Multi-Modal Few-Shot Temporal Action Detection via Vision-Language Meta-Adaptation

- Under Review

UNIVERSITY OF SURREY

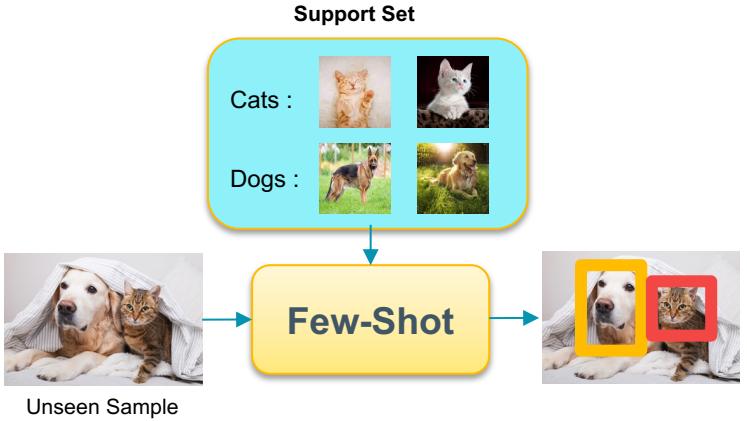
People-Centred AI
UNIVERSITY OF SURREY

PROBLEM DEFINITION

CVSSP
Centre for Vision, Speech and Signal Processing

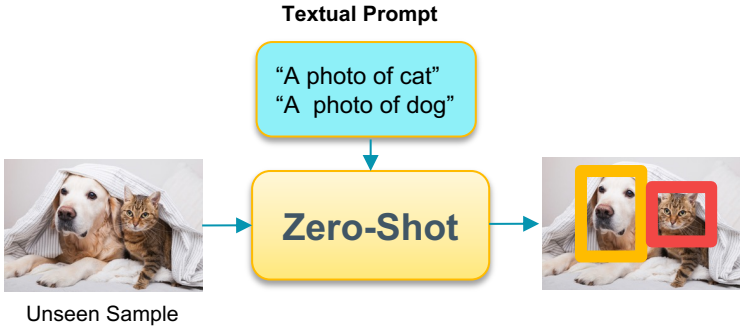
Few-Shot Learning

Detect a unseen novel class given few annotated support examples



Zero-Shot Learning

Detect a unseen novel class without any annotation



PROBLEM DEFINITION:
MULTI-MODAL FEW-SHOT LEARNING

CVSSP
Centre for Vision, Speech and Signal Processing

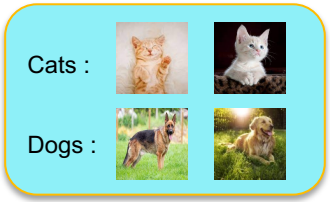
Few-Shot Learning + Zero-Shot Learning = Multi-Modal Few-Shot Learning

PROBLEM DEFINITION:
MULTI-MODAL FEW-SHOT LEARNING

CVSSP
Centre for Vision, Speech and Signal Processing

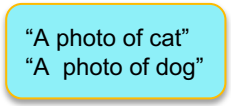
Few-Shot Learning + Zero-Shot Learning = Multi-Modal Few-Shot Learning

Support Set



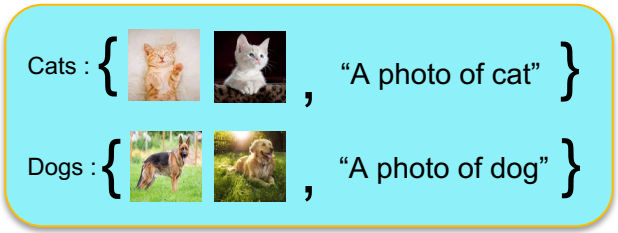
Images only

Textual Prompt



Text only

Support Set



Images and Text

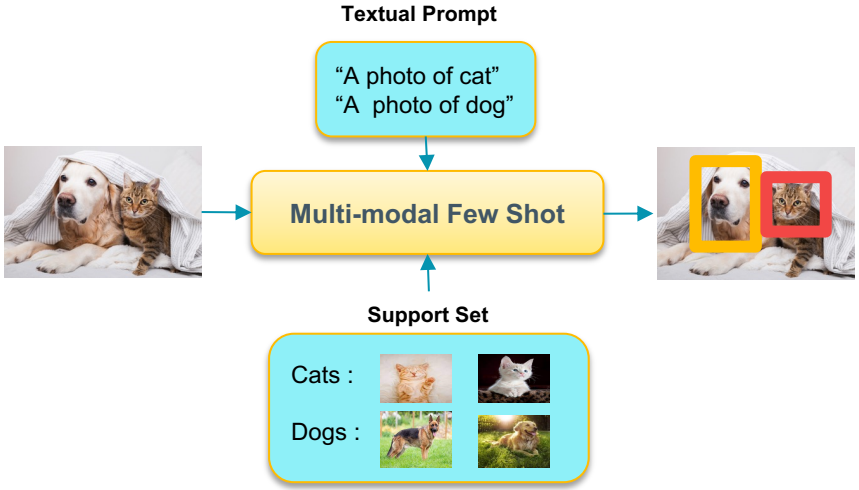
PROBLEM DEFINITION:
MULTI-MODAL FEW-SHOT LEARNING

CVSSP
Centre for Vision, Speech and Signal Processing

→ We are the first to propose this **multi-modal few-shot setting** for dense downstreams

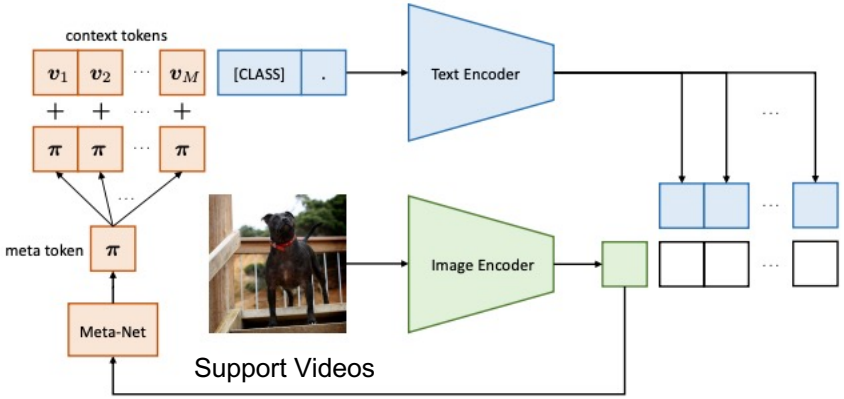
→ This is a **natural extension** of Few-Shot Learning using Vision-Language

→ **Stronger** Few-Shot performance because of large CLIP pre-training



VISION-LANGUAGE MODELS:
HOW TO MODEL MULTI-MODAL FEW-SHOT ?

