The Unreasonable Effectiveness of Large Language-Vision Models for Video Domain Adaptation

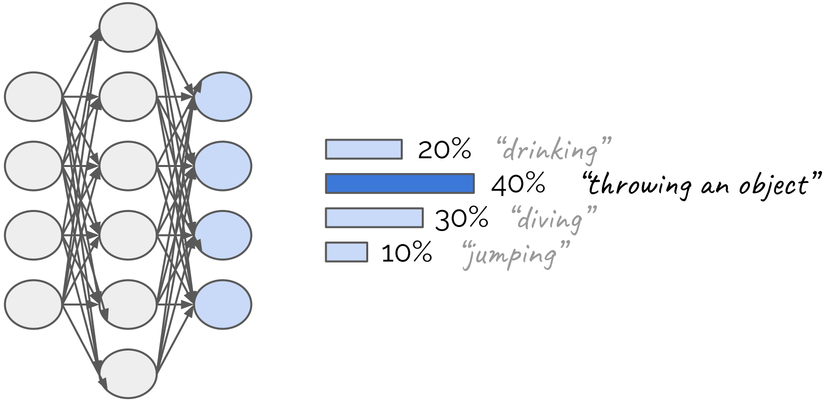
Elisa Ricci

UNIVERSITÀ DI TRENTO

FONDAZIONE BRUNO KESSLER

Action Recognition

$$(X, y)$$



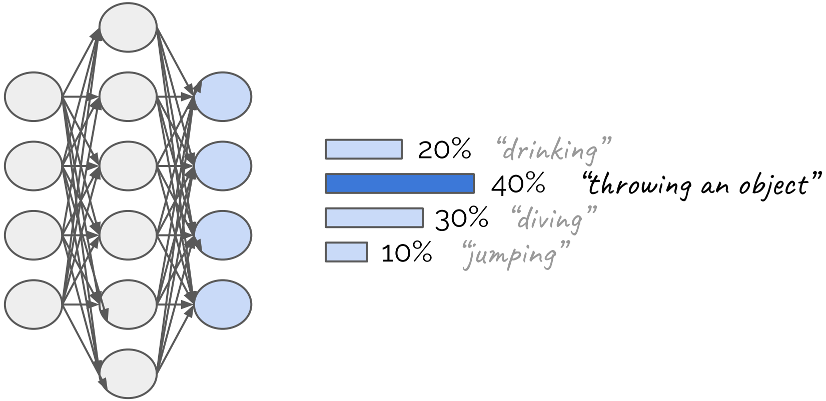
Goal: Learn to recognize human actions from labelled data.

Challenges

- Leveraging the temporal dimension
  - How to effectively model spatio-temporal data?
- Complexity
  - Impact on storage and computational cost
- Annotated large-scale datasets availability

Action Recognition

(X, y)



Goal: Learn to recognize human actions from labelled data.

Downside: Expensive and time-consuming to collect **annotations**.

Solution: Leverage **unlabelled** data.

Challenge: Domain Shift

- Unlabelled (or *target* domain) videos exhibit **domain shift**.

$$p(\mathcal{X}^S) \neq p(\mathcal{X}^T)$$

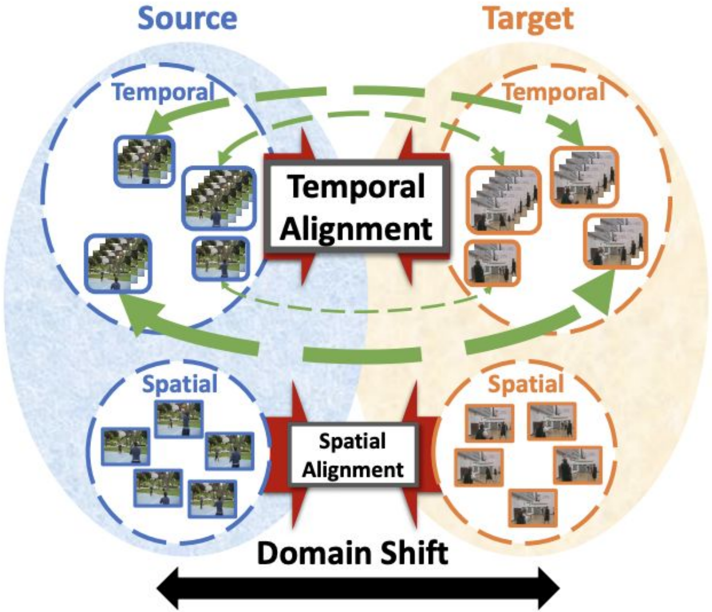
- Domain shift can arise due to several **factors**:
  - lighting
  - resolution
  - environment
  - camera position

(X^S, y)

$(X^T, ?)$

Unsupervised Domain Adaptation (UDA) with Attention

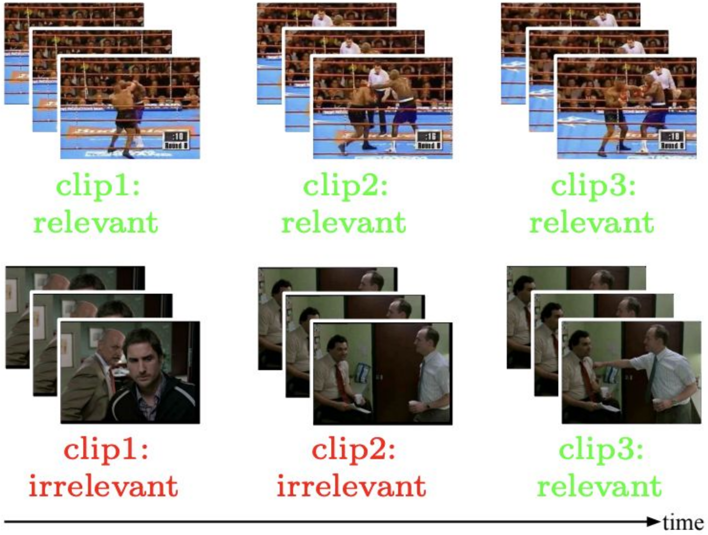
Attention mechanism to effectively align the temporal representations
Domain **adversarial loss** at spatio-temporal levels



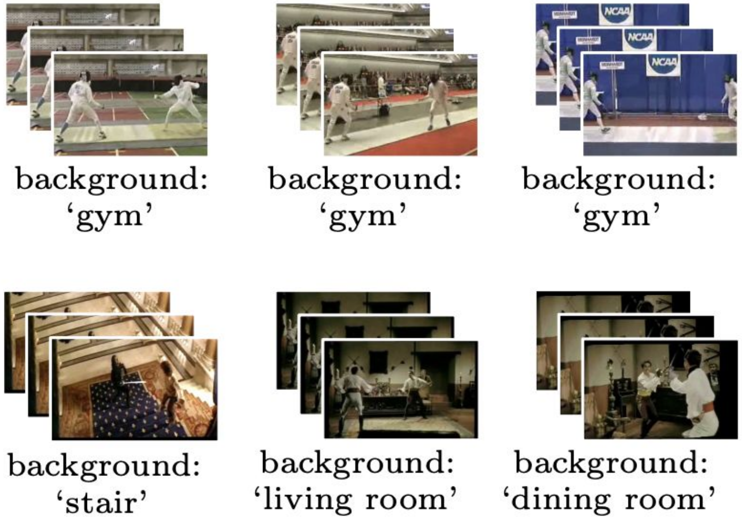
[Chen et al. ICCV 2019] [Pan et al. AAAI 2020]

Pretext Tasks for UDA

Clip Attention



Clip Ordering Prediction



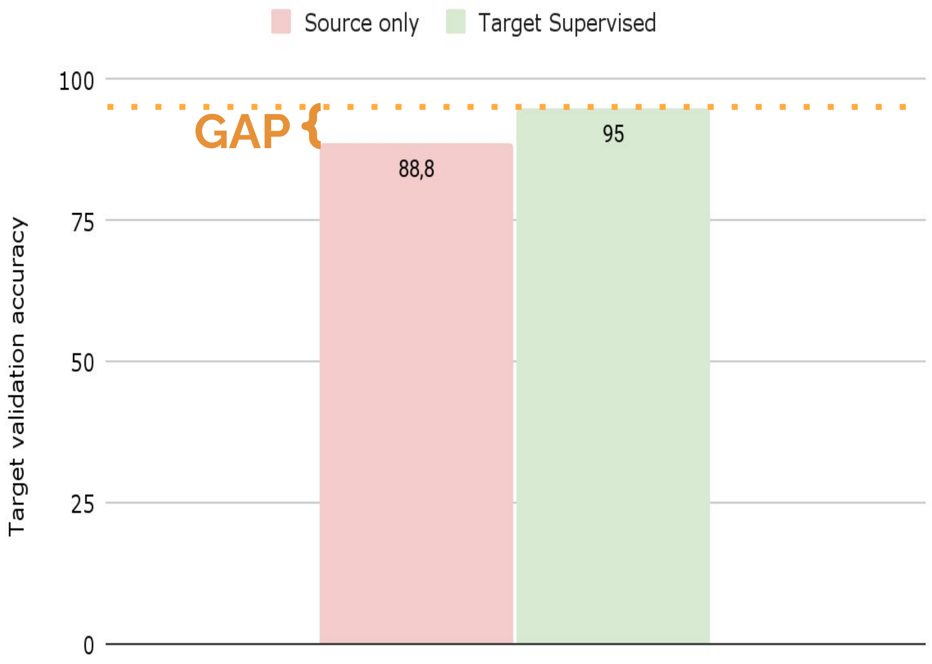
[Choi et al. ECCV2020]

Measure the gap

HMDB51

"climb"

"golf"



UCF101

"playing guitar"

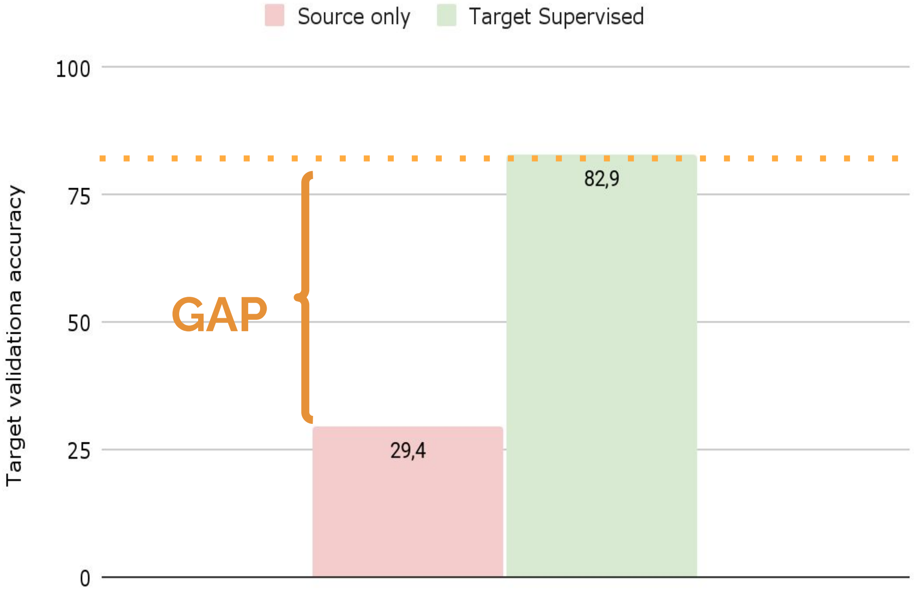
"walking the dog"

Measure the gap

Kinetics

"jogging"

"punching person"



NEC drone

"jumping"

"drinking from a bottle"

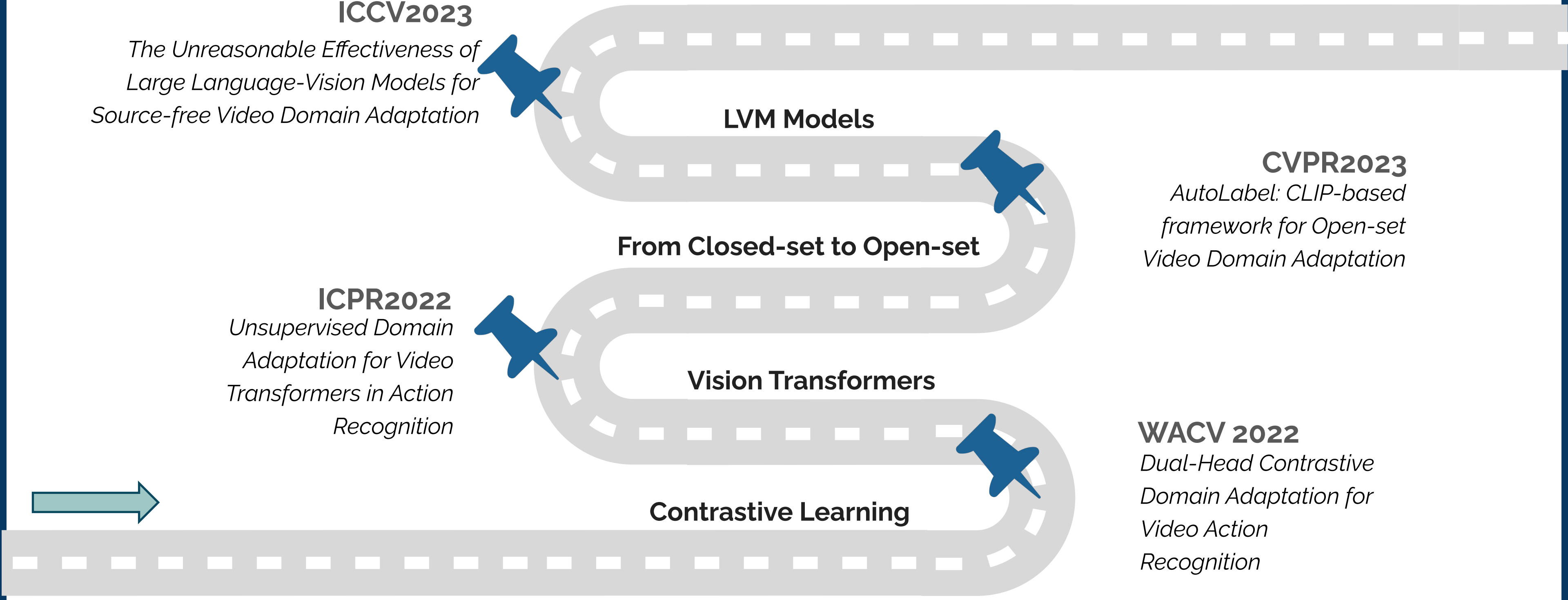
Our Journey

ICCV2023
The Unreasonable Effectiveness of Large Language-Vision Models for Source-free Video Domain Adaptation

