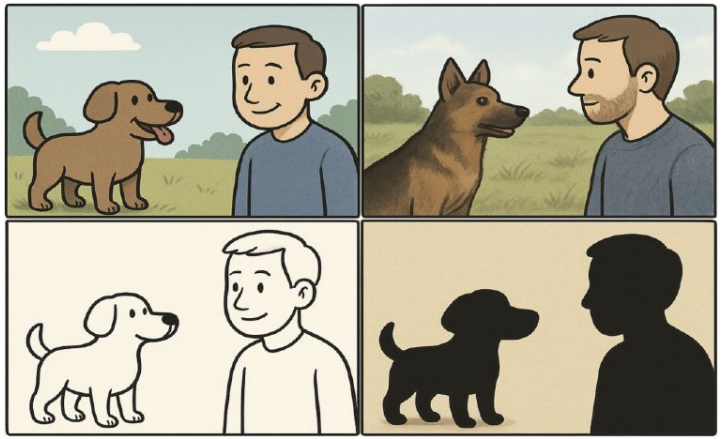
On the Quest of Generalizable Video Understanding

David Pujol Perich
PhD defense
Supervisors: Sergio Escalera and Albert Clapés

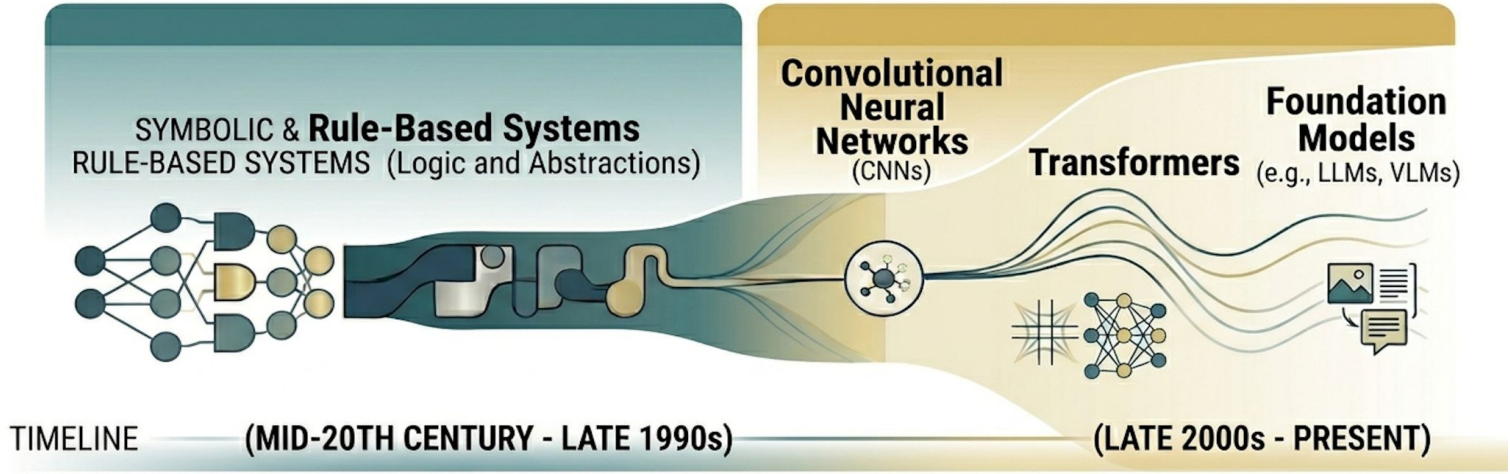
Generalization in human intelligence

➔ Humans are able to reuse knowledge in multiple tasks and environments.



Evolution of artificial intelligence systems

➔ Artificial intelligence has gone a long way.



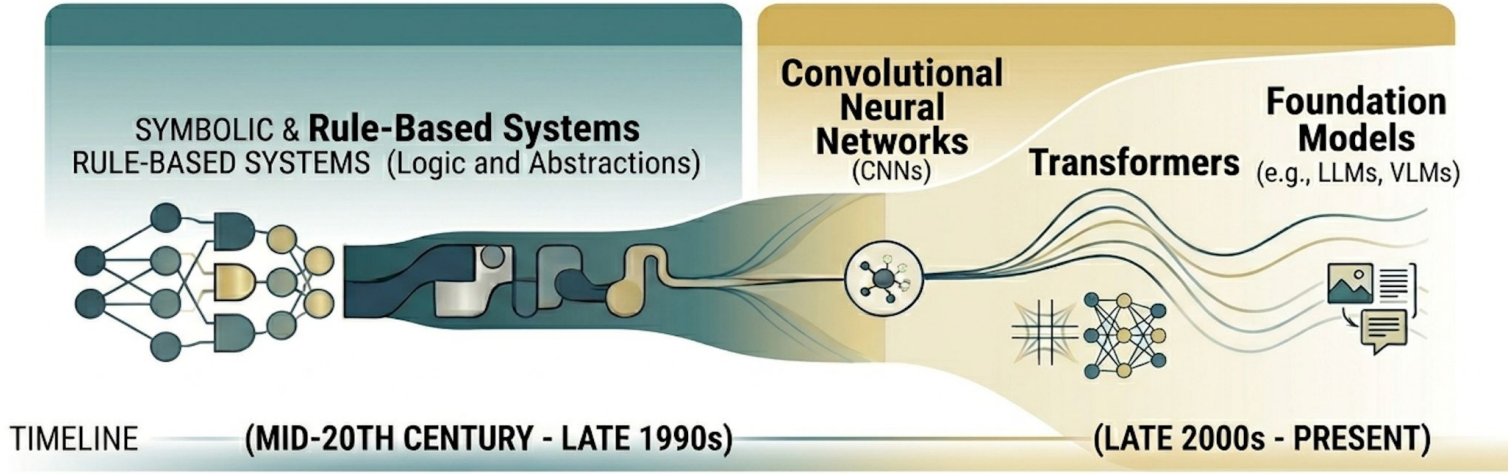
Video understanding as the next frontier

- ➔ **Multimodal video understanding** models are systems that reason about videos, complemented by other modalities
- ➔ It is essential for applications like surveillance, search or assistance.

Evolution of artificial intelligence systems

➔ Artificial intelligence has gone a long way.

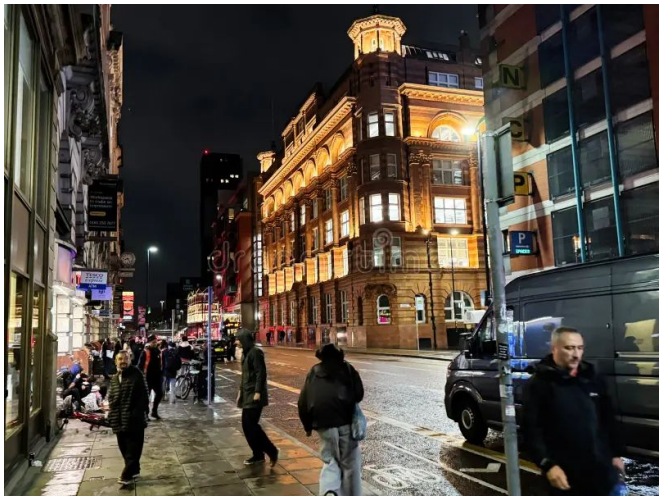
➔ How well do **multimodal video understanding** models generalize?



Are video understanding models able to generalize?

➔ Unfortunately not... models are often unable to work on diverse, more realistic scenarios.

Training



Inference

What are we trying to address?

- ➔ **Generalization (OOD):** Ability to maintain predictive performance when test-time distributions differ from training distributions.
- ➔ **Transferability:** Efficiency in adapting pre-trained representations to related, yet different target tasks *T*.

What are we trying to address?

- ➔ **Generalization (OOD):** Ability to maintain predictive performance when test-time distributions differ from training distributions.
- ➔ **Transferability:** Efficiency in adapting pre-trained representations to related, yet different target tasks *T*.

Visual OOD	Linguistic OOD	Transferability
Can models generalize to new visual domains?	**Can models generalize to new linguistic domains?**	**Can pre-trained models transfer efficiently to new tasks?**

Chapter 1: Visual OOD generalization

Temporal action localization: What happens, and when?

- ➔ TAL predicts the time interval of the set of actions $\mathcal{C}$ of a given video.
- ➔ Actions are often sparse, overlapping, and ambiguous.
- ➔ It is the core task behind video search, summarization, and event detection.

Definitions: Unsupervised domain adaptation

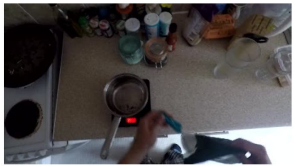
➔ **Unsupervised Domain Adaptation (UDA):** The task of training a model on $\mathscr{S}$ and deploying it on $\mathscr{T}$, despite the distribution shift between the two.