CVSSP
Centre for Vision, Speech and Signal Processing



CoCOOP, CVPR21

Baseline Multi-Modal Few Shot

- Aligns Vision and Textual Modality
- Use Support Videos to learn meta-network
- Still needs CLIP[1] Tokenizer for visual samples

[1]Radford et al. , “Learning Transferable Visual Models From Natural Language Supervision".

UNSOLVED QUESTION:
HOW TO MODEL MULTI-MODAL FEW-SHOT ?

CVSSP
Centre for Vision, Speech and Signal Processing

Q1) Can we learn **task-specific parameters** instead of **full fine-tuning**

Q2) Can we better use the visual samples **without using the CLIP tokenizer**

Q3) Can we reduce the **intra-class variance problem**

Q4) **Which modality** to meta-learn ?

LITERATURE:
MULTI-MODAL FEW-SHOT

CVSSP
Centre for Vision, Speech and Signal Processing

Q1) “**Multimodal Few-Shot Learning with Frozen Language Models**” , DeepMind 21

→ Uses vision as prefix to frozen auto-regressive language models

Q2) “**Conditional Prompt Learning for Vision-Language Models**” , CVPR 22

→ Uses meta-network to project vision embedding to language space

Q3) “**Meta Learning to Bridge Vision and Language Models for Multimodal Few-Shot Learning**” , ICLR 23

→ Uses meta-mapper to project vision embedding to language space as prefix

TASK DEFINATION:
TEMPORAL ACTION DETECTION (TAD)

CVSSP
Centre for Vision, Speech and Signal Processing

Untrimmed Video

What is the Activity ?
("Playing Ice Hockey")

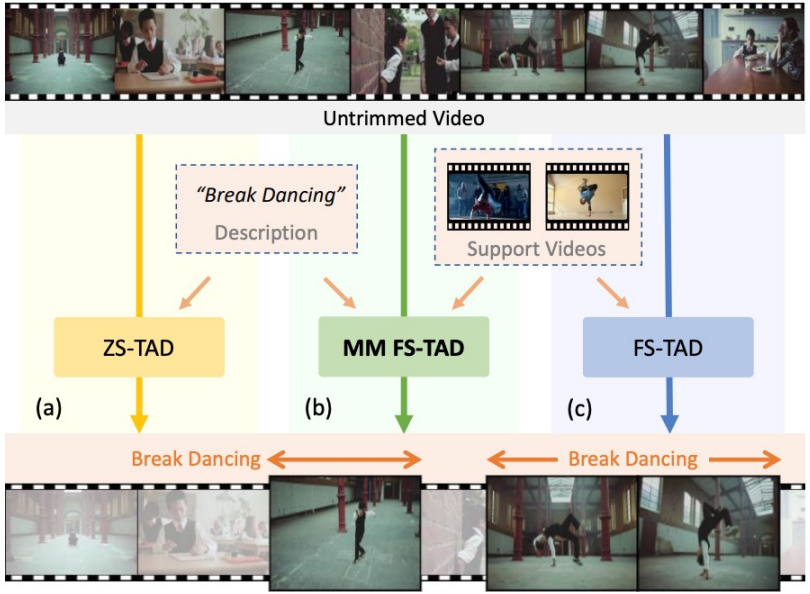
When is the Activity Occuring ?
(13 s – 28 s)

Sub-Task 1 :
Action **Classification**

Sub-Task 2 :
Temporal **Regression**

PROBLEM DEFINITION:
MULTI-MODAL FEW-SHOT TAD (MMFS-TAD)