CVPR2023
AutoLabel: CLIP-based framework for Open-set Video Domain Adaptation

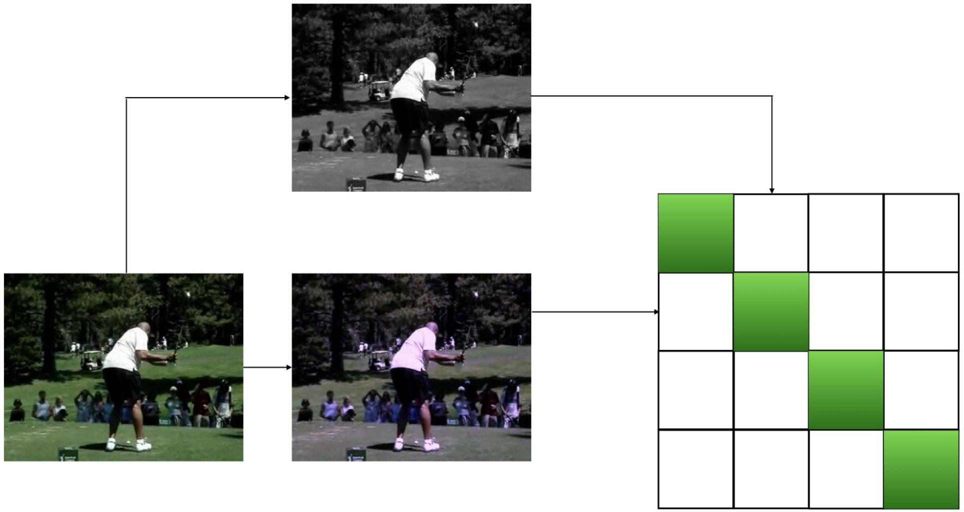
ICPR2022
Unsupervised Domain Adaptation for Video Transformers in Action Recognition

WACV 2022
Dual-Head Contrastive Domain Adaptation for Video Action Recognition



No pretext task, instead Contrastive Learning

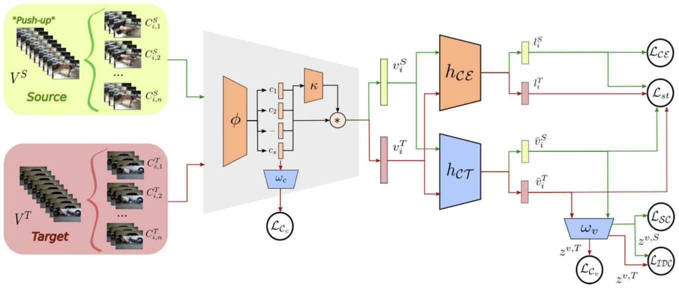
Contrastive Learning: Self-supervised feature representation learning make model prediction robust to domain shift

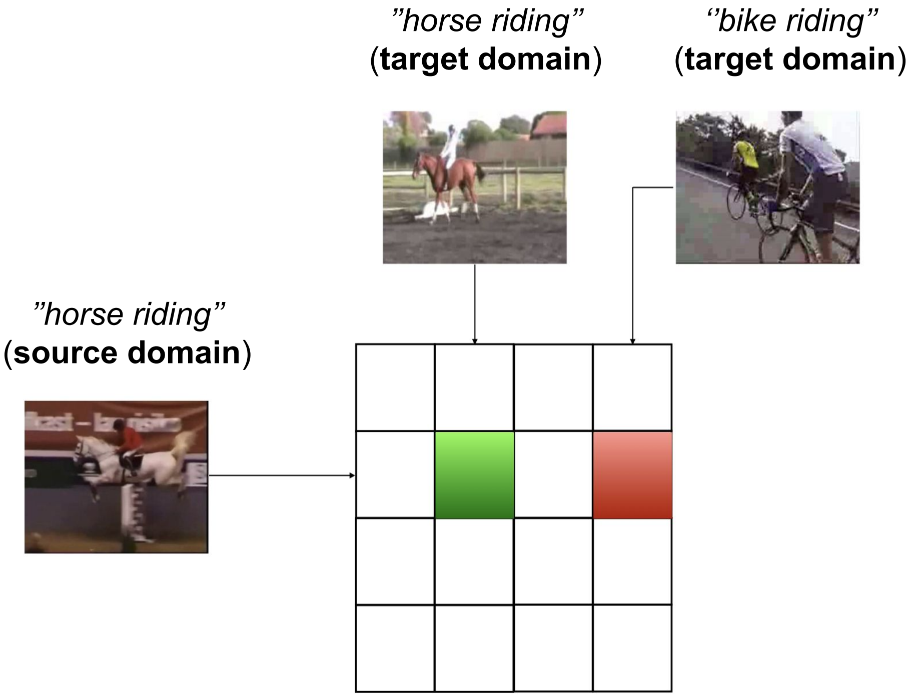


[Chen et al. ICML2020]

Supervised cross-domain representation learning

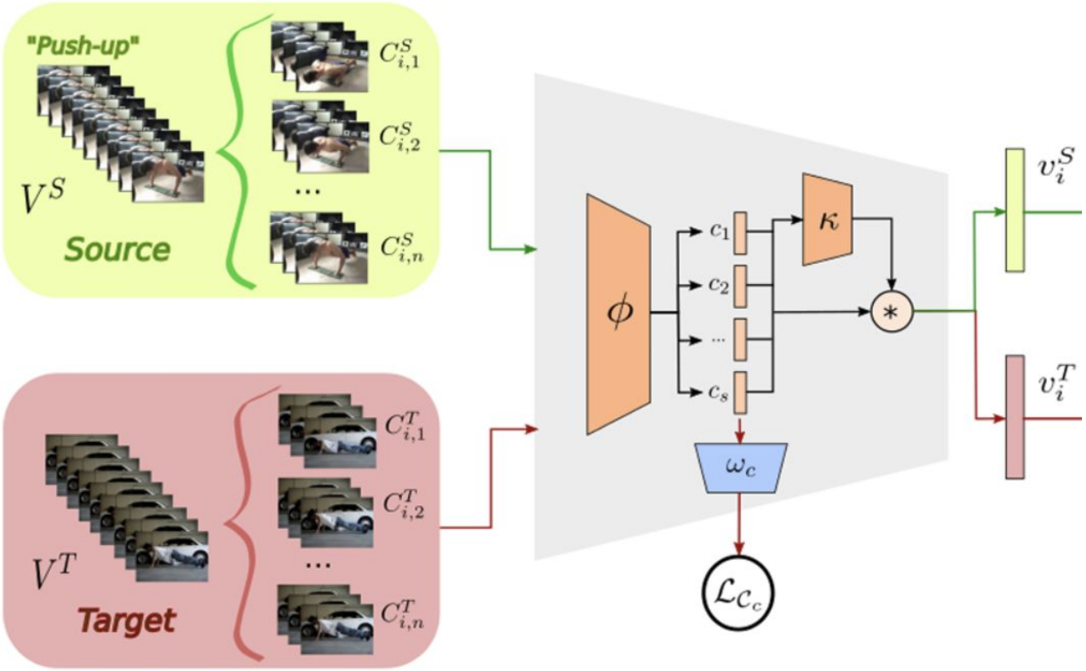
- **Pull together** video representations from different domains belonging the same class
- **Push apart** video representations from different domains belonging different classes





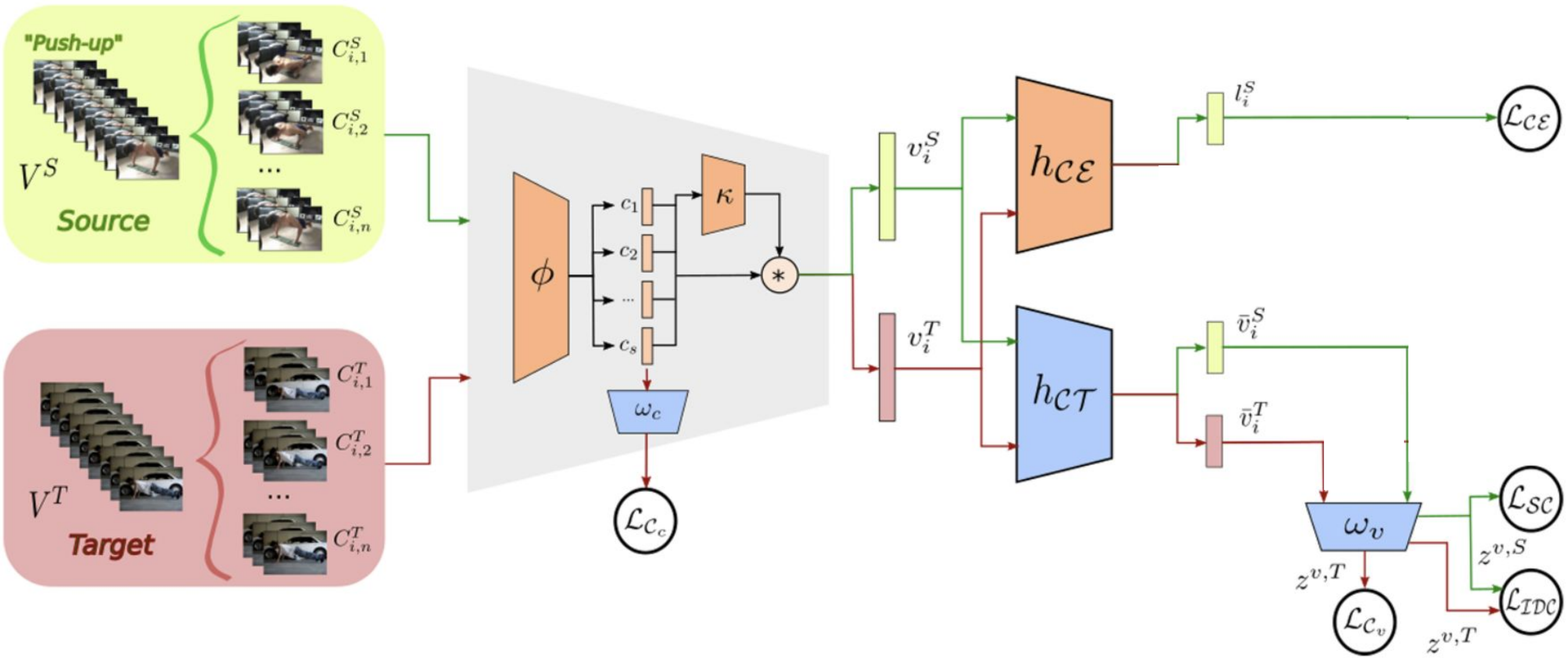
[Turrisi et al. WACV2022]

