Data-Efficient Learning On Structured Output Data

Raghav Goyal

Advised by Prof. Leonid Sigal

Committee members: Mark Schmidt and Kwang Moo Yi

External examiner: Dima Damen

University examiners: Purang Abolmaesumi and Michiel van de Panne

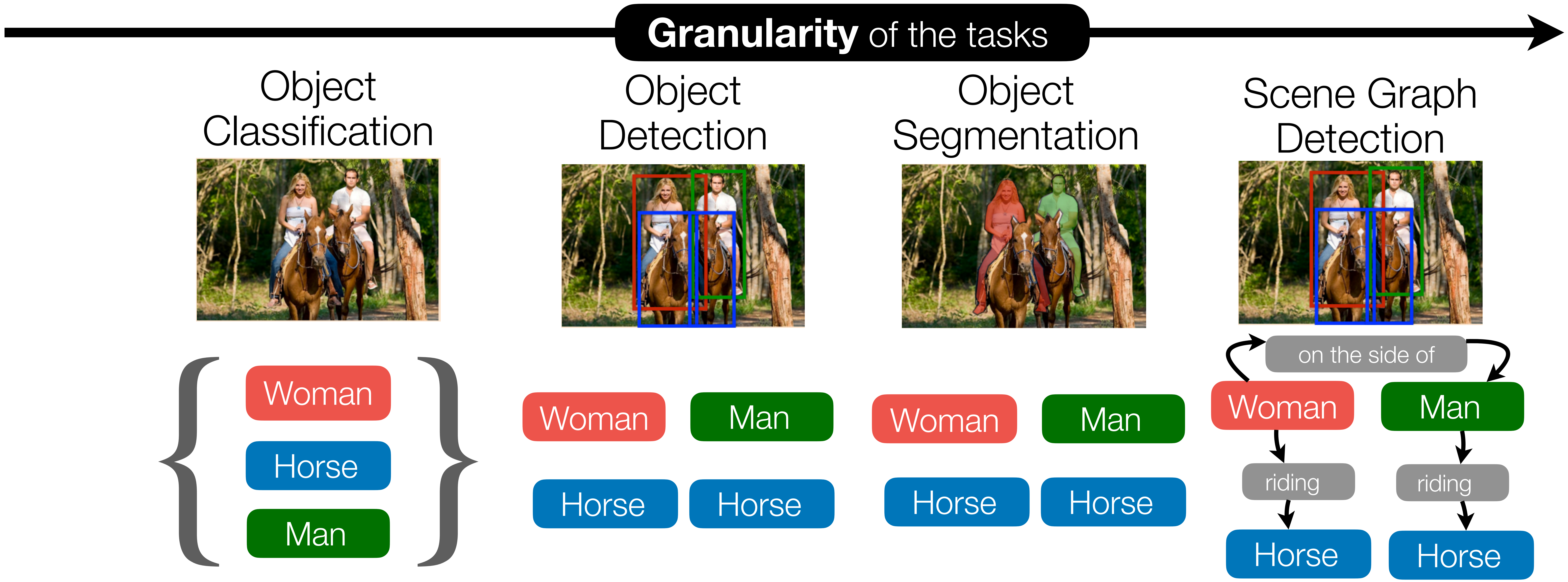
Chair: Christoph Ortner

Nov 19, 2024

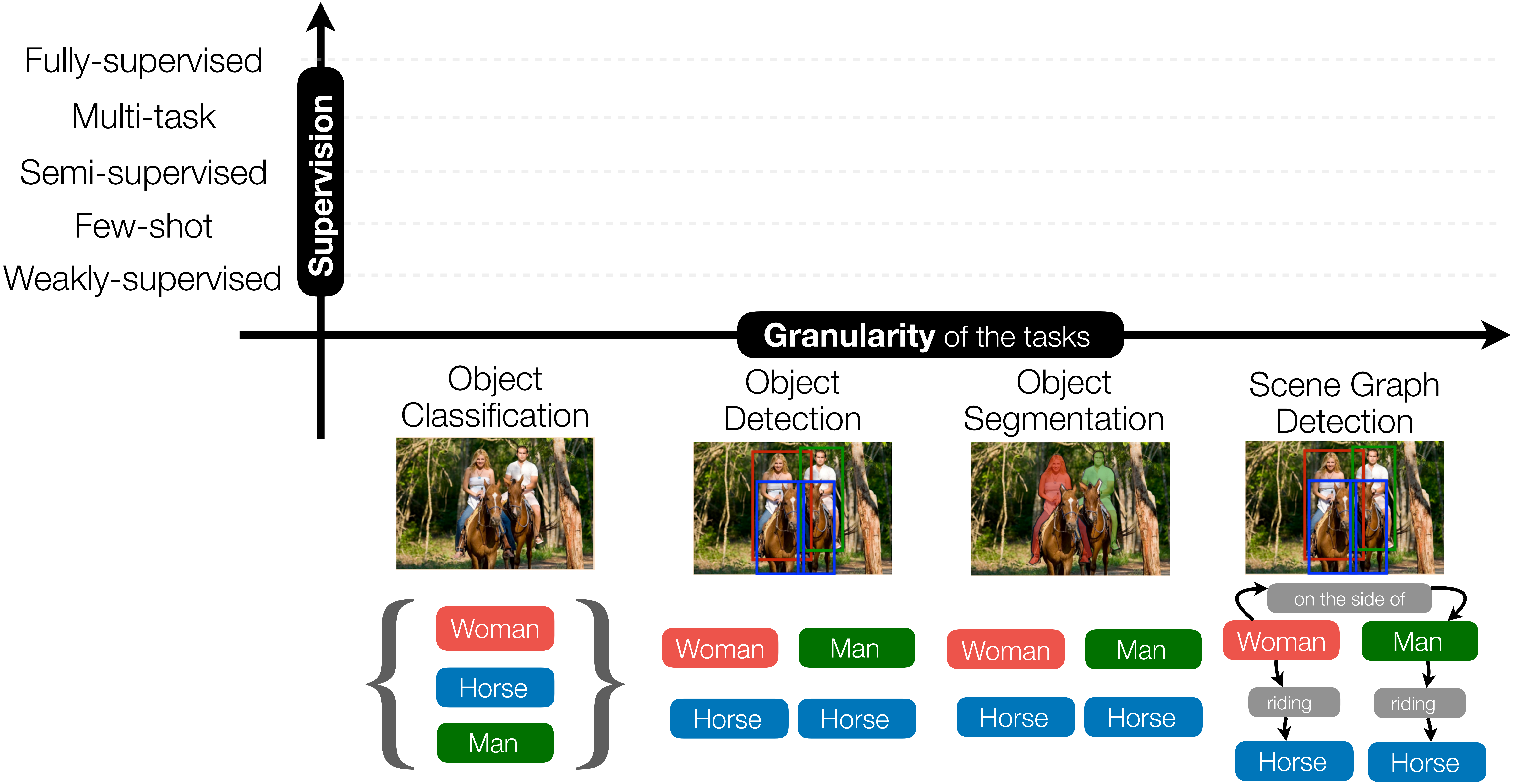
Thesis Statement

We explore **data-efficient learning** approaches for **visual structured prediction tasks**

Positioning: Image understanding



Positioning: Image understanding



Positioning: Image understanding

Fully-supervised

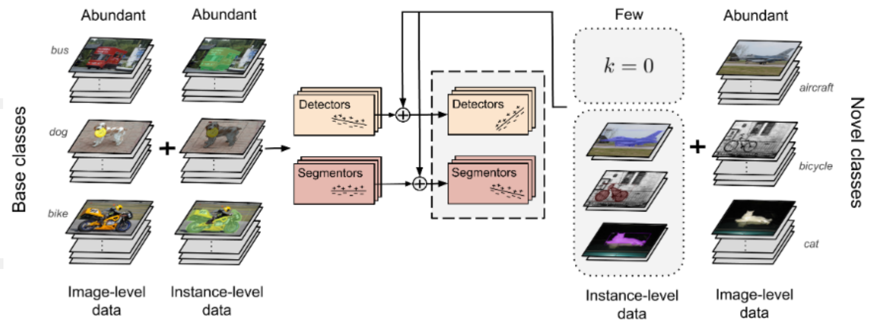
Multi-task

Semi-supervised

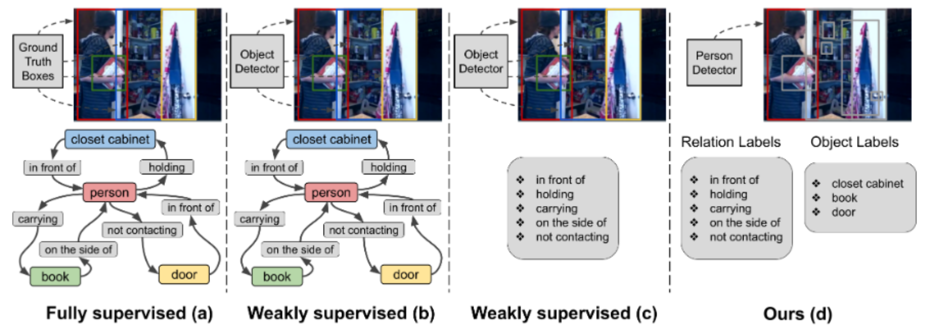
Few-shot

Weakly-supervised

Supervision



S Khandelwal*, **R Goyal*** and L Sigal. "UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation". In CVPR 2021.



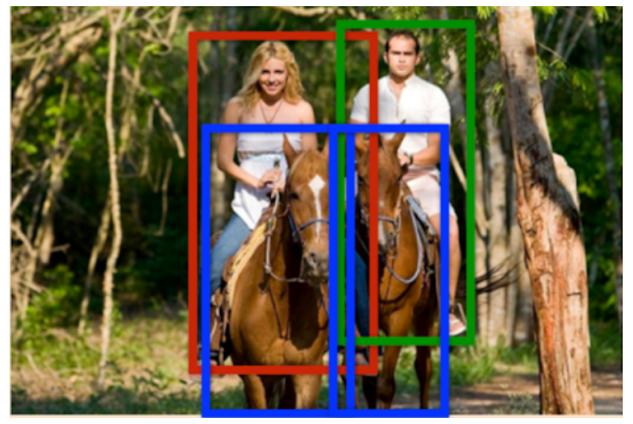
R Goyal and L Sigal. "A Simple Baseline for Weakly-Supervised Human-centric Relation Detection". In BMVC 2021.

Granularity of the tasks

Object Classification

{ Woman, Horse, Man }

Object Detection



Woman Man

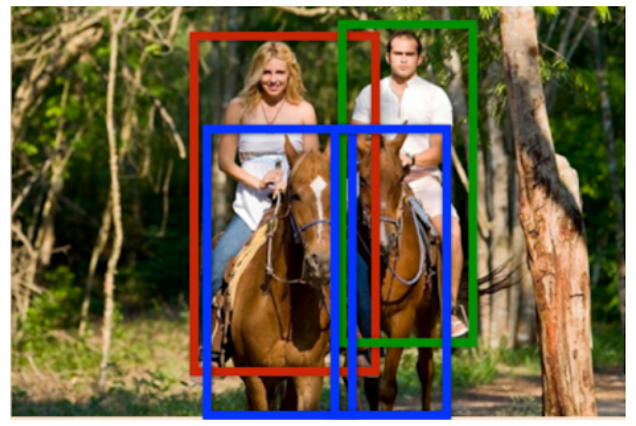
Horse Horse

Object Segmentation

Woman Man

Horse Horse

Scene Graph Detection



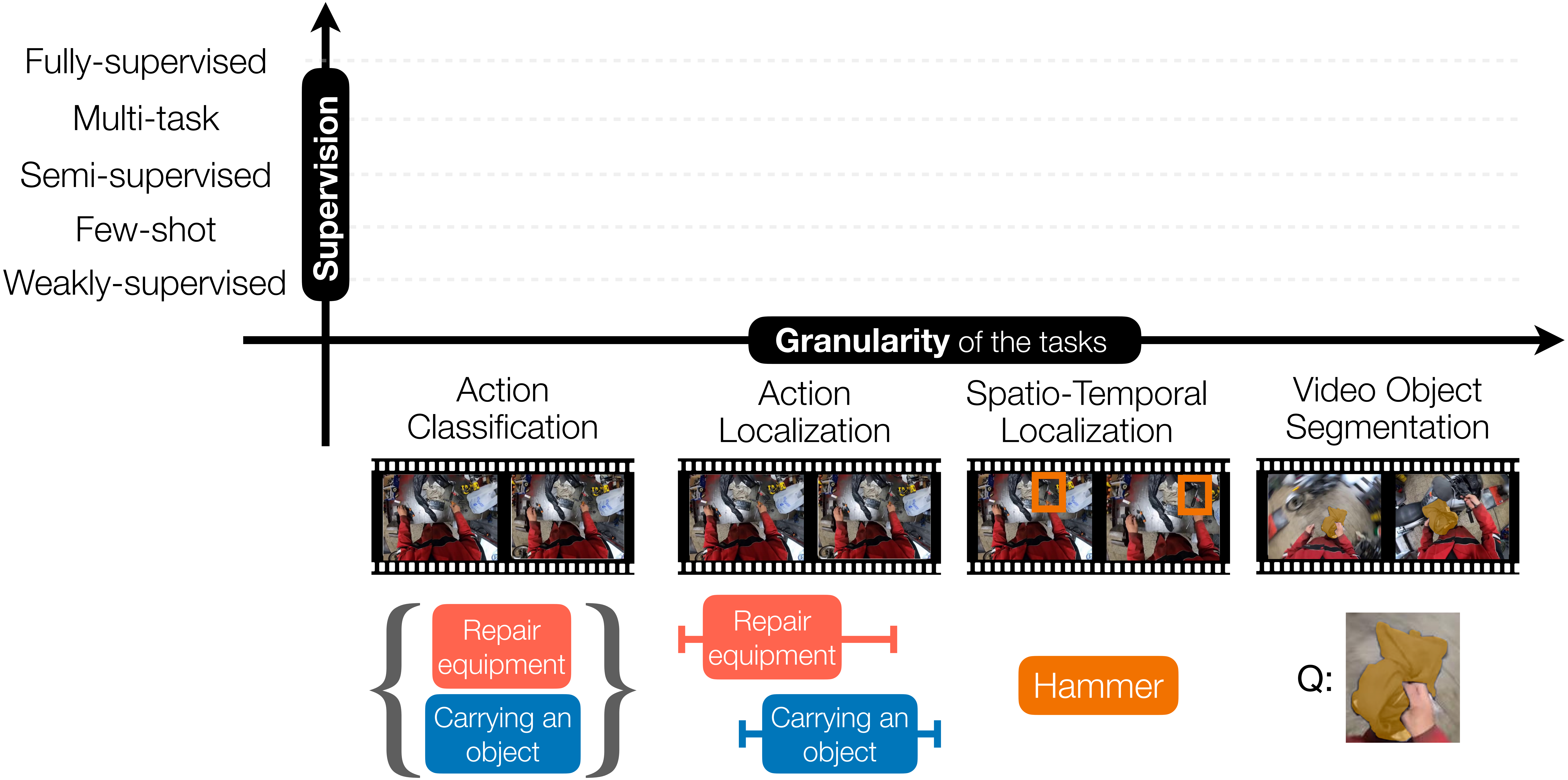
on the side of

Woman Man

riding riding

Horse Horse

Positioning: Video understanding



Positioning: Video understanding

Fully-supervised

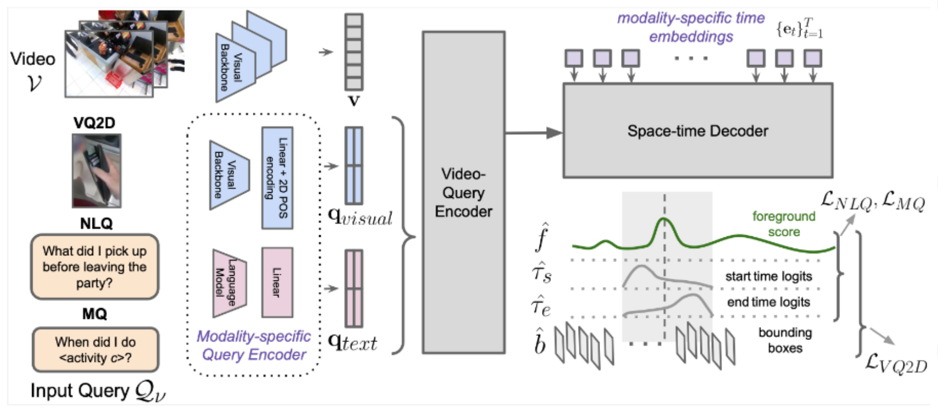
Multi-task

Semi-supervised

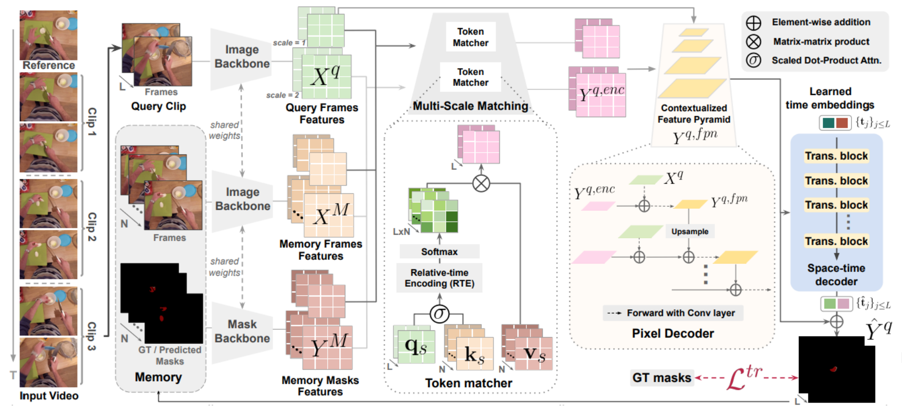
Few-shot

Weakly-supervised

Supervision

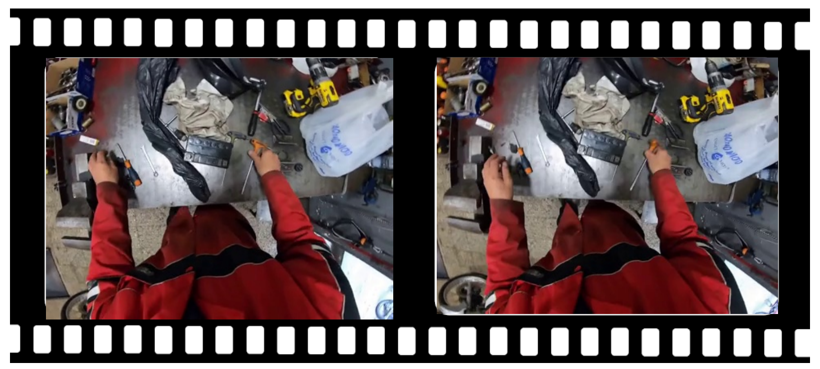
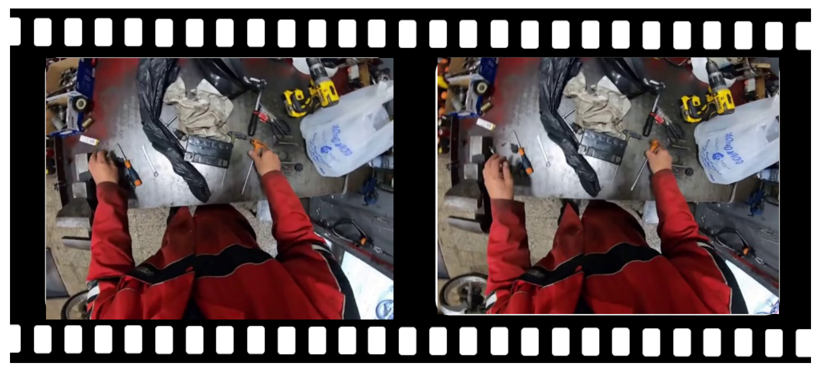


R Goyal, E Mavroudi, X Yang, S Sukhbaatar, L Sigal, M Feiszli, L Torresani, D Tran . "MINOTAUR: Multi-task Video Grounding From Multimodal Queries". arXiv. 2302.08063.



R Goyal*, WC Fan*, M Siam, L Sigal, . "TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking". In WACV 2025.