CVSSP
Centre for Vision, Speech and Signal Processing



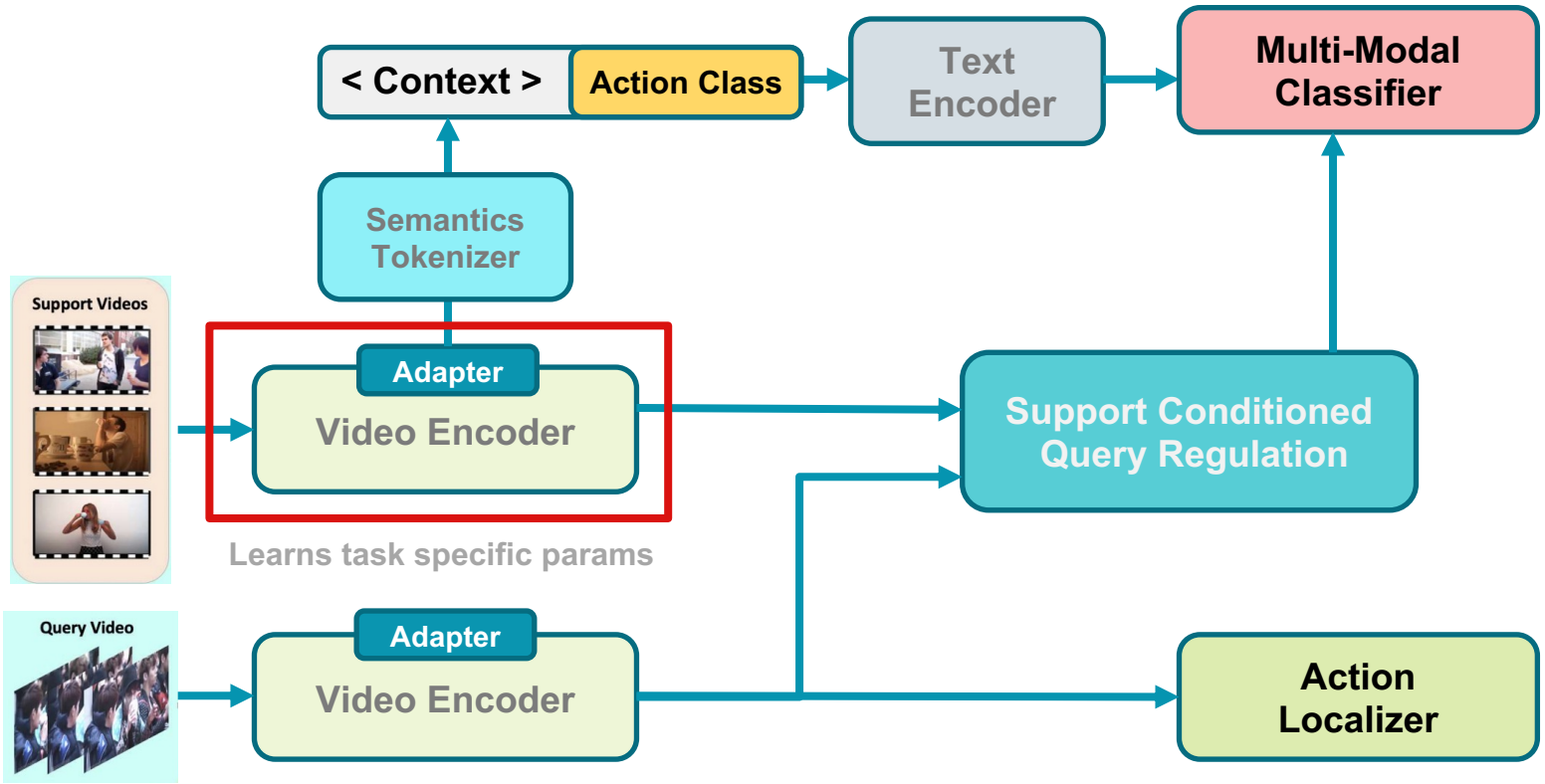
Unseen Video, Few Annotation

Semantic Context e.g, Text
+
Visual Context e.g, Support Video

Localize Unseen Action
in
Unseen Video

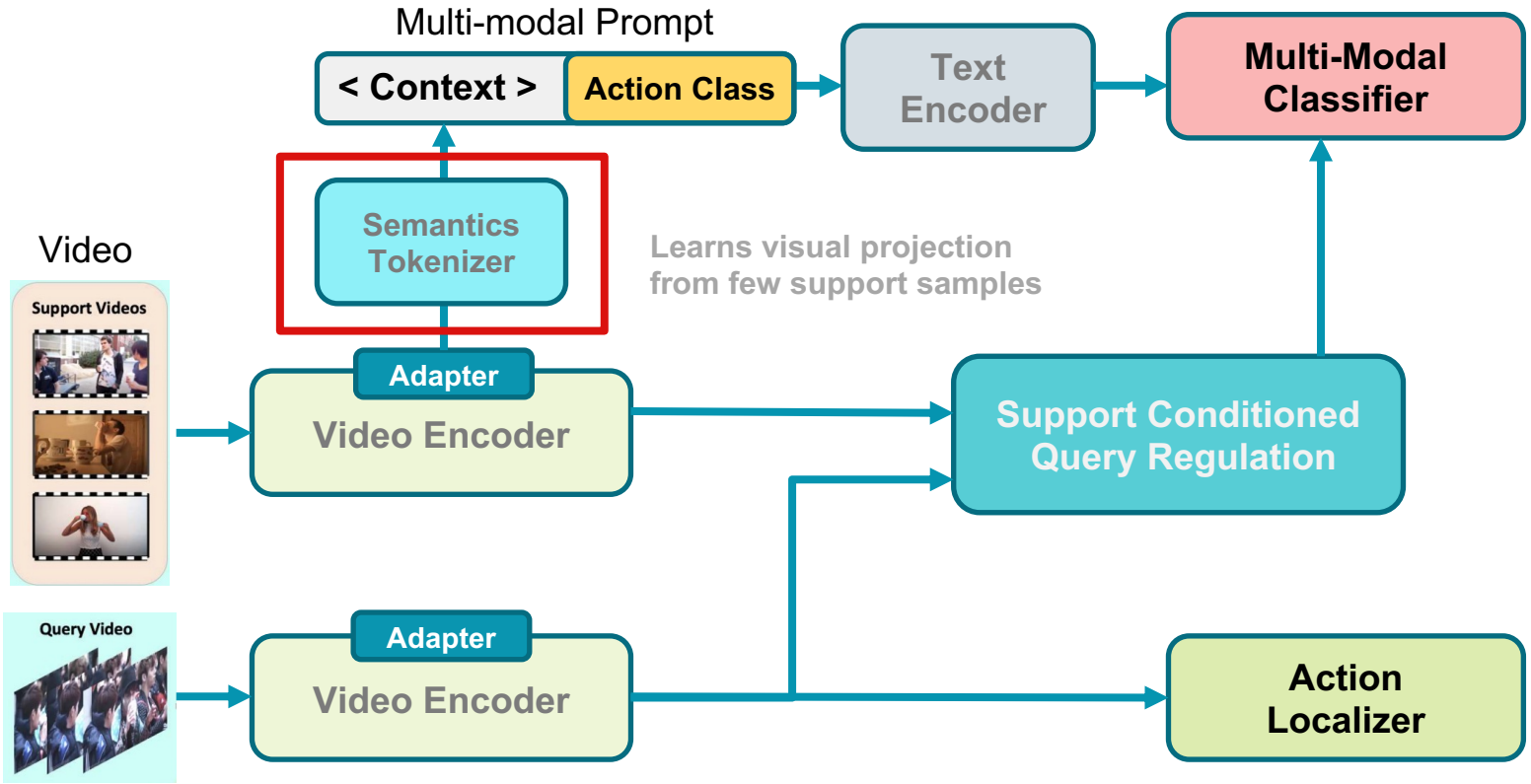
MUPPET:
MULTI-MODAL FEW SHOT ACTION DETECTION

CVSSP
Centre for Vision, Speech and Signal Processing



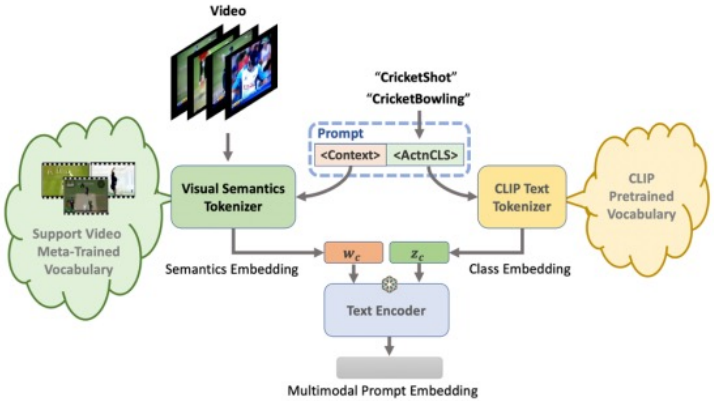
MUPPET:
MULTI-MODAL FEW SHOT ACTION DETECTION

CVSSP
Centre for Vision, Speech and Signal Processing

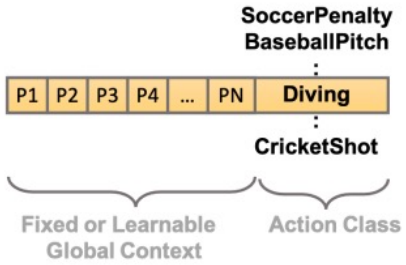


MUPPET:
VISUAL SEMANTICS TOKENIZER

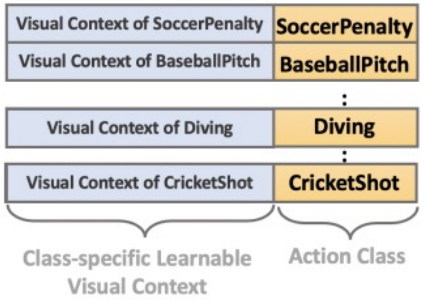
CVSSP
Centre for Vision, Speech and Signal Processing