Proposed architecture



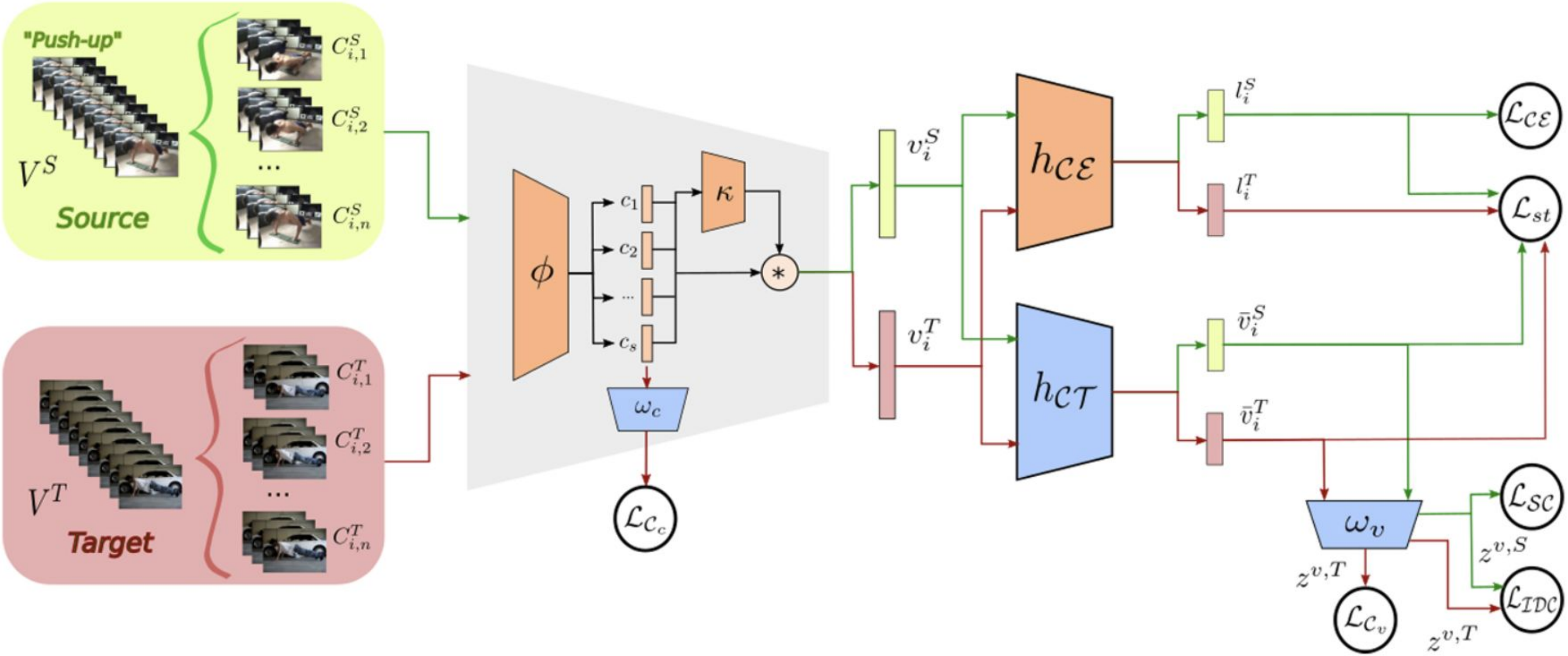
[Turrisi et al. WACV2022]

Proposed architecture



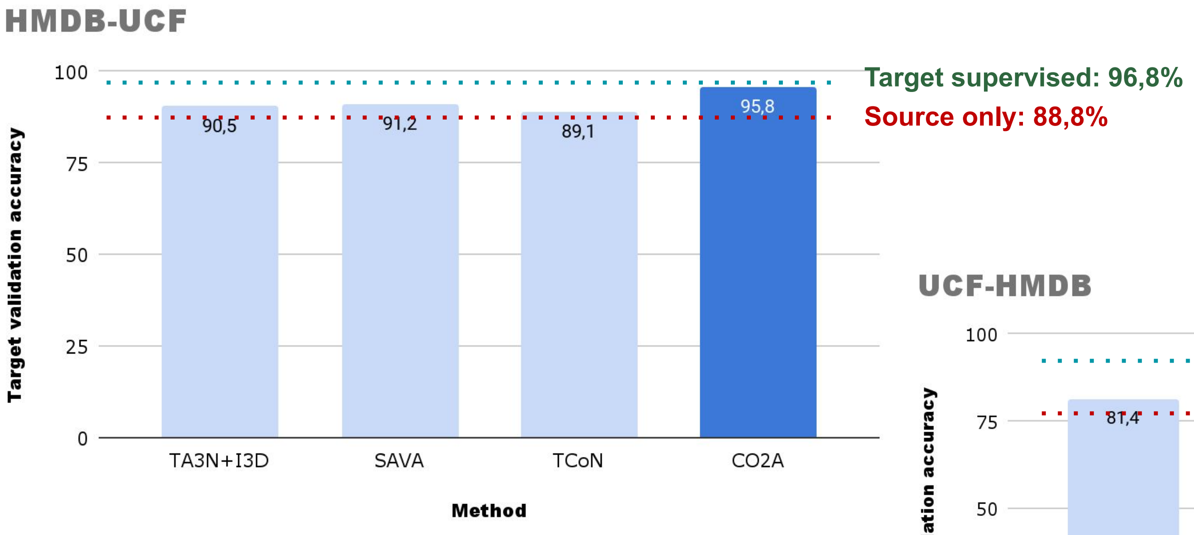
[Turrisi et al. WACV2022]

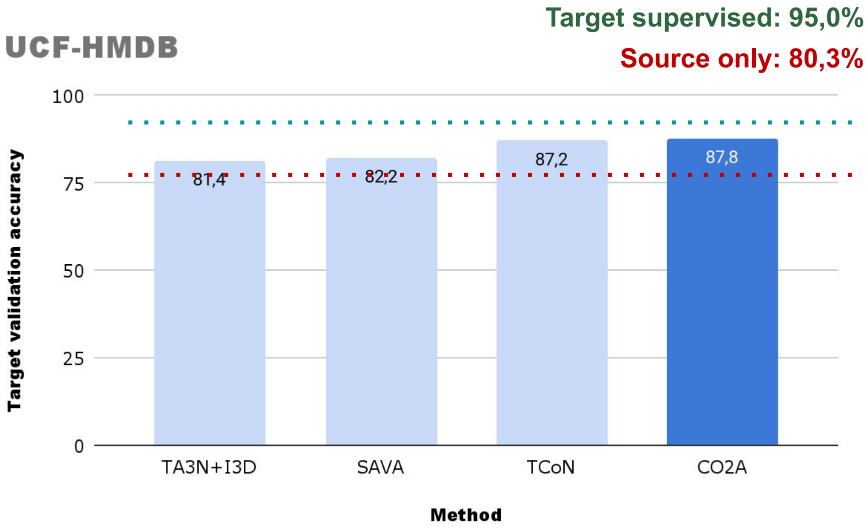
Proposed architecture



[Turrisi et al. WACV2022]

Results UCF ↔ HMDB

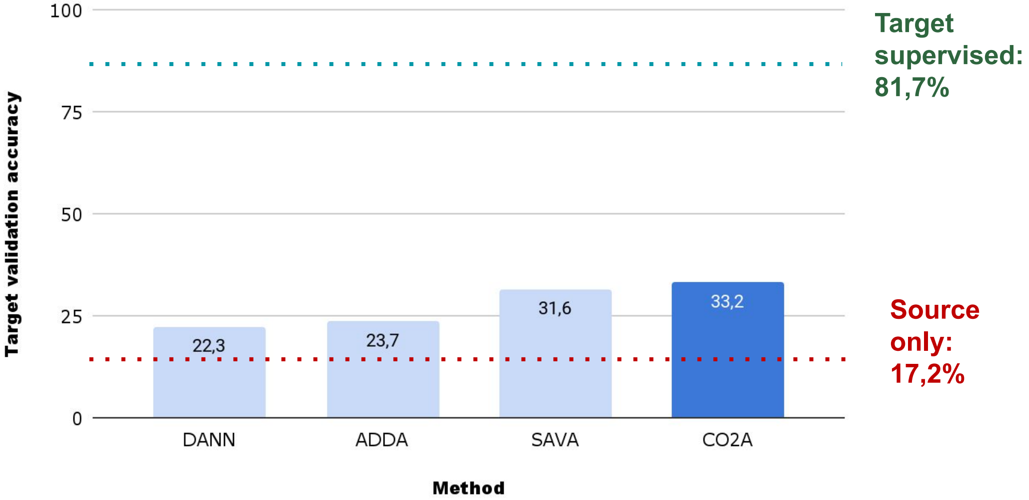




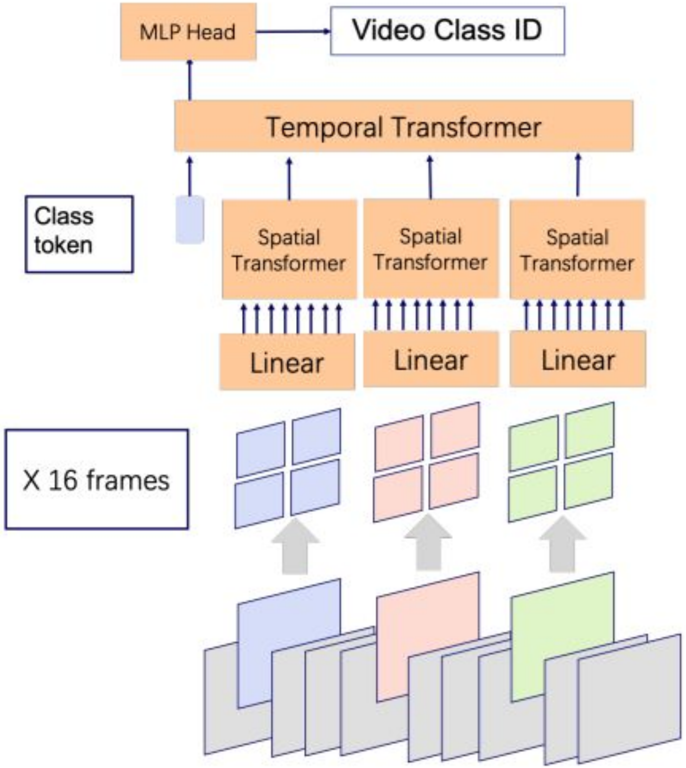
Results on Kinetics → NEC-Drone

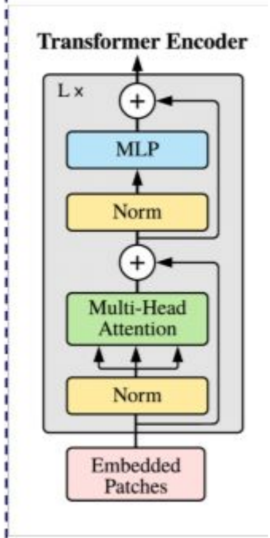
Kinetics

NEC drone



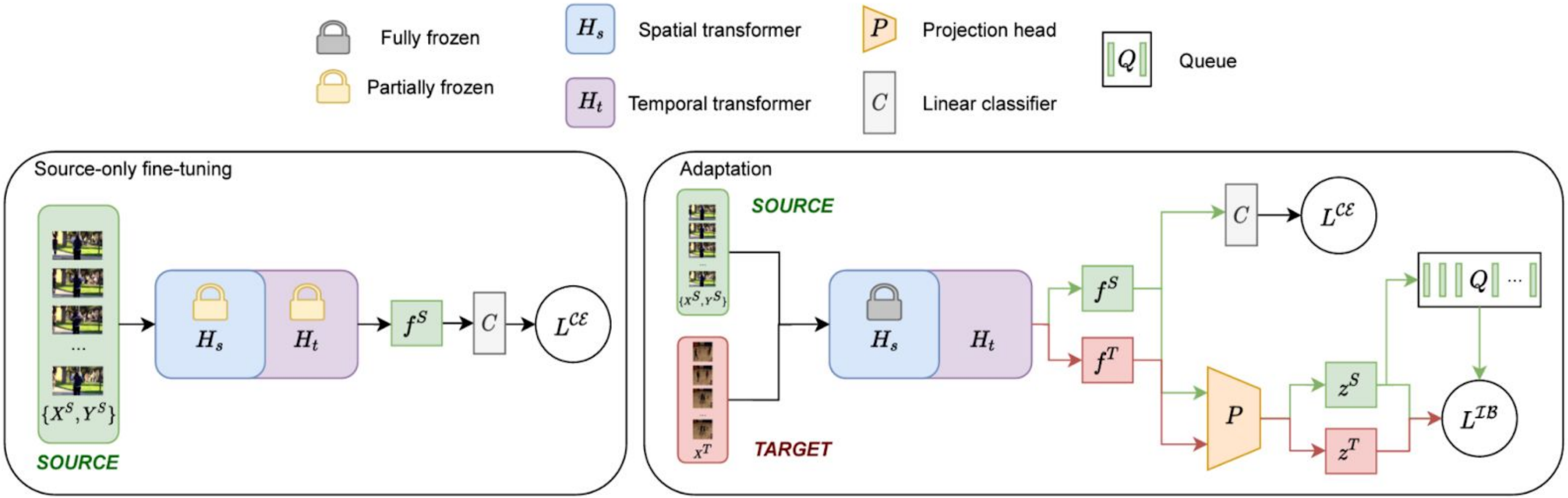
Video Transformers





[Shair et al. 2021]

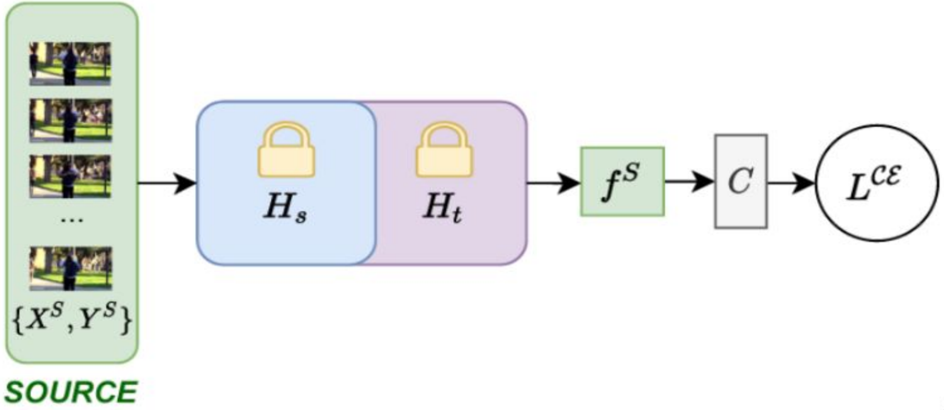
Cross-Domain Video Transformers

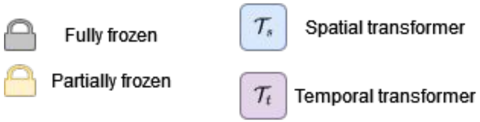


[Turrisi et al. ICPR2022]