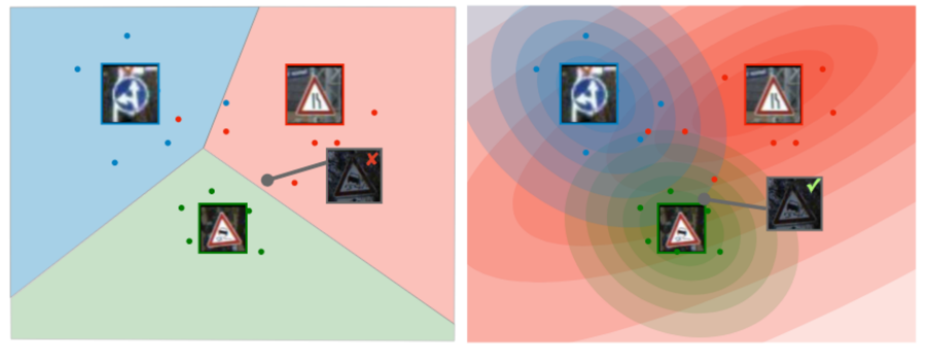
Granularity of the tasks

Action Classification	Action Localization	Spatio-Temporal Localization	Video Object Segmentation
			
{ Repair equipment, Carrying an object }	Repair equipment; Carrying an object	Hammer	Q: 

Overview of the presentation

Image tasks

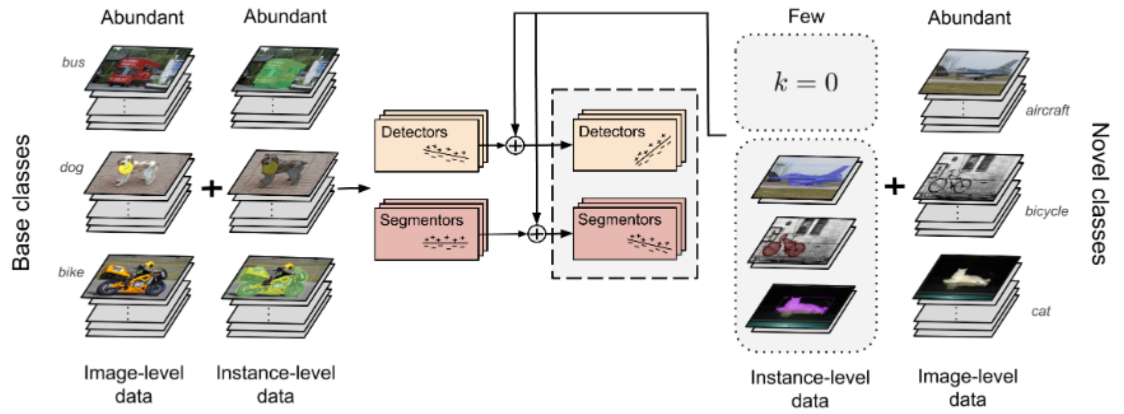
0



(a) Squared Euclidean Distance (b) Squared Mahalanobis Distance

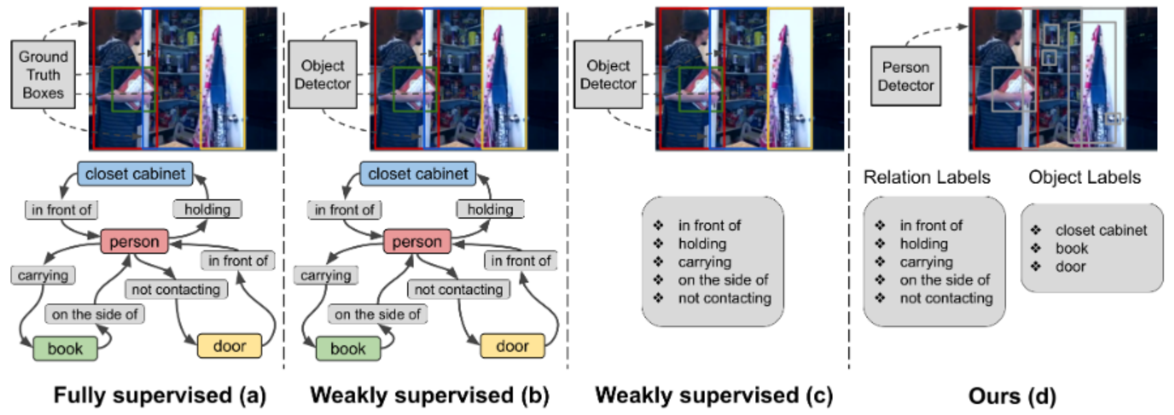
P Bateni, **R Goyal**, V Masrani, F Wood and L Sigal. "Improved Few-Shot Visual Classification". In CVPR 2020.

I



S Khandelwal*, **R Goyal*** and L Sigal. "UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation". In CVPR 2021.

II

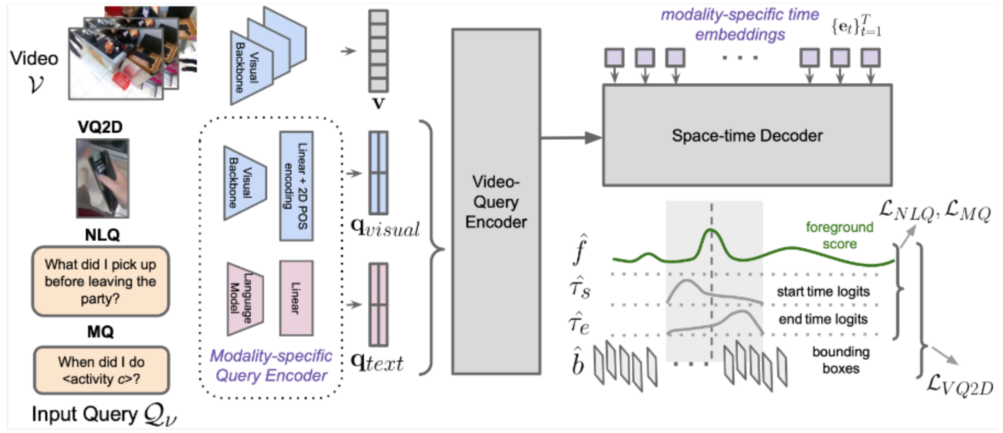


R Goyal and L Sigal. "A Simple Baseline for Weakly-Supervised Human-centric Relation Detection". In BMVC 2021.

** denotes equal contribution*

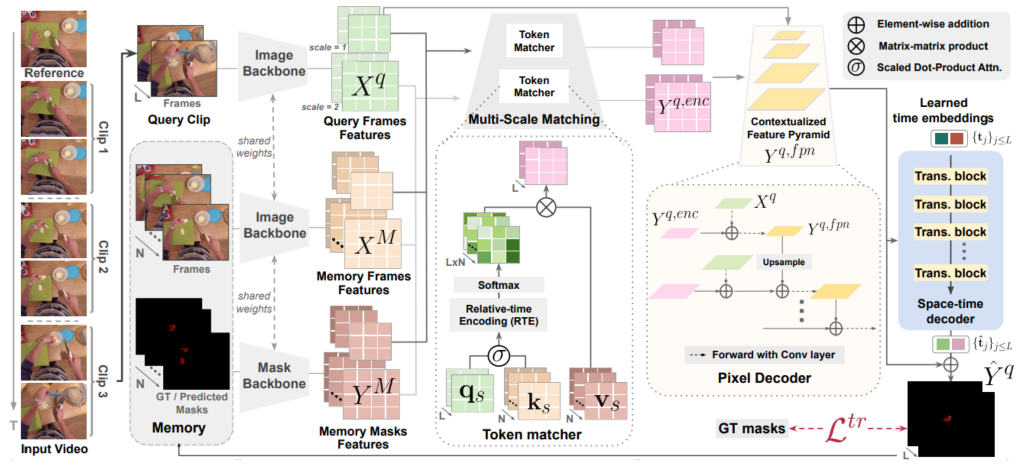
Video tasks

III



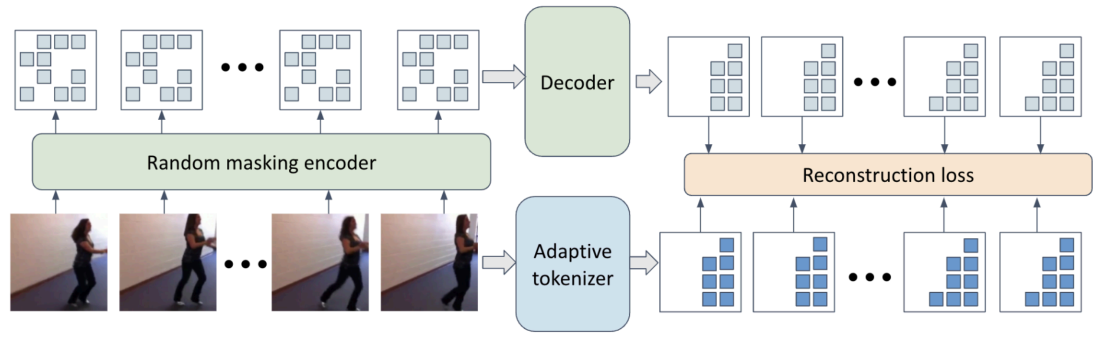
R Goyal, E Mavroudi, X Yang, S Sukhbaatar, L Sigal, M Feiszli, L Torresani, D Tran . "MINOTAUR: Multi-task Video Grounding From Multimodal Queries". arXiv. 2302.08063.

IV



R Goyal*, WC Fan*, M Siam, L Sigal, . "TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking". In WACV 2025.

V



NB Gundavarapu*, L Friedman*, **R Goyal***, C Hegde*, E Agustsson, S M Waghmare, M Sirotenko, MH Yang, T Weyand, B Gong, L Sigal . "Extending Video Masked Autoencoders to 128 frames". In NeurIPS 2024.

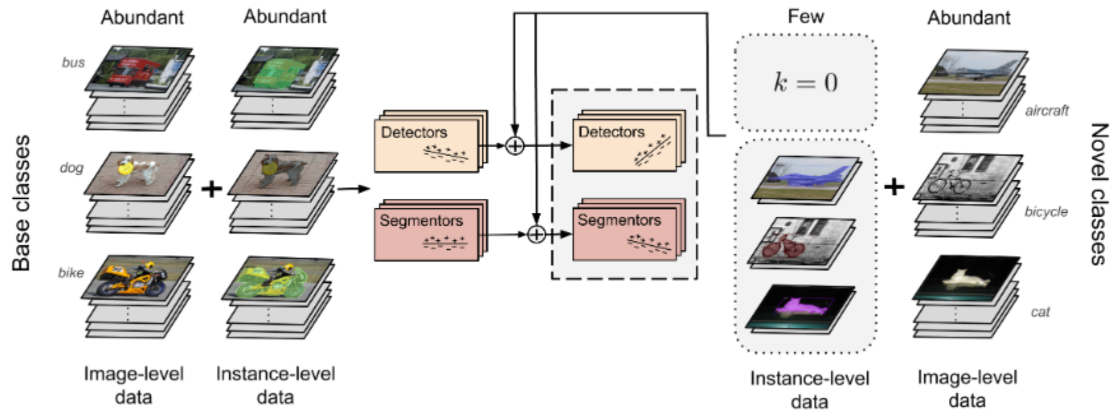
Overview of the presentation

Image tasks

0

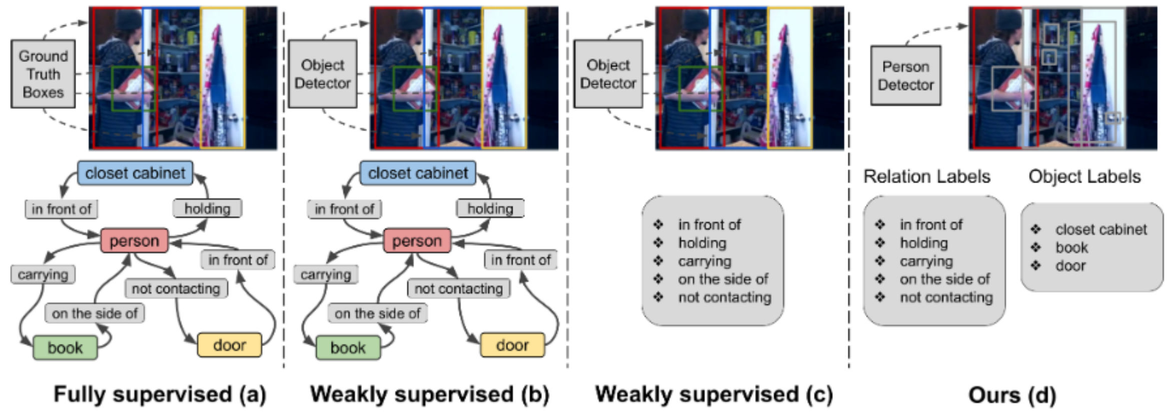
P Bateni, **R Goyal**, V Masrani, F Wood and L Sigal. “Improved Few-Shot Visual Classification”. In CVPR 2020.

I



S Khandelwal*, **R Goyal*** and L Sigal. “UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation”. In CVPR 2021.

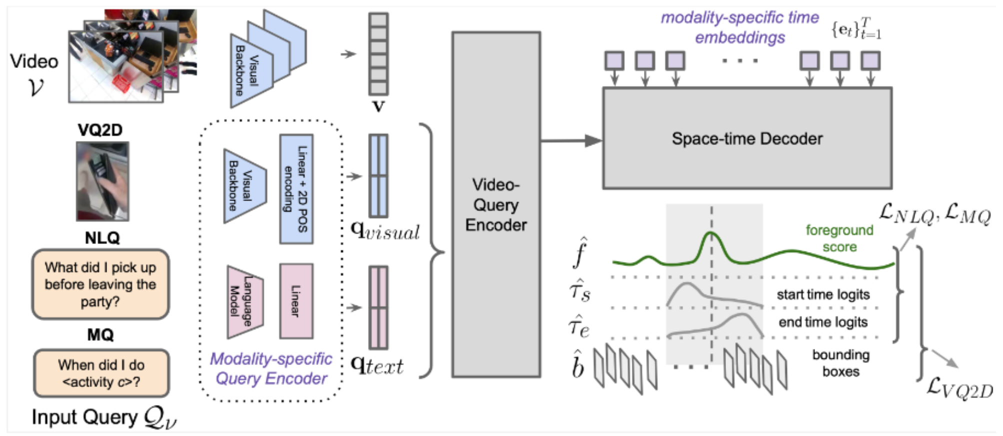
II



R Goyal and L Sigal. “A Simple Baseline for Weakly-Supervised Human-centric Relation Detection”. In BMVC 2021.

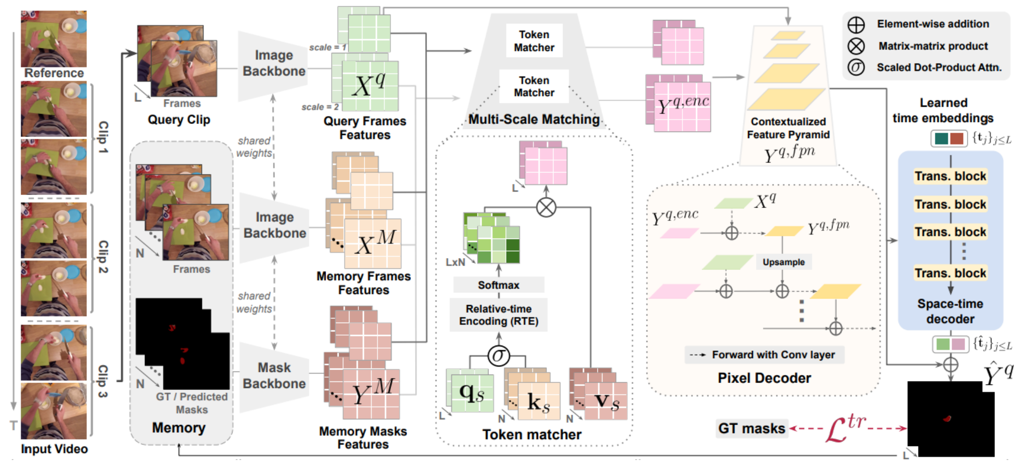
Video tasks

III



R Goyal, E Mavroudi, X Yang, S Sukhbaatar, L Sigal, M Feiszli, L Torresani, D Tran . “MINOTAUR: Multi-task Video Grounding From Multimodal Queries”. arXiv. 2302.08063.

IV



R Goyal*, WC Fan*, M Siam, L Sigal, . “TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking”. In WACV 2025.

V

NB Gundavarapu*, L Friedman*, **R Goyal***, C Hegde*, E Agustsson, S M Waghmare, M Sirotenko, MH Yang, T Weyand, B Gong, L Sigal . “Extending Video Masked Autoencoders to 128 frames”. In NeurIPS 2024.

* *denotes equal contribution*

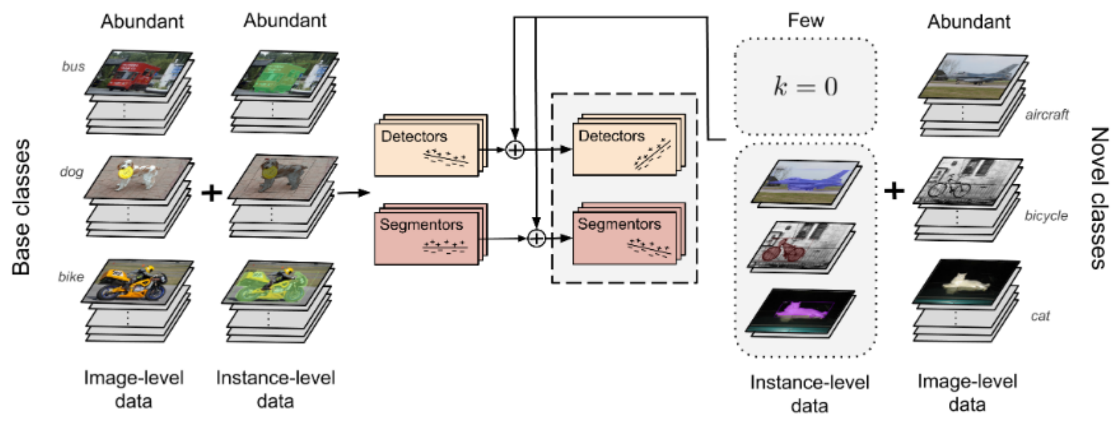
Chapter I

Image tasks

0

P Bateni, **R Goyal**, V Masrani, F Wood and L Sigal. “Improved Few-Shot Visual Classification”. In CVPR 2020.

I



S Khandelwal*, **R Goyal*** and L Sigal. “UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation”. In CVPR 2021.

II

R Goyal and L Sigal. “A Simple Baseline for Weakly-Supervised Human-centric Relation Detection”. In BMVC 2021.

Video tasks

III

R Goyal, E Mavroudi, X Yang, S Sukhbaatar, L Sigal, M Feiszli, L Torresani, D Tran . “MINOTAUR: Multi-task Video Grounding From Multimodal Queries”. arXiv. 2302.08063.

IV

R Goyal*, WC Fan*, M Siam, L Sigal, . “TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking”. In WACV 2025.

V

NB Gundavarapu*, L Friedman*, **R Goyal***, C Hegde*, E Agustsson, S M Waghmare, M Sirotenko, MH Yang, T Weyand, B Gong, L Sigal . “Extending Video Masked Autoencoders to 128 frames”. In NeurIPS 2024.

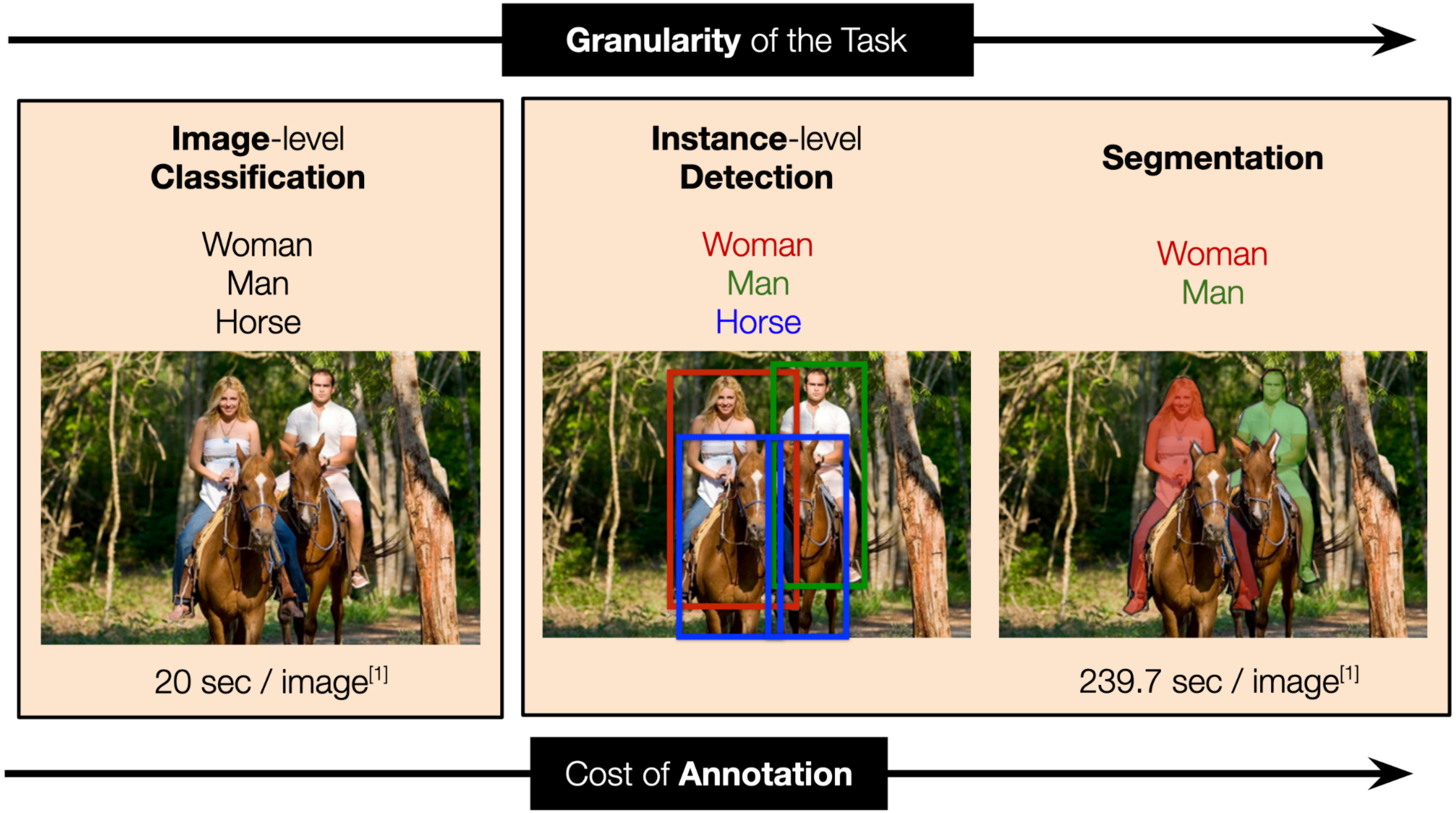
** denotes equal contribution*

I. Contributions

- First method that seamlessly incorporates **zero- to few-shot supervision** in a **single framework**
- Explored **effectiveness** of different (strong / weak) forms of supervision in a **limited budget setting**

I. **Complexity** of Annotation

Annotation is costlier for granular instance-level tasks like object detection and segmentation.



[1] Bearman et al., “What’s the point: Semantic segmentation with point supervision”, ECCV, 2016.