(a) Multi-Modal Prompting



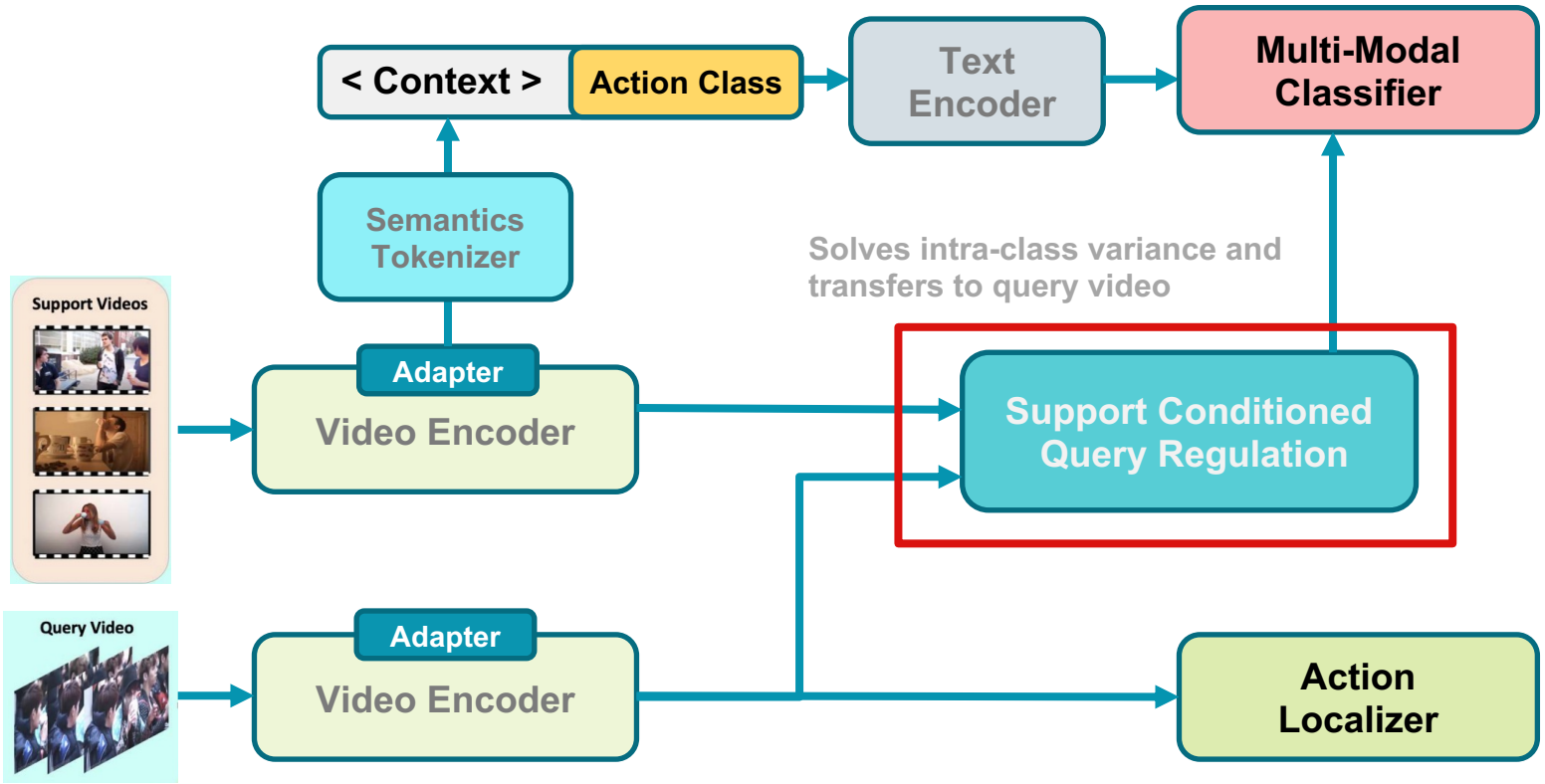
(b) Existing Prompt Design



(c) Our Prompt Design

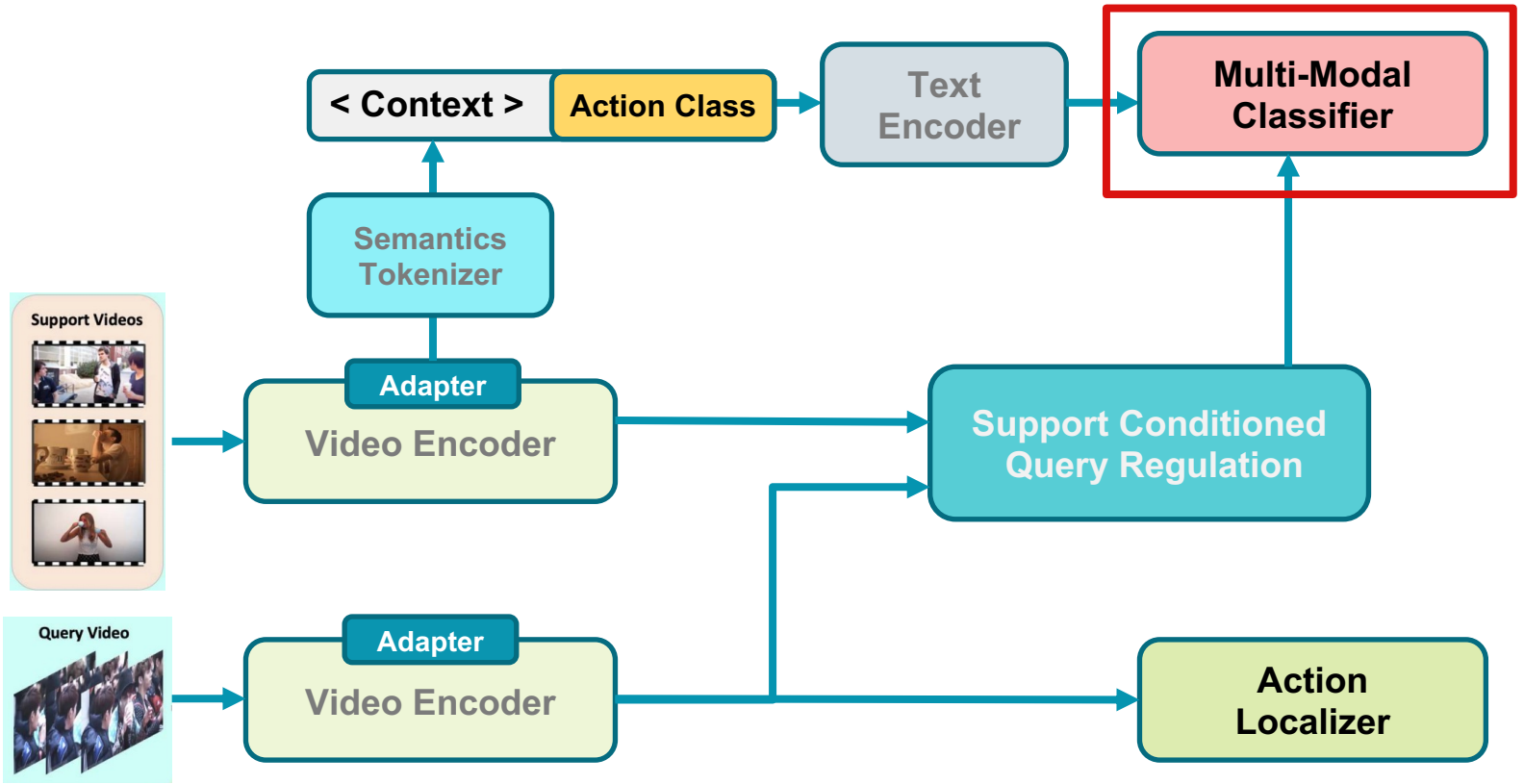
MUPPET:
MULTI-MODAL FEW SHOT ACTION DETECTION