Cross-Domain Video Transformers

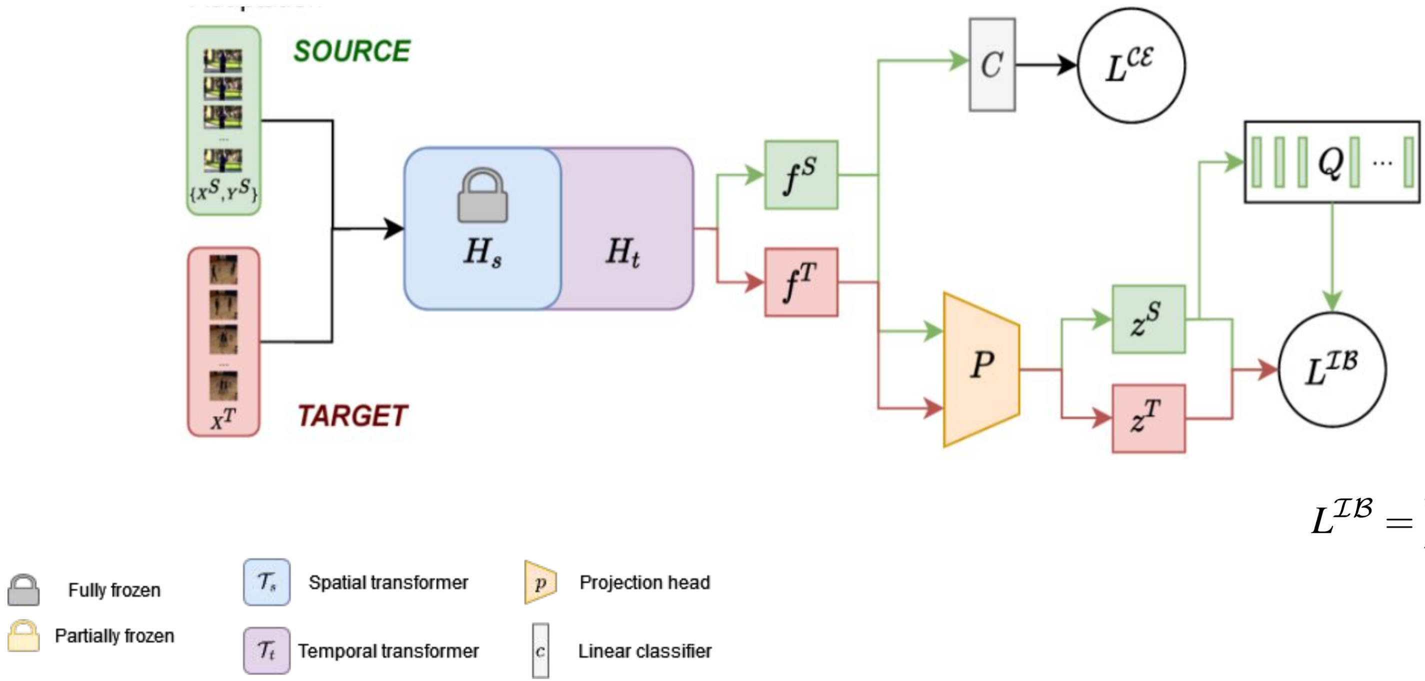
Step 1: source-only fine tuning

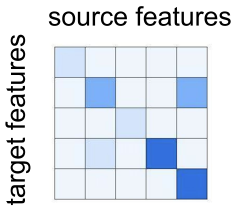




Cross-Domain Video Transformers

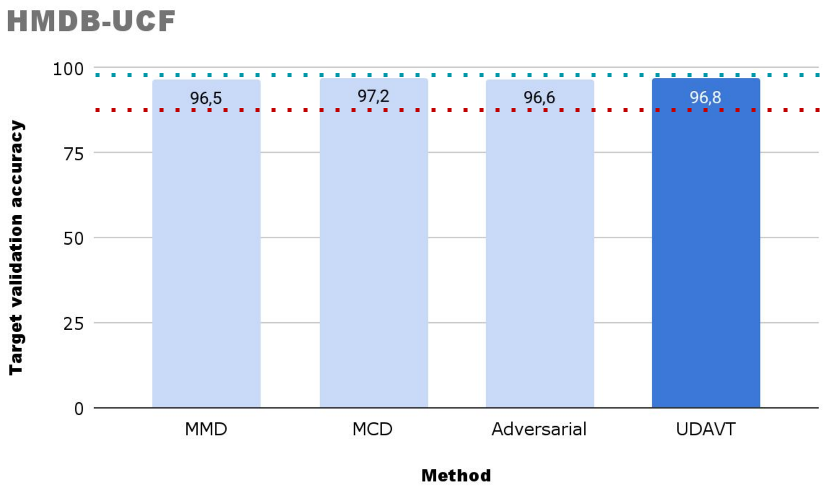
Step 2: adaptation





$$L^{\mathcal{IB}} = \sum_{i}^{d}(1 - C_{ii})^2 + \lambda \sum_{i}^{d}\sum_{j \neq i}^{d}(C_{ij})^2$$

Results UCF ↔ HMDB

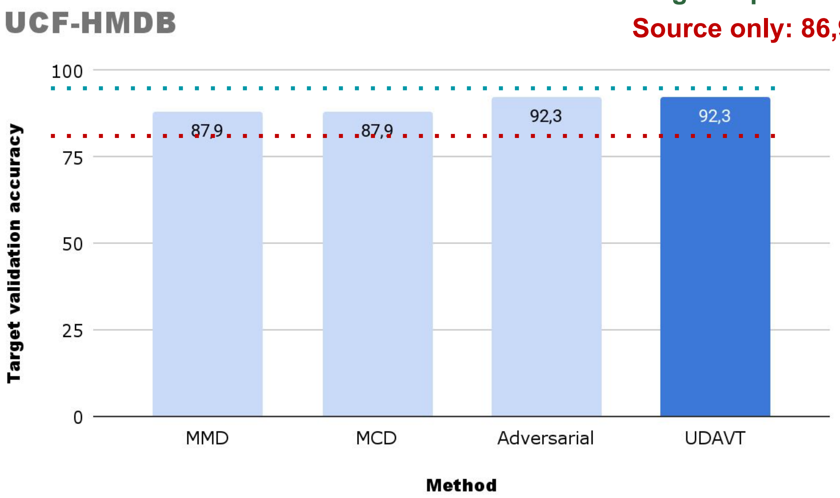


Target supervised: 97,9%

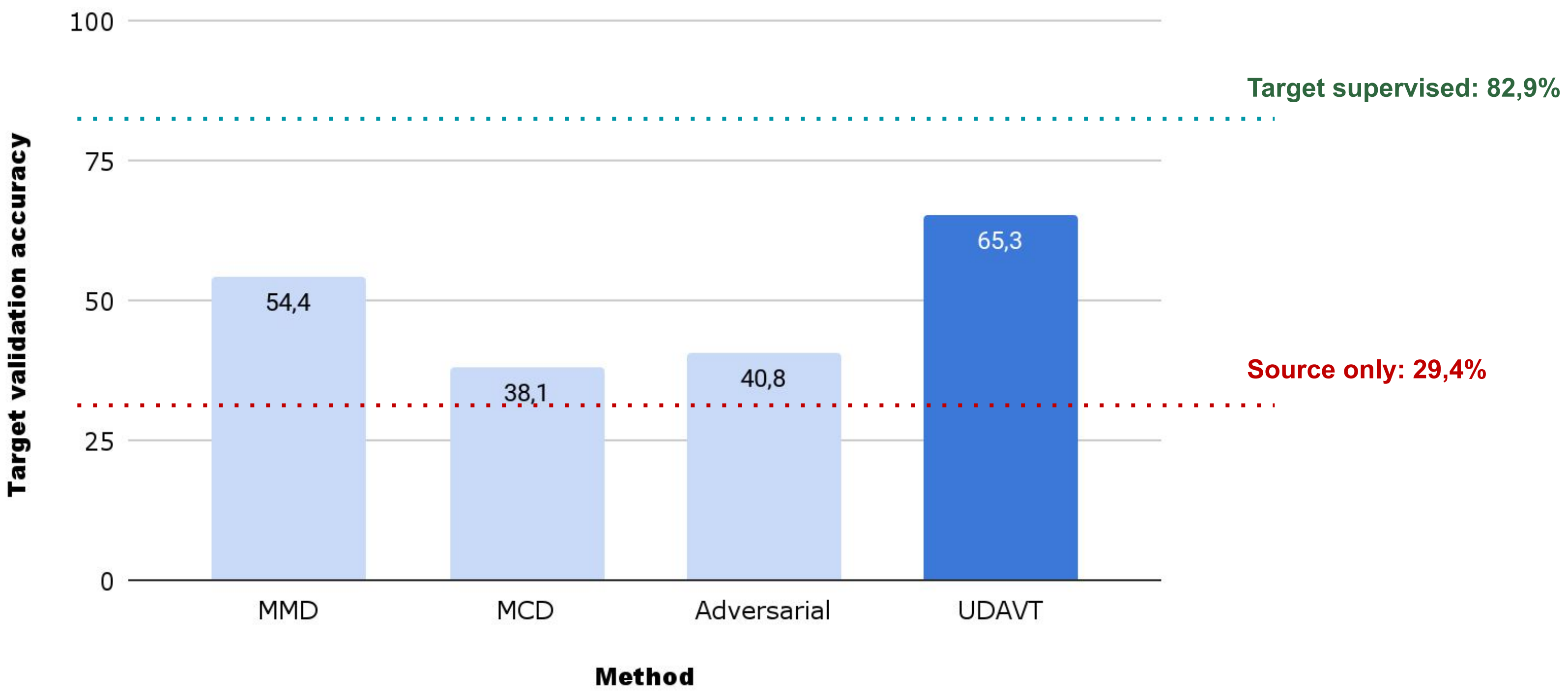
Source only: 93.7%

Target supervised: 95,8%

Source only: 86,9%



Results on Kinetics → NEC-Drone



Results on Kinetics → NEC-Drone

What we learned

- Methods from **self-supervised learning** can be adapted for cross-domain feature alignment
- **Video Transformers** are more robust to domain shift but they need to be adapted
- Domain shift is a severe issue also in the **egocentric setting**: EPIC-Kitchens Unsupervised Domain Adaptation Challenge [1]

[1] https://epic-kitchens.github.io/2023#challenge-domain-adaptation

So far: Closed-set Domain Adaptation

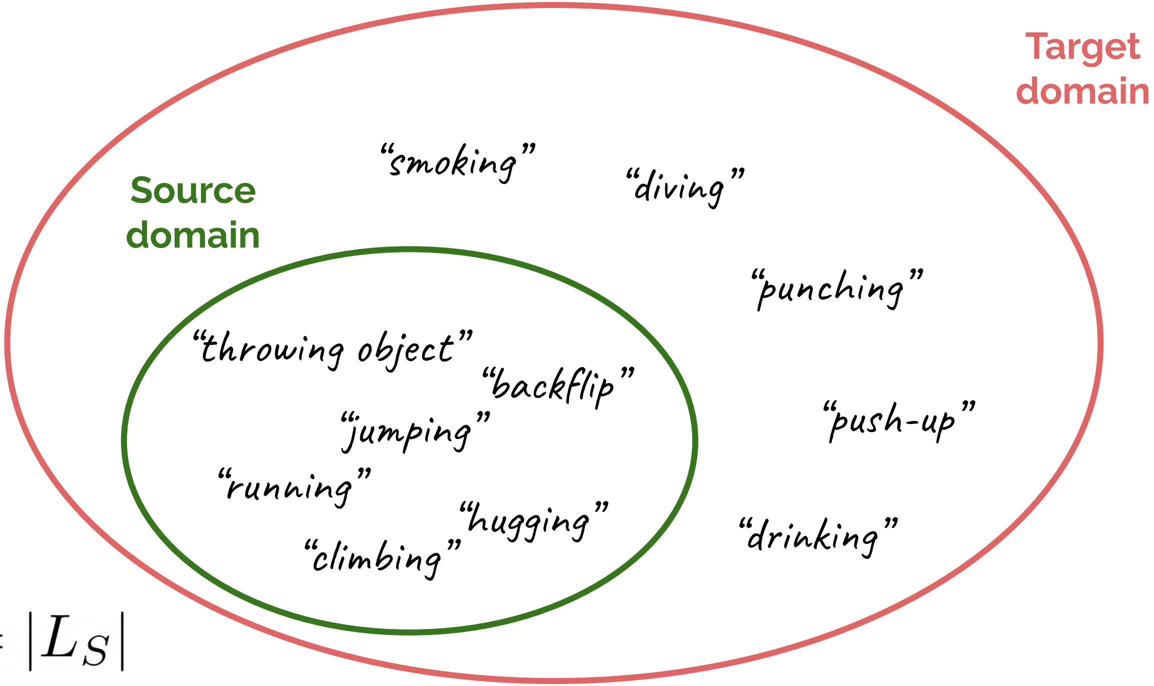
Source and **target** label sets are the same