I. **Any-shot** object detection / segmentation

Base classes

Abundant

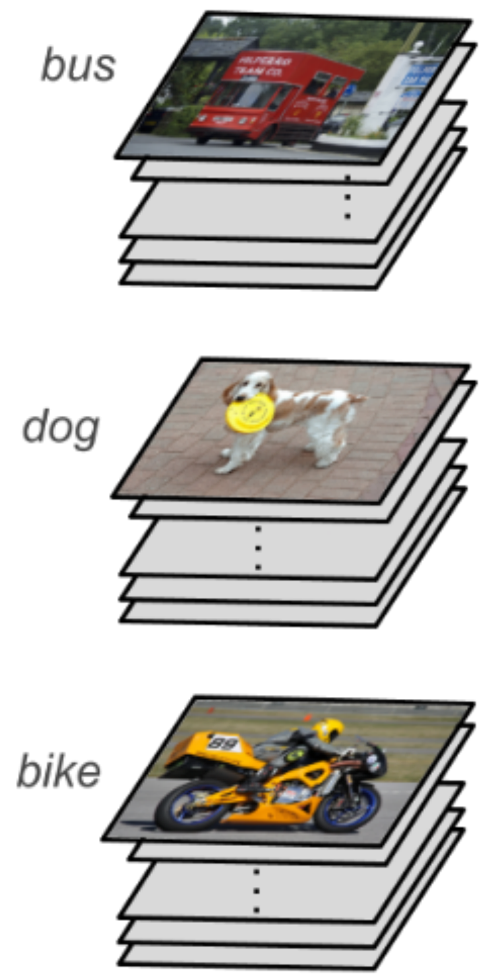


Image-level data

Image-level object data gives us **weak detectors**

Abundant

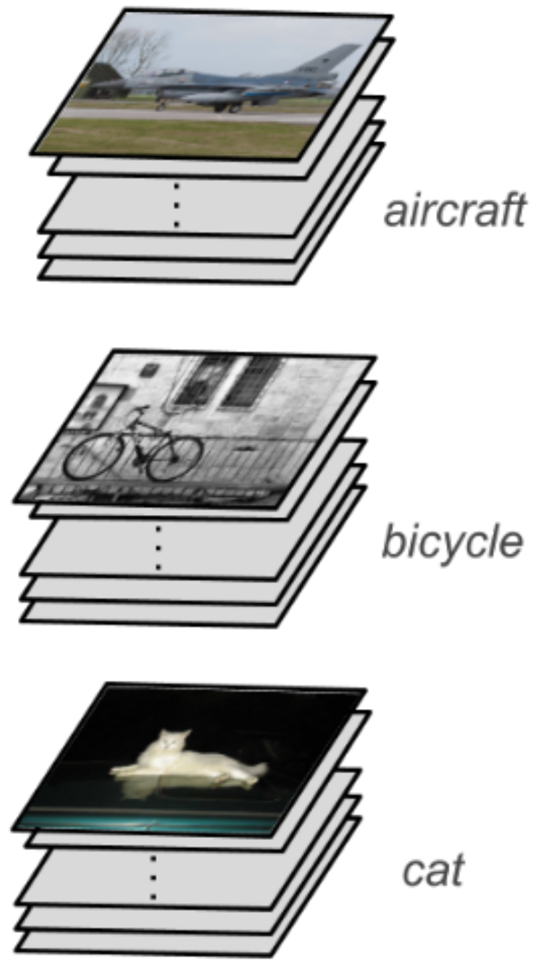


Image-level data

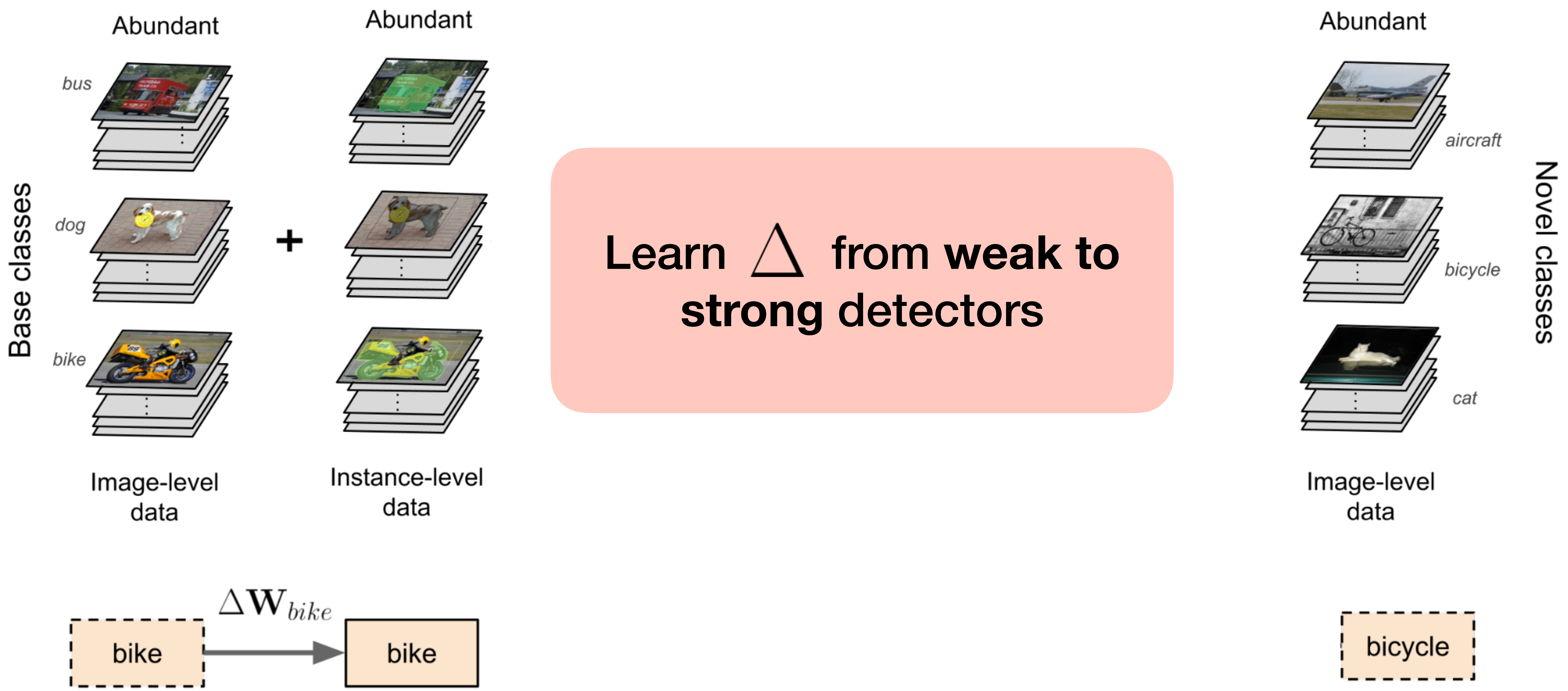
Novel classes

bike

bicycle

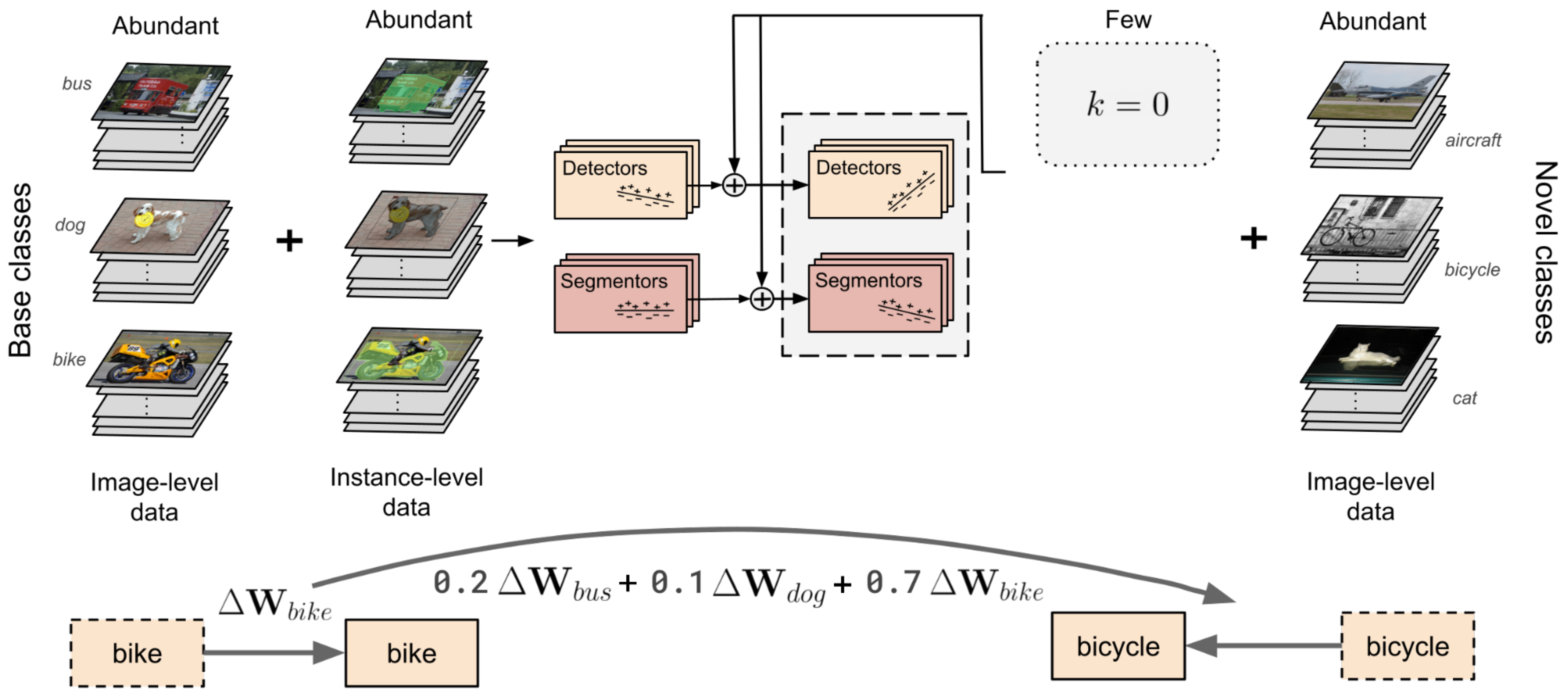
Hoffman et al. "LSDA: Large scale detection through adaptation." In NeurIPS 2014.
Tang et al. "Large scale semi-supervised object detection using visual and semantic knowledge transfer." In ICCV 2016.
Kumar Singh et al. “Dock: Detecting objects by transferring common-sense knowledge”. In ECCV 2018.

I. **Any-shot** object detection / segmentation



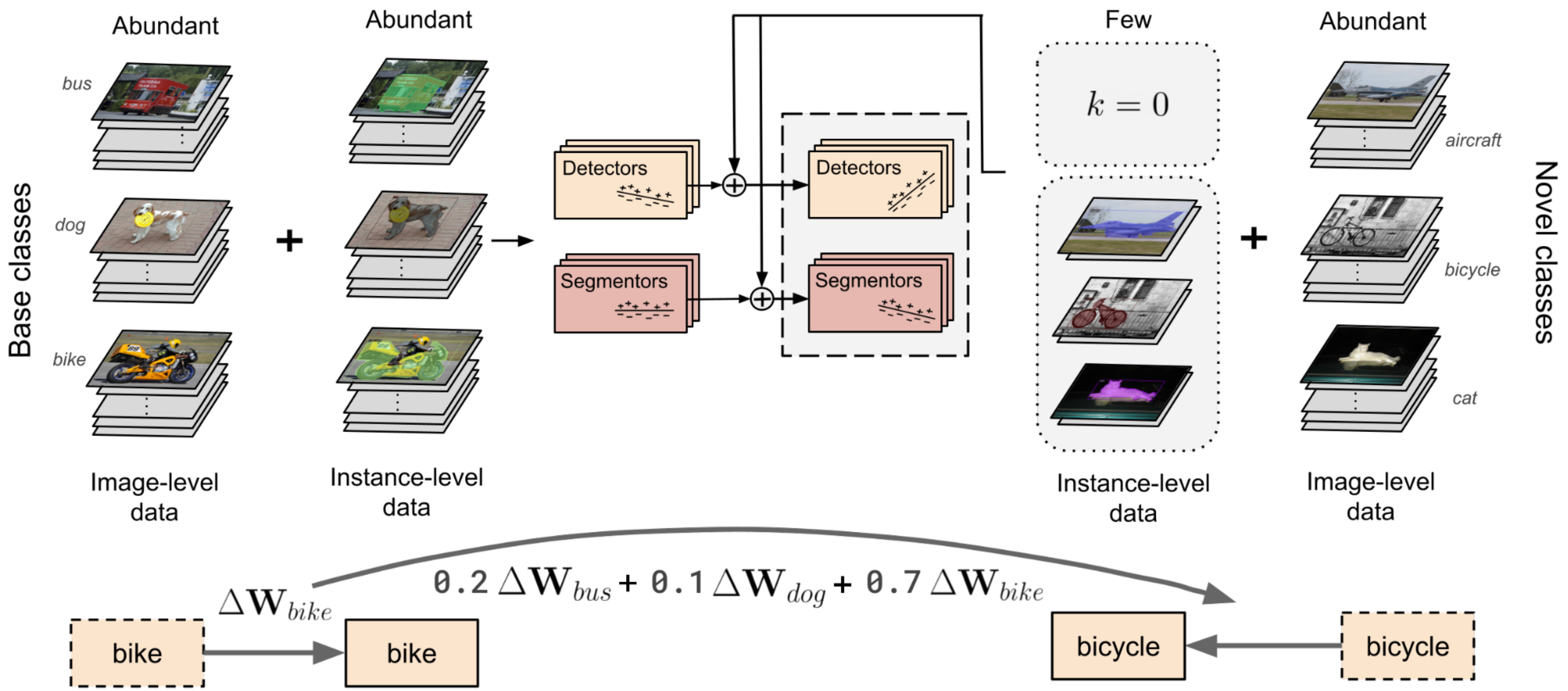
Hoffman et al. "LSDA: Large scale detection through adaptation." In NeurIPS 2014.
Tang et al. "Large scale semi-supervised object detection using visual and semantic knowledge transfer." In ICCV 2016.
Kumar Singh et al. “Dock: Detecting objects by transferring common-sense knowledge”. In ECCV 2018.

I. **Any-shot** object detection / segmentation



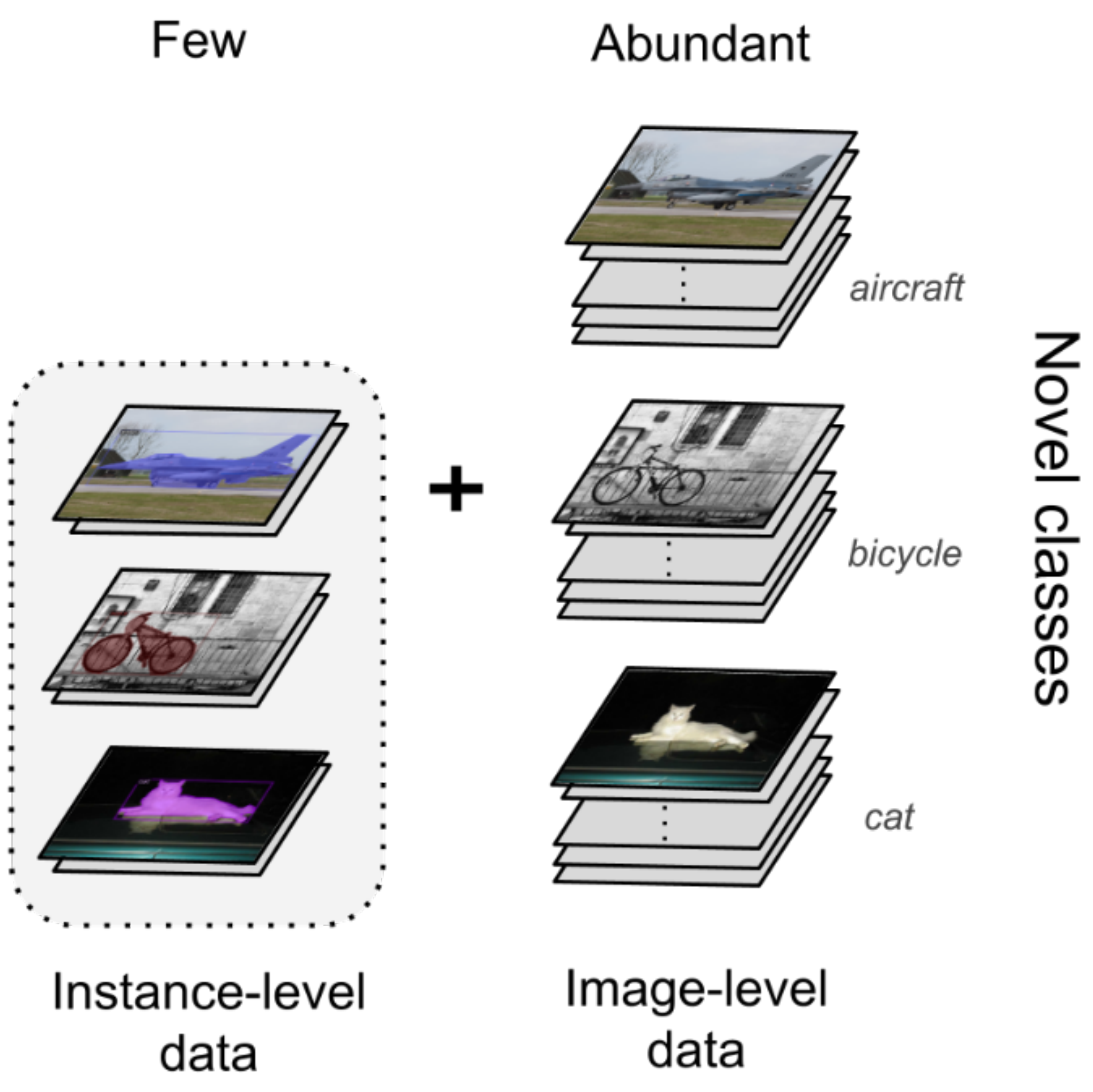
Hoffman et al. "LSDA: Large scale detection through adaptation." In NeurIPS 2014.
Tang et al. "Large scale semi-supervised object detection using visual and semantic knowledge transfer." In ICCV 2016.
Kumar Singh et al. “Dock: Detecting objects by transferring common-sense knowledge”. In ECCV 2018.

I. **Any-shot** object detection / segmentation



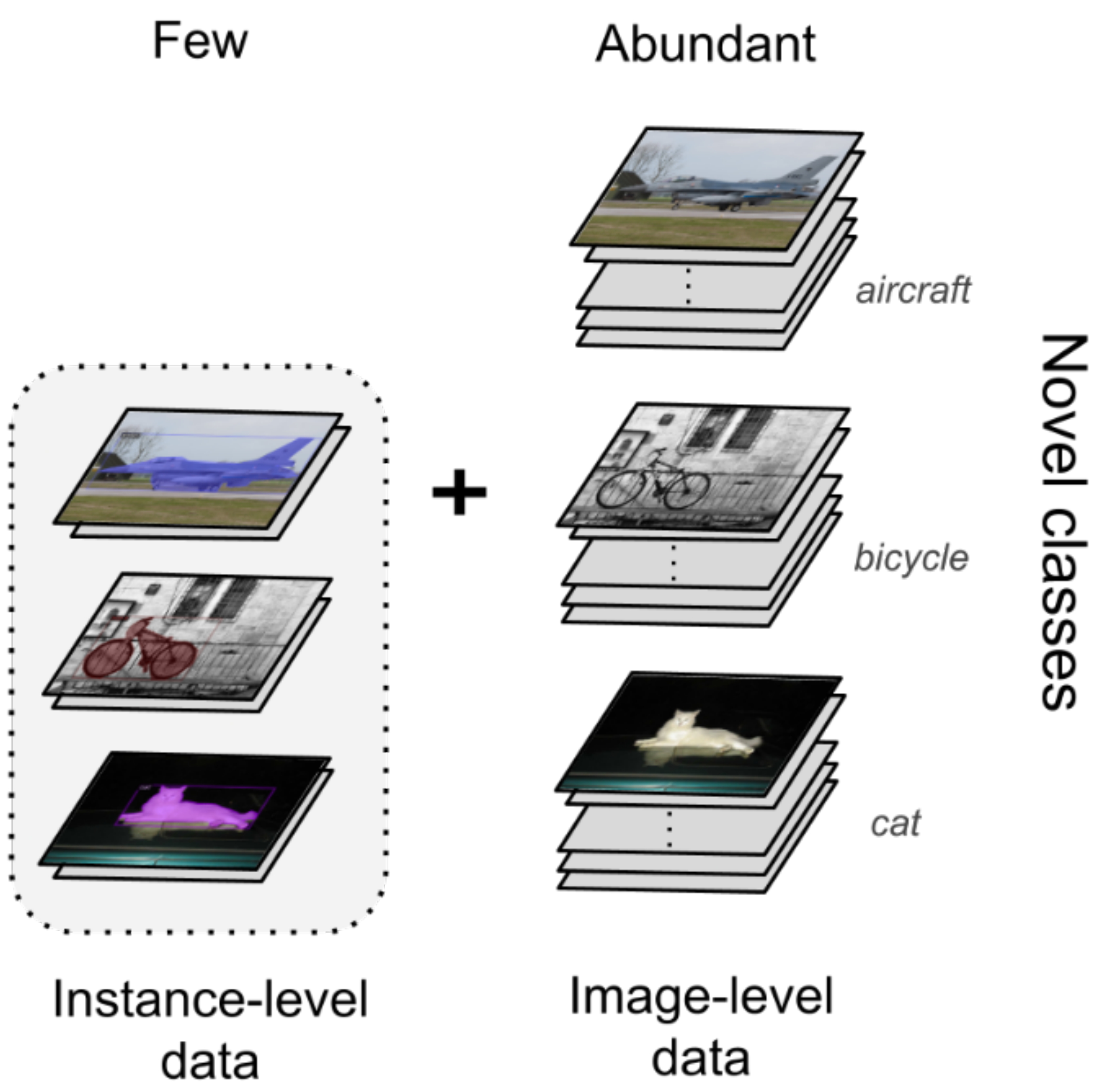
Hoffman et al. "LSDA: Large scale detection through adaptation." In NeurIPS 2014.
Tang et al. "Large scale semi-supervised object detection using visual and semantic knowledge transfer." In ICCV 2016.
Kumar Singh et al. “Dock: Detecting objects by transferring common-sense knowledge”. In ECCV 2018.