CVSSP
Centre for Vision, Speech and Signal Processing



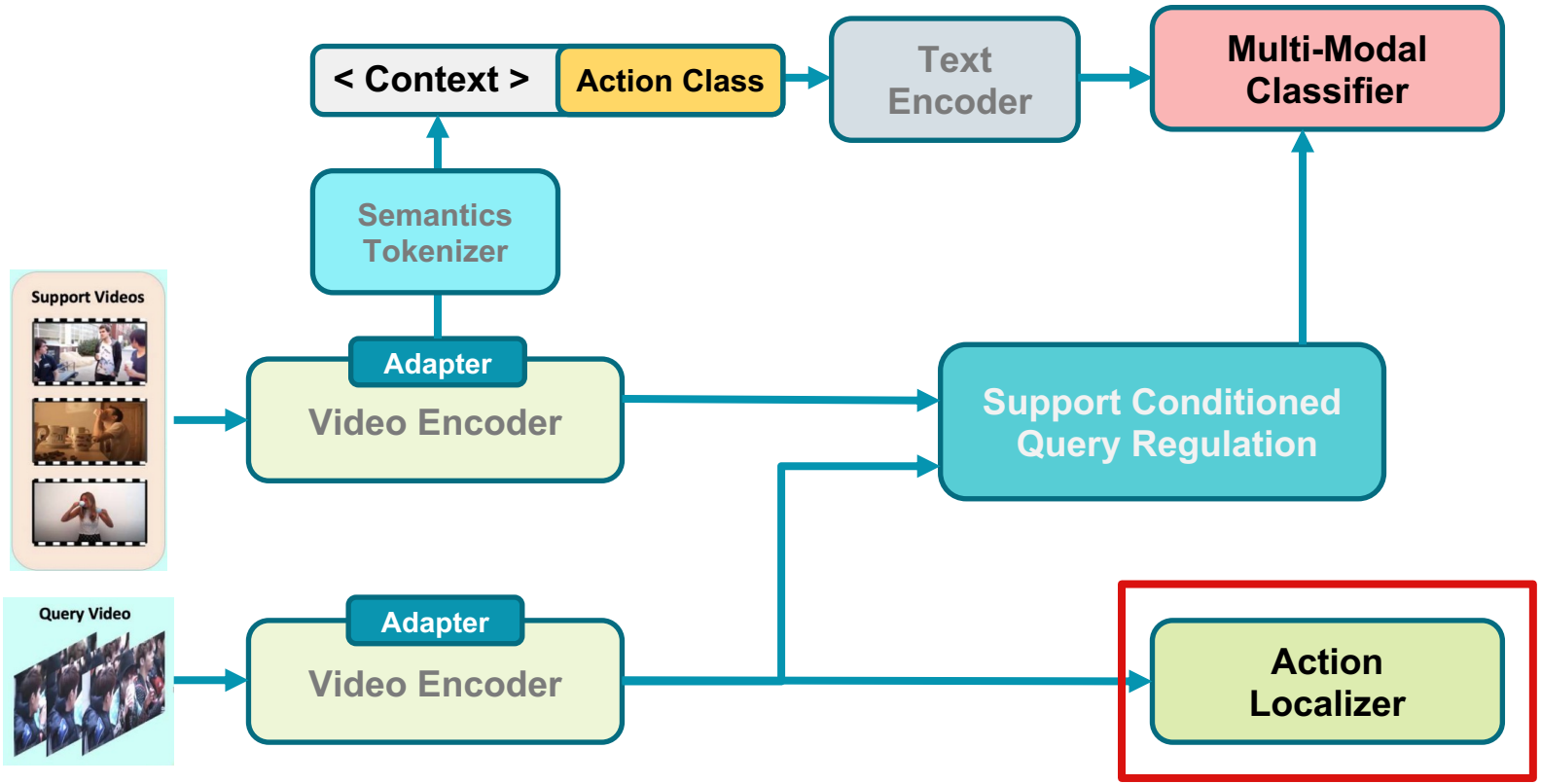
MUPPET:
MULTI-MODAL FEW SHOT ACTION DETECTION