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

AutoLabel: CLIP-based framework for **Open-set** Video Domain Adaptation

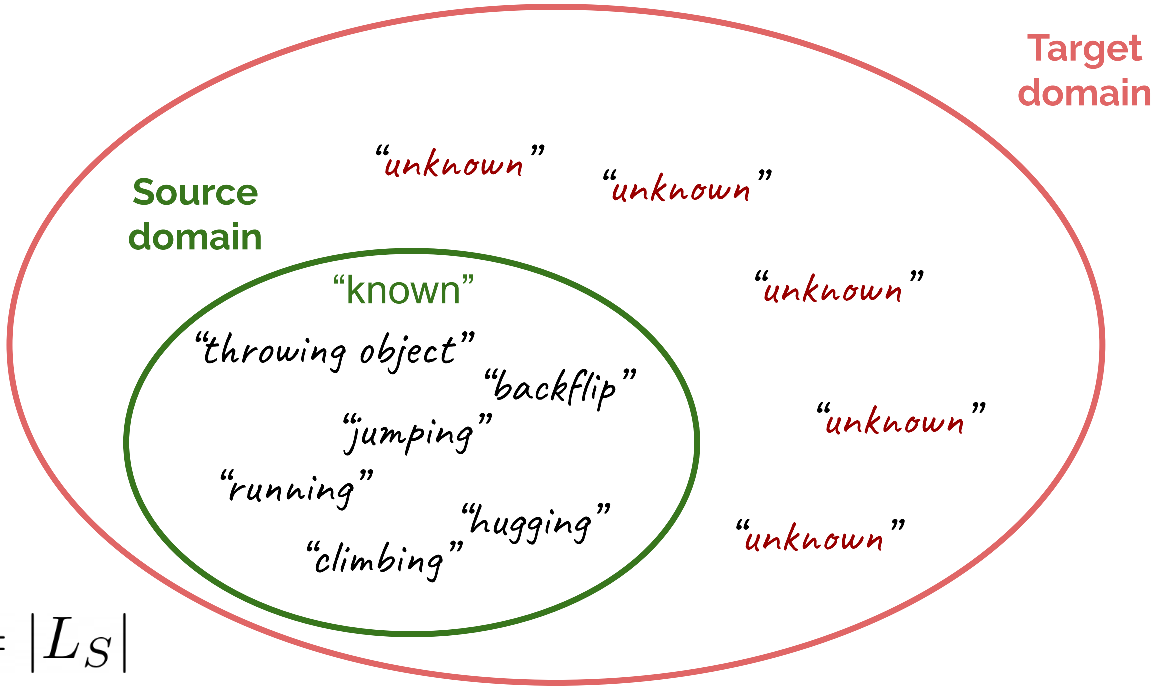
Giacomo Zara, Subhankar Roy, Paolo Rota, Elisa Ricci

Challenge: Open-set classes in Target



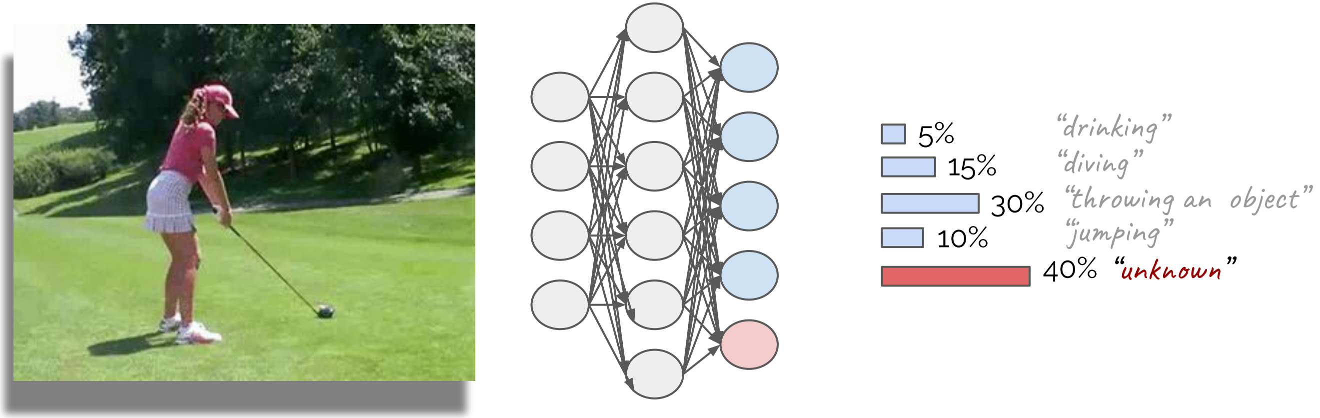
$$|L_S \cap L_T| = |L_S|$$

Challenge: Open-set classes in Target



$$|L_S \cap L_T| = |L_S|$$

Open-set Video Domain Adaptation



$(X^T, ?)$

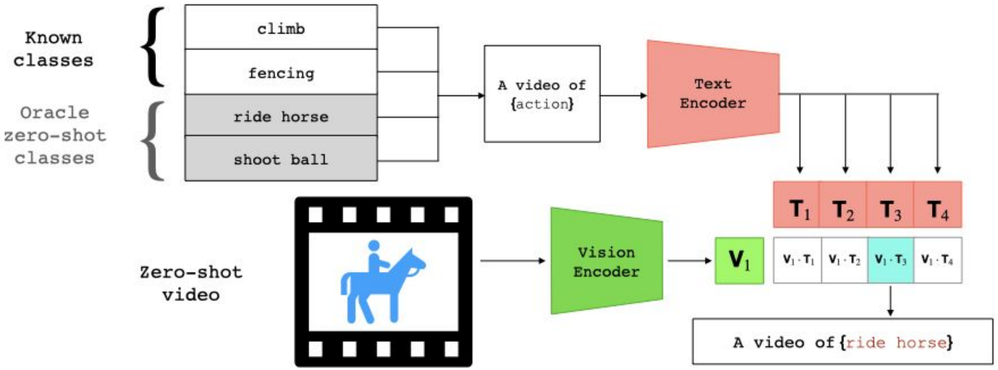
Goal: Adapt a model to the target domain that can:

- classify a sample to one of the 'known' classes in L_S
- reject the 'unknown' sample belonging to L_T / L_S

CLIP: Large Language & Vision Models

Why CLIP[1]?

- **robust** to domain shifts due to web-scale pre-training
- enables **zero-shot classification**

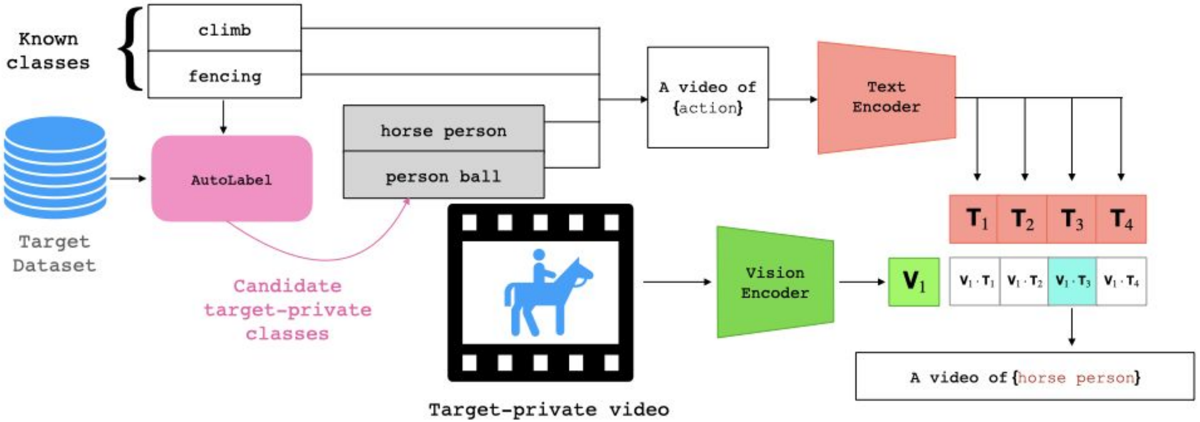


Downside: CLIP assumes knowledge of the **class names** in order to carry out zero-shot classification.

How to leverage CLIP *without* any a priori knowledge of the 'unknown' class names?

[1]Radford et al., "Learning Transferable Visual Models From Natural Language Supervision". In ICML, 2021.

Proposed Method: AutoLabel



Automatically discover the 'unknown' (or **target private**) class names and **extend** the 'known' classes label set.

Intuition behind AutoLabel

Discovering unknown class names

- An action can be *loosely* defined by:
    - object(s)
    - actor(s)
    - environment
- We aim to **discover** the *candidate* 'unknown' class names by finding **attributes** that appear in the video sequences.

$$(X^T, ?)$$

captioner

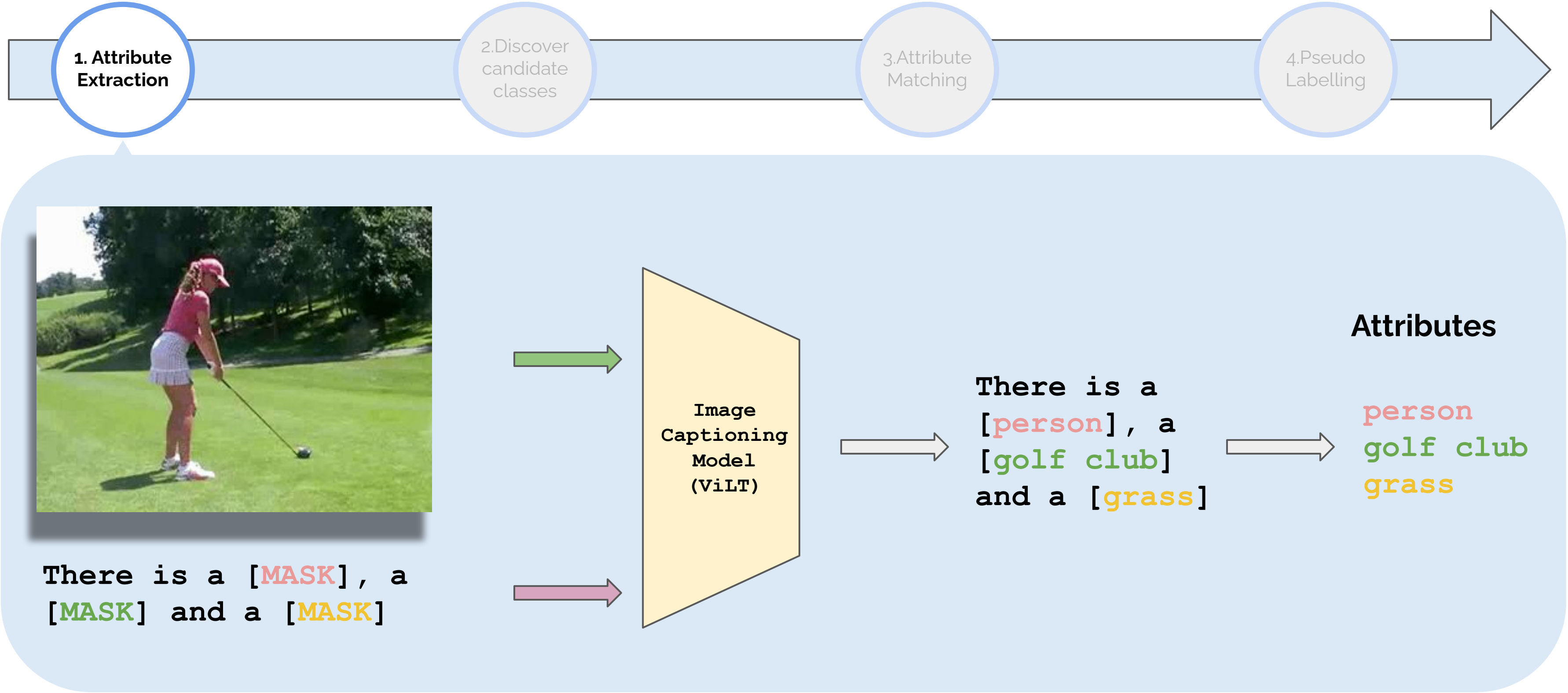
Attributes

- person
- golf club
- grass
- trees
- ...

Image captioning models[2] can serve the purpose.