I. Limited budget for **novel** classes



[1] VOC dataset has *20 object classes* and *2.8 objects on avg. per image*
[2] Su, Hao, Jia Deng, and Li Fei-Fei. "Crowdsourcing annotations for visual object detection." In AAAI 2012.

I. **Limited budget** for **novel** classes



Instance-level
bounding box annotations
(1 image)

137.2 sec / image[1,2]

7x →

Image-level
weak annotations
(7 images)

20 sec / image[1,2]

[1] VOC dataset has *20 object classes* and *2.8 objects on avg. per image*
[2] Su, Hao, Jia Deng, and Li Fei-Fei. "Crowdsourcing annotations for visual object detection." In AAAI 2012.

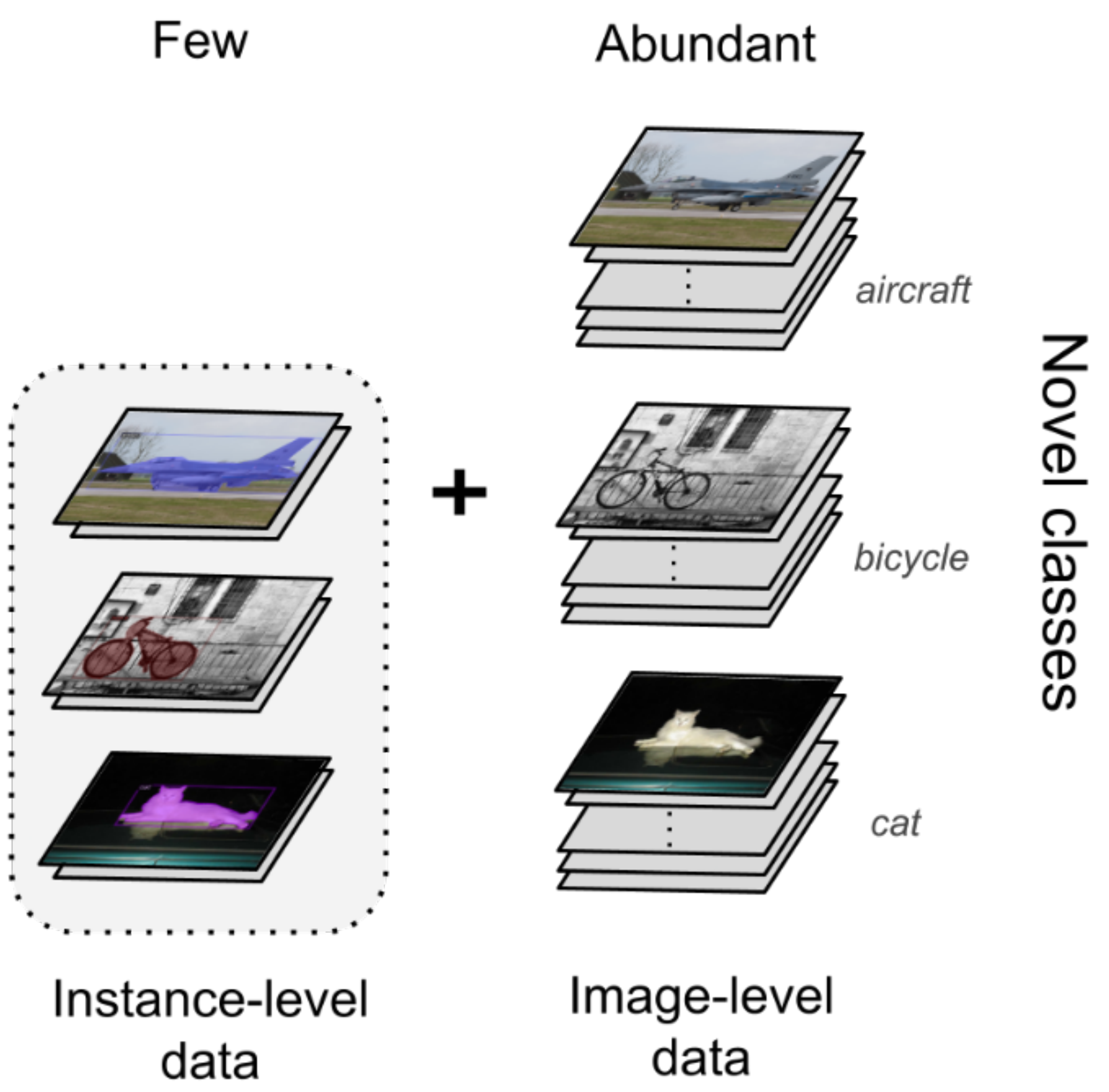
I. Limited budget for novel classes

More budget towards **weak annotations**

Budget: 10 instance-level annotations

%		AP_{50}
Instance	Weak	
100	0	N/A
90	10	49.2 ± 0.6
50	50	54.0 ± 0.8
0	100	$\mathbf{59.0 \pm 1.5}$

Higher performance



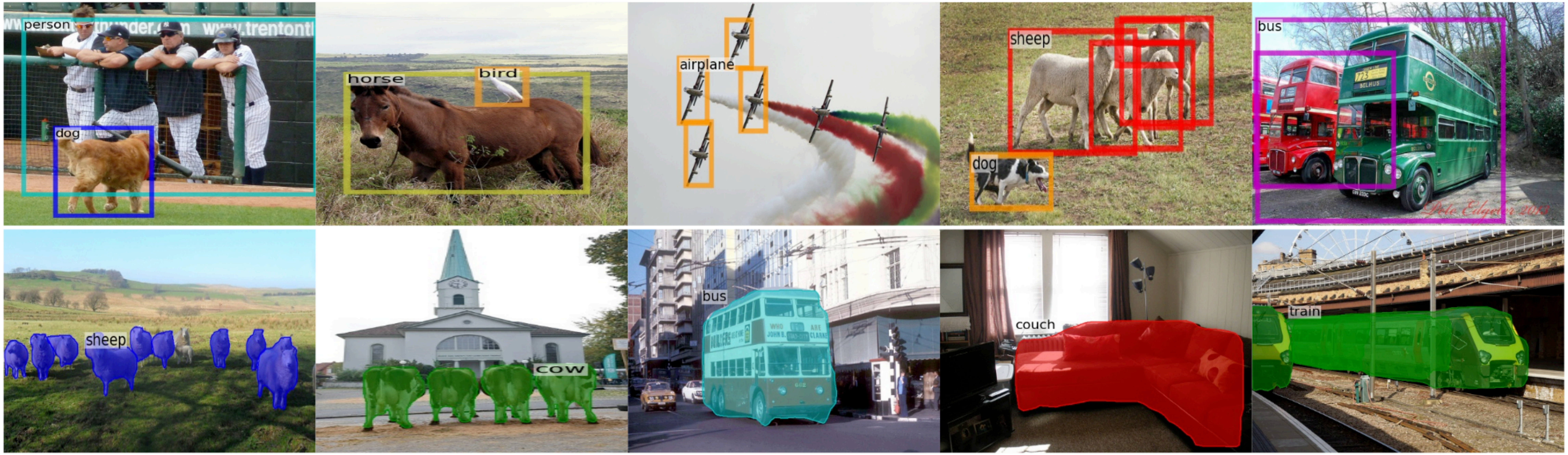
Instance-level bounding box annotations (1 image) → **7x** → **Image-level** weak annotations (7 images)

137.2 sec / image[1,2]

20 sec / image[1,2]

[1] VOC dataset has *20 object classes* and *2.8 objects on avg. per image*
[2] Su, Hao, Jia Deng, and Li Fei-Fei. "Crowdsourcing annotations for visual object detection." In AAAI 2012.

Examples for **weakly-supervised zero-shot (*k* = *0*)**



Chapter II

Image tasks

0

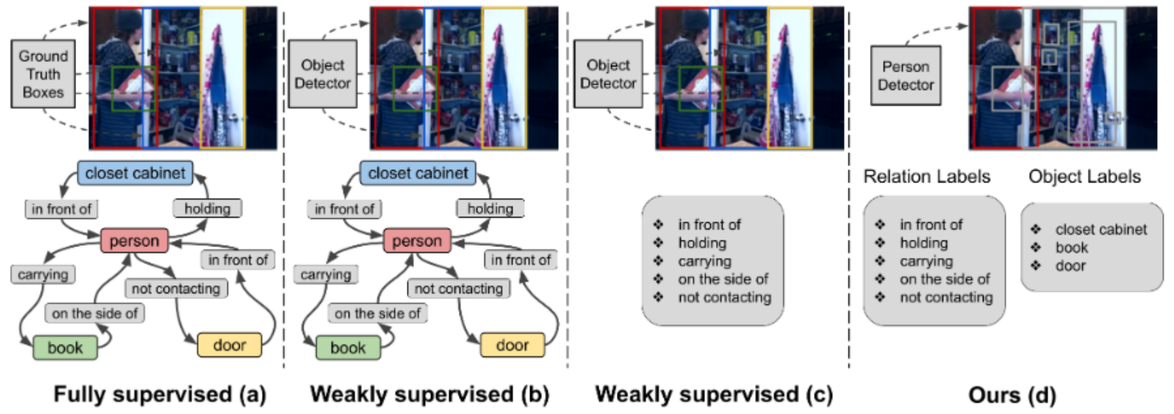
(a) Squared Euclidean Distance (b) Squared Mahalanobis Distance

P Bateni, **R Goyal**, V Masrani, F Wood and L Sigal. "Improved Few-Shot Visual Classification". In CVPR 2020.

I

S Khandelwal*, **R Goyal*** and L Sigal. "UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation". In CVPR 2021.

II



R Goyal and L Sigal. "A Simple Baseline for Weakly-Supervised Human-centric Relation Detection". In BMVC 2021.

Video tasks

III

R Goyal, E Mavroudi, X Yang, S Sukhbaatar, L Sigal, M Feiszli, L Torresani, D Tran . "MINOTAUR: Multi-task Video Grounding From Multimodal Queries". arXiv. 2302.08063.

IV

R Goyal*, WC Fan*, M Siam, L Sigal, . "TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking". In WACV 2025.

V

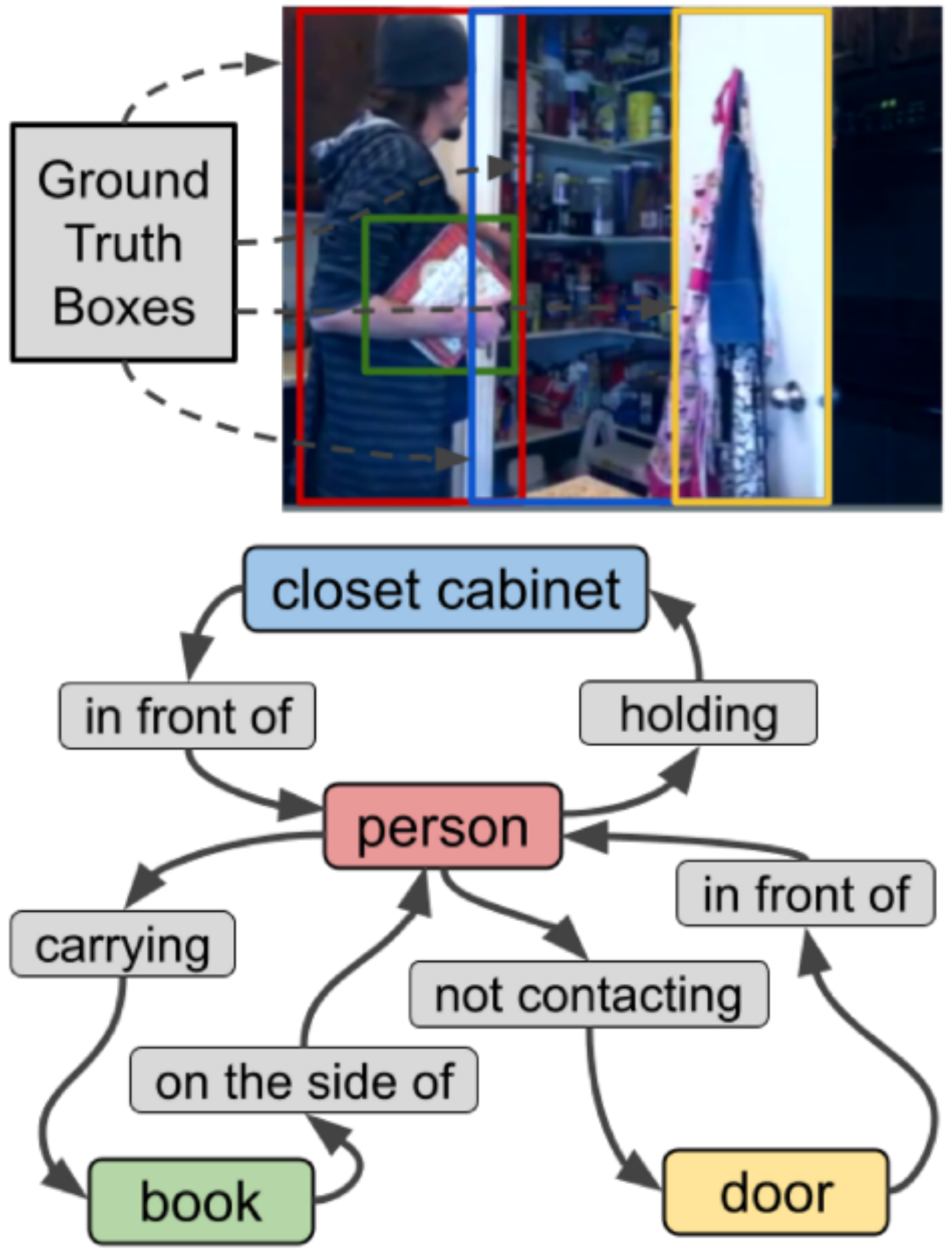
NB Gundavarapu*, L Friedman*, **R Goyal***, C Hegde*, E Agustsson, S M Waghmare, M Sirotenko, MH Yang, T Weyand, B Gong, L Sigal . "Extending Video Masked Autoencoders to 128 frames". In NeurIPS 2024.

** denotes equal contribution*

II. Contributions

- **Weaker supervision** for human-object interaction (**scene-graphs**) leads to competitive results compared with stronger supervision

II. **Spectrum** of weak-supervision for scene-graph

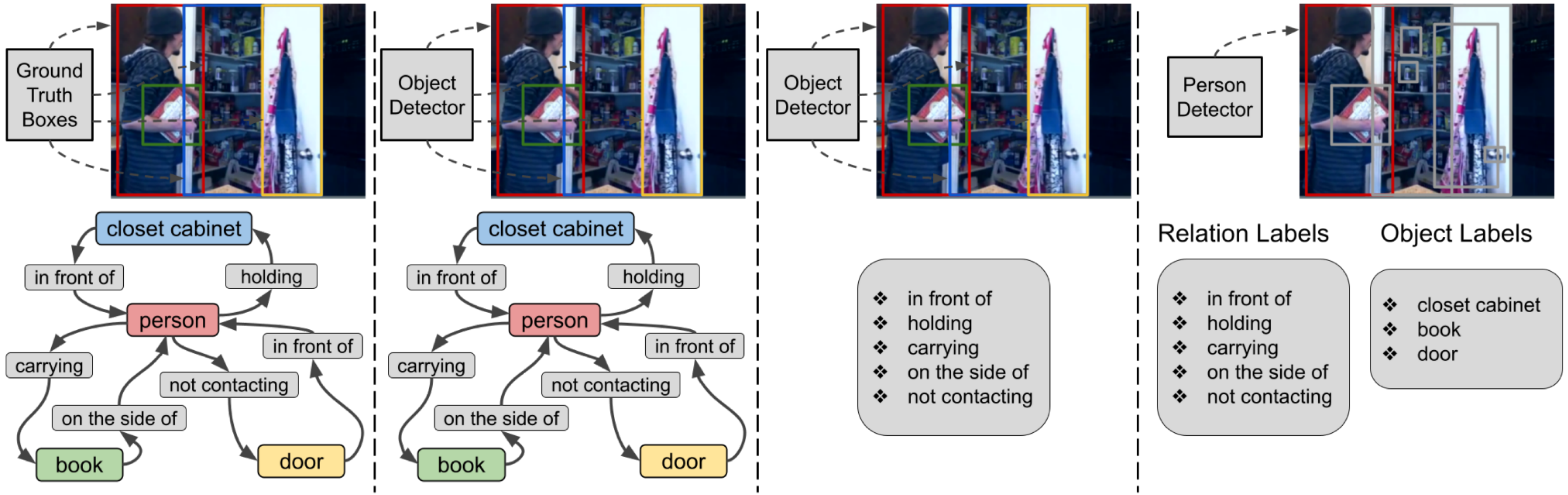


Fully supervised

Object bounding boxes and their labels
Grounded relation triplets `<sub,pred,obj>`

Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In CVPR 2017.
Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In ECCV 2016.
Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In CVPR 2017.
Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In CVPR 2018.

II. **Spectrum** of weak-supervision for scene-graph



Fully supervised | **Weakly supervised** | **Weakly supervised** | **Ours**

Object bounding boxes and their labels
Grounded relation triplets `<sub,pred,obj>`

Assumes object detectors

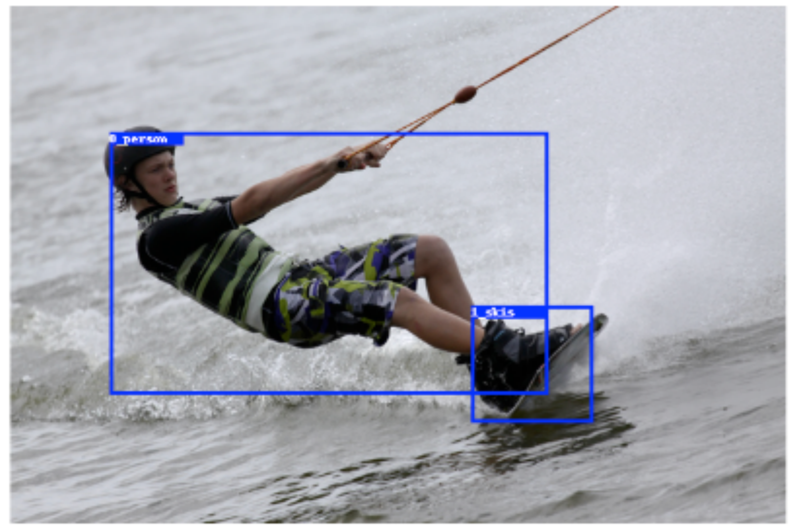
Person detector
List of objects and relations

Weaker supervision

(b) Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In ICCV 2017.
(b) Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In ICCV 2017.
(c) Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In ECCV 2020.