CVSSP
Centre for Vision, Speech and Signal Processing



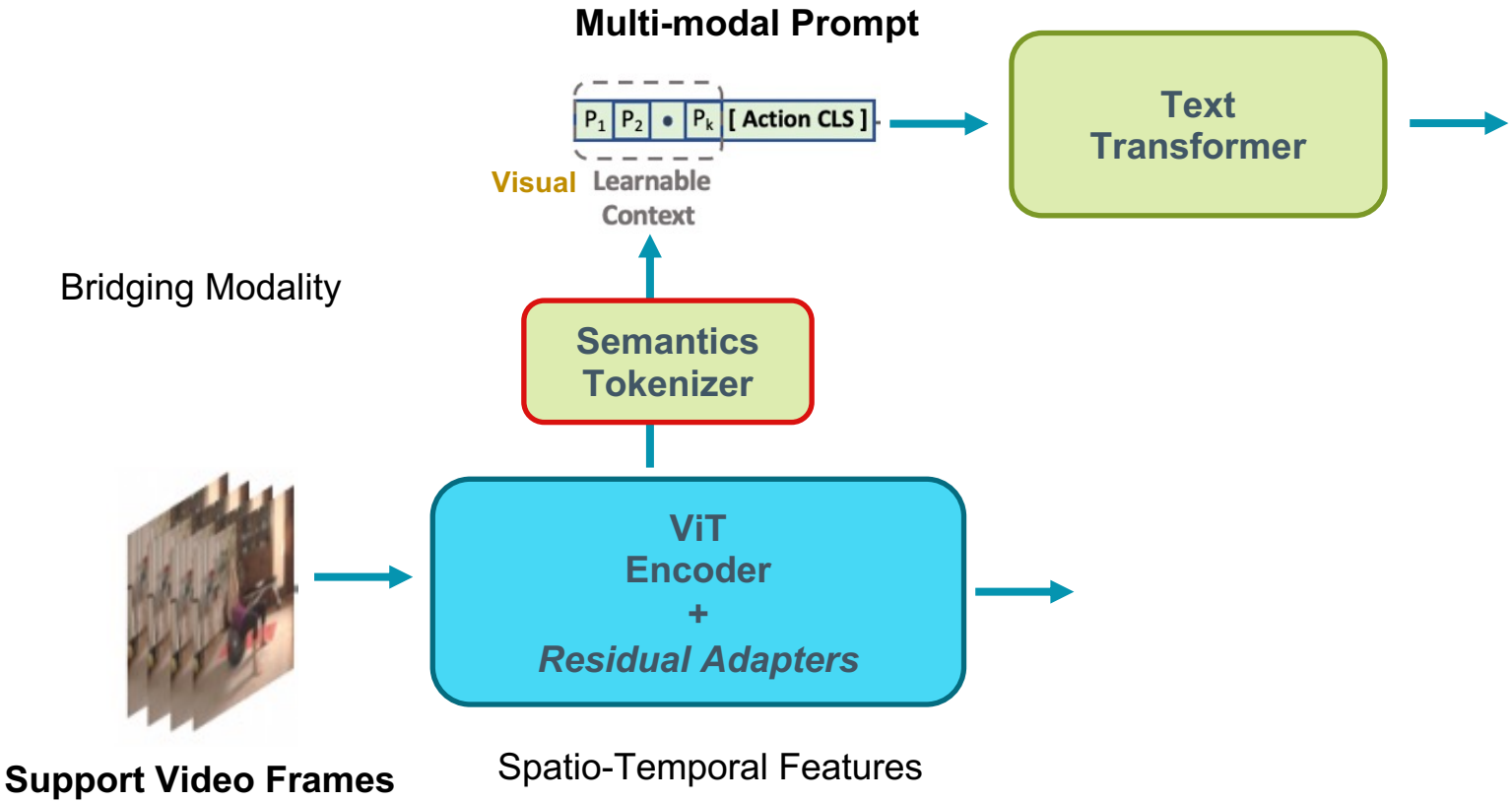
MUPPET:
MULTI-MODAL FEW SHOT ACTION DETECTION