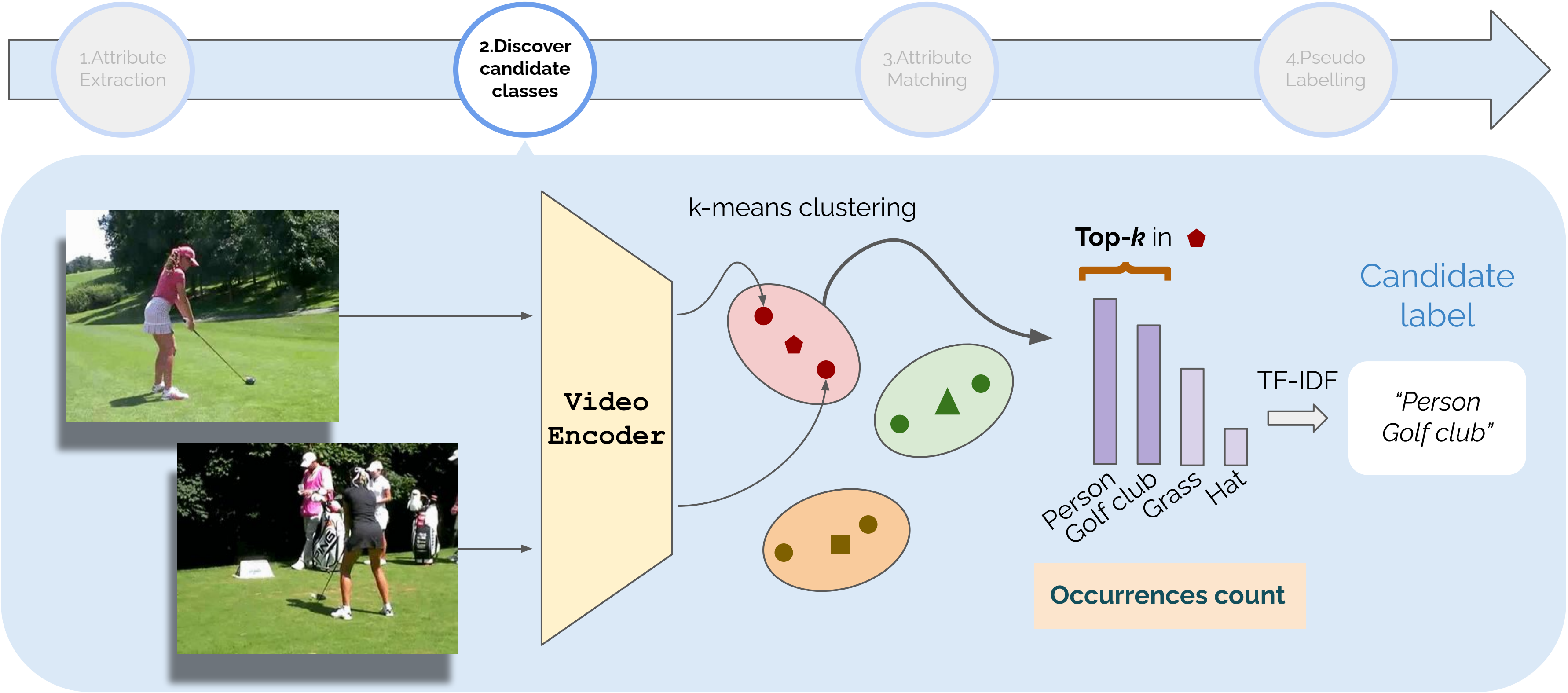
[2]Kim et al., "Vilt: Vision and-language transformer without convolution or region supervision". In ICML, 2021.

AutoLabel

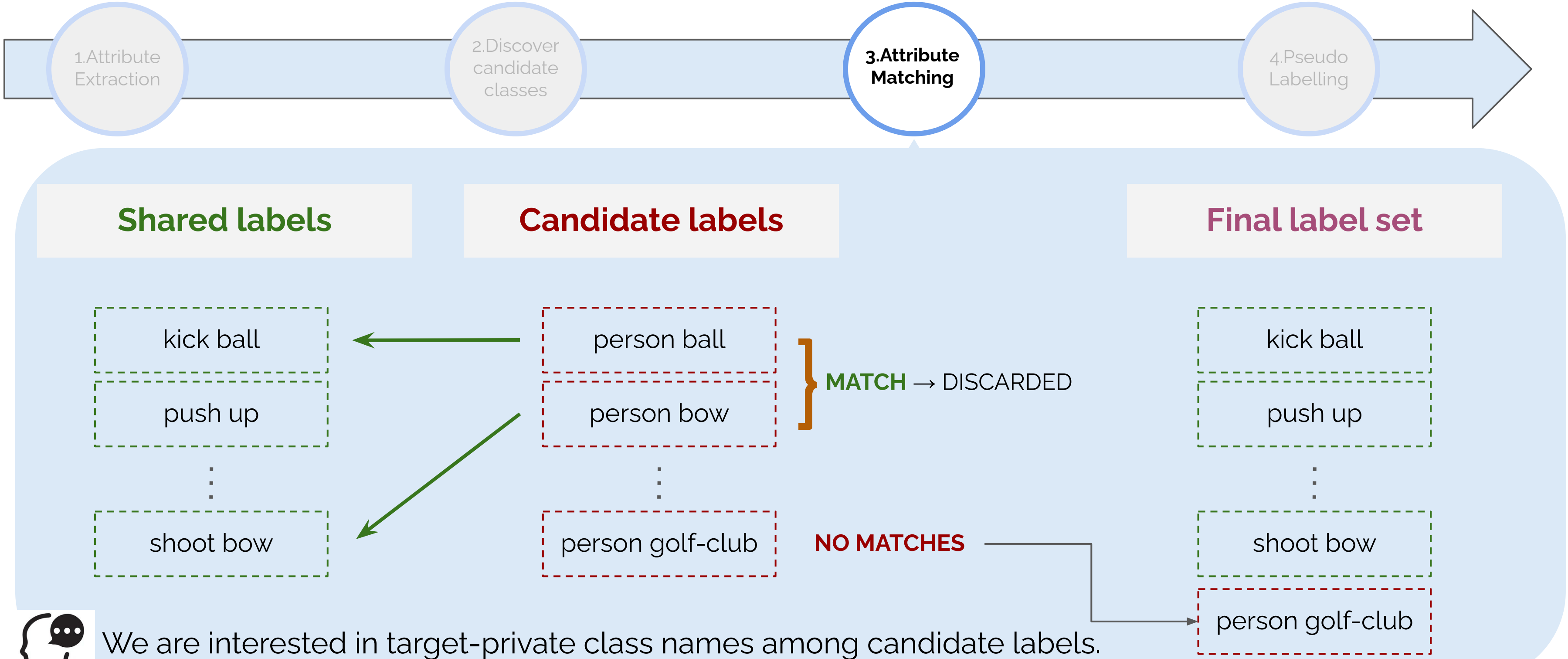


Kim et al., "Vilt: Vision and-language transformer without convolution or region supervision". In ICML, 2021.

AutoLabel

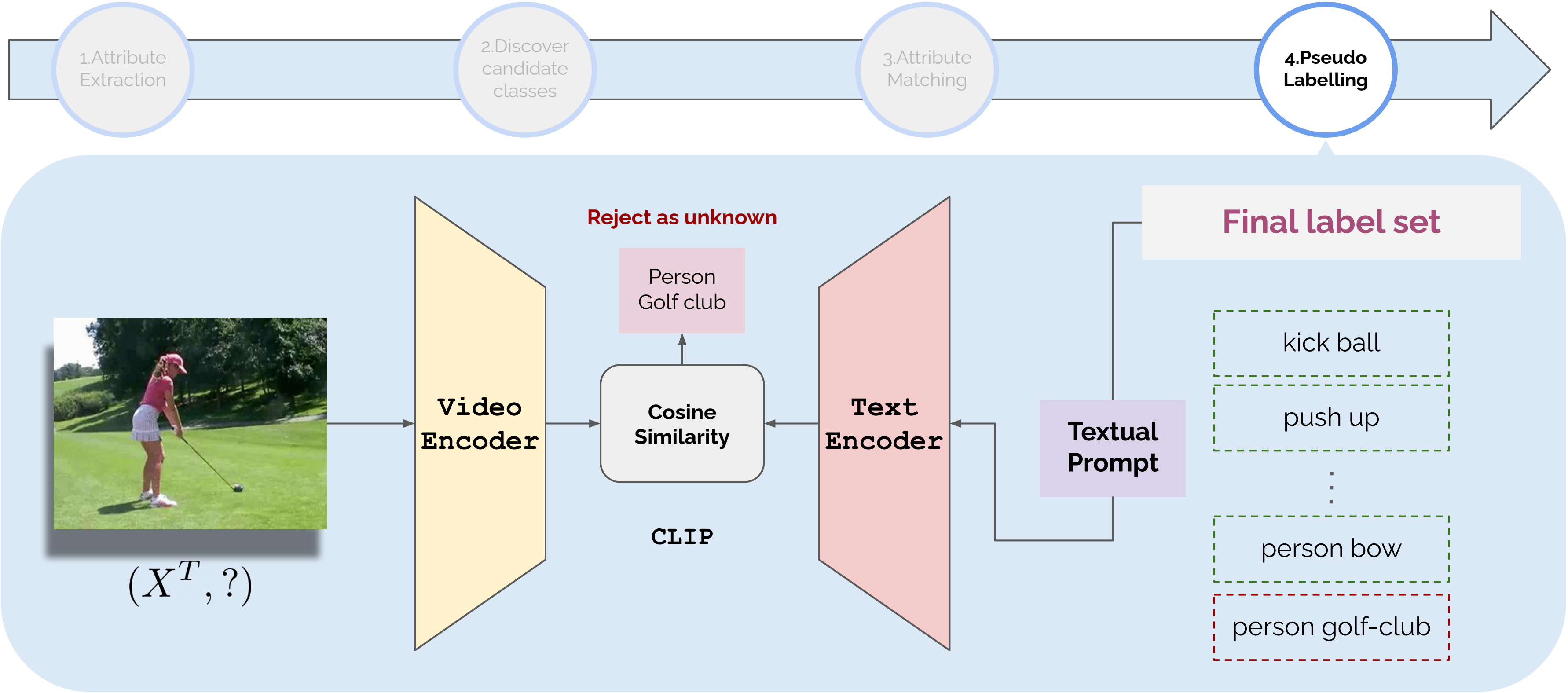

1.Attribute Extraction
2.Discover candidate classes
3.Attribute Matching
4.Pseudo Labelling
Video Encoder
k-means clustering
Top-*k* in
Person
Golf club
Grass
Hat
Occurrences count
TF-IDF
Candidate label
"Person Golf club"

AutoLabel



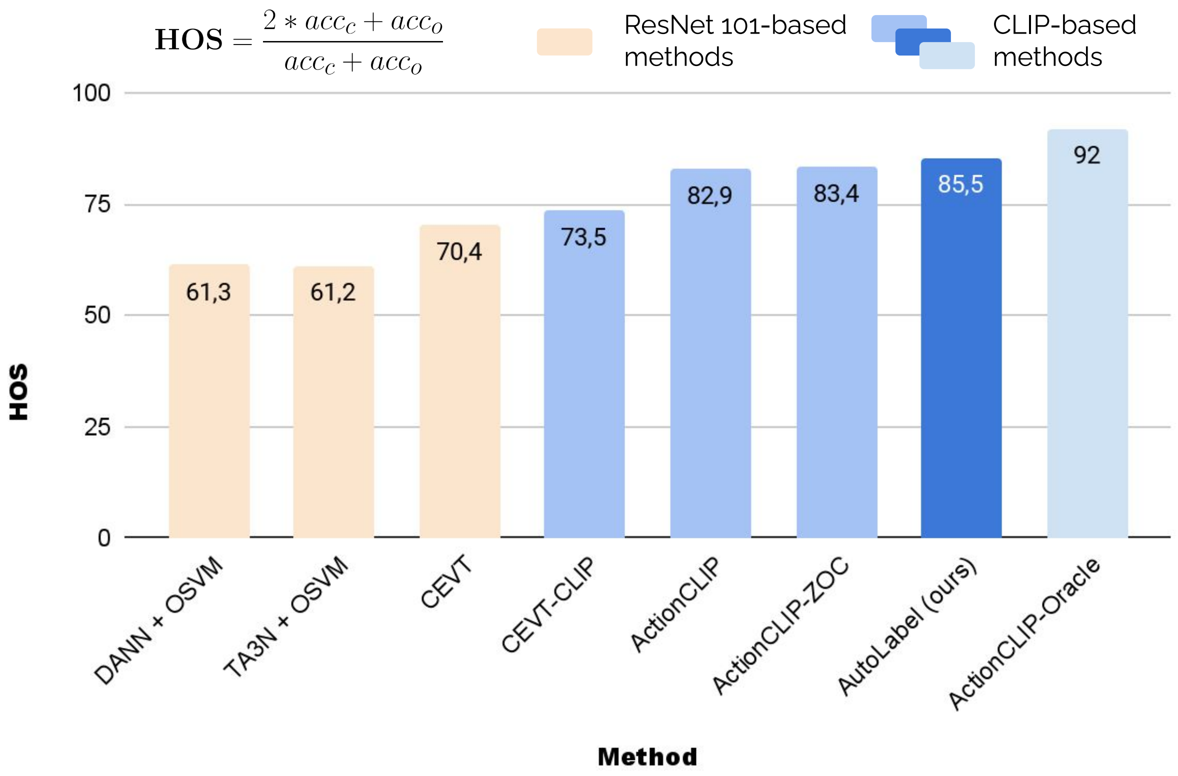
We are interested in target-private class names among candidate labels.