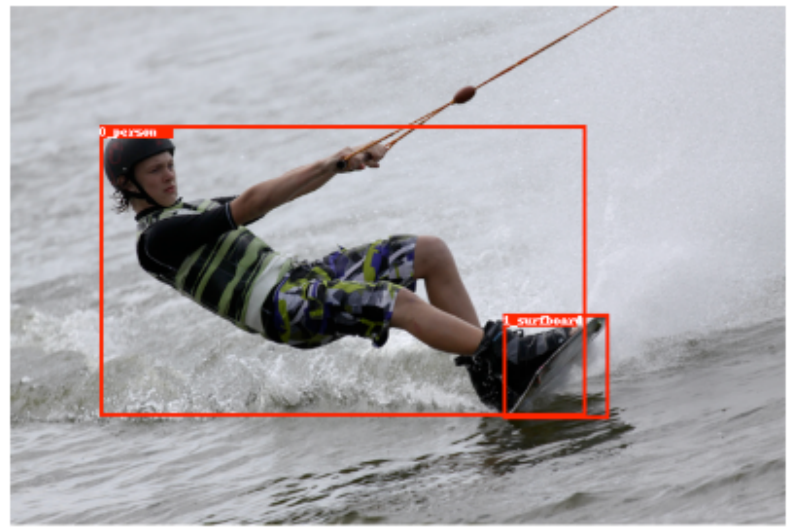
II. Weak supervision

Ground-truth



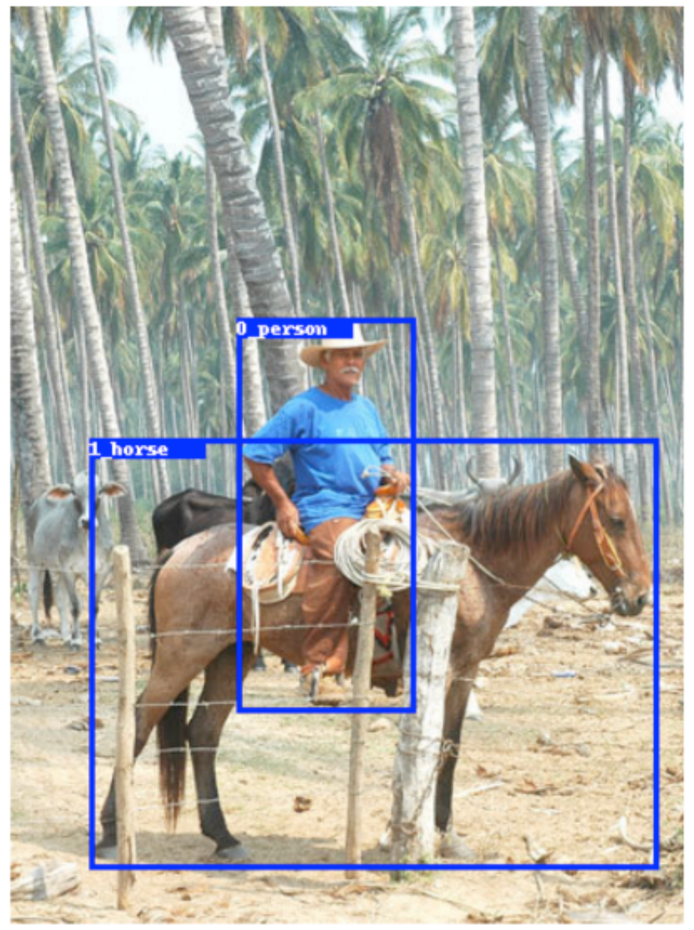
person ride skis

Prediction



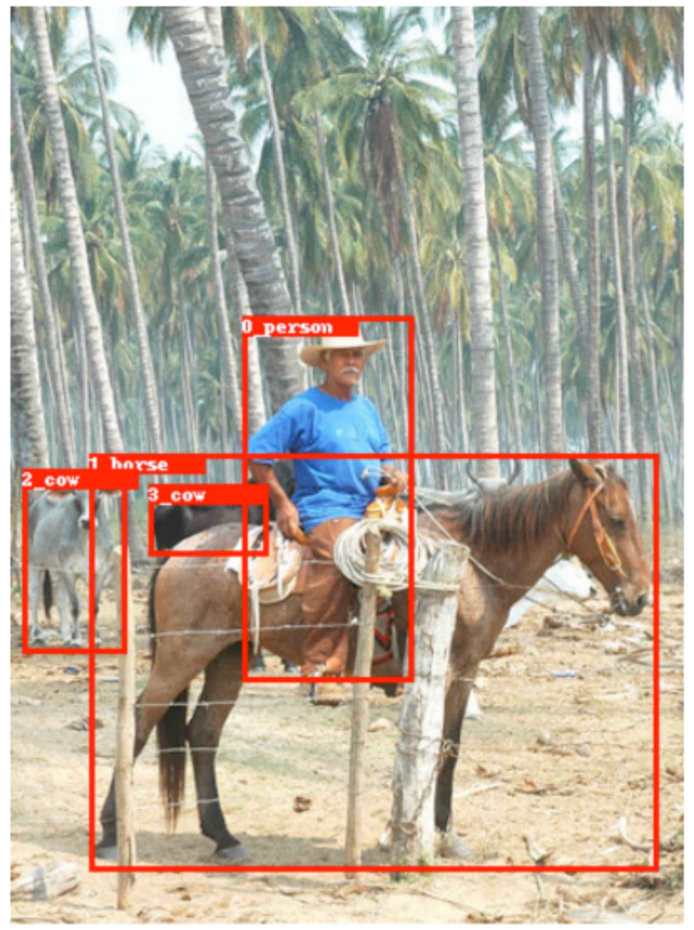
person stand on surfboard

Ground-truth



person straddle horse

Prediction



person straddle horse
person lasso cow
person hold cow

Method	R@20	R@50
Fully-supervised		
VRD [22]	10.28	10.94
Freq Prior [47]	24.03	24.87
IMP [43]	23.88	25.52
MSDN [17]	24.00	25.64
Graph R-CNN [44]	24.12	25.77
RelDN [50]	<u>25.00</u>	26.21
Ours	**27.93**	**30.42**
Weakly-supervised		
Ours	23.21	<u>27.24</u>

Action Genome dataset

Weak supervision can be **competitive** with fully-supervised approaches

Chapter III

Image tasks

0

(a) Squared Euclidean Distance (b) Squared Mahalanobis Distance

P Bateni, **R Goyal**, V Masrani, F Wood and L Sigal. “Improved Few-Shot Visual Classification”. In CVPR 2020.

I

S Khandelwal*, **R Goyal*** and L Sigal. “UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation”. In CVPR 2021.

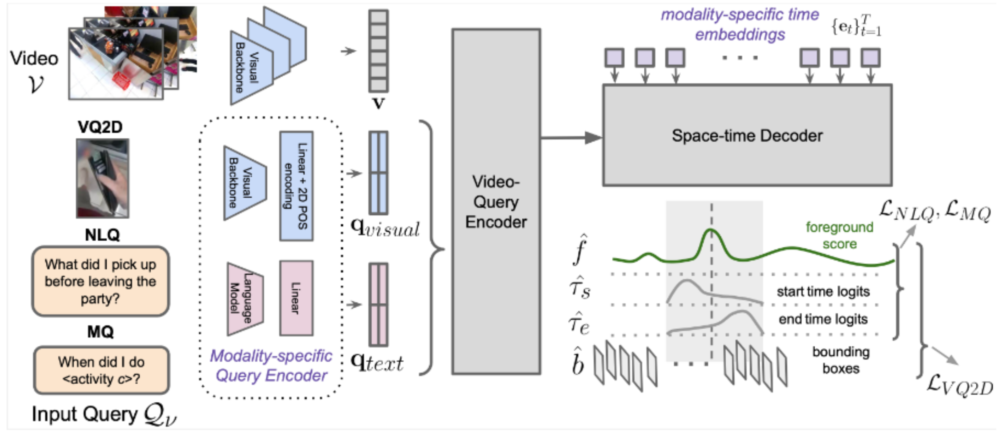
II

Fully supervised (a) Weakly supervised (b) Weakly supervised (c) Ours (d)

R Goyal and L Sigal. “A Simple Baseline for Weakly-Supervised Human-centric Relation Detection”. In BMVC 2021.

Video tasks

III



R Goyal, E Mavroudi, X Yang, S Sukhbaatar, L Sigal, M Feiszli, L Torresani, D Tran . “MINOTAUR: Multi-task Video Grounding From Multimodal Queries”. arXiv. 2302.08063.

IV

R Goyal*, WC Fan*, M Siam, L Sigal, . “TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking”. In WACV 2025.

V

NB Gundavarapu*, L Friedman*, **R Goyal***, C Hegde*, E Agustsson, S M Waghmare, M Sirotenko, MH Yang, T Weyand, B Gong, L Sigal . “Extending Video Masked Autoencoders to 128 frames”. In NeurIPS 2024.

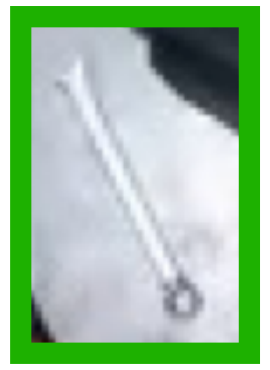
** denotes equal contribution*

III. Contributions

- We study heterogenous video tasks under a **single, unified model**
- We find **multi-task learning** leads to **cross-task transfer** and **generalization to unseen tasks**

III. **Synergy** among video tasks

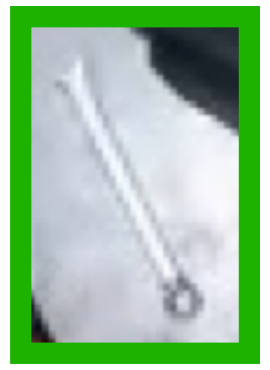
Let's find answers in the video!

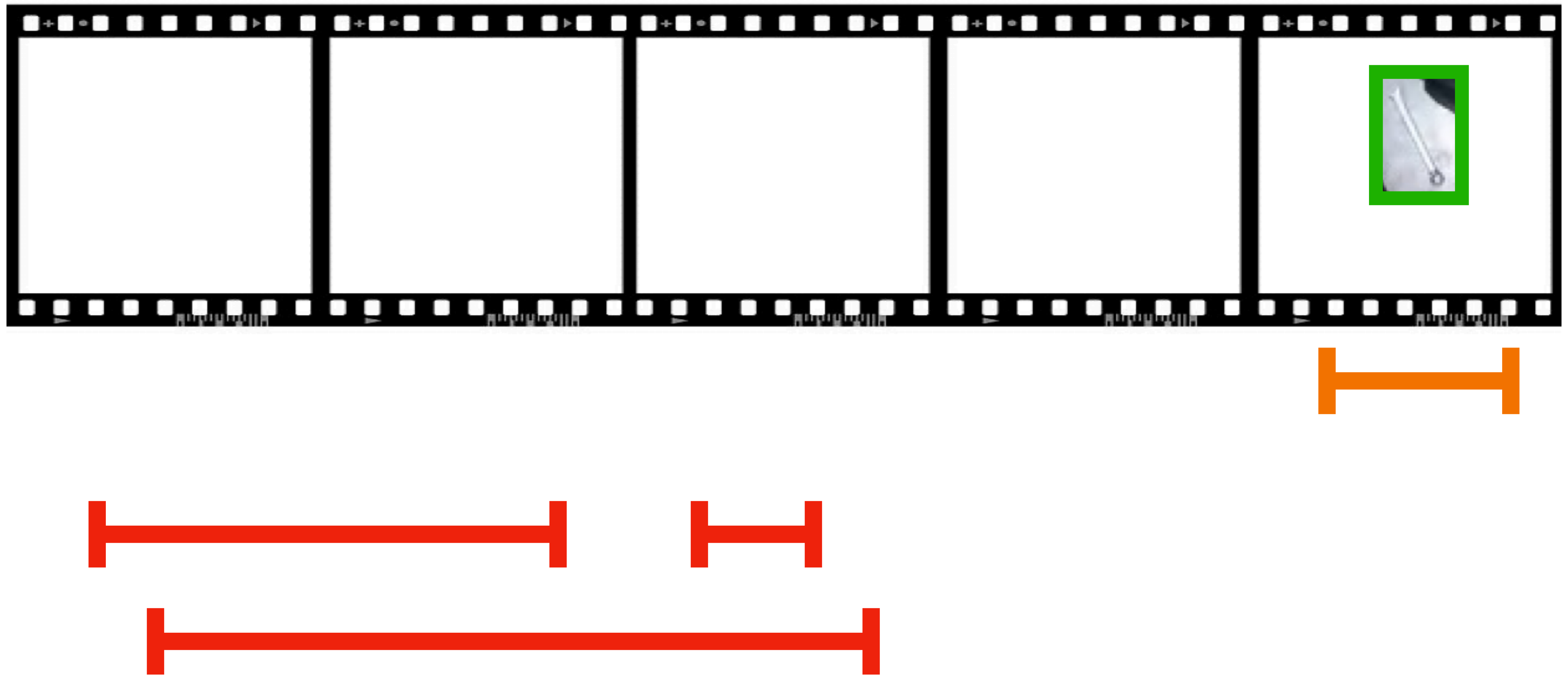
When and **where** did I last see  ?

When did I put the ***spanner***?

When did I ***repair small equipment***?

III. **Synergy** among video tasks

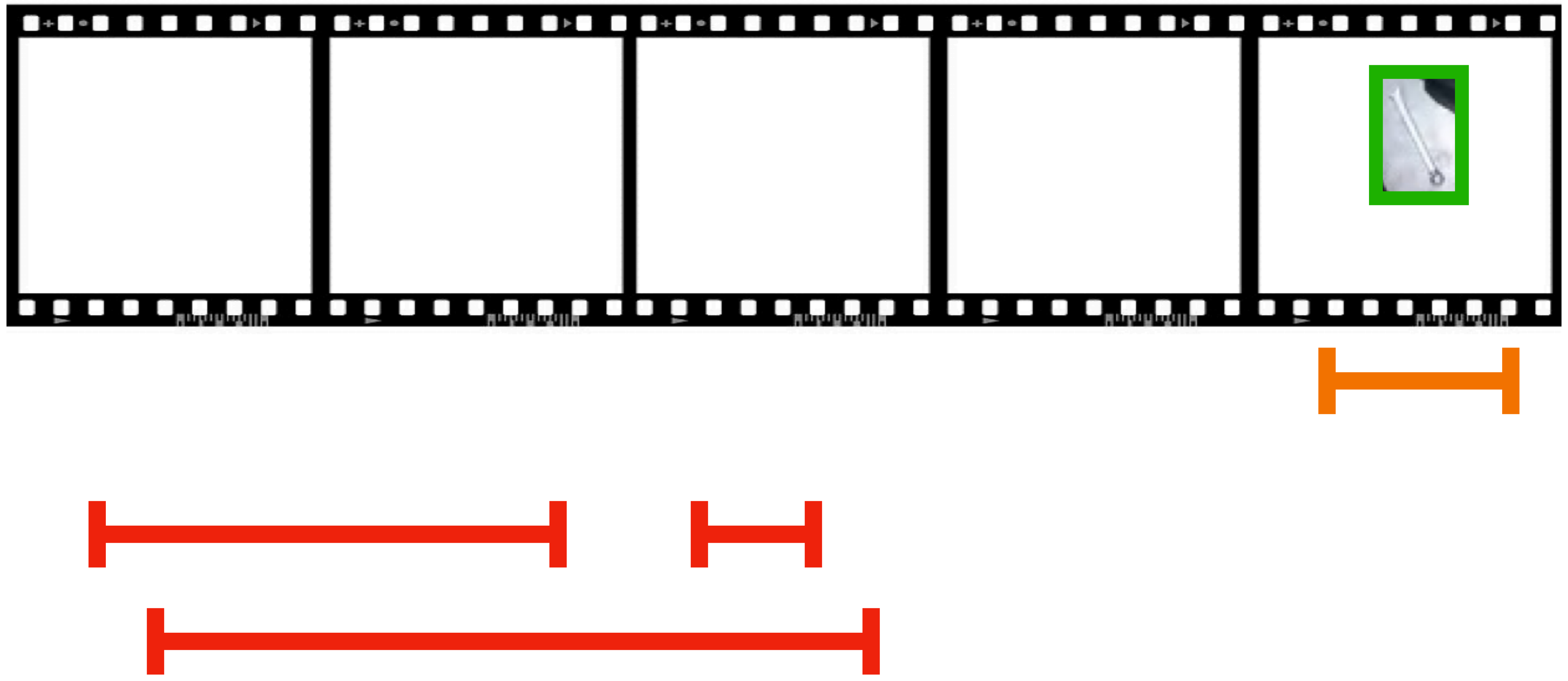
When and **where** did I last see  ?



When did I put the ***spanner***?

When did I ***repair small equipment***?

III. **Synergy** among video tasks



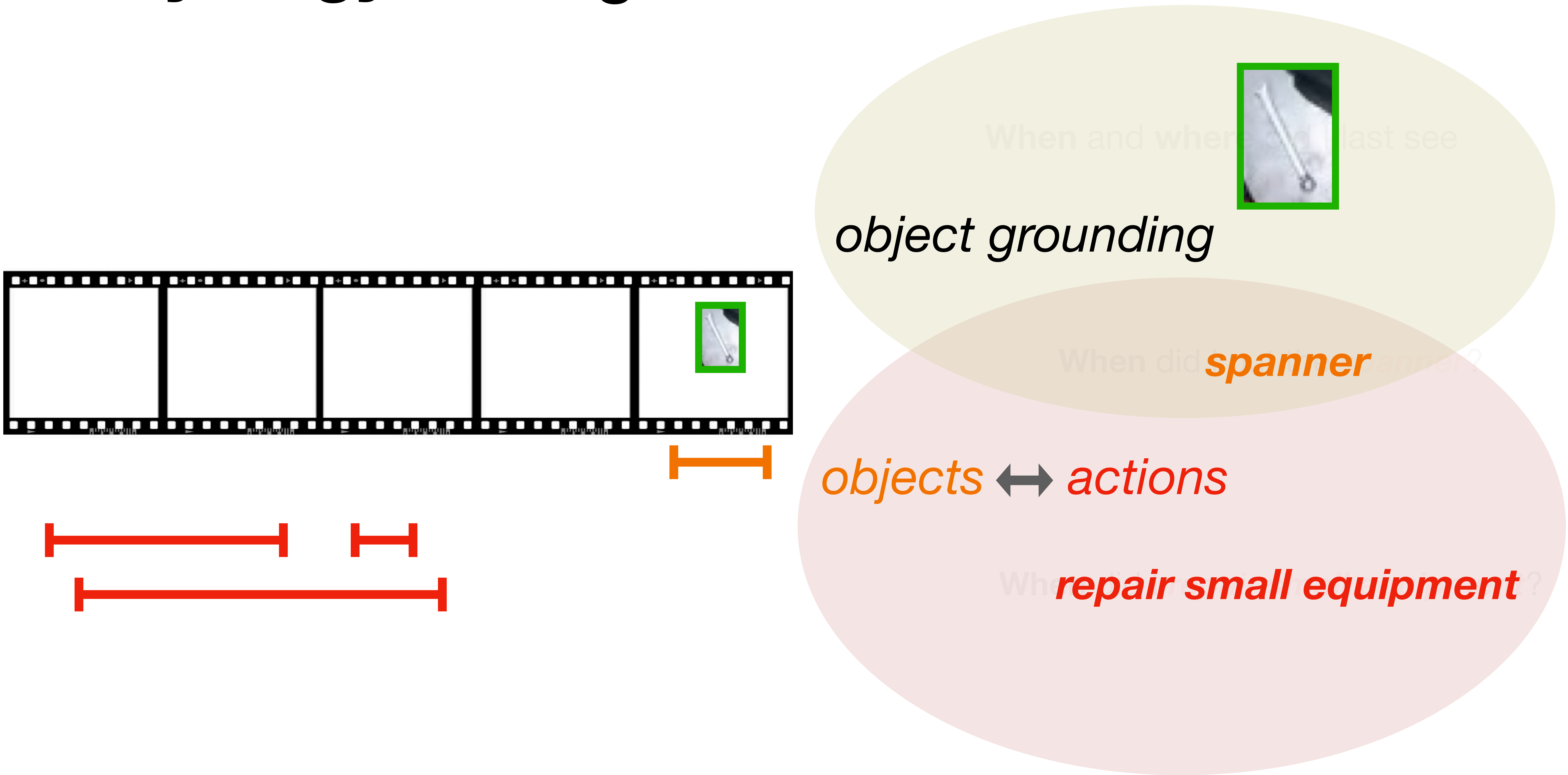
spanner

repair small equipment

Heterogeneous nature of **output**

Multi-modal nature of **input queries**

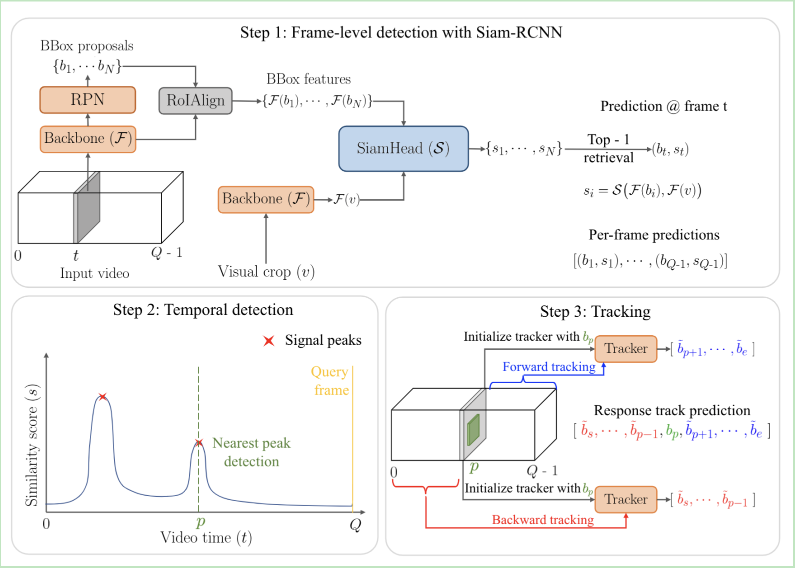
III. Synergy among video tasks



III. **Synergy** among video tasks

Video object tracking

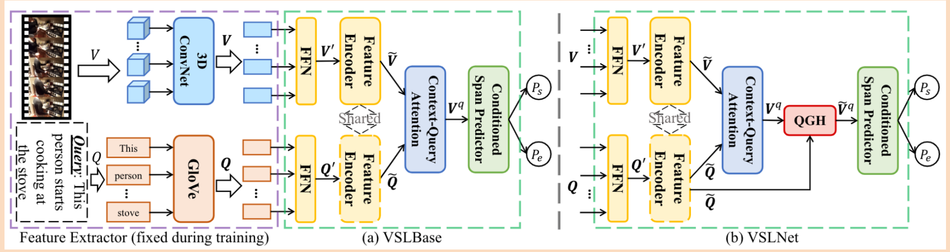
Grauman et al. 2022
Xu et al. 2022



Grauman, Kristen, et al. "Ego4d: Around the world in 3,000 hours of egocentric video." In CVPR 2022.

Video language grounding

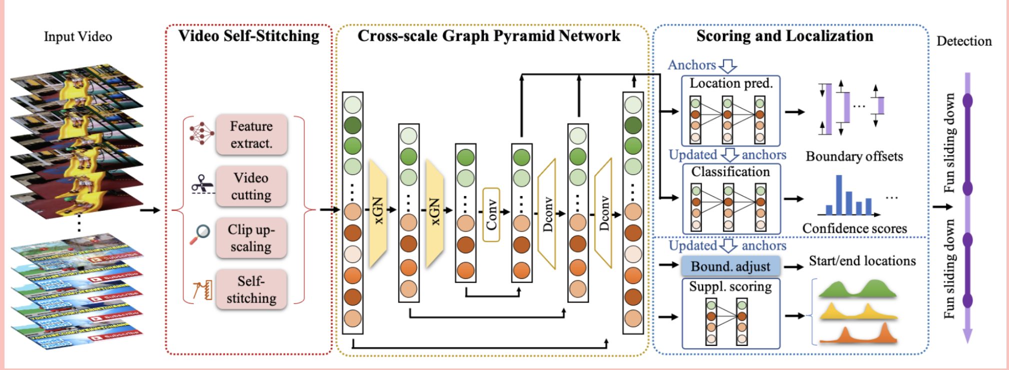
Liu et al. 2022
Zhang et al. 2021
Yang et al. 2022



Zhang, Hao, et al. "Span-based localizing network for natural language video localization." arXiv:2004.13931 (2020).