	Source domain ($\mathscr{S}$)	Target domain ($\mathscr{T}$)
Labels $\mathcal{C}$ available?	✓	✗
Used for training?	✓	✓
Used for evaluation?	✗	✓

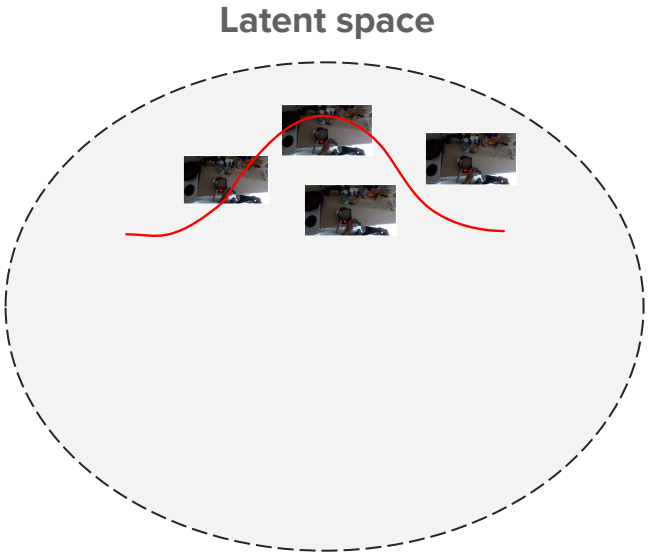
Motivation: Domain shifts break TAL models

➔ *What happens when we train a model on a source domain, but evaluate on a different target domain?*

Training

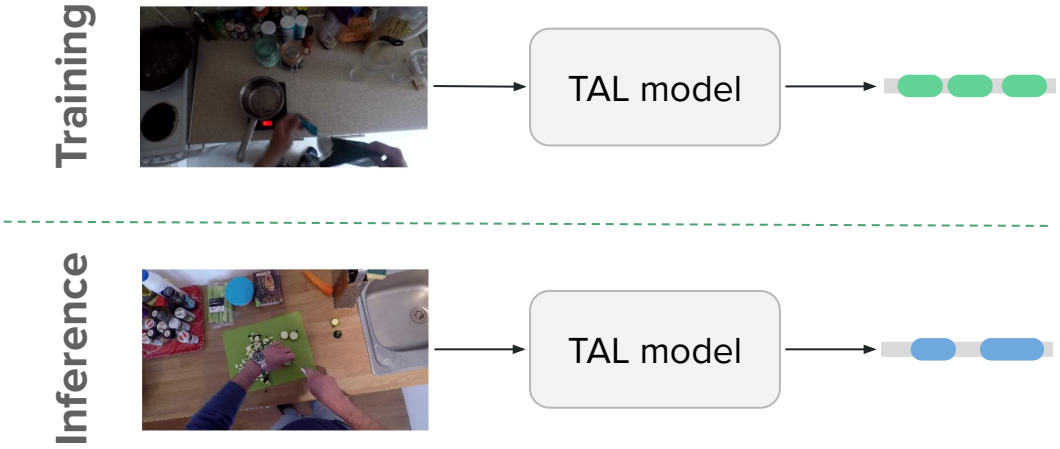


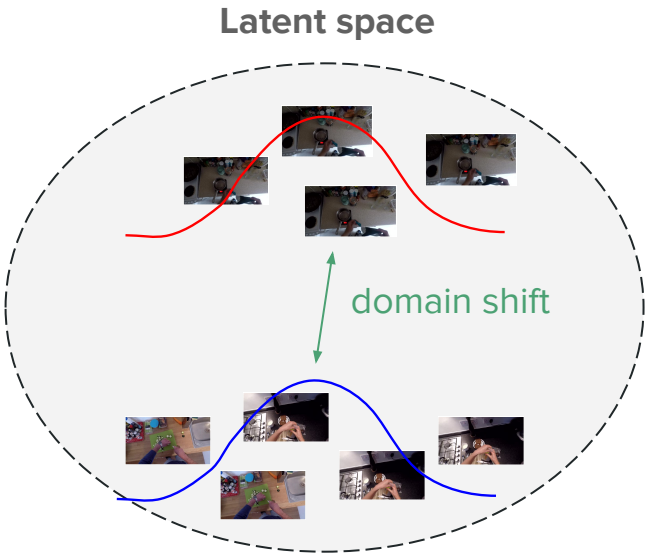
TAL model



Motivation: Domain shifts break TAL models

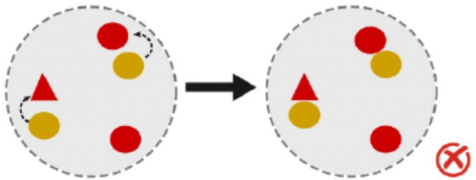
➔ *What happens when we train a model on a source domain, but evaluate on a different target domain?*





Existing UDA models are not semantically aware

- ➔ Most DA methods align the latent space of source and target domains globally.
- ➔ **Problem:** This can merge embeddings of "*cutting vegetables*" with "*washing hands*".

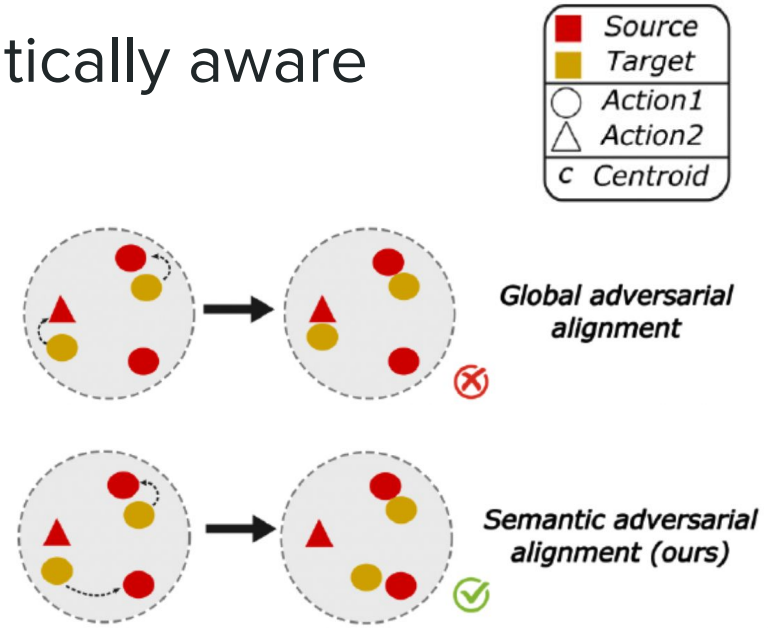


Global adversarial alignment

Pujol-Perich, David et al. "SADA: Semantic Adversarial Unsupervised Domain Adaptation for Temporal Action Localization." *IEEE/CVF Winter Conference on Applications of Computer Vision* 2025.

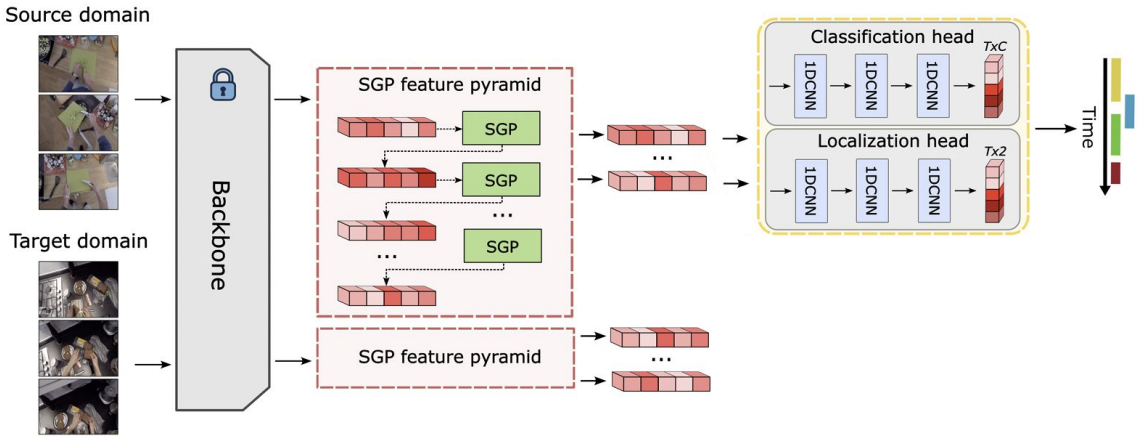
Existing UDA models are not semantically aware

- ➔ Most DA methods align the latent space of source and target domains globally.
- ➔ **Problem:** This can merge embeddings of "*cutting vegetables*" with "*washing hands*".
- ➔ **Solution:** We introduce SADA, a semantically consistent adversarial loss.



SADA: Semantic adversarial domain adaptation

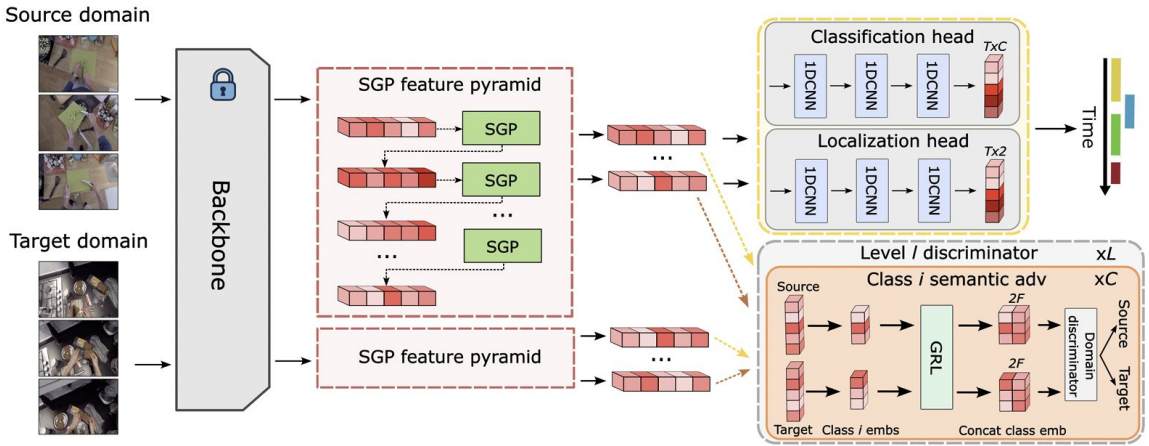
1) Apply a shared backbone B for both source and target videos V^S and V^T.
2) Apply the classification and localization to the source only (labeled).