AutoLabel


1.Attribute Extraction
2.Discover candidate classes
3.Attribute Matching
4.Pseudo Labelling
Reject as unknown
Person Golf club
Video Encoder
Cosine Similarity
Text Encoder
CLIP
$(X^T, ?)$
Textual Prompt
Final label set
kick ball
push up
person bow
person golf-club

Experimental Results

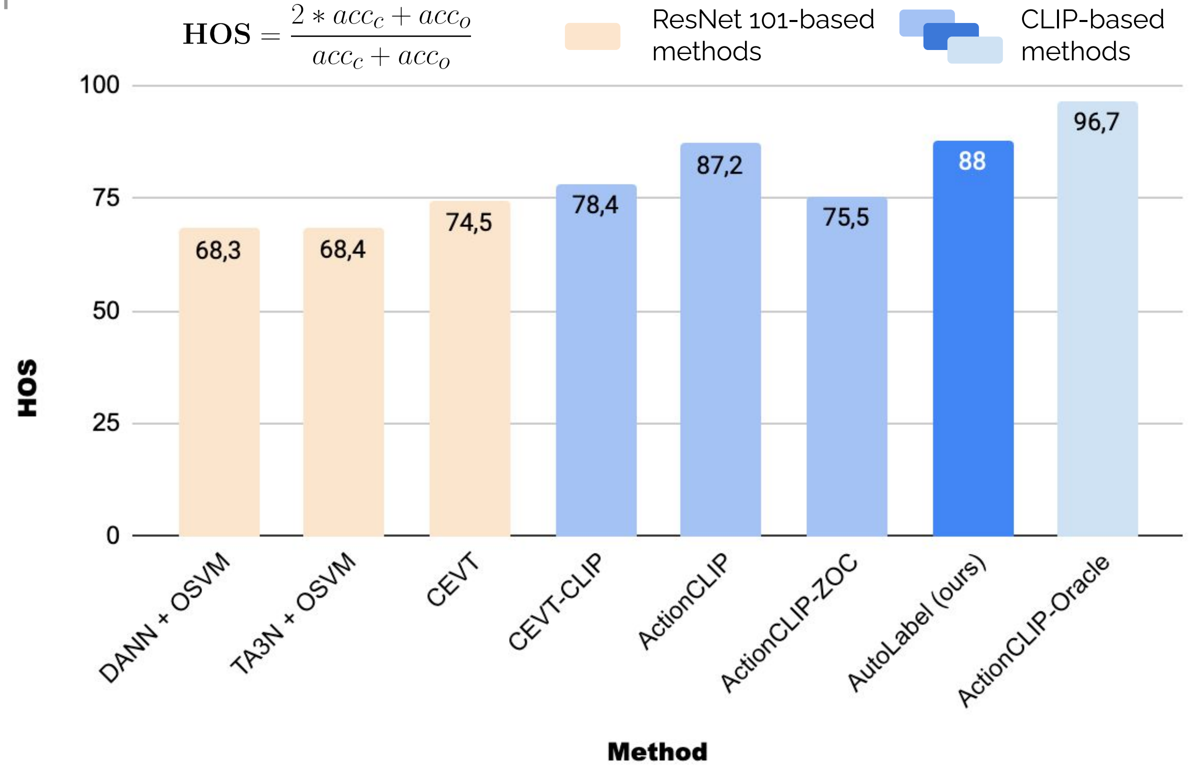
UCF→HMDB



Experimental Results

HMDB→UCF

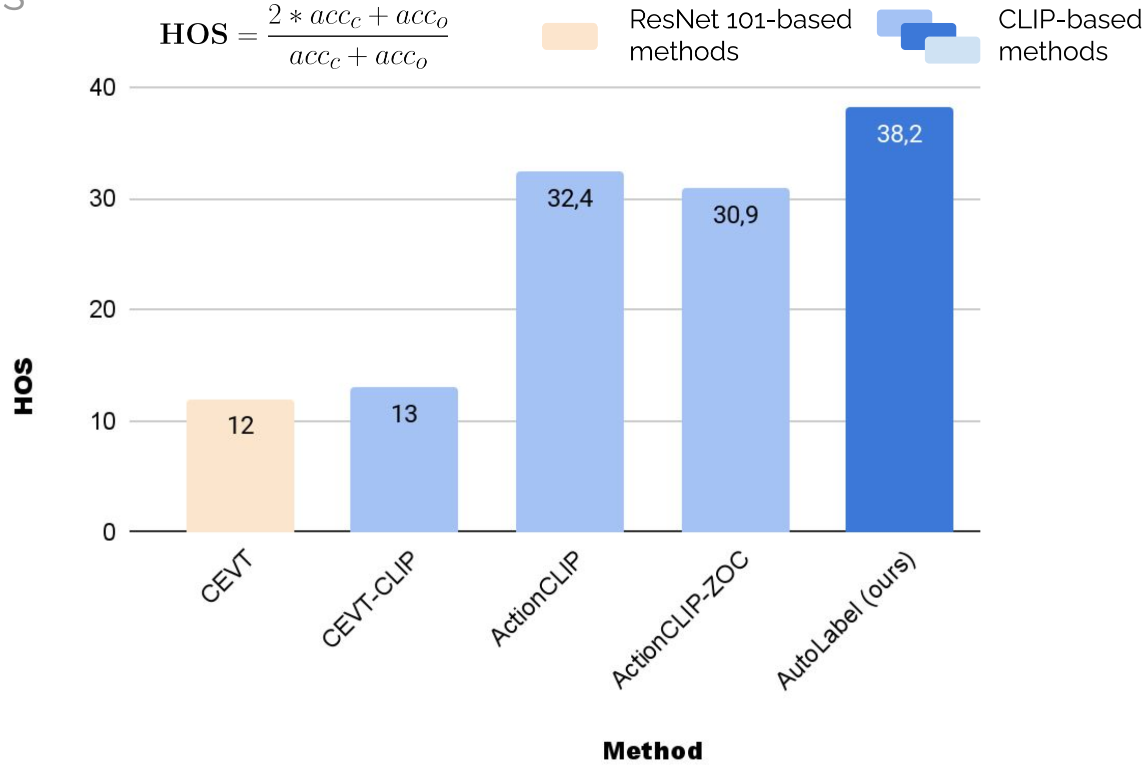
$$\textbf{HOS} = \frac{2 * acc_c + acc_o}{acc_c + acc_o}$$



Experimental Results

Epic-Kitchens

$$\mathbf{HOS} = \frac{2 * acc_c + acc_o}{acc_c + acc_o}$$



Take home messages

- **CLIP-based framework** can be devised for addressing open-set unsupervised video domain adaptation.
- AutoLabel enhances the **zero-shot** prediction capabilities of CLIP without knowing *a priori* the ‘unknown’ class names.
- We leverage a simple yet powerful idea that ***actions can be described by attributes***.
- Image captioning models were used to extract attributes, which are then processed by AutoLabel to zero in the ‘unknown’ class names.
- We obtain **state-of-the-art** results in open-set unsupervised video domain adaptation.

ICCV23
PARIS

The Unreasonable Effectiveness of Large Language-Vision Models for **Source-free** Video Domain Adaptation

Giacomo Zara, Alessandro Conti, Subhankar Roy,
Stéphane Lathuilière, Paolo Rota, Elisa Ricci

Moving one step further: source-free

Source data

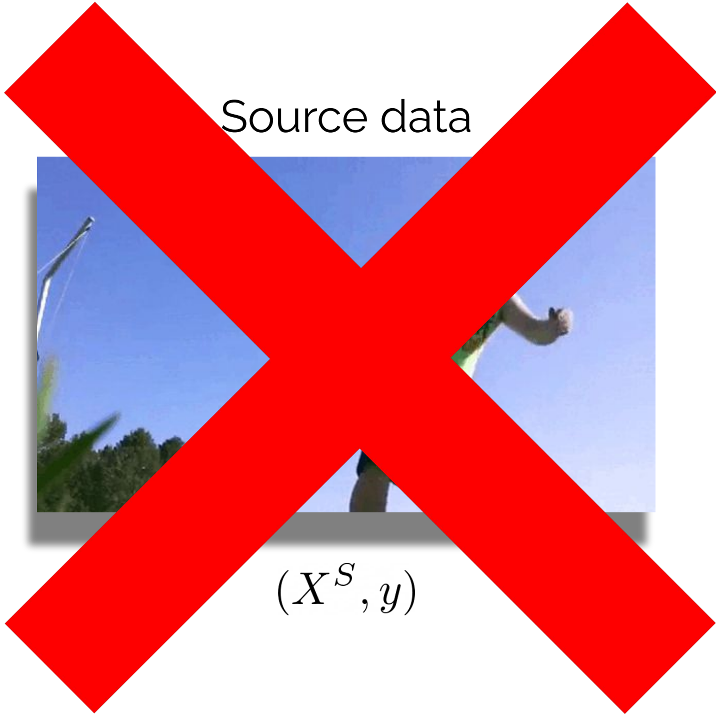
(X^S, y)

Target data

$(X^T, ?)$

Moving one step further: source-free

Source data



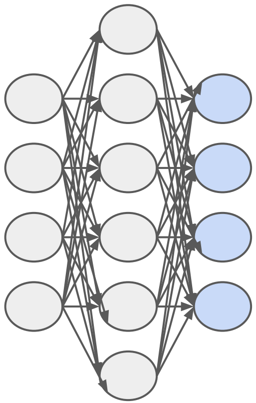
(X^S, y)

Target data

$(X^T, ?)$

Moving one step further: source-free video domain adaptation (SFVUDA)

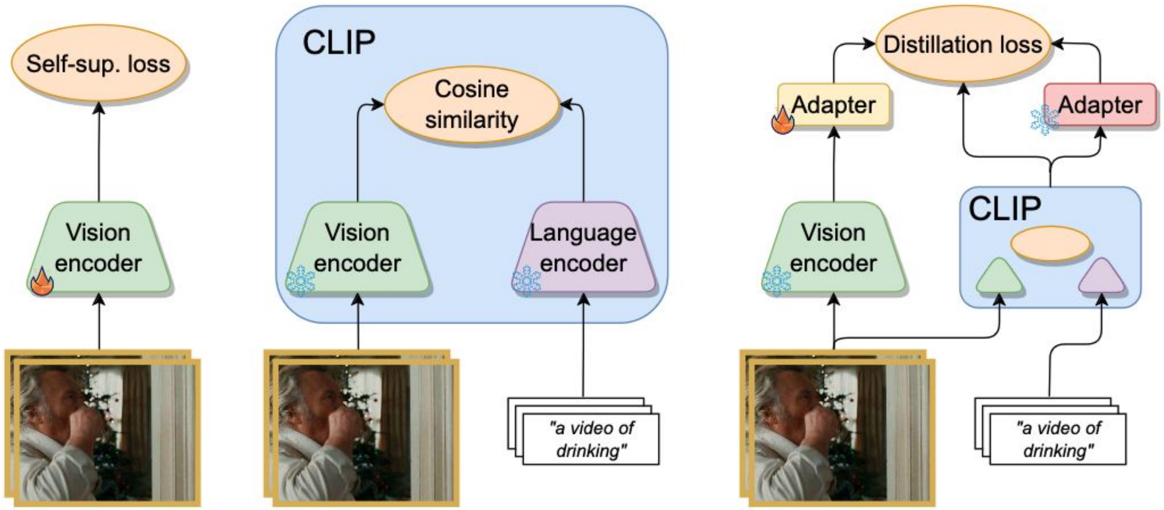
Source pretrained model



Target data

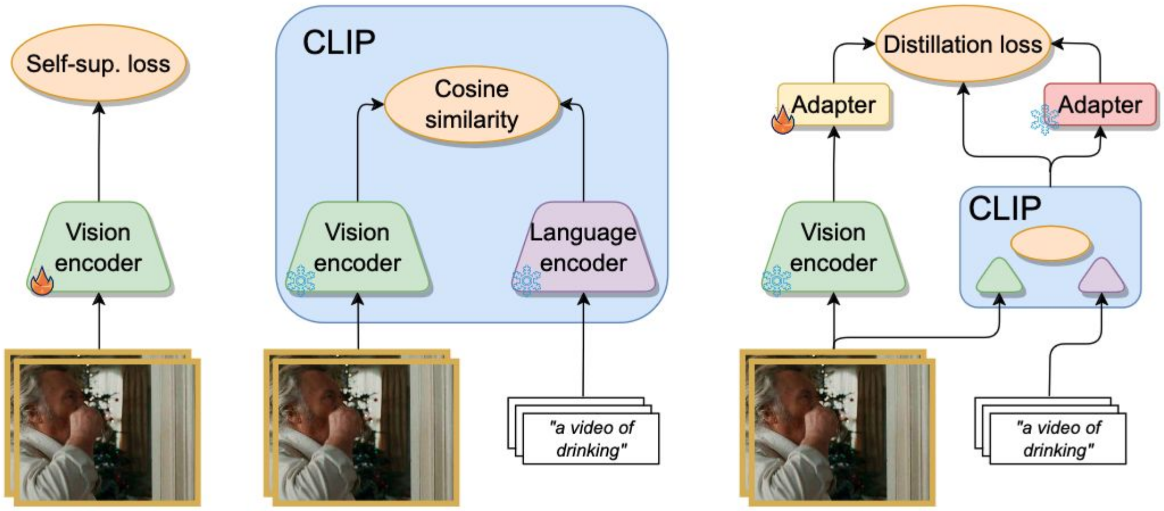
$(X^T, ?)$

General intuition



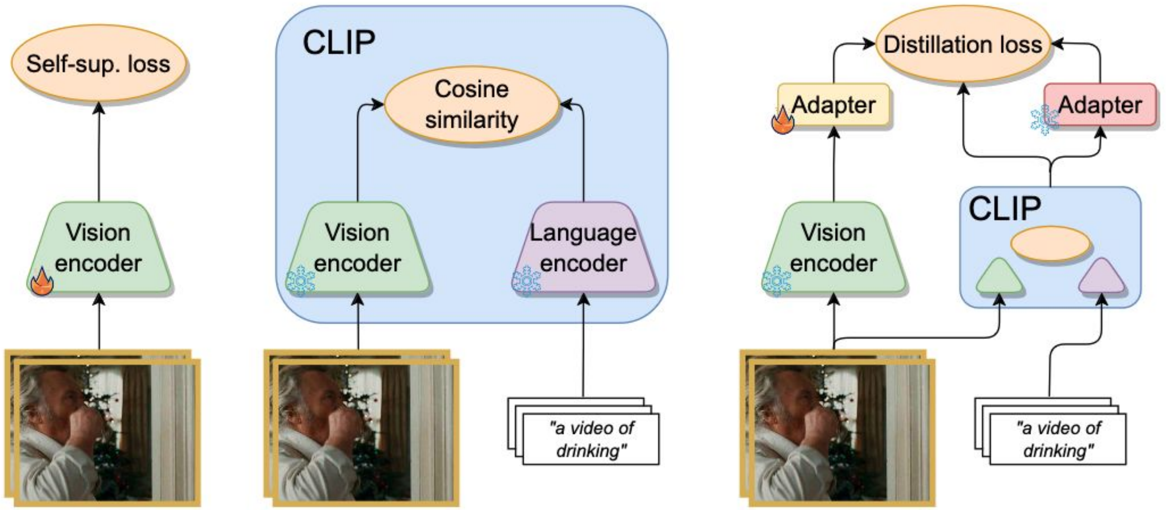
Traditional SFVUDA The source model is fine-tuned on the target domain by means of a self-supervised loss