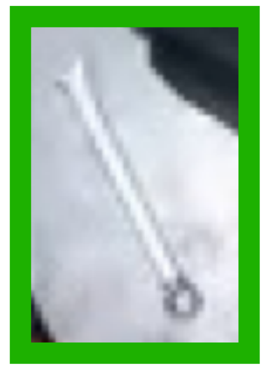
Action detection

Yang et al. 2020
Zhao et al. 2021
Liu et al. 2022
Zhang et al. 2022



Zhao, Chen, Ali K. Thabet, and Bernard Ghanem. "Video self-stitching graph network for temporal action localization." In ICCV 2021.

Visual Query (VQ2D)

When and **where** did I last see  ?

Natural Language Query (NLQ)

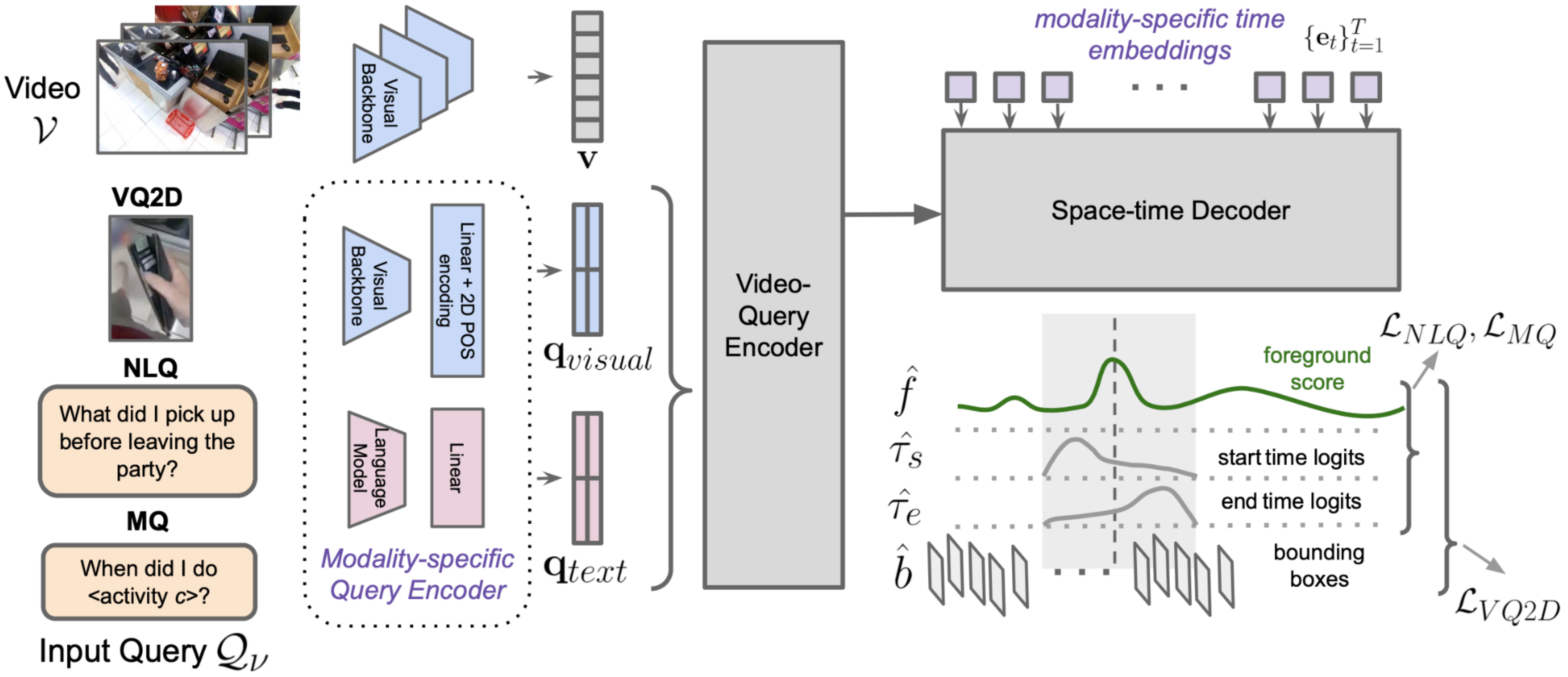
When did I put the ***spanner***?

Moment Query (MQ)

When did I ***repair small equipment***?

III. A **Unified** approach for heterogenous tasks

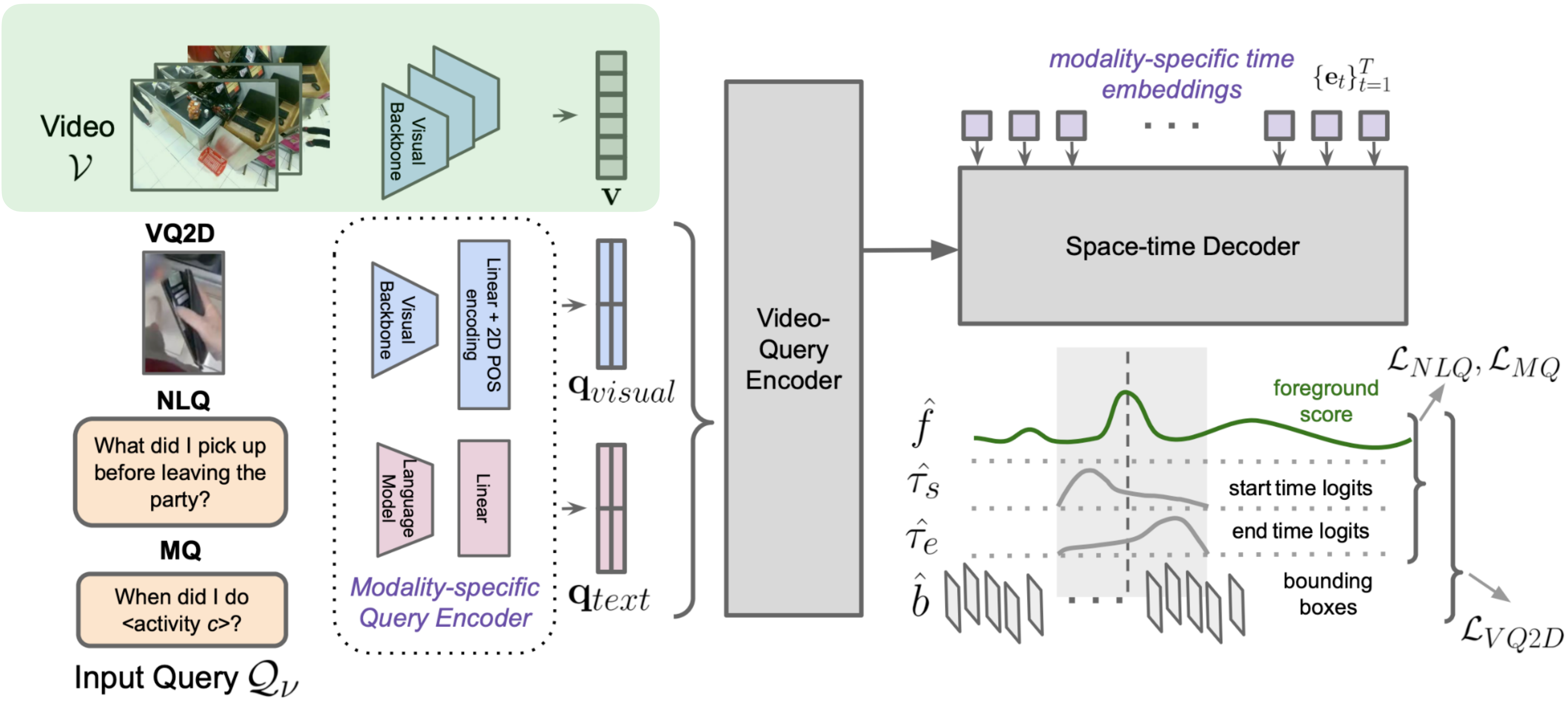
Transformer-based encoder-decoder architecture



III. A **Unified** approach for heterogenous tasks

Transformer-based encoder-decoder architecture

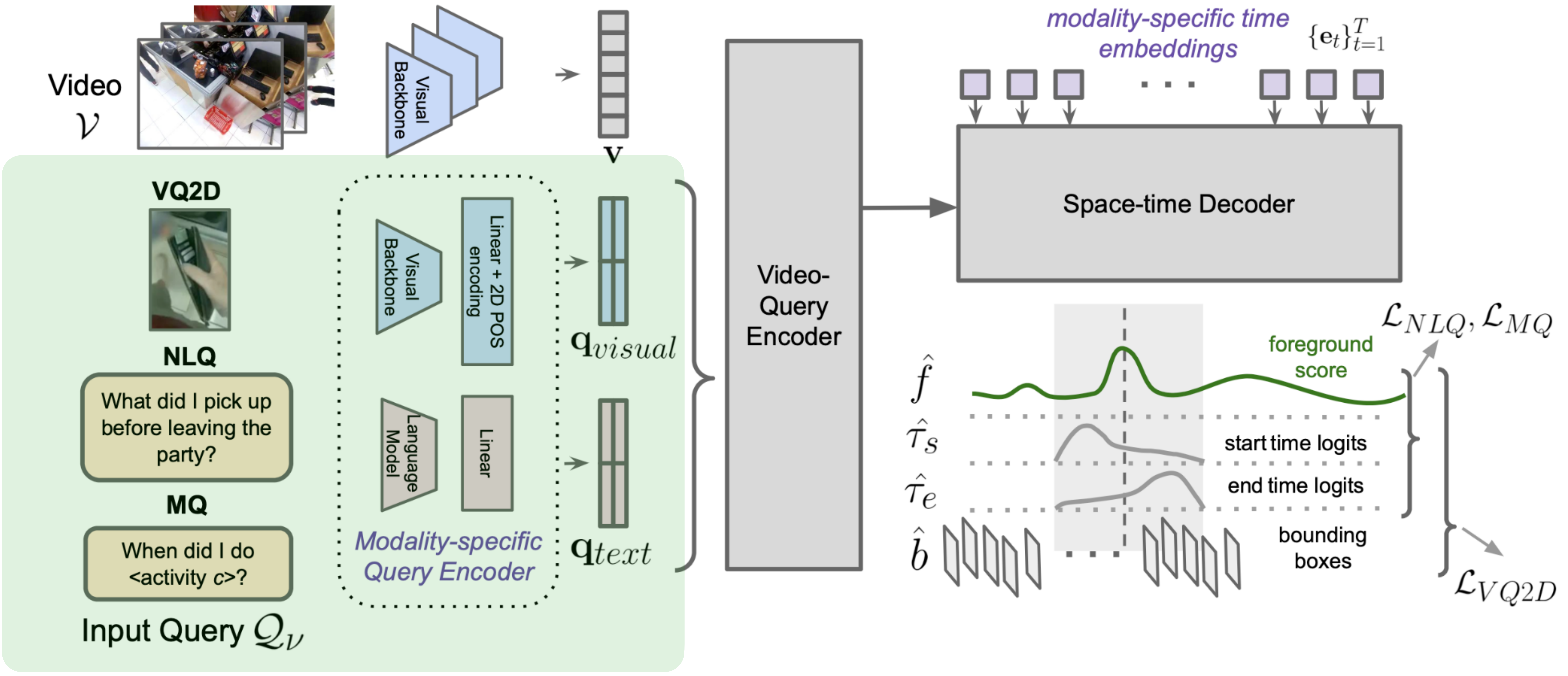
2D CNN



III. A **Unified** approach for heterogenous tasks

Transformer-based encoder-decoder architecture

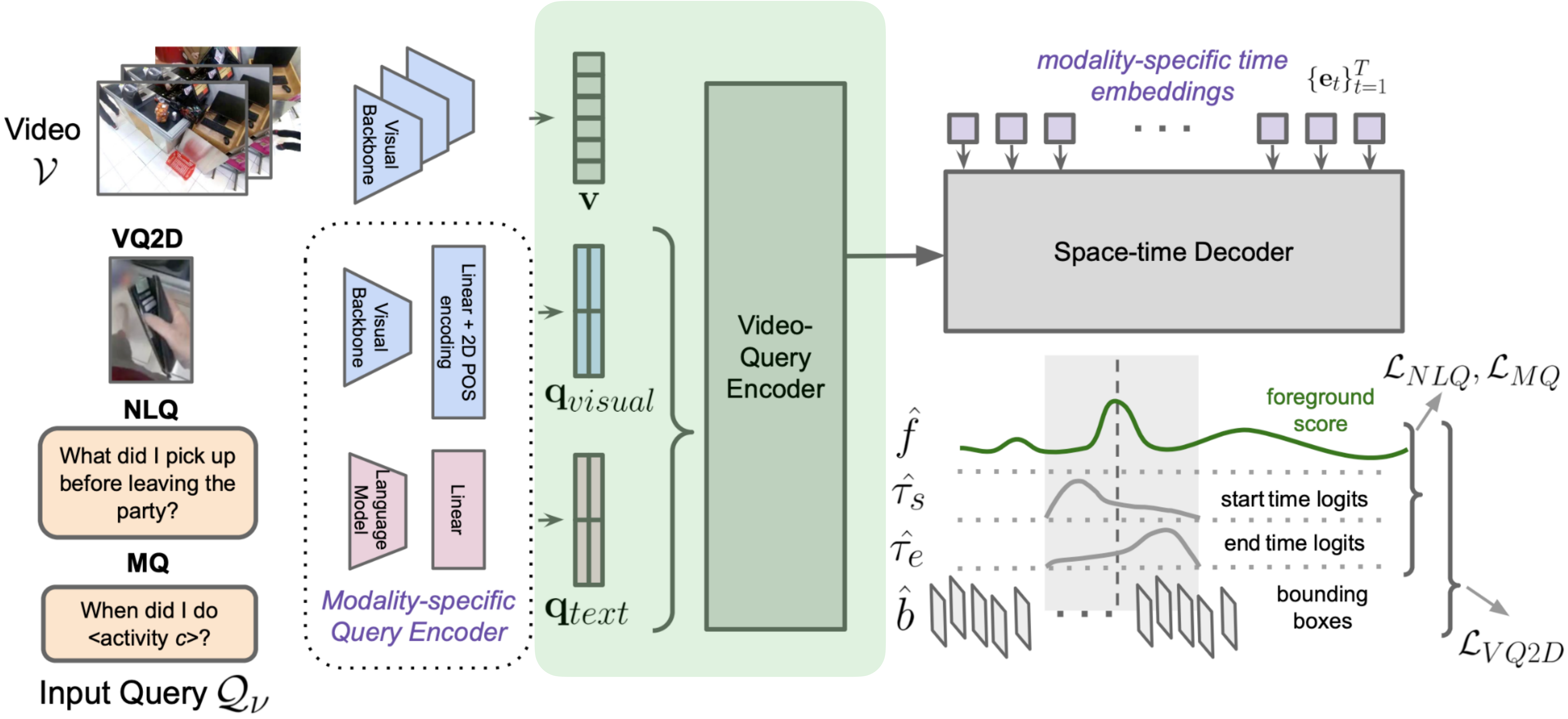
2D CNN



III. A **Unified** approach for heterogenous tasks

Transformer-based encoder-decoder architecture

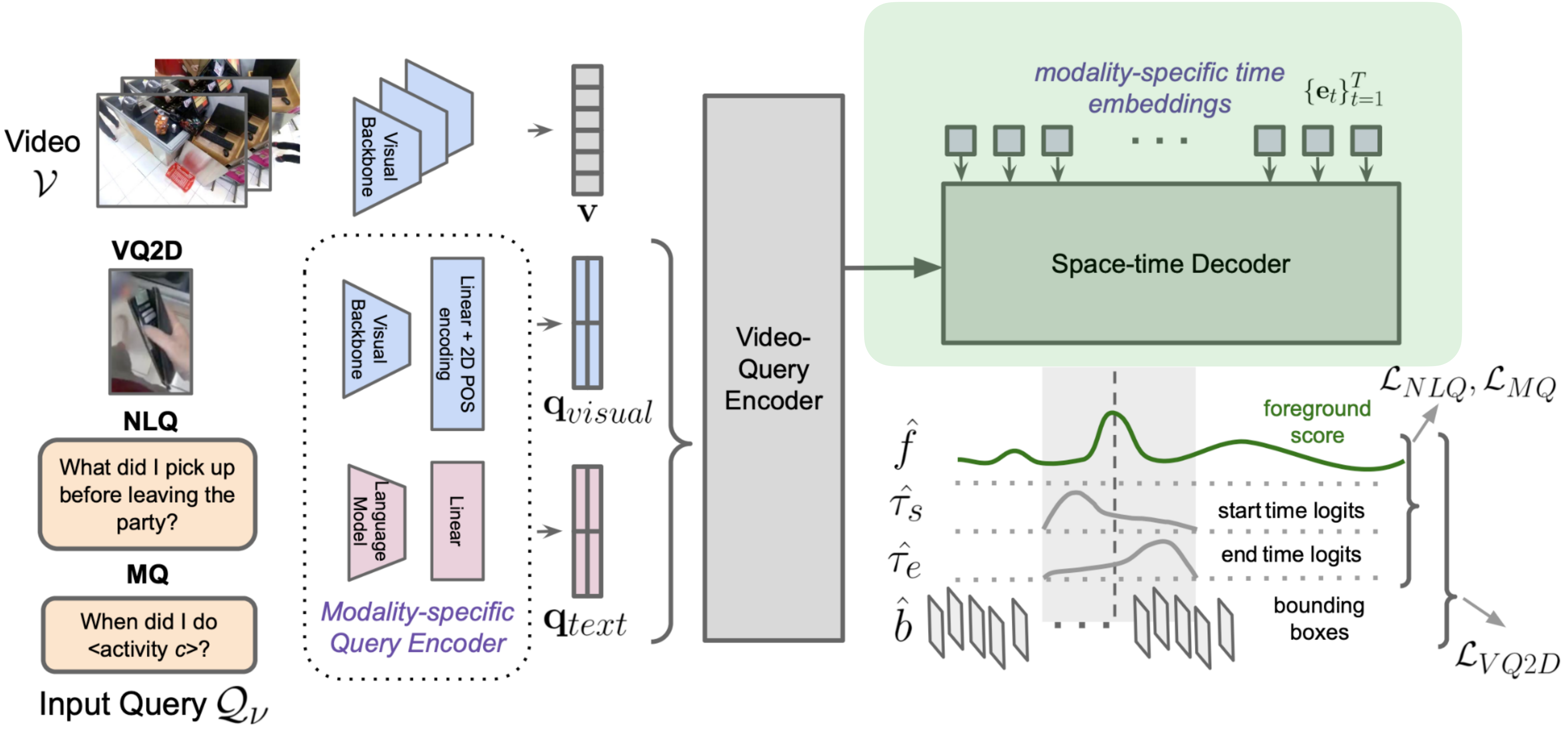
2D CNN



III. A **Unified** approach for heterogenous tasks

Transformer-based encoder-decoder architecture

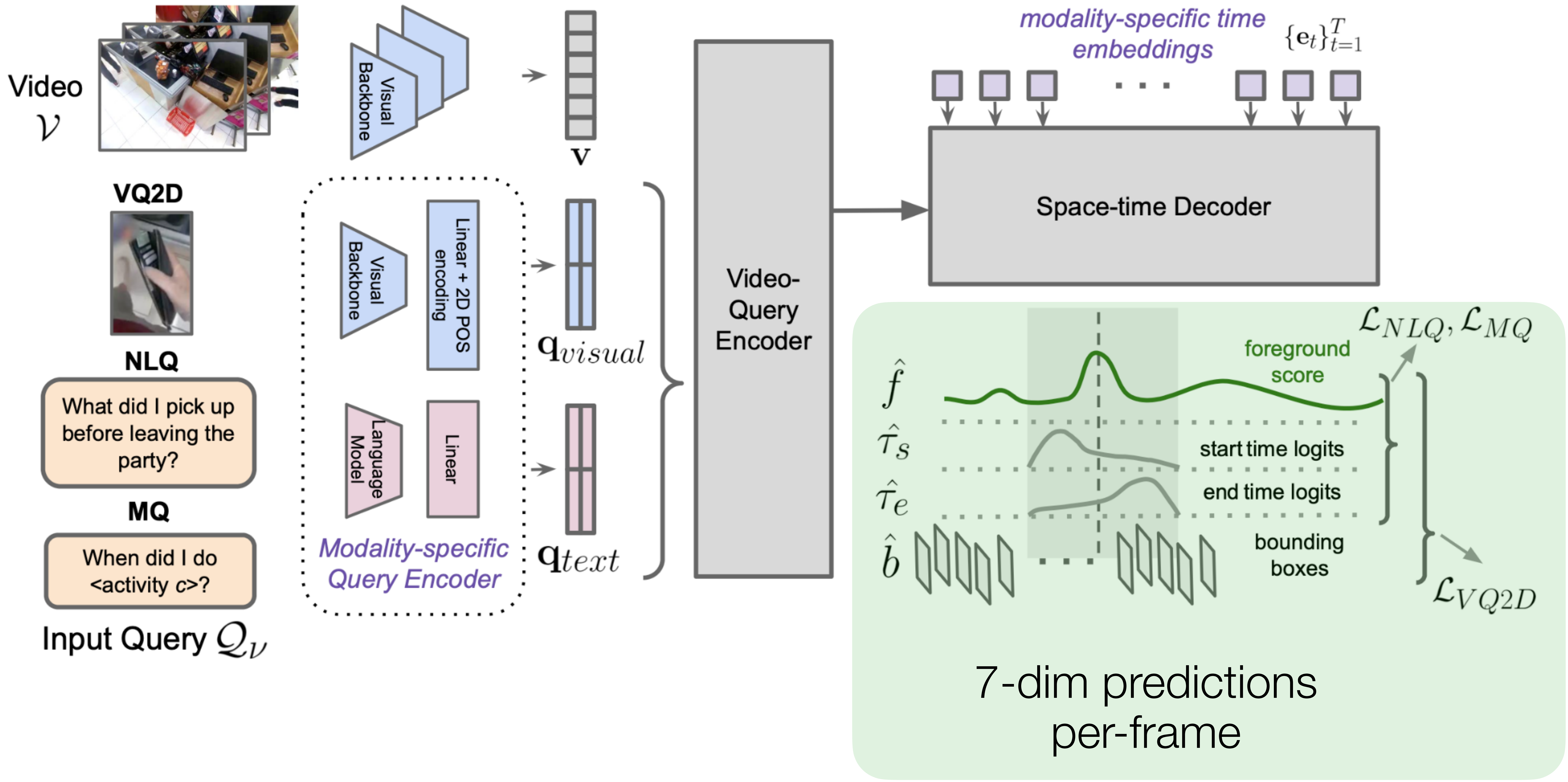
2D CNN



III. A **Unified** approach for heterogenous tasks

Transformer-based encoder-decoder architecture

2D CNN



III. Multi-task learning leads to *cross-task transfer*

We train on three tasks using multi-task learning

III. Multi-task learning leads to *cross-task transfer*

We train on three tasks using multi-task learning

		R@5, tIoU=0.5		
Category	Template	NLQ only	All-Tasks	Gain (in %)
Objects	Where is object X before / after event Y?	6.21	7.30	+17.50
	Where is object X?	10.29	13.42	+30.43
	What did I put in X?	5.43	7.67	+41.18
	How many X's? (quantity)	17.67	23.67	+33.96
	What X did I Y?	9.94	13.78	+38.71
	In what location did I see object X?	10.24	11.95	+16.67
	What X is Y?	10.13	12.42	+22.58
	State of an object	**11.31**	**22.02**	**+94.74**
	Where is my object X?	**6.49**	**11.69**	**+80.00**
Place	Where did I put X?	5.43	7.67	+41.18
People	Who did I interact with when I did activity X?	12.75	11.76	-7.69
	Who did I talk to in location X?	15.66	16.87	+7.69
	When did I interact with person with role X?	4.00	4.00	0.00