SADA: Semantic adversarial domain adaptation

3) Compute pseudo-labels for each of the unlabelled target embeddings.

$$\hat{c}_z = \begin{cases} \underset{i}{\text{argmax}}\ P_l[z,i] & \text{if } P_l[z,i] > \alpha \\ 0 & \text{otherwise,} \end{cases}$$



SADA: Semantic adversarial domain adaptation

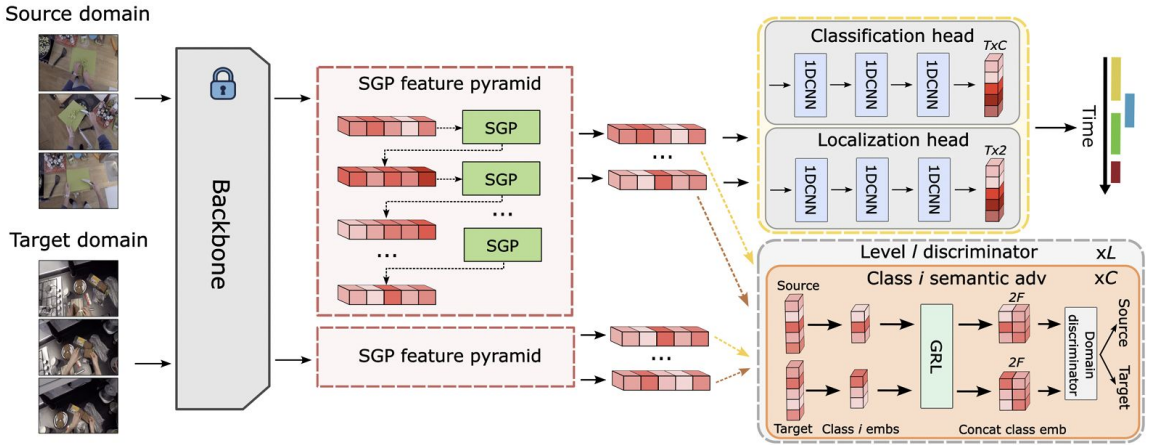
3) Compute pseudo-labels for each of the unlabelled target embeddings.

4) Align source and target embeddings of class *i*.

$$A_i^l = \{Z_l^{\mathcal{S}}[z] \ : \ c_z = i\}_{z \in T_l}$$

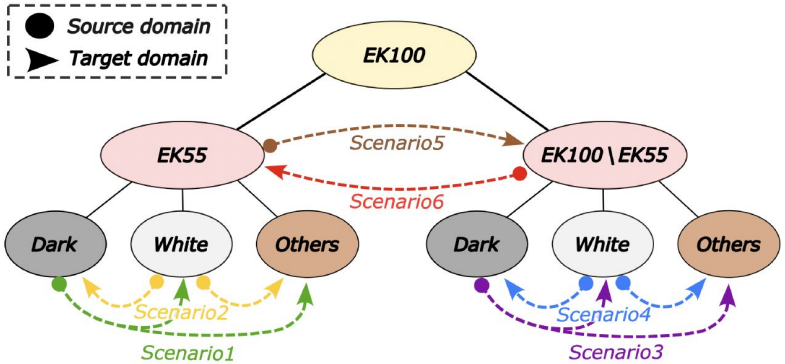
$$B_i^l = \{Z_l^{\mathcal{T}}[z] \ : \ \hat{c}_z = i\}_{z \in T_l}$$

$$\mathcal{L}_{local}^l = \sum_{i=1}^{C} (\mathcal{L}_{\text{BCE}}(D(A_i^l \,||\, E_i), d_{\mathcal{S}})) + (\mathcal{L}_{\text{BCE}}(D(B_i^l \,||\, E_i), d_{\mathcal{T}}))$$

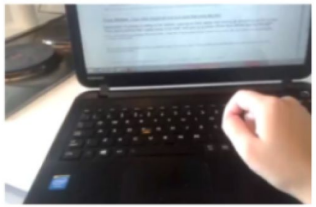


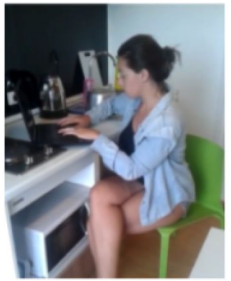
7 new cross-domain benchmarks for sparse TAL

Epic Kitchens100 (6 benchmarks)

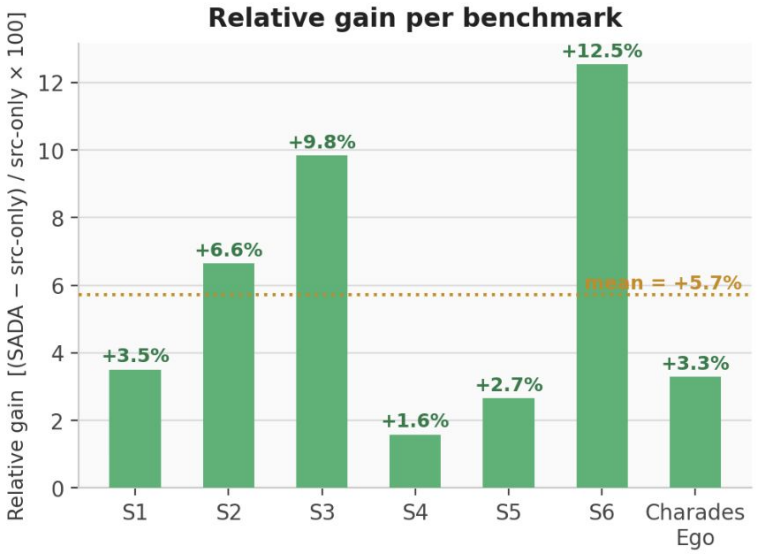


Charades to CharadesEgo



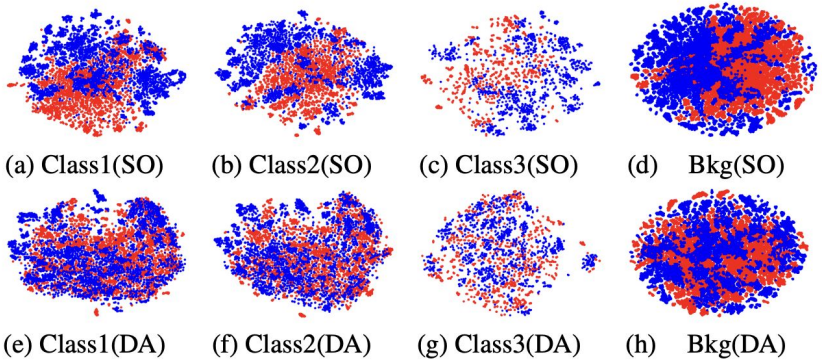


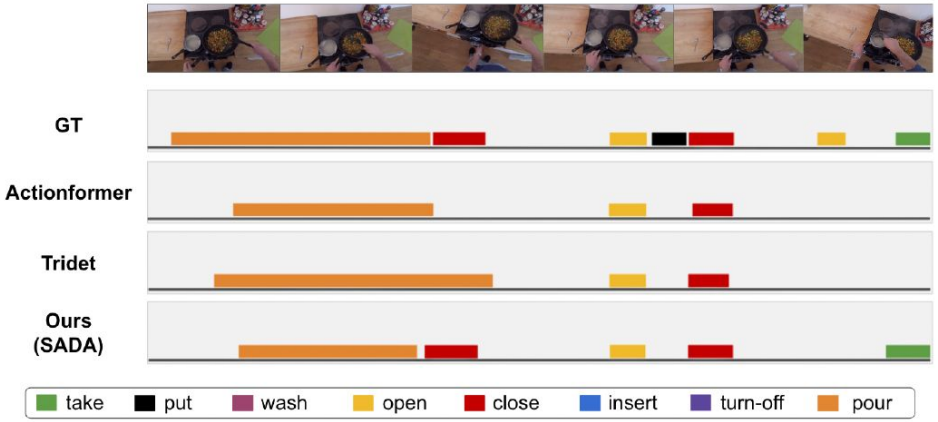
Main results



Ablations

1. We attain a considerable improvement in per-class latent space alignment.
2. Our results show that SADA consistently improves other UDA methods.





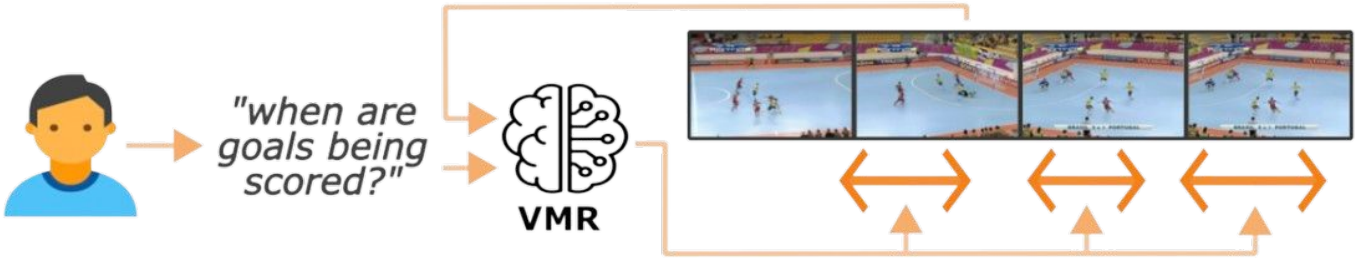
Conclusions

- ➔ **SADA** is the first UDA method specifically designed for **sparse TAL**.
- ➔ We introduce **7 new cross-domain TAL benchmarks**
- ➔ SADA achieves **+1.14 mAP avg** over the best competing UDA baseline.