CVSSP
Centre for Vision, Speech and Signal Processing

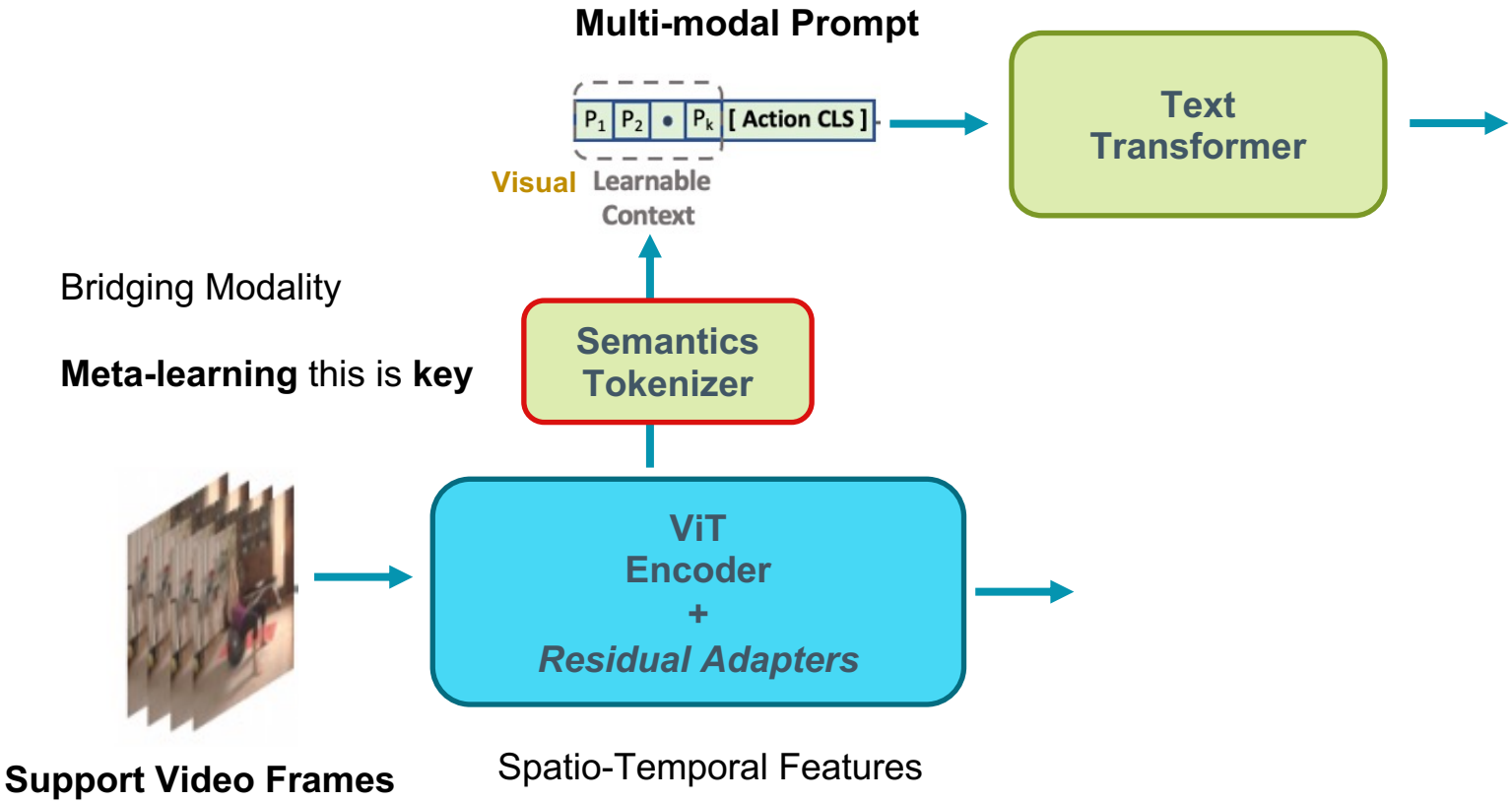


MUPPET:
FEATURE EXTRACTION

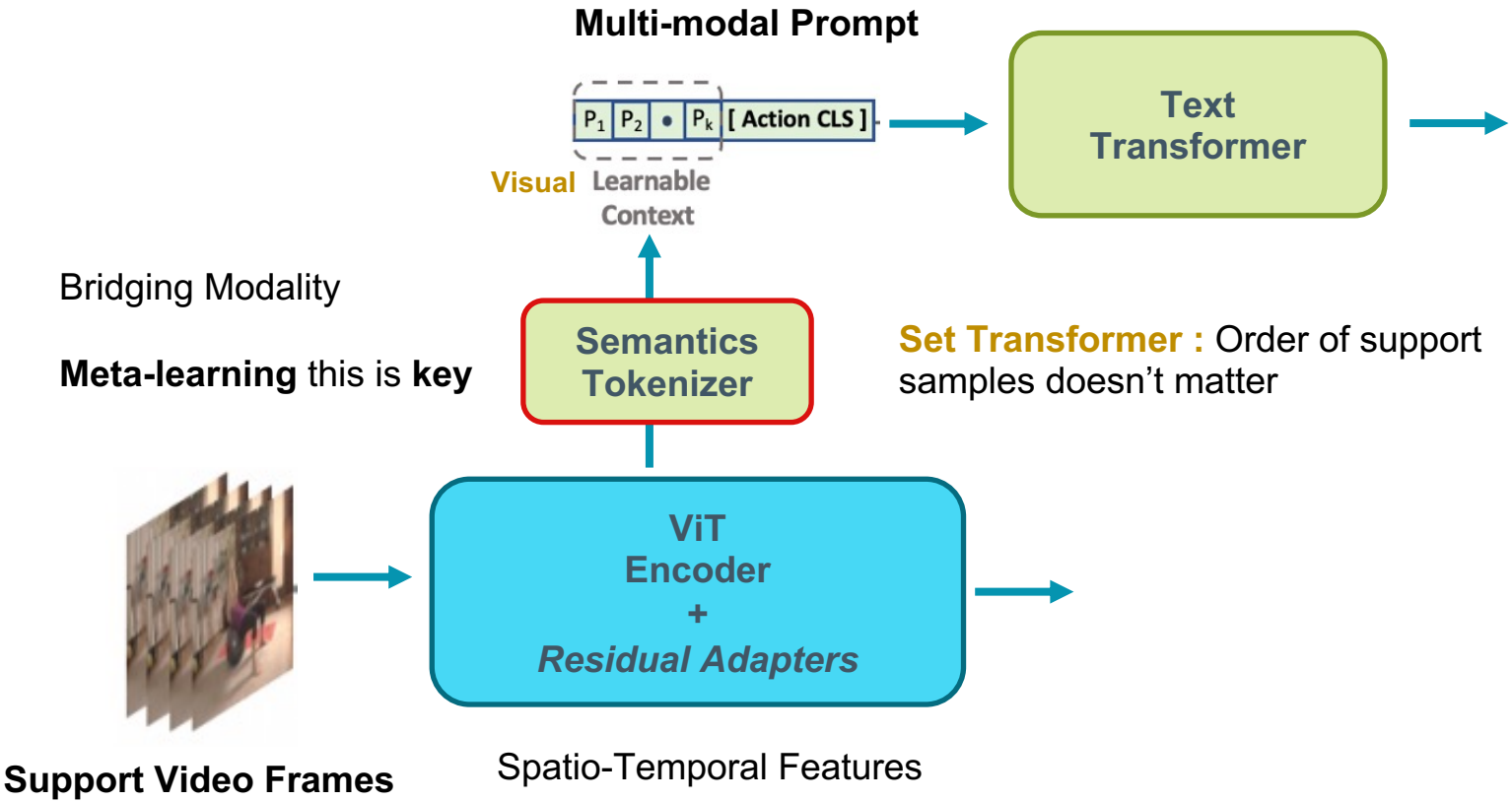
CVSSP
Centre for Vision, Speech and Signal Processing



MUPPET:
FEATURE EXTRACTION

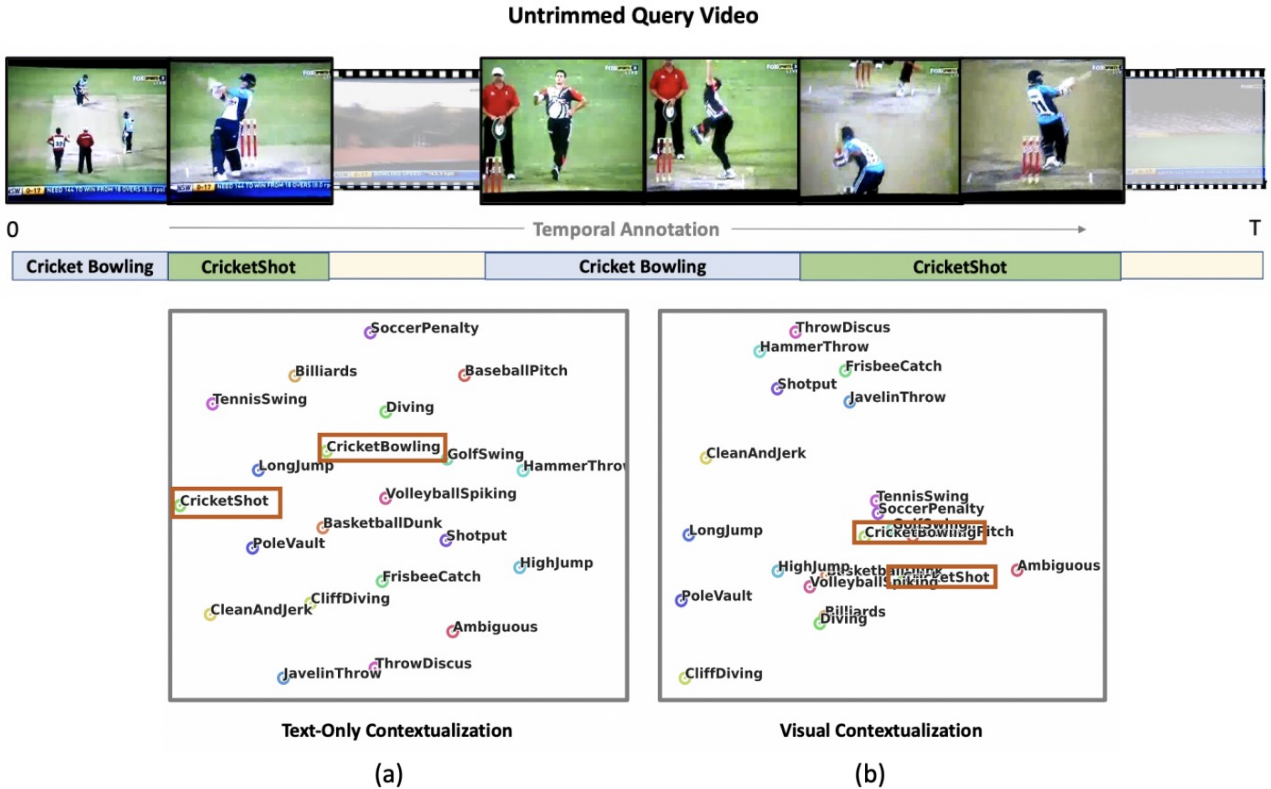


MUPPET:
FEATURE EXTRACTION



MUPPET:
VISUAL SEMANTICS TOKENIZER

CVSSP
Centre for Vision, Speech and Signal Processing



EXPERIMENTAL RESULTS:

