp contrast to 2D/3D baselines



Full model gets the highest performance

Table 3: Results of temporal-related and spatial (non-temporal) related triplet mAP. T%/S% means relative temporal/spatial mAP change compared to 2D model [39].

		Temporal	T%	Spatial	S%
Oracle	2D model [39]	8.3	-	18.6	-
	3D model	7.7	-7.2	20.9	12.3
	Ours-T	**14.4**	**73.5**	24.7	32.8
	Ours-T+V	13.6	63.9	24.6	32.3
	Ours-T+P	12.9	55.4	**25.0**	**34.4**
	Ours-T+V+P	**14.4**	**73.5**	**25.0**	**34.4**
Detection	2D model [39]	1.5	-	2.7	-
	3D model	1.6	6.7	2.9	7.4
	Ours-T	1.8	20.0	**3.3**	**23.6**
	Ours-T+V	1.8	20.0	**3.3**	**23.6**
	Ours-T+P	1.8	20.0	**3.3**	**23.6**
	Ours-T+V+P	**1.9**	**26.7**	**3.3**	**23.6**

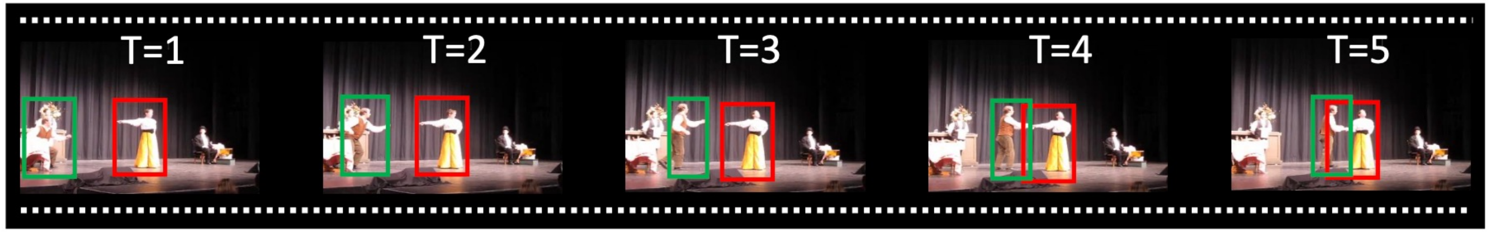
Qualitative Results

- Compared to the 2D baseline, our model predicts more accurate HOIs (*e.g. hold_hand_of* in T4 and T5 of the upper example and *lift* in T1 of the lower example).
- ST-HOI also produces less false positives in both examples.

?

	2D-baseline					ST-HOI Full				
	T=1	T=2	T=3	T=4	T=5	T=1	T=2	T=3	T=4	T=5
next_to	O	O	O	O	O	O	O	O	O	O
watch	O	O	O	O	O	O	O	O	O	O
towards	O	O	O	O	O	O	O	O	O	O
hold_hand_of	-	-		X	X	-	-	-	O	O
in_front_of	O	O	O	O	O	O	O	O	O	O
behind	O	O	O	O	O	O	O	O	O	O
hold	-	O	-	-	-	-	-	-	-	-
lean_on	-	O	-	-	-	-	-	-	-	-
hug	-	O	-	-	-	-	-	-	-	-
away	-	-	O	-	O	-	O	O	O	O

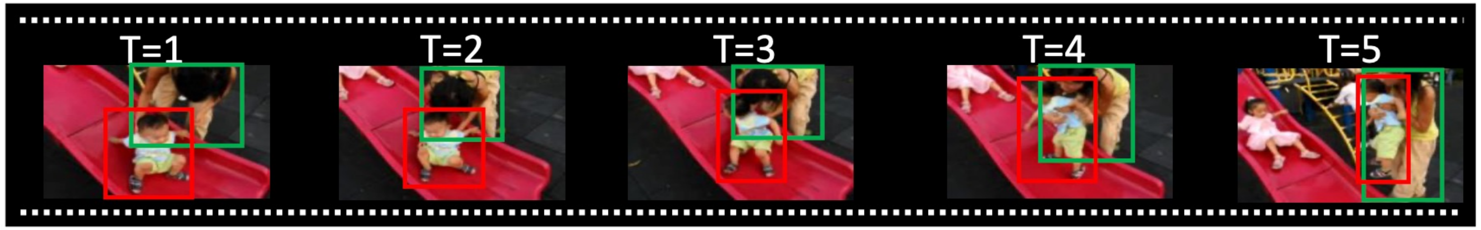
O: TP
O: FP
X: FN
- : TN



?

	2D-baseline					ST-HOI Full				
	T=1	T=2	T=3	T=4	T=5	T=1	T=2	T=3	T=4	T=5
next_to	O	O	O	O	O	O	O	O	O	O
behind	O	O	O	O	O	O	O	O	O	O
lift	X	X	X	X	-	O	X	X	X	-
in_front_of	O	O	O	O	O	O	O	O	O	O
hug	-	O	X	O	O	O	O	X	O	O
above	O	O	-	O	O	O	-	-	-	-
watch	O	O	O	O	O	O	O	O	O	O
hold	-	O	-	O	O	-	O	-	O	O
lean_on	O	O	-	O	O	O	-	-	-	-

O: TP
O: FP
X: FN
- : TN



Conclusion

- In this work, we addressed the inability of conventional HOI approaches to recognize temporal-aware HOIs by re-focusing on neighboring video frames
- We discussed the existing problems in conventional VideoHOI:
  - the lack of a suitable setting and dataset;
  - feature-inconsistency problem due to the improper order of RoI/temporal pooling
- We established a video HOI benchmark **VidHOI**. We then proposed a spatial-temporal baseline **ST-HOI** which exploits trajectory-based temporal features
- We showed that our model provides a huge performance boost compared to both the 2D and 3D baselines and is effective in differentiating temporal-related HOIs.

Thank you for your attention! ☺

Code and dataset available at https://github.com/coldmanck/VidHOI