Chapter 2: Linguistic OOD generalization

Video Moment Retrieval (VMR)

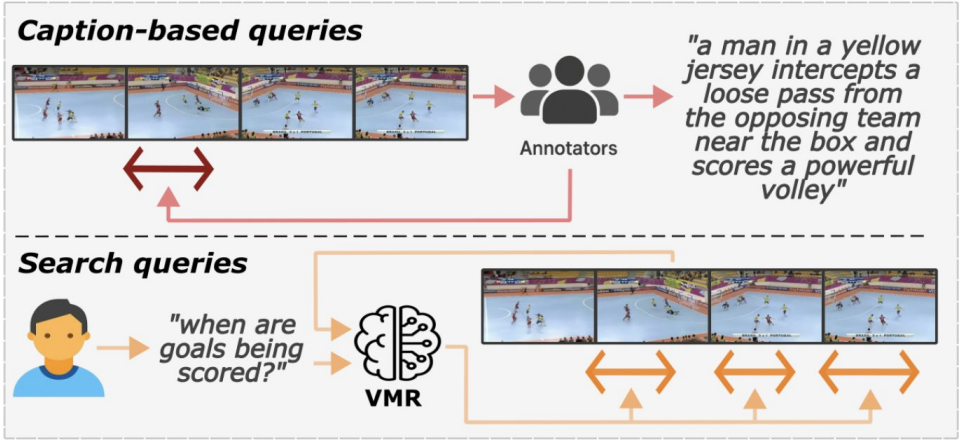
- ➔ VMR extends TAL, by expecting a natural-language query instead of a predefined action class.
- ➔ This represents a more realistic problem setup.



Training queries and real queries are different

➔ Annotators watch the video **before** writing a query—producing detailed, visually-informed captions.

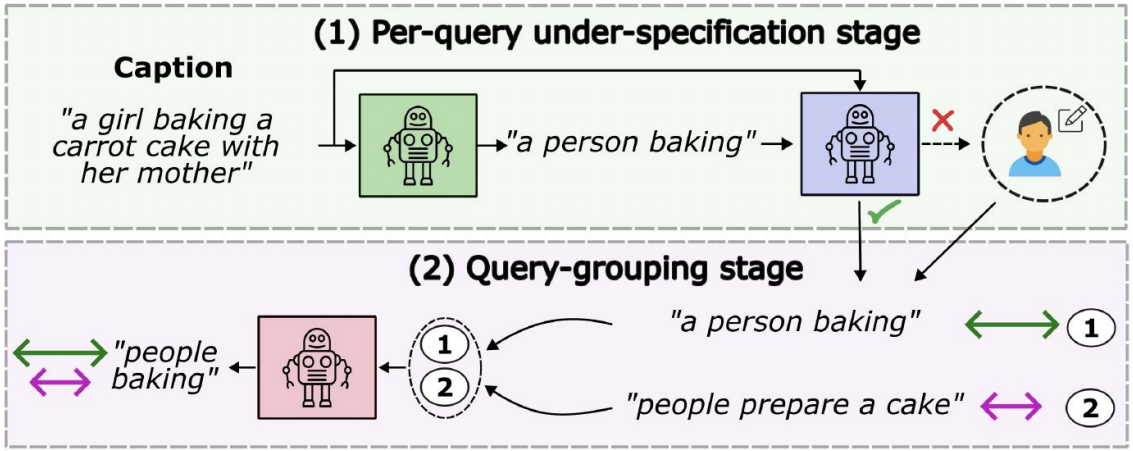
➔ Real users search **without** watching first, relying on short descriptions.



Pujol-Perich, David, et al. "Beyond Caption-Based Queries in Video Moment Retrieval." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2026.

Search-query benchmarks: Our LLM-based pipeline

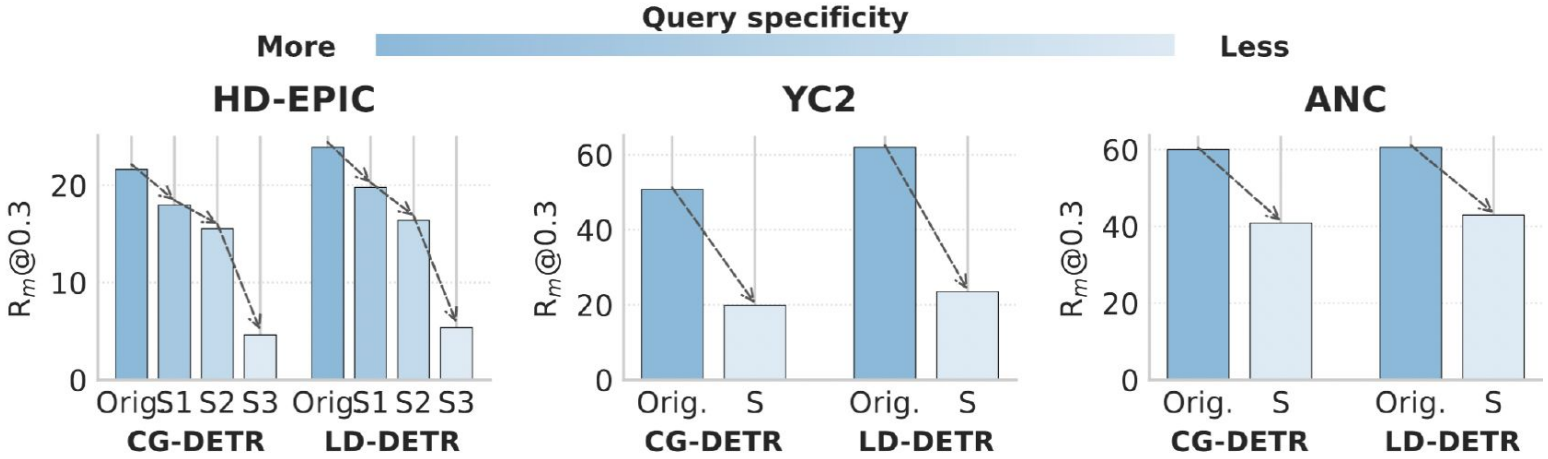
1. An LLM agent progressively removes details from captions.
2. A validator agent catches inconsistencies.
3. A grouping step merges equivalent queries into multi-moment instances.
4. This yields 5 new benchmarks — HD-EPIC-S1/2/3, YC2-S, ANC-S.



Performance drops very significantly…

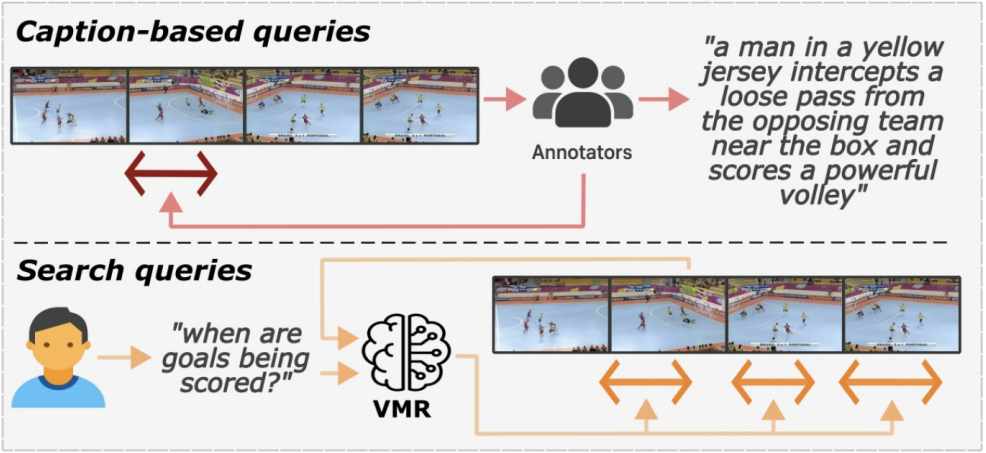
➔ The more under-specified the search queries are, the bigger the gap with caption-based performance.

➔ <u>Training on captions does not guarantee good performance on real queries.</u>



Two distinct failure modes

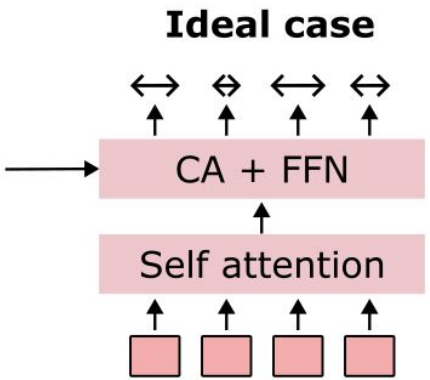
- ➔ **Language gap:** Reflects the vocabulary and specificity difference between captions and search queries.
- ➔ **Multi-moment gap:** Arises because an under-specified query matches multiple clips, but models trained on captions only ever predict one.

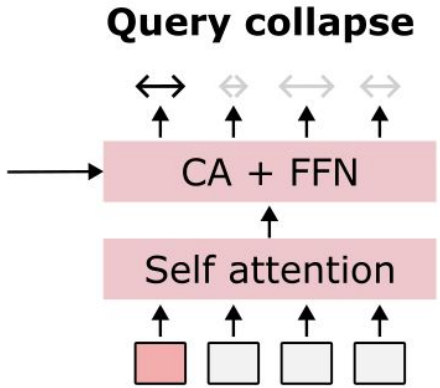


DETR decoder query collapse: Single-moment bias

➔ Ideally, the *M* decoder queries yield high quality predictions of the GT moments.