FEW-SHOT, MULTI-MODAL FEW-SHOT, ZERO-SHOT

CVSSP
Centre for Vision, Speech and Signal Processing

Few-Shot

Multimodal Few-Shot

Zero-Shot

Method		N-way	Modality		ActivityNetv1.3				THUMOS14			
			Visual	Text	0.5	0.75	0.95	Avg	0.3	0.5	0.7	Avg
FS	FS-Trans [51]	1	✗	✗	42.2	24.8	5.2	25.6	42.6	25.7	8.2	25.5
	QAT [31]				44.6	26.4	4.9	26.9	38.7	24.4	7.5	24.3
	MUPPET				**45.4**	**28.1**	**5.6**	**27.8**	**44.1**	**26.2**	**8.5**	**26.1**
	Feat-RW [20]	5			30.7	16.6	2.9	17.1	35.3	19.6	6.8	20.1
	Meta-DETR [54]				32.9	20.3	4.6	19.4	37.5	20.7	7.5	21.9
	FSVOD [9]				34.5	18.9	5.1	21.6	37.9	23.8	7.3	22.8
	MUPPET				**36.9**	**22.2**	**5.9**	**23.0**	**41.2**	**25.7**	**8.5**	**24.9**
MMFS	OV-DETR [52]	1	✗	✓	44.2	27.9	6.3	28.7	46.1	29.7	9.0	30.4
	Owl-Vit [27]				43.7	27.0	6.0	27.2	45.2	29.0	9.0	30.2
	EffPrompt [19]				45.9	27.9	5.2	29.4	47.2	30.4	9.8	31.1
	STALE [30]				47.7	29.3	7.6	30.3	48.9	32.1	10.3	32.0
	Baseline-I		✓		46.9	28.6	6.9	29.7	47.3	30.5	9.2	31.8
	MUPPET				**49.7**	**32.9**	**9.2**	**32.7**	**50.6**	**33.5**	**11.2**	**33.8**
	OV-DETR [52]	5	✗		39.8	22.3	5.4	23.1	40.4	23.9	7.5	24.0
	Owl-Vit [27]				37.9	20.3	5.6	21.9	38.3	21.9	7.7	22.6
	EffPrompt [19]				41.1	21.6	5.4	23.8	39.5	23.5	7.6	24.8
	STALE [30]				42.3	22.9	6.8	24.5	40.7	24.9	7.1	25.4
	Baseline-I		✓		42.1	22.7	6.0	24.0	40.2	24.7	7.0	25.0
	MUPPET				**45.3**	**25.6**	**6.3**	**26.2**	**42.3**	**27.2**	**7.8**	**27.5**
ZS	EffPrompt [19]	All	✗	✓	32.0	19.3	2.9	19.6	37.2	21.6	7.2	21.9
	STALE [30]		✗		32.1	20.7	5.9	20.5	38.3	21.2	7.0	22.2
	Baseline-I		✓		30.6	18.0	4.1	18.7	35.8	20.5	7.1	20.8
	MUPPET		✗		**33.5**	**21.9**	**6.7**	**22.0**	**40.1**	**22.8**	**8.1**	**24.8**

EXPERIMENTAL RESULTS:
ABLATION STUDIES

CVSSP
Centre for Vision, Speech and Signal Processing

Table 2. Prompt learning design on ActivityNet. Setting: 5-way.

Design	Shots	Prompt style		mAP	
		Learnable	Context	0.5	Avg
LPS	-	✗	-	18.4	13.6
LVP	5	✓	Visual	43.2	25.0
LTP	5	✓	Text	42.7	24.7
Ours	**1**	✓	**Visual**	**43.7**	**25.1**
Ours	**5**	✓	**Visual**	**45.3**	**26.2**

Visual Projection is better

Table 3. Design of visual semantics tokenizer on ActivityNet. Setting: 5-way 5-shot. #T/C: Tokens per Class.

Network	**Meta-Learn**	**#T/C**	**mAP**	
			0.5	**Avg**
1D-CNN	✗	20	37.4	21.3
	✓	20	40.8	23.0
	✓	1	39.7	22.5
Set Transformer [22]	✗	1	43.8	24.7
	✓	1	**45.3**	**26.2**
	✓	20	44.7	25.6

Only **single** <Context> Token is enough to optimize

EXTENSION OF MUPPET:
OBJECT DETECTION

CVSSP
Centre for Vision, Speech and Signal Processing

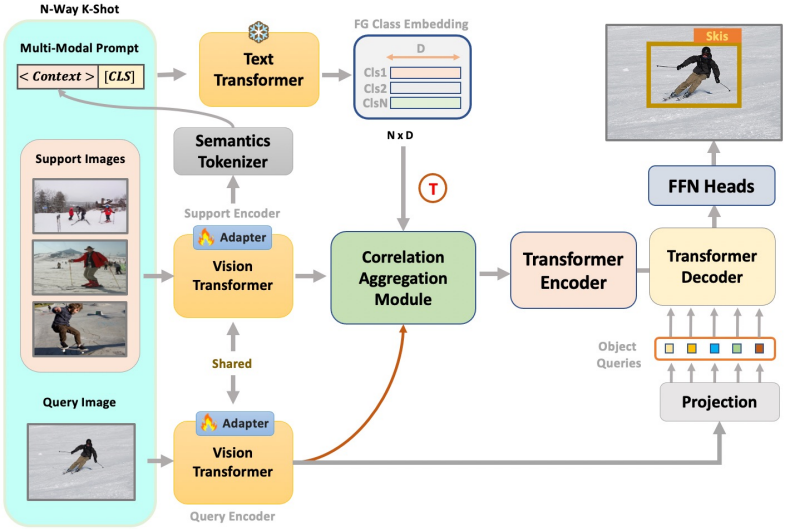


Table 10. Comparing our adapted MUPPET with existing Few-Shot Object Detection methods on COCO dataset.

Method	5-Shot			10-Shot		
	AP	AP_{50}	AP_{75}	**AP**	AP_{50}	AP_{75}
FRCN [34]	4.6	8.7	4.4	5.5	10.0	5.5
TFA w/ cos [44]	7.0	13.3	6.5	9.1	17.1	8.8
Deform-DETR [59]	7.4	12.3	7.7	11.7	19.6	12.1
FSOD [10]	-	-	-	12.0	22.4	11.8
QA-FewDet [15]	9.7	20.3	8.6	11.6	23.9	9.8
META-DETR [54]	15.4	25.0	**15.8**	19.0	30.5	19.7
MUPPET	**15.9**	**26.4**	14.8	**20.1**	**32.3**	**19.9**

Can also be plugged into any existing Object Detection module

SUMMARY:
MULTI-MODAL FEW-SHOT

CVSSP
Centre for Vision, Speech and Signal Processing

- **1st work** on **Multi-Modal Few Shot Learning** for **Video Domain**
- **One Design Three Setting** can be solved : Few-Shot, Multi-modal Few Shot, Zero-Shot
- **Meta-Learning Visual Projection** is key to multi-modal few-shot
- Can be **extended to Object Detection** as well

THANKS!

Any questions?