Language grounding (NLQ)
benefits
from Video tracking (VQ2D) task

III. Multi-task learning generalizes to *unseen* tasks

Input	**Output**	**Task**
Visual	Spatio-Temporal	VQ2D
Language / Class-label	Temporal	NLQ / MQ
Language	Spatio-Temporal	No supervision

III. Multi-task learning generalizes to *unseen* tasks

Input	Output	Task
Visual	Spatio-Temporal	VQ2D
Language / Class-label	Temporal	NLQ / MQ
Language	Spatio-Temporal	No supervision

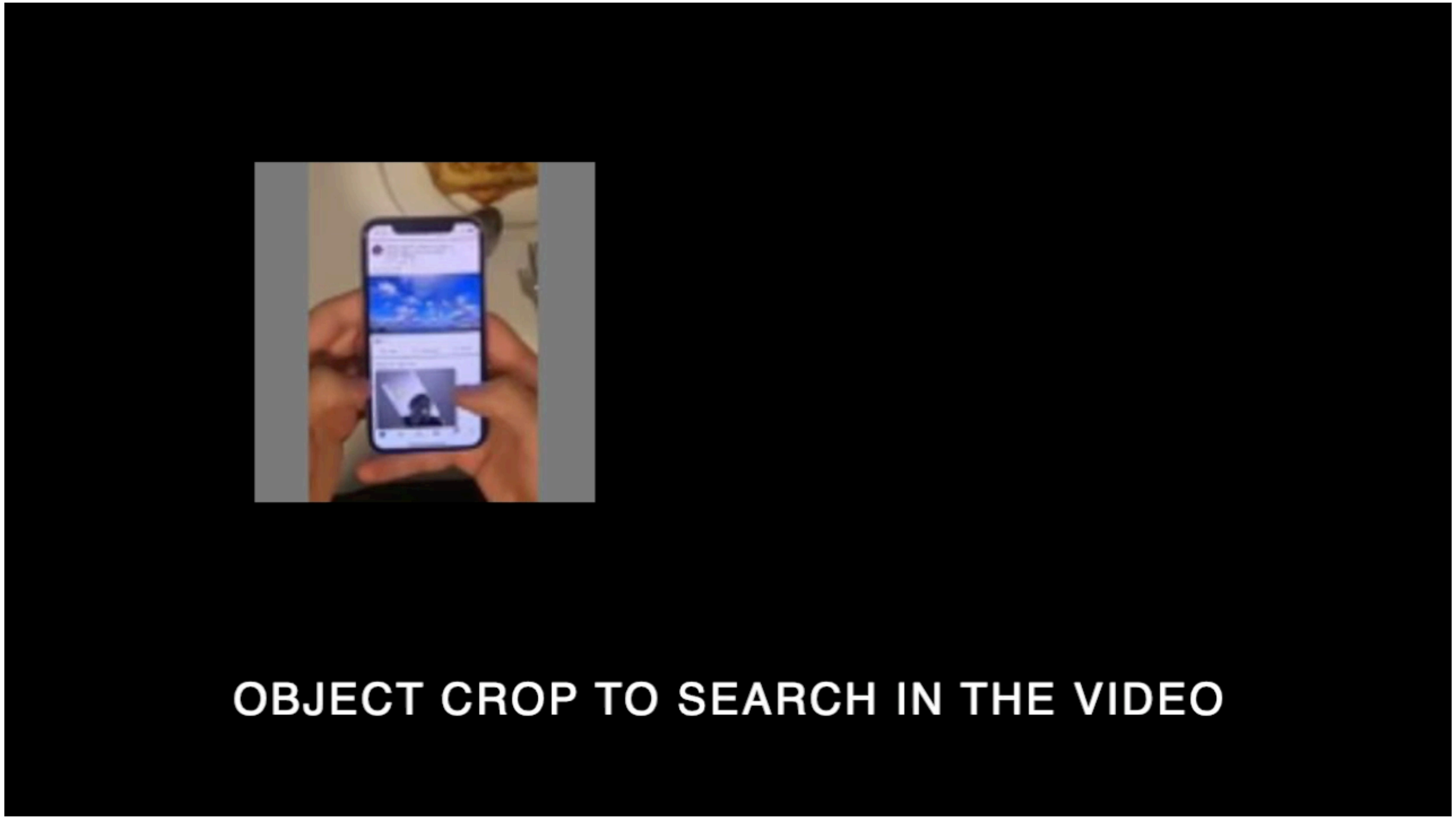
Evaluated on **spatially annotated** subset of validation data (=10 videos / question)

Model	Spatio-temporal				Temporal
	spatial branch	stIoU=0.3 R@1	stIoU=0.3 R@5	mean stIoU	mean tIoU
NLQ-only	N/A	-	-	-	5.35
MINOTAUR (All-Tasks)	random boxes	0 ± 0	0 ± 0	0.40 ± 0.04	**8.35**
	random centered boxes	0 ± 0	0.47 ± 0.38	1.25 ± 0.03	
	All-Tasks	**2.33**	**4.65**	**2.27**	

Zero-shot spatio-temporal localization on NLQ task

Meaningful performance on ***unseen*** task

III. A video example result



Chapter IV

Image tasks

0

P Bateni, **R Goyal**, V Masrani, F Wood and L Sigal. “Improved Few-Shot Visual Classification”. In CVPR 2020.

I

S Khandelwal*, **R Goyal*** and L Sigal. “UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation”. In CVPR 2021.

II

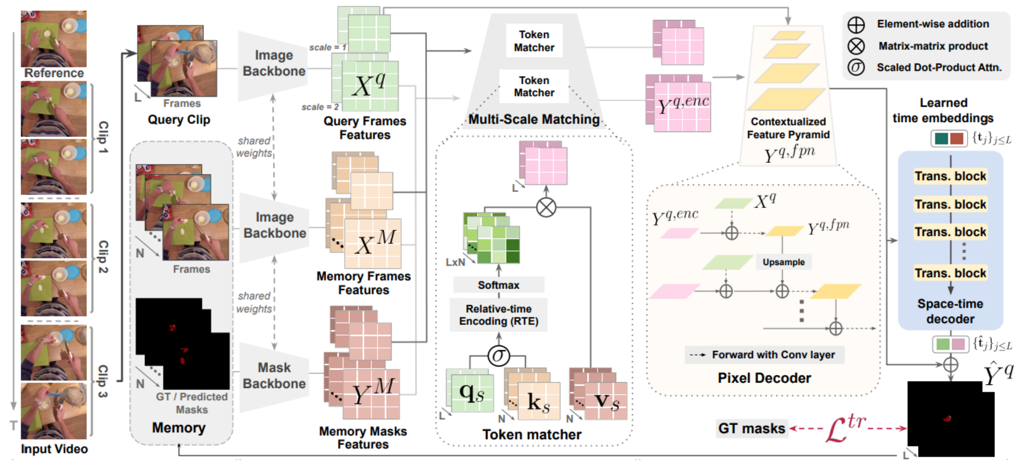
R Goyal and L Sigal. “A Simple Baseline for Weakly-Supervised Human-centric Relation Detection”. In BMVC 2021.

Video tasks

III

R Goyal, E Mavroudi, X Yang, S Sukhbaatar, L Sigal, M Feiszli, L Torresani, D Tran . “MINOTAUR: Multi-task Video Grounding From Multimodal Queries”. arXiv. 2302.08063.

IV



R Goyal*, WC Fan*, M Siam, L Sigal, . “TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking”. In WACV 2025.

V

NB Gundavarapu*, L Friedman*, **R Goyal***, C Hegde*, E Agustsson, S M Waghmare, M Sirotenko, MH Yang, T Weyand, B Gong, L Sigal . “Extending Video Masked Autoencoders to 128 frames”. In NeurIPS 2024.

** denotes equal contribution*

IV. Contributions

- We explore spatio-temporal **video object segmentation** with **deformations** on **long videos**
- We find **time-coded memory** and **transformation-aware loss** to be crucial components

IV. Video Object Segmentation under **Transformations**

Large deformations

State changes and/or multiple instances

Small objects
(<1% relative area)

Objects can get lost at a typical feature map resolution (=1/16th)

Long videos (>20 secs)

Drift in tracking

IV. Video Object Segmentation under **Transformations**

Large deformations

State changes and/or multiple instances

Dense propagation that takes *semantics* into account

Yang, Z., et al. Associating objects with transformers for video object segmentation. NeurIPS 2021
Oh, S.et al. Video object segmentation using space-time memory networks. ICCV 2019

Small objects (<1% relative area)

Objects can get lost at a typical feature map resolution (=1/16th)

Multi-scale feature maps

Karim, Rezaul, et al. "MED-VT: Multiscale encoder-decoder video transformer with application to object segmentation." CVPR 2023.
Seong, H., et al. Hierarchical memory matching network for video object segmentation. ICCV 2021.

Long videos (>20 secs)

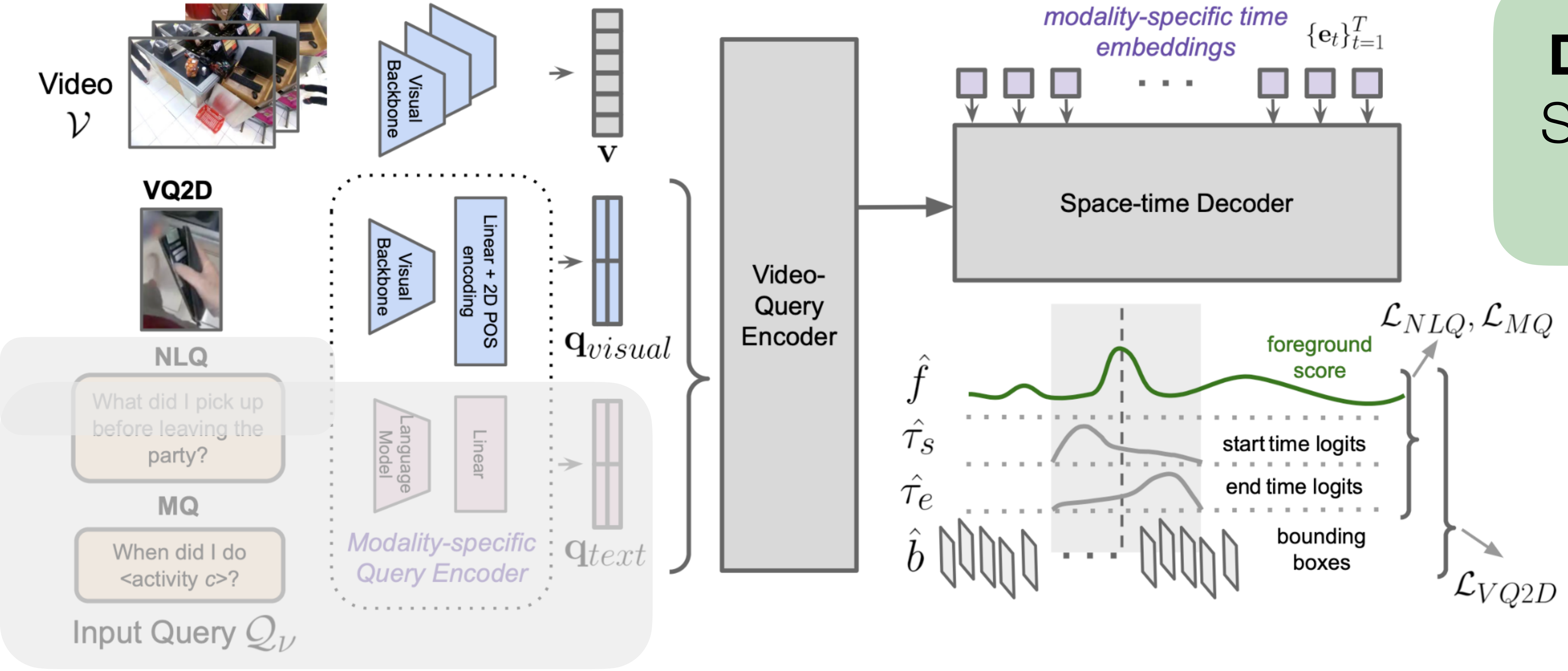
Drift in tracking

Robust **memory** module to track changes *long-term*

Cheng, Ho Kei, et al. "Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model." ECCV 2022.
Hong, Lingyi, et al. "Lvos: A benchmark for long-term video object segmentation." ICCV 2023.

IV. Comparison to the previous chapter

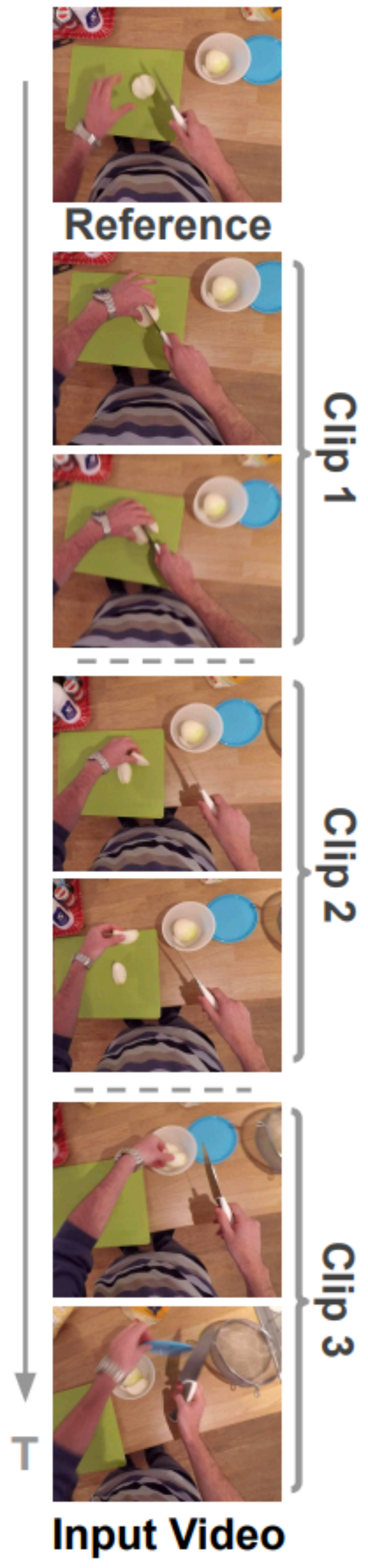
Visual query: Segmentation mask



Decode only Segmentation mask

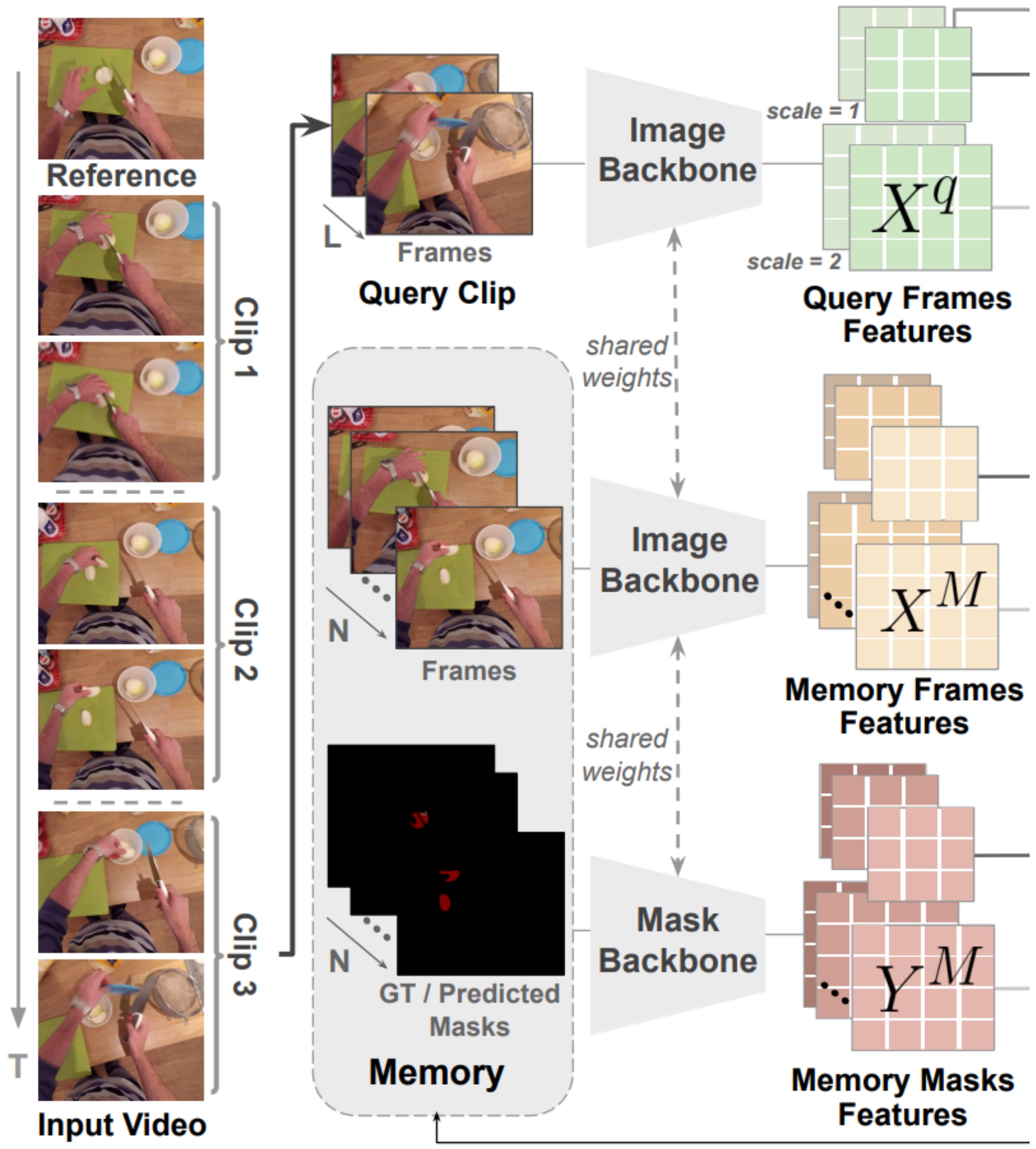
+ Visual encoding at **multiple scales**
+ **Memory module** to store past predictions
+ **Dense matching** to **propagate memory**

IV. Multi-scale encoder-decoder design



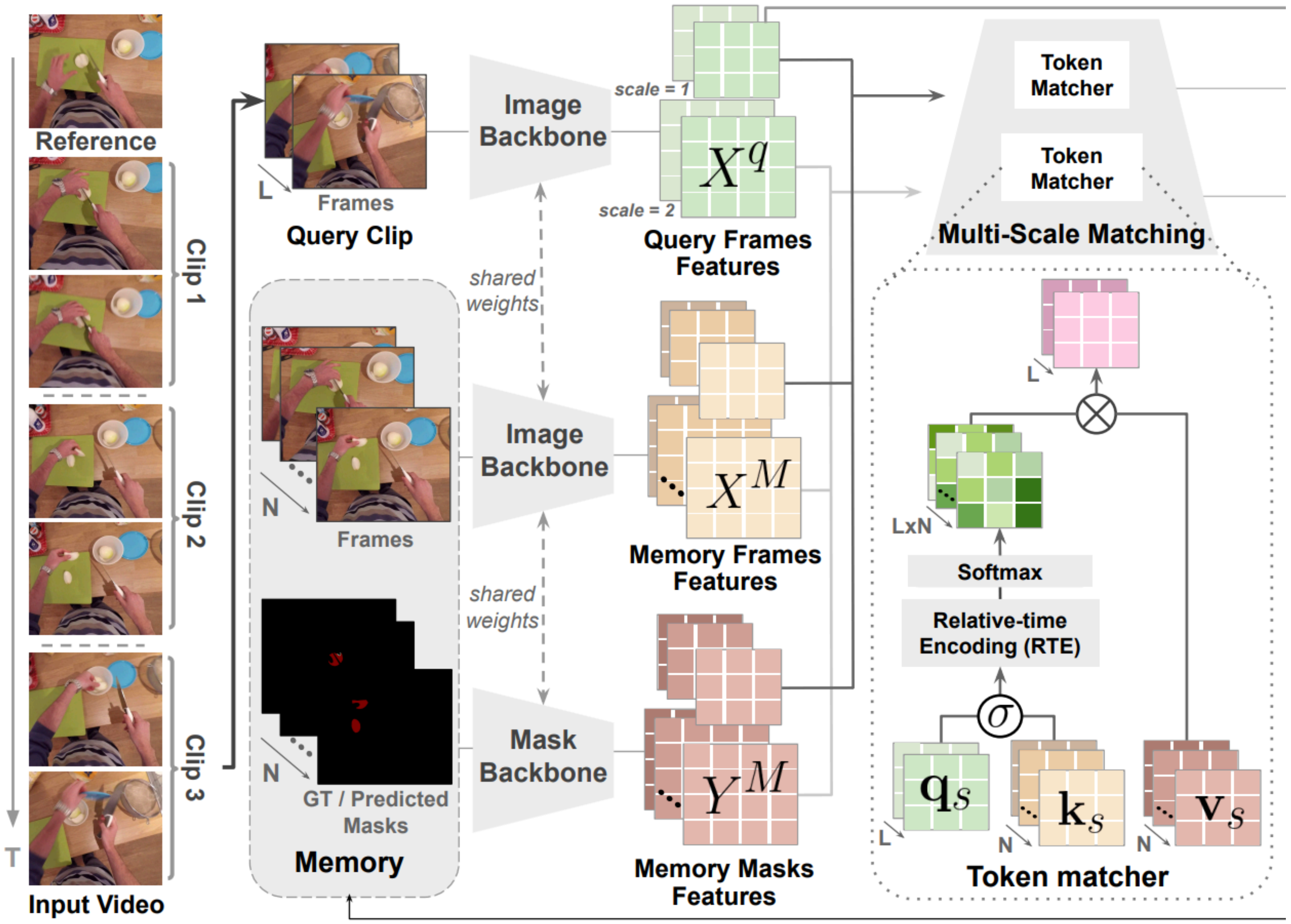
Divide input video non-overlapping **clips** (each of length **L frames**)