➔ Since during training, there is only 1 GT, almost all predictions result in “garbage”.

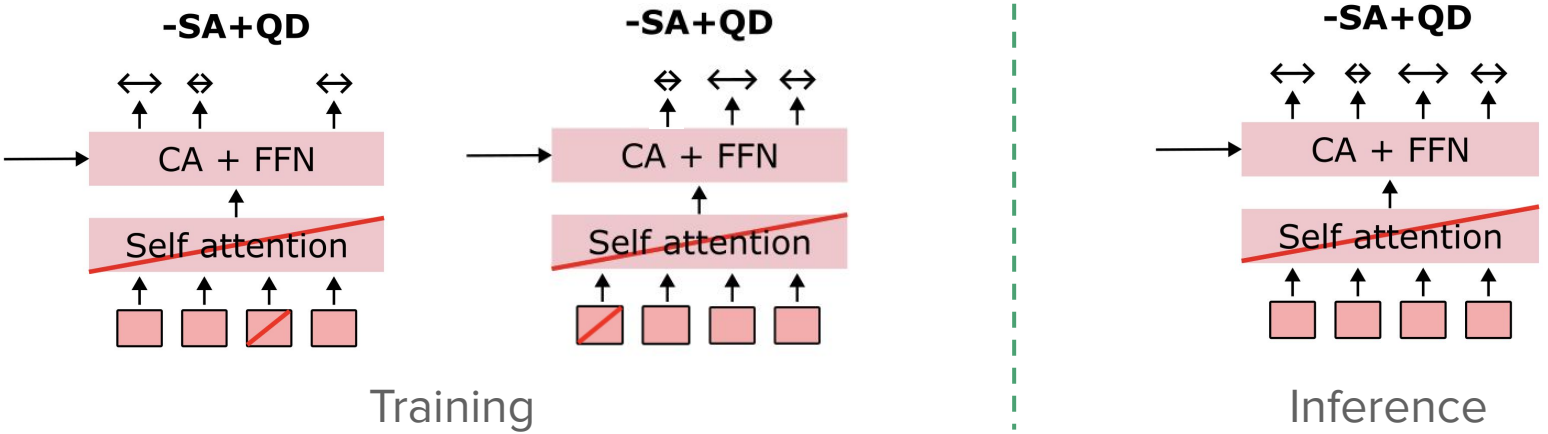




Mitigating this collapse (-SA+QD)

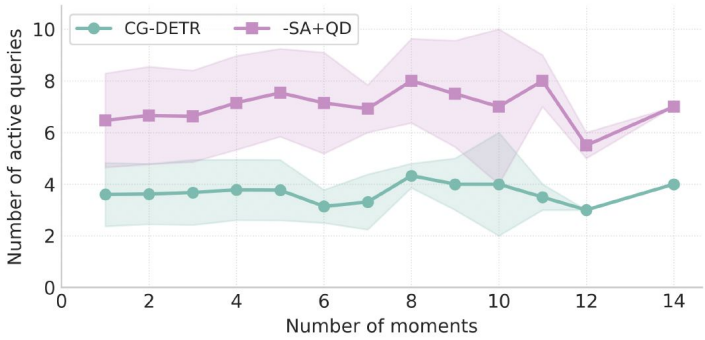
➔ We remove the self-attention to avoid communication across queries (-SA).

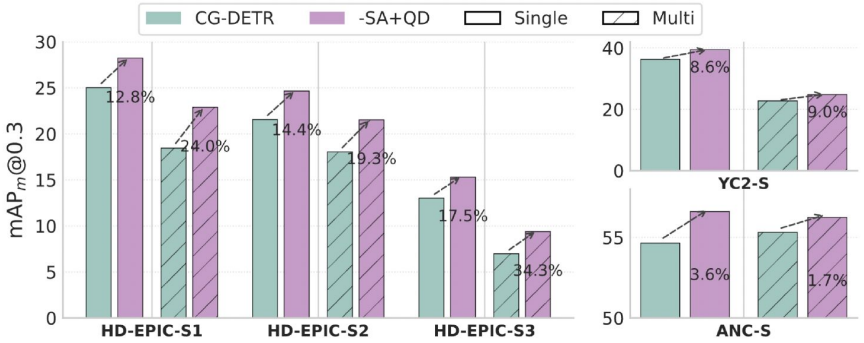
➔ We introduce query dropout to prevent the learnable queries from learning which one to always activate.



Main results

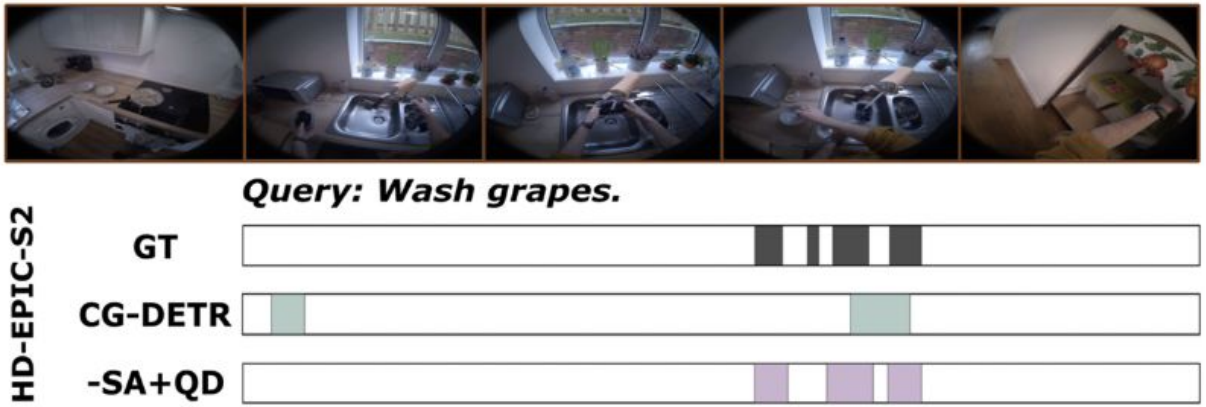
1. Our modifications are able to activate more decoder queries.
2. This results in a considerable performance boost in both single and multi-moment instances.





Ablations

➔ Qualitative results also indicate an improved detection of multi-moment instances.



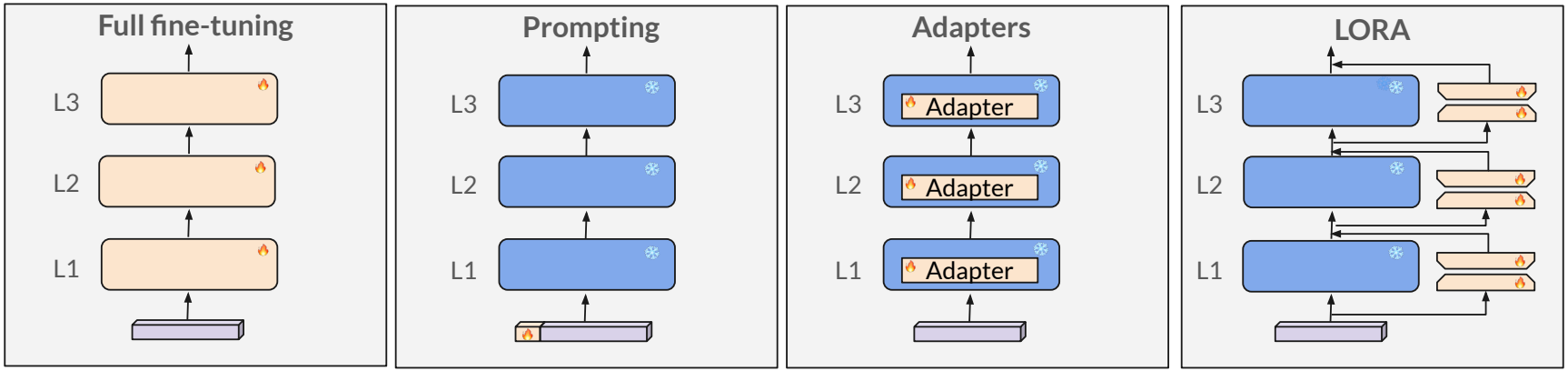
Conclusions

- ➔ We reveal a systematic **linguistic OOD gap**
- ➔ We identify **two distinct failure modes**: the *language and multi-moment gap*.
- ➔ We propose 5 new benchmarks.
- ➔ Our architectural modifications yield gains of **+12–34%** across all benchmarks.

Chapter 3: Knowledge transferability

Limitations of existing parameter-efficient fine-tuning

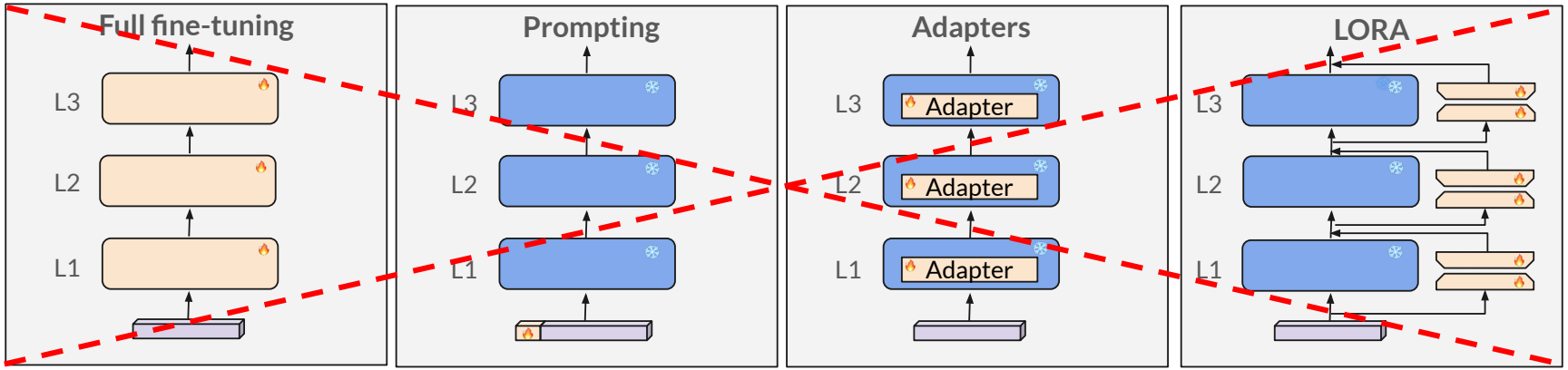
- ➔ Large pretrained VLMs contain rich video representations
- ➔ Updating all parameters per task is computationally infeasible.
- ➔ Existing parameter-efficient fine-tuning (PEFT) methods are memory inefficient.



Limitations of existing parameter-efficient fine-tuning

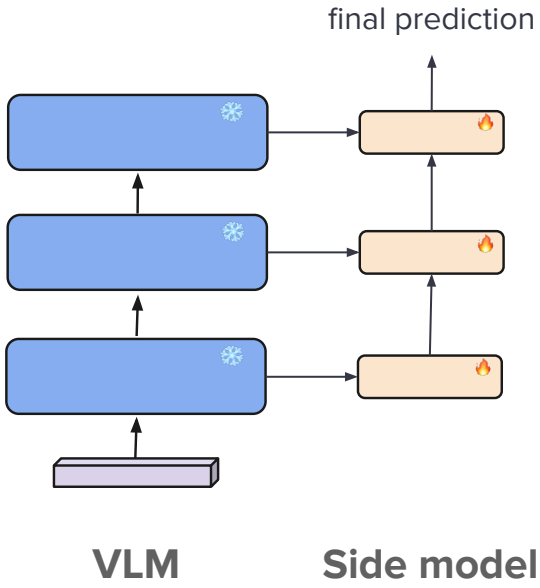
- ➔ Large pretrained VLMs contain rich video representations
- ➔ Updating all parameters per task is computationally infeasible.
- ➔ Existing parameter-efficient fine-tuning (PEFT) methods are memory inefficient.

all require full backpropagation



A parameter and memory efficient alternative: Side tuning

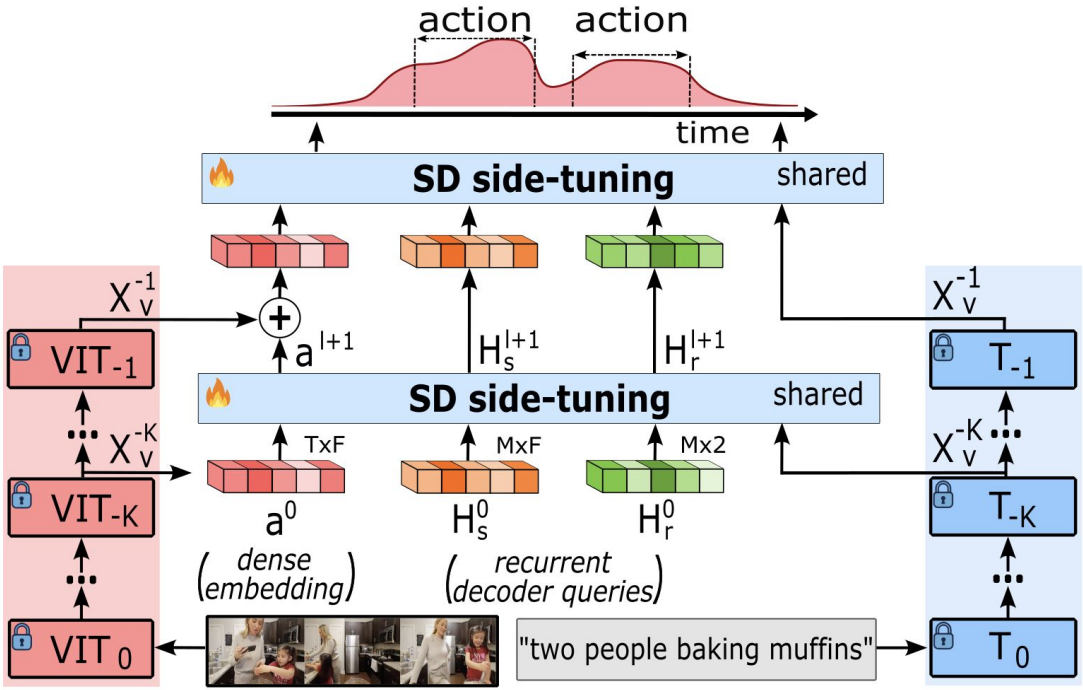
- ➔ Combine intermediate frozen representations into final representation.
- ➔ Leverage this representation for the final task prediction.



- ➔ **Parameter efficient:** Only a small additional learnable module needed.
- ➔ **Memory efficient:** Backpropagation only on the parallel stream.

SDST: General overview

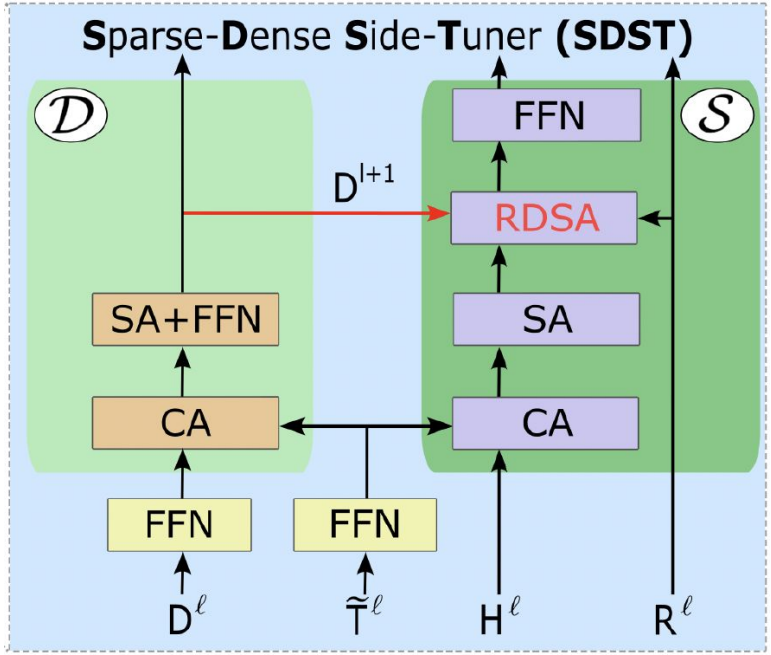
- ➔ A frozen backbone extracts *K* intermediate multi-modal embeddings.
- ➔ Our shared SDST module combines them using a sparse and a dense stream.



Pujol-Perich, David et al.. "Sparse-dense side-tuner for efficient video temporal grounding." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025.

SDST: General overview

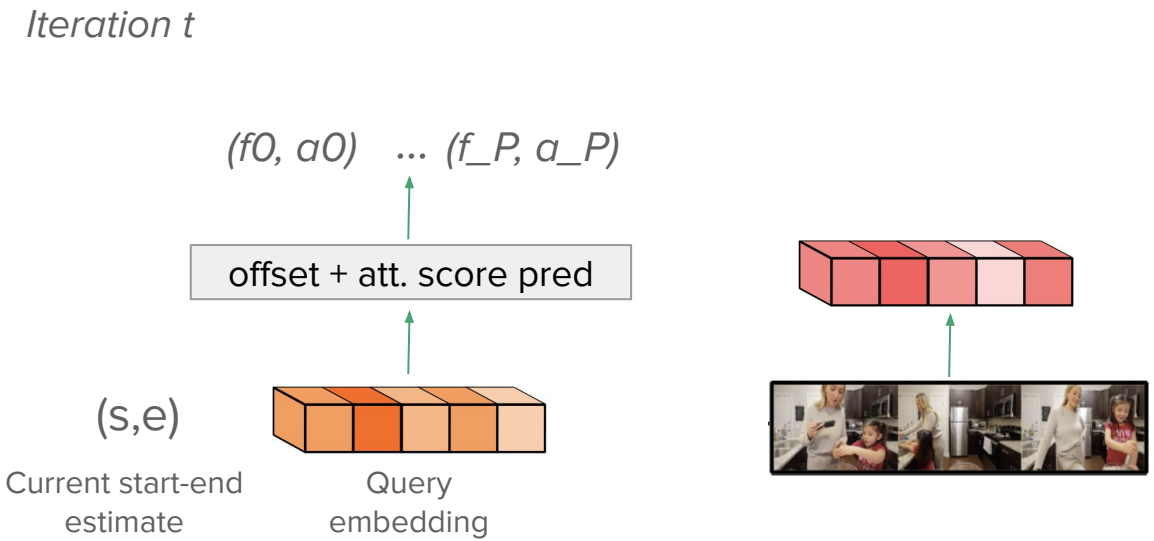
- ➔ The dense stream refines frame embeddings $\boldsymbol{D}$.
- ➔ The sparse stream uses recurrent decoder queries $\boldsymbol{H}$.
- ➔ The RDSA module is traditionally replaced by a Deformable Cross Attention.



Standard deformable attention

➔ Efficient alternative to full self-attention.