IV. Multi-scale encoder-decoder design



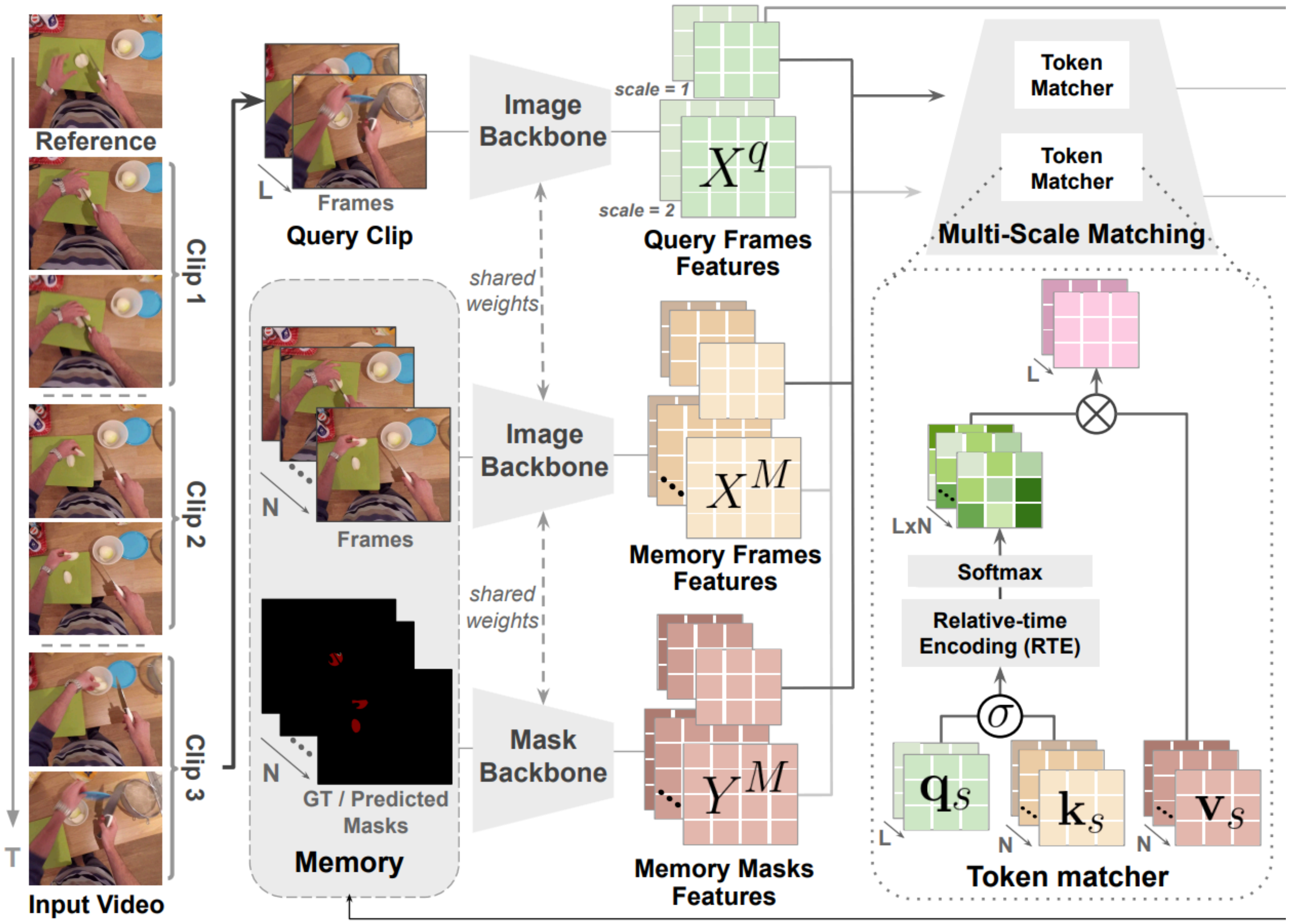
Encode a query clip (=***L***) using image backbone and retrieve Memory (=***N***)

IV. Multi-scale encoder-decoder design



Perform **dense matching** b/w Query clip (=***L***) and Memory (=***N***) using attention over multiple-scales

IV. Multi-scale encoder-decoder design



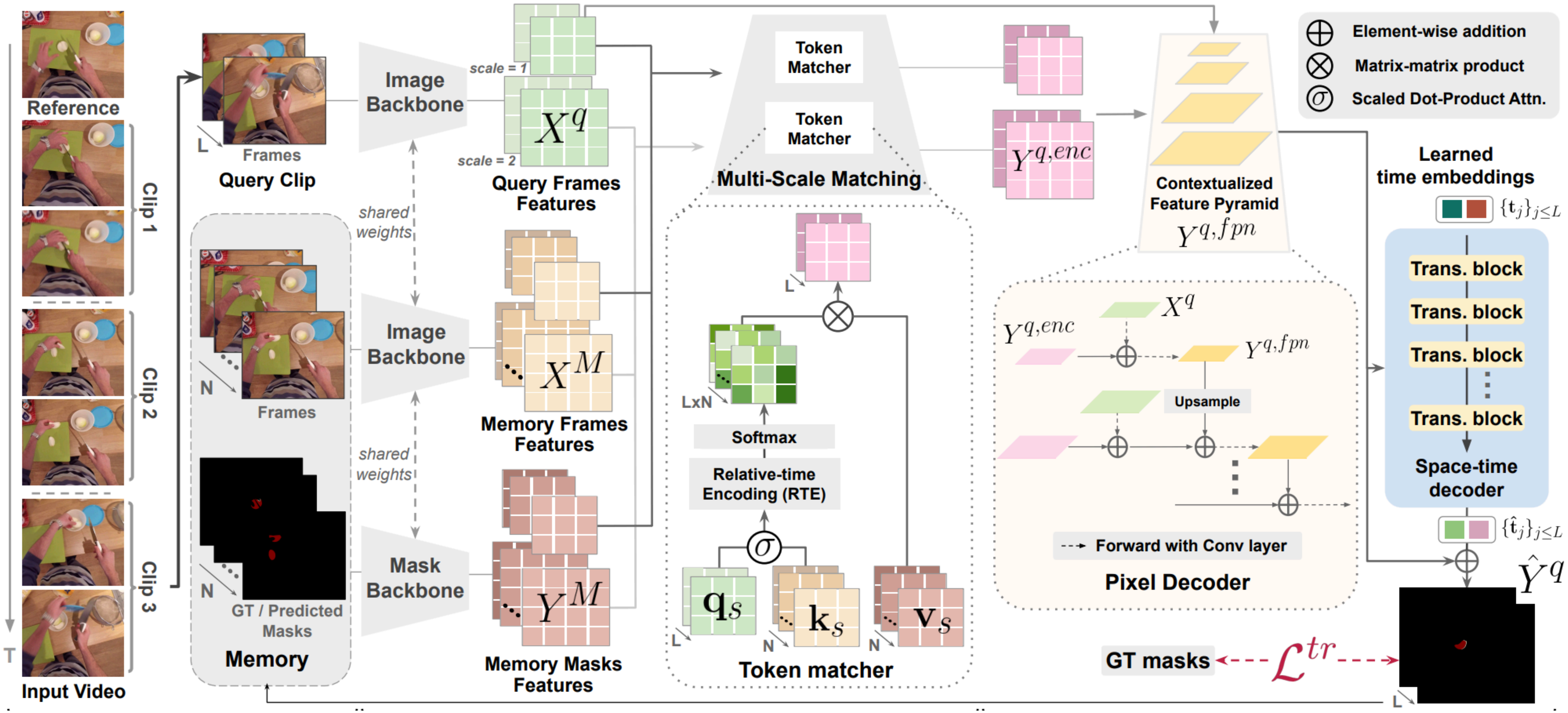
#1 Relative Time Encoding (RTE)

Idea: Modulate association from memory based on **time / recency**

#2 Transformation-aware Loss

Idea: Place greater emphasis on frames that contains objects **undergoing transformations**

IV. Multi-scale encoder-decoder design



Use Pixel Decoder to obtain **Contextualized Feature Pyramid**, followed by Space-Time Decoder to obtain mask predictions

IV. Results

Approach	Pre-training	VOST	
		$\mathcal{J}_{tr}$	$\mathcal{J}$
OSMN-Match [50]	Static + DAVIS [38]	7.0	8.7
OSMN-Tune [50]	Static + DAVIS	17.6	23.0
CRW [23]	IN1K [14] + DAVIS	13.9	23.7
HODOR-Img [1]	COCO [32] + DAVIS	13.9	26.2
HODOR-Vid [1]	COCO + DAVIS	25.4	37.1
CFBI [51]	IN1K + COCO + DAVIS	32.0	45.0
CFBI+ [53]	Static + DAVIS	32.6	46.0
XMem [8]	Static + DAVIS	33.8	44.1
AOT† [52]	Static	35.1	47.1
AOT [52]	Static + DAVIS	36.4	48.7
TAM-VT(Ours)	Static	36.5	48.2
TAM-VT(Ours)	Static + DAVIS	**37.7**	**49.3**

(a) Val performance on VOST

Outperforms prior approaches

IV. Results

Approach	Pre-training	VOST	
		$\mathcal{J}_{tr}$	$\mathcal{J}$
OSMN-Match [50]	Static + DAVIS [38]	7.0	8.7
OSMN-Tune [50]	Static + DAVIS	17.6	23.0
CRW [23]	IN1K [14] + DAVIS	13.9	23.7
HODOR-Img [1]	COCO [32] + DAVIS	13.9	26.2
HODOR-Vid [1]	COCO + DAVIS	25.4	37.1
CFBI [51]	IN1K + COCO + DAVIS	32.0	45.0
CFBI+ [53]	Static + DAVIS	32.6	46.0
XMem [8]	Static + DAVIS	33.8	44.1
AOT† [52]	Static	35.1	47.1
AOT [52]	Static + DAVIS	36.4	48.7
TAM-VT(Ours)	Static	36.5	48.2
TAM-VT(Ours)	Static + DAVIS	**37.7**	**49.3**

(a) Val performance on VOST

	OSMN Tune [50]	CFBI+ [53]	HODOR Vid [1]	AOT [52]	TAM-VT(Ours) (diff with AOT [52])
All	17.6	32.6	25.4	36.4	**37.7** (+1.3)
LNG	12.4	30.4	25.0	34.7	**41.9** (+7.2)
MI	14.7	26.4	20.6	27.2	**29.2** (+2.0)
SM	14.4	23.3	16.6	24.7	**28.4** (+3.7)

(b) Quantitative analysis of factors on VOST.

LNG: Long videos (>20 sec)
MI: Multiple instances
SM: Small objects (<0.5% rel. area)

Outperforms prior approaches

Meaningful gains over **Long videos (LNG)** and **Small objects (SM)**

IV. Results

Approach	Pre-training	VOST $\mathcal{J}_{tr}$	VOST $\mathcal{J}$
OSMN-Match [50]	Static + DAVIS [38]	7.0	8.7
OSMN-Tune [50]	Static + DAVIS	17.6	23.0
CRW [23]	IN1K [14] + DAVIS	13.9	23.7
HODOR-Img [1]	COCO [32] + DAVIS	13.9	26.2
HODOR-Vid [1]	COCO + DAVIS	25.4	37.1
CFBI [51]	IN1K + COCO + DAVIS	32.0	45.0
CFBI+ [53]	Static + DAVIS	32.6	46.0
XMem [8]	Static + DAVIS	33.8	44.1
AOT† [52]	Static	35.1	47.1
AOT [52]	Static + DAVIS	36.4	48.7
TAM-VT(Ours)	Static	36.5	48.2
TAM-VT(Ours)	Static + DAVIS	**37.7**	**49.3**

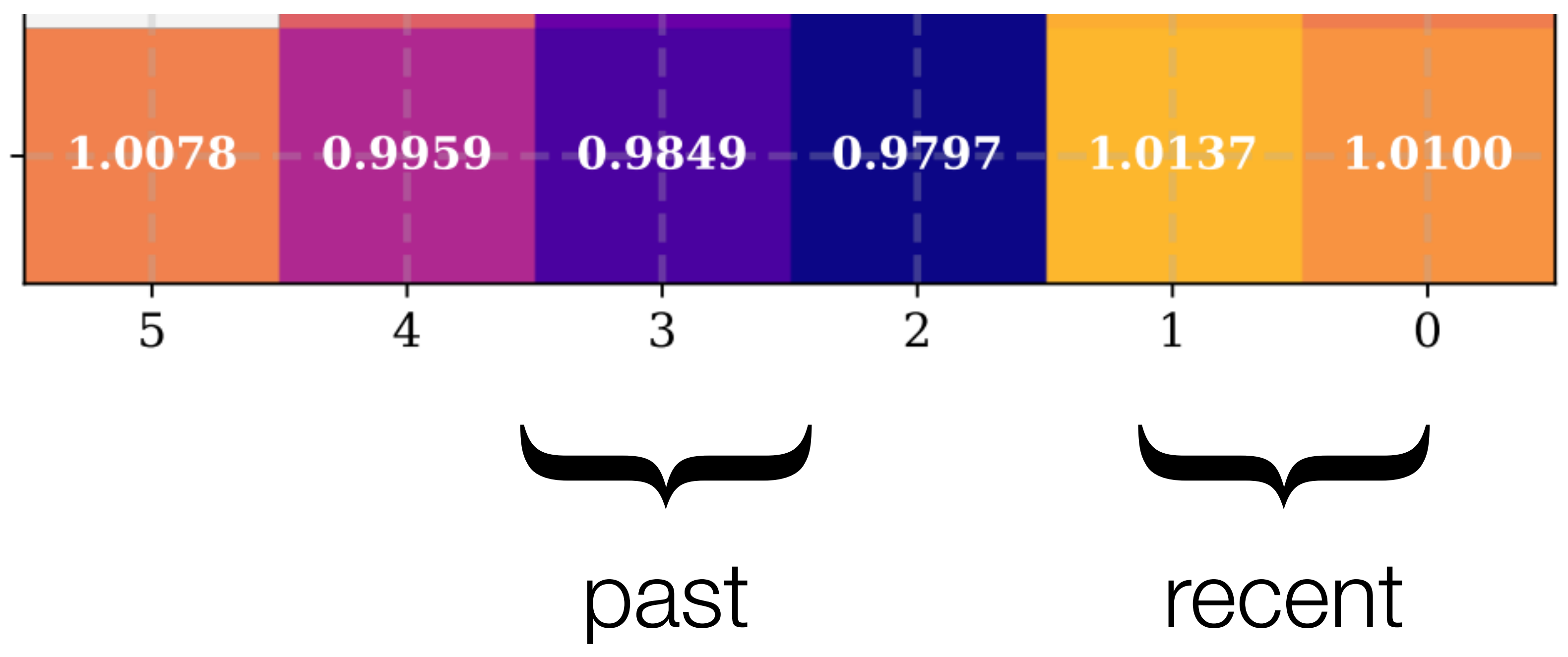
(a) Val performance on VOST

	OSMN Tune [50]	CFBI+ [53]	HODOR Vid [1]	AOT [52]	TAM-VT(Ours) (diff with AOT [52])
All	17.6	32.6	25.4	36.4	**37.7** (+1.3)
LNG	12.4	30.4	25.0	34.7	**41.9** (+7.2)
MI	14.7	26.4	20.6	27.2	**29.2** (+2.0)
SM	14.4	23.3	16.6	24.7	**28.4** (+3.7)

(b) Quantitative analysis of factors on VOST.

LNG: Long videos (>20 sec)
MI: Multiple instances
SM: Small objects (<0.5% rel. area)

Relative Time Encoding (RTE)



Outperforms prior approaches

Meaningful gains over **Long videos (LNG)** and **Small objects (SM)**

RTE learns **higher weights** for **recent** and **first** frame

IV. A video example result

Conclusion

Fully-supervised

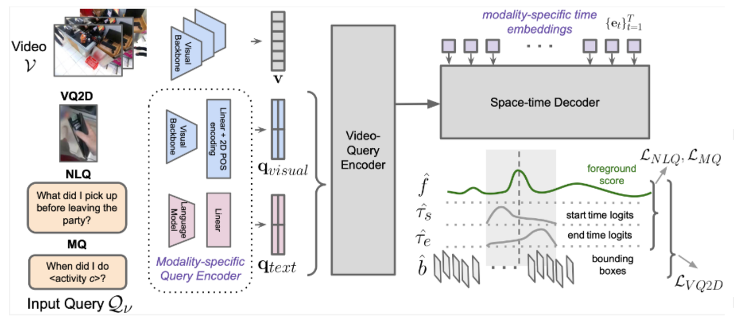
Multi-task

Semi-supervised

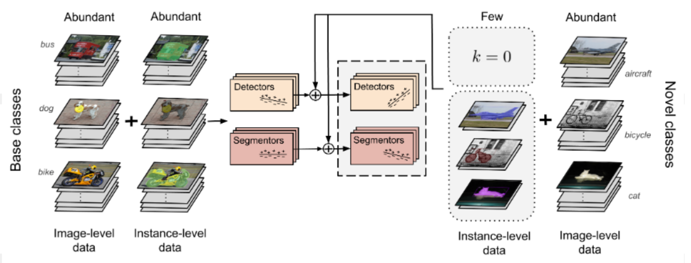
Few-shot

Weakly-supervised

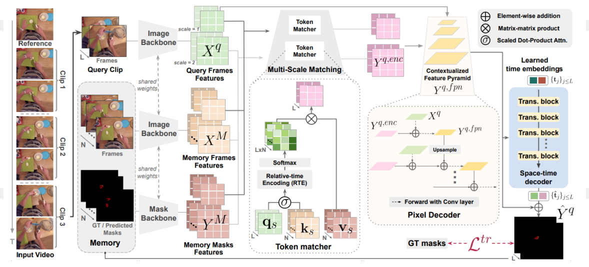
Supervision



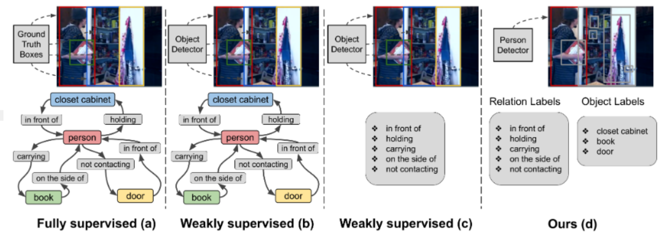
R Goyal, E Mavroudi, X Yang, S Sukhbaatar, L Sigal, M Feiszli, L Torresani, D Tran . "MINOTAUR: Multi-task Video Grounding From Multimodal Queries". arXiv. 2302.08063.



S Khandelwal*, **R Goyal*** and L Sigal. "UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation". In CVPR 2021.



R Goyal*, WC Fan*, M Siam, L Sigal, . "TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking". In WACV 2025.



R Goyal and L Sigal. "A Simple Baseline for Weakly-Supervised Human-centric Relation Detection". In BMVC 2021.

Granularity of the tasks

We explore **data-efficient learning** approaches for **visual structured prediction tasks**

Questions?

Thanks to everyone with whom I had the pleasure of collaborating* during my PhD

UBC

PhD Supervisor


Leonid Sigal

University collaborators


Frank Wood


Peyman Bateni


Siddhesh Khandelwal


Wan-Cyuan Fan


Mennatullah Siam


Tanzila Rahman

Meta AI

Internship, 2022

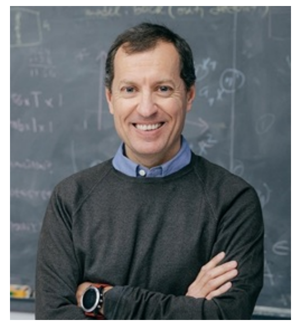

Effrosyni Mavroudi


Xitong Yang


Matt Feiszli


Sainbayar Sukhbaatar


Du Tran

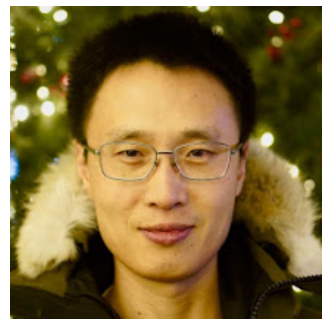

Lorenzo Torresani

Google Research

Internship, 2023


Nitesh B. Gundavarapu


Luke Friedman


Boqing Gong


Tobias Weyand


Mikhail Sirotenko


Ming-Hsuan Yang

** the list is not exhaustive, so please excuse me if I unintentionally excluded someone*