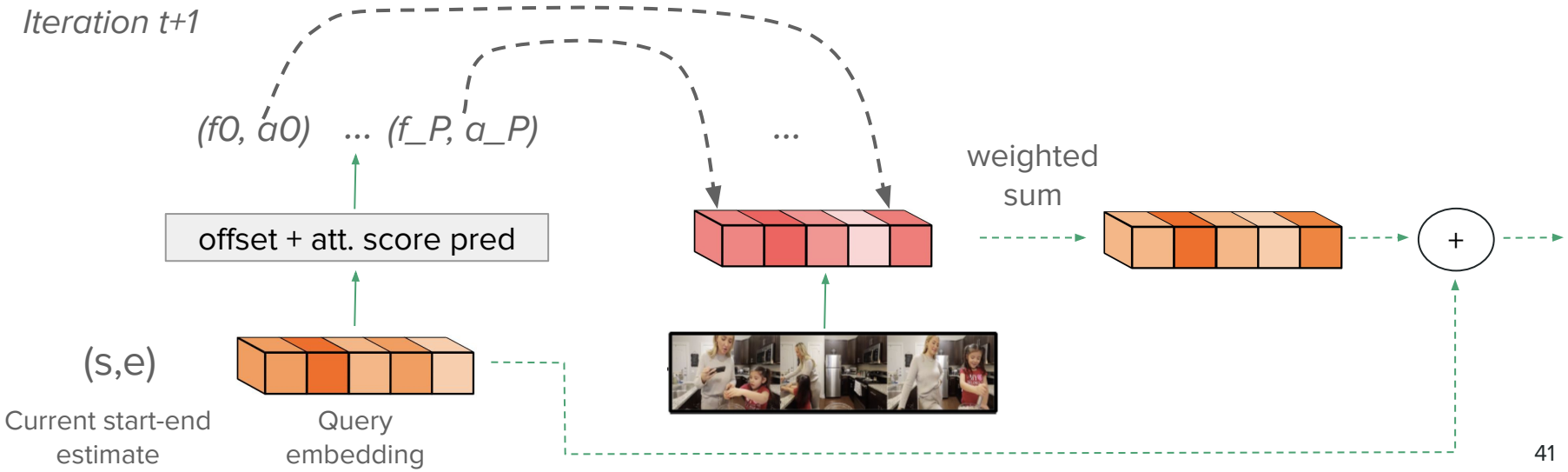
➔ Every query predicts *P* positions to attend to, and then aggregates them.



Standard deformable attention

➔ Efficient alternative to full self-attention.

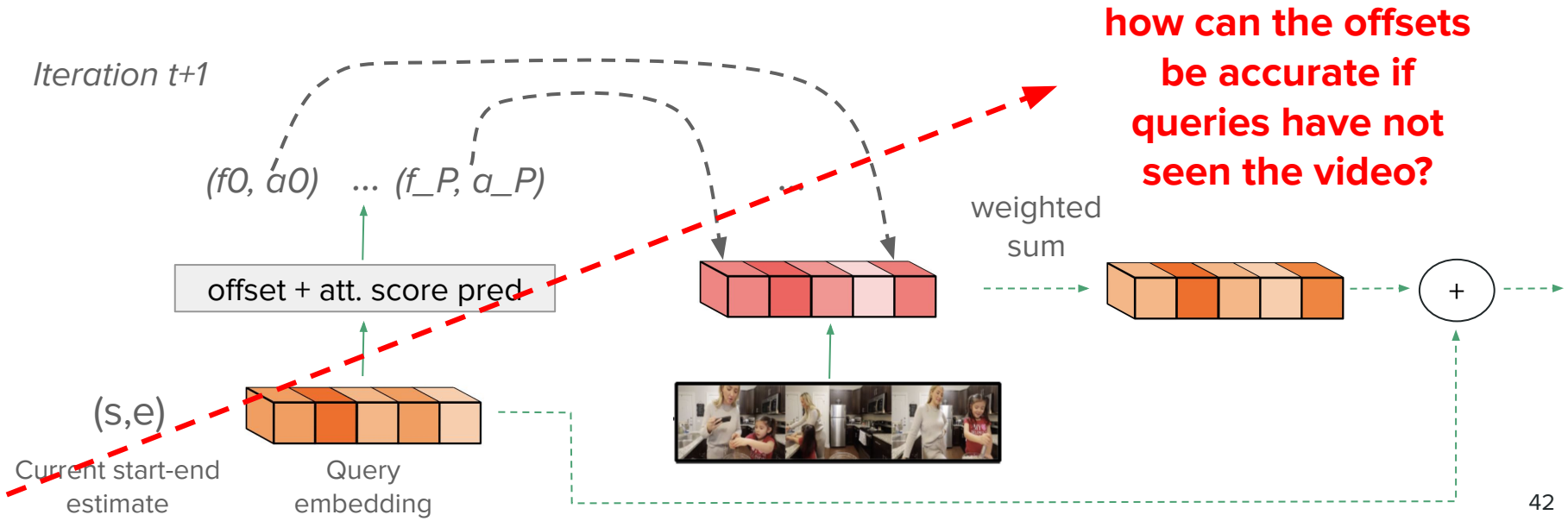
➔ Every query predicts *P* positions to attend to, and then aggregates them.



Standard deformable attention

➔ Efficient alternative to full self-attention.

➔ Every query predicts *P* positions to attend to, and then aggregates them.

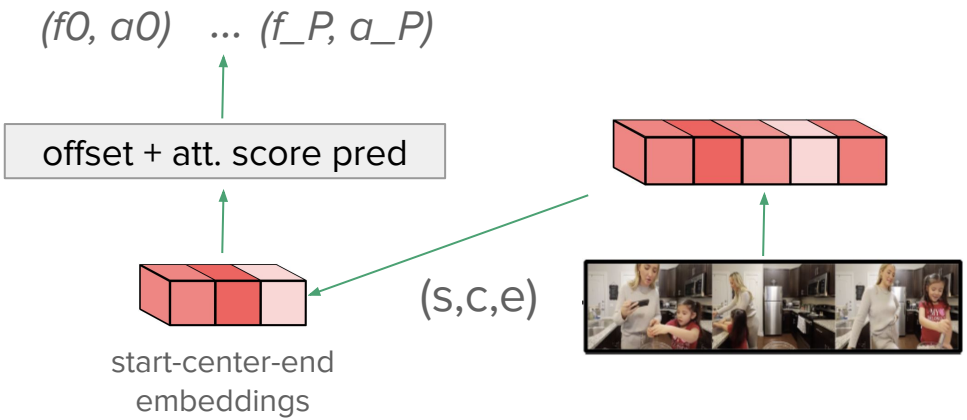


RDSA: Fixing the deformable attention collapse

➔ We propose the **RDSA**.

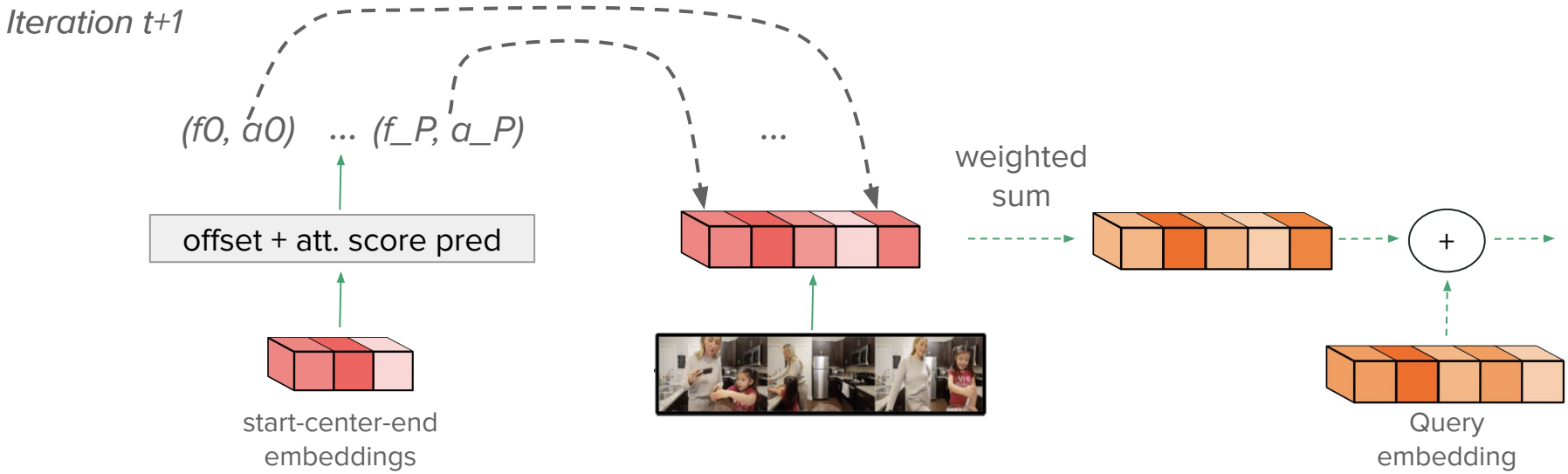
➔ Leverage the estimate of the start, center and end for the offset prediction.

Iteration t



RDSA: Fixing the deformable attention collapse

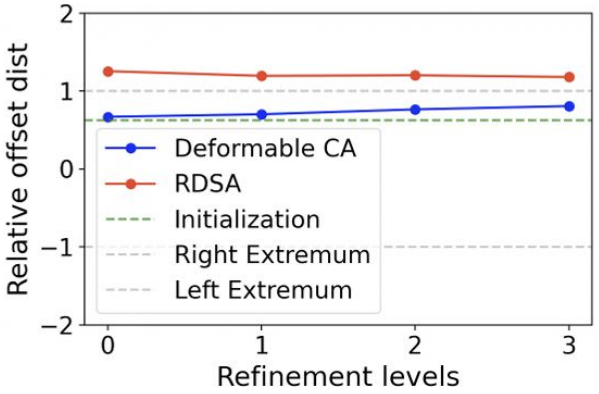
- ➔ We propose the **RDSA**.
- ➔ Leverage the estimate of the start, center and end for the offset prediction.



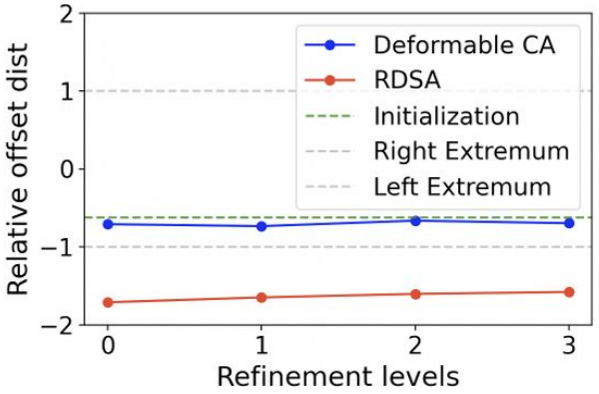
Fixing the deformable attention collapse

➔ This overcomes the previous collapse around the initialization.

➔ RDSA looks for more complex patterns beyond the current action boundaries.



Head 0

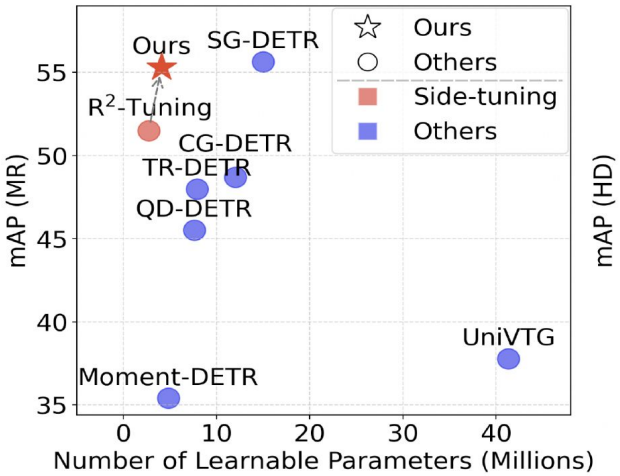


Head 1

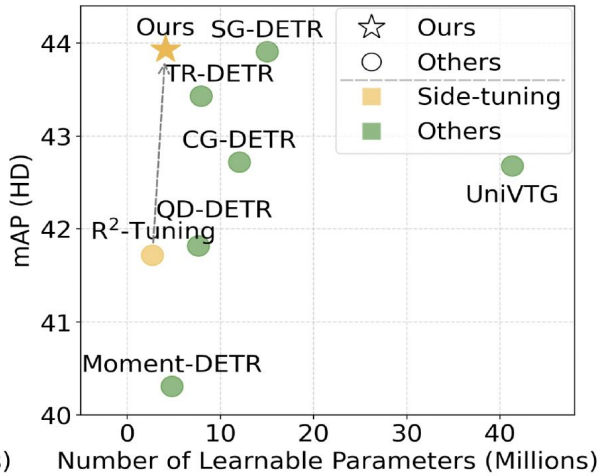
Main results

➔ Our SDST attains SOTA performance with minimal number of parameters.

Sparse detection



Highlight detection



Ablations

1. SDST is suitable for memory constrained scenarios.
2. Leveraging the sparse-dense nature of the task poses an important advantage.

	# Params (M)	Memory (GB)	Runtime (it/s)
Moment-DETR	4.8	1.54	7.45
R2-Tuning	2.7	2.4	5.55
TR-DETR	7.9	1.76	4.75
HL-CLIP	2.0	22.98	0.64
Llava-MR	17.0	$\approx 80 \times 8$	-
MR.Blip	19.0	$\approx 80 \times 8$	-
SG-DETR	15.0	-	-
Flash-VTG	10.9	2.3	5.2
Ours	4.1	3.4	4.16

Method	#Params (M)	Memory (GB)	MR R1@0.5	MR R1@0.7	MR mAP	HD mAP	HD HIT@1
w/o Tuning	2.70	2.35	66.97	51.10	46.19	41.45	67.23
E^3VA [47]	2.57	2.96	68.97	53.16	47.68	41.04	68.13
LoSA [9]	6.40	**2.39**	72.13	58.32	53.73	41.82	68.19
LST [38]	**2.04**	2.49	70.32	55.55	50.59	41.53	69.48
R^2-Tuning[25]	2.70	2.44	70.84	55.35	51.30	41.64	69.74
Ours	4.10	3.40	**73.68**	**60.90**	**55.60**	**44.00**	**72.00**

Conclusions

- ➔ **SDST** is the first side-tuning approach tailored for sparse-dense tasks like VMR.
- ➔ The novel **RDSA** solves the deformable attention collapse problem.
- ➔ **SOTA** performance while being parameter and memory efficient.

Conclusions:

How did we advance generalization and transferability?

- ➔ **Chapter 1:** We showcase the importance of semantic-aware visual alignment in unsupervised domain adaptation.
- ➔ **Chapter 2:** Identify and address a critical linguistic generalization gap in VMR.
- ➔ **Chapter 3:** Show that parameter-and-memory efficient fine-tuning can be competitive.

Thesis conclusions

1. Benchmark success often does **not equal generalization.**
2. True progress **requires probing** *how* models fail under domain shift, linguistic ambiguity.
3. **Without** effective **transferability**, massive foundation models will **not meaningfully impact** day-to-day applications.

Limitations and future directions

1. SADA's pseudo-labels **introduce noise** under severe domain shift.
2. Our linguistic OOD benchmarks use LLM-simplified captions, **not real** user logs — the true deployment gap may be larger.
3. Human-centered retrieval requires **interactive** query-refinement-based VMR.

Thesis publications

Main publications:

- Pujol-Perich, David et al. "SADA: Semantic Adversarial Unsupervised Domain Adaptation for Temporal Action Localization." *IEEE/CVF Winter Conference on Applications of Computer Vision* 2025.
- Pujol-Perich, David et al.. "Sparse-dense side-tuner for efficient video temporal grounding." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025.
- Pujol-Perich, David, et al. "Beyond Caption-Based Queries in Video Moment Retrieval." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2026.

Collaborations:

- Barsbey, Melih, et al. "Higher-order molecular learning: The cellular transformer." *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*. 2025.

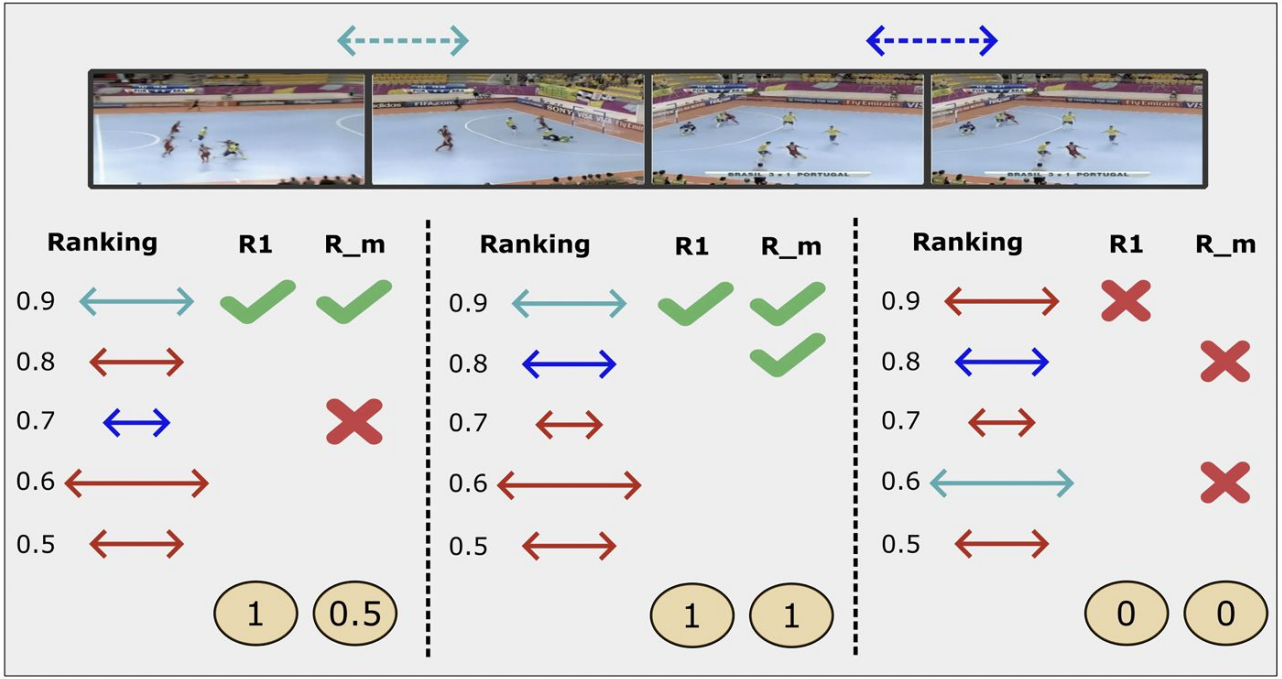
Thank you

Questions?

David Pujol Perich
Supervisors: Sergio Escalera and Albert Clapés

Appendix:

A. Metrics: R1_m



A. Metrics: mAP_m