General intuition



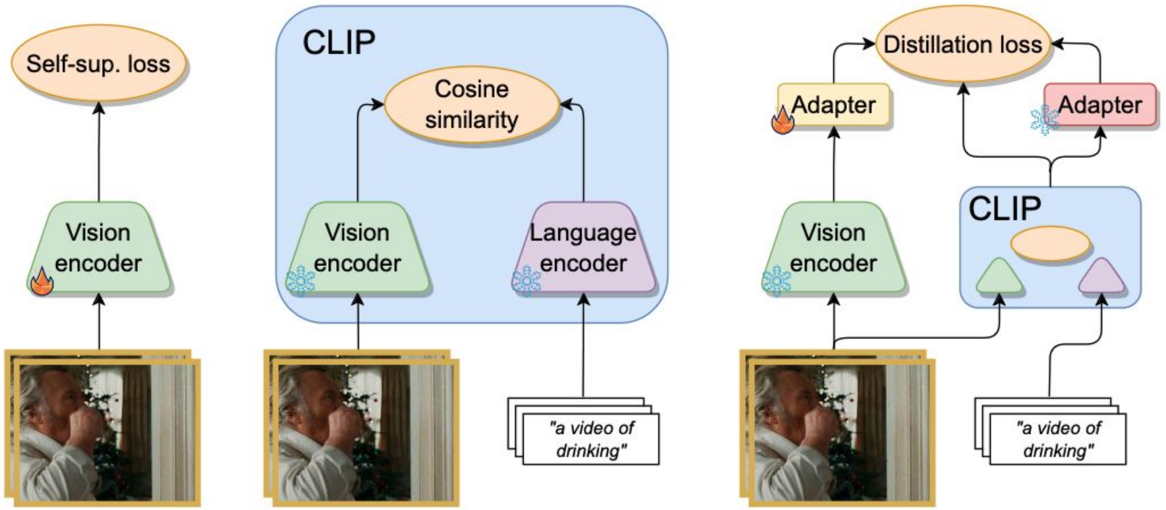
ISSUE Useful knowledge from the source domain may be overwritten by the tuning process on the target domain

General intuition



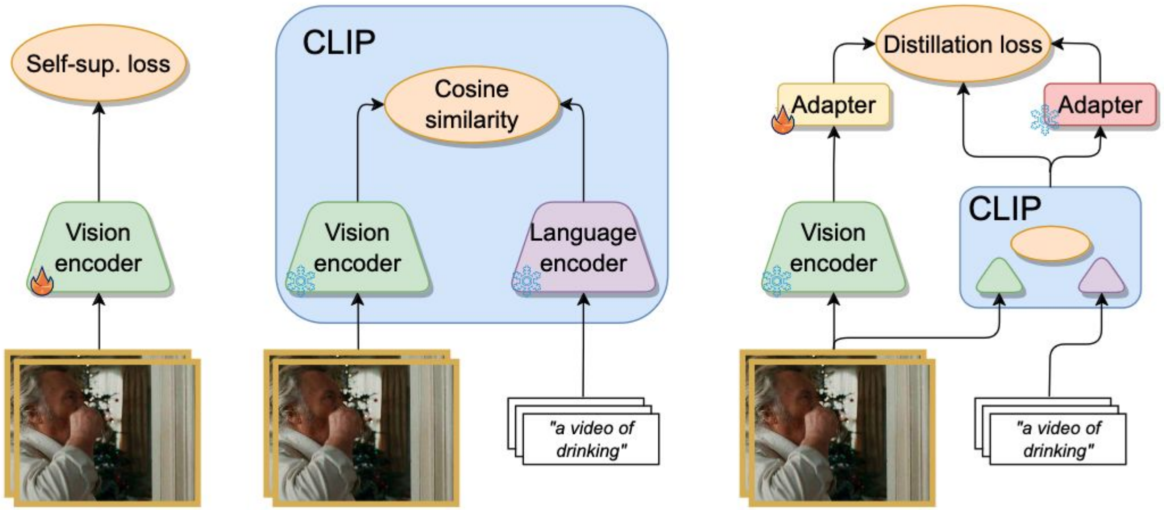
Zero-shot CLIP The CLIP model is used for inference off-the-shelf, without further tuning, leveraging it generalization capabilities

General intuition



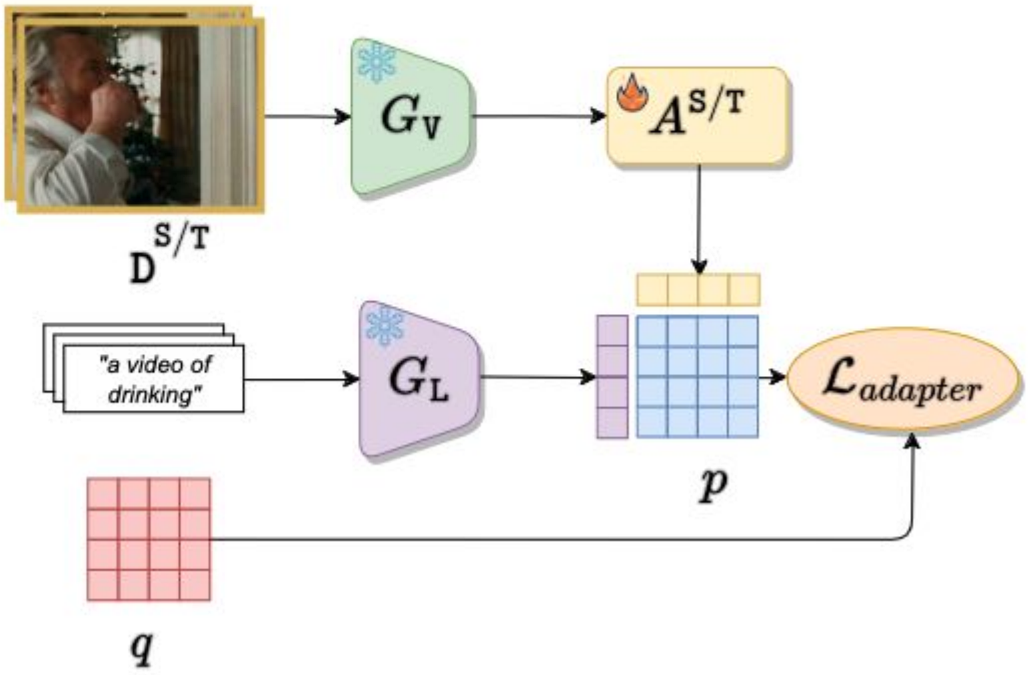
ISSUE This approach does not leverage useful domain-specific knowledge from either domain, relying solely on CLIP generalization

General intuition



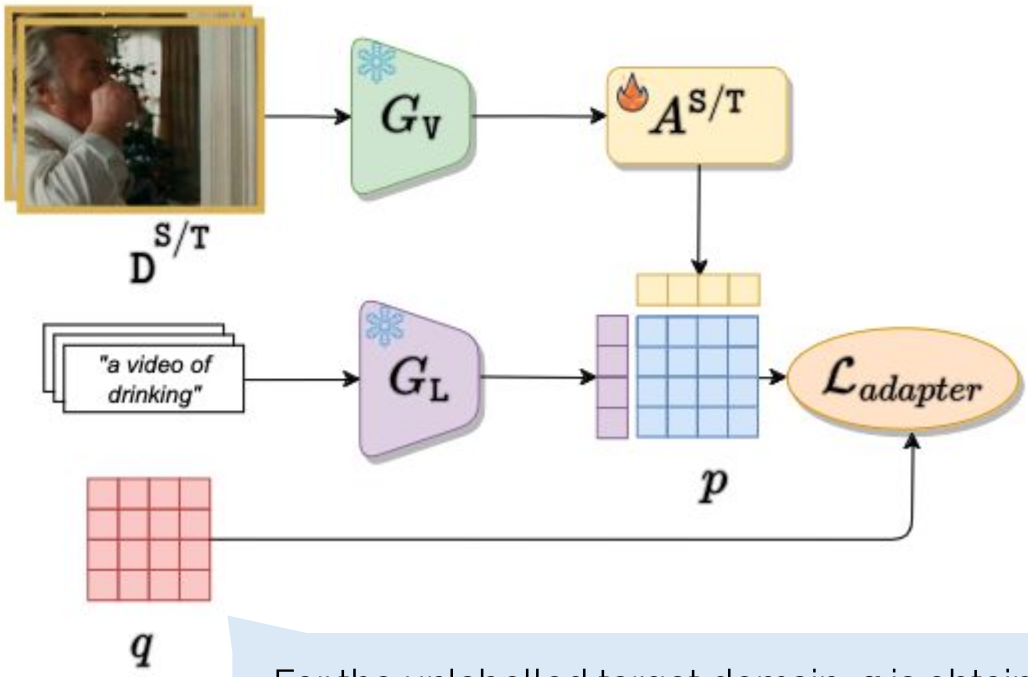
Our Solution Leveraging the complementarity of general CLIP knowledge and domain specific information through a distillation process, by only learning an adapter

Adapter training



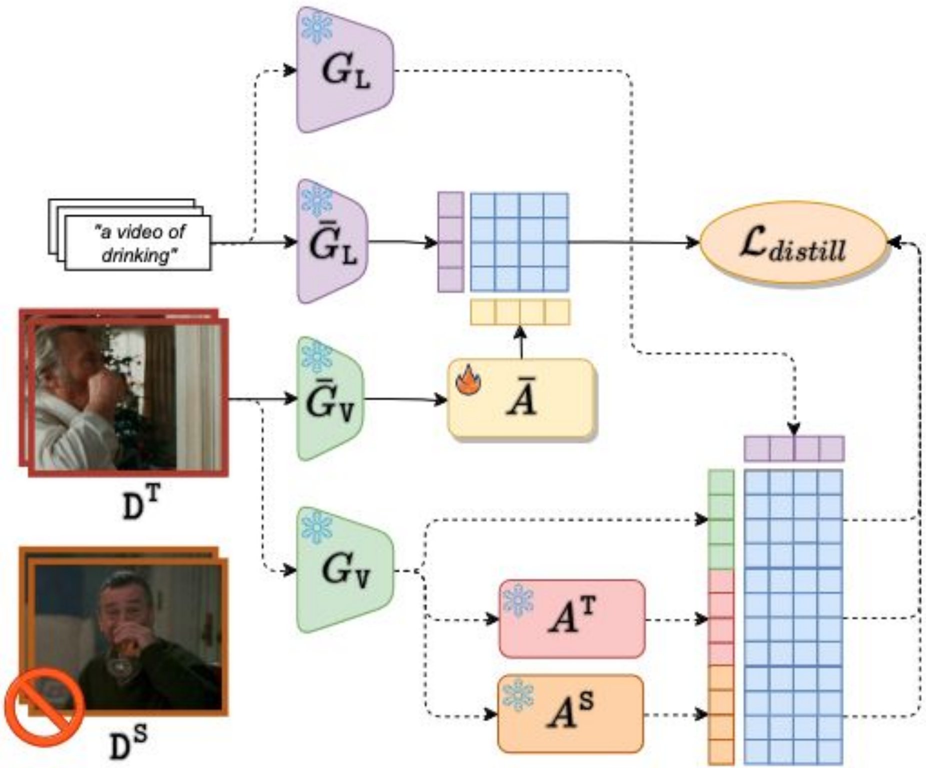
The training process of the adapters is governed by a standard Language&Vision loss

Adapter training



For the unlabelled target domain, $\boldsymbol{q}$ is obtained by pseudo-labelling with CLIP (ViT/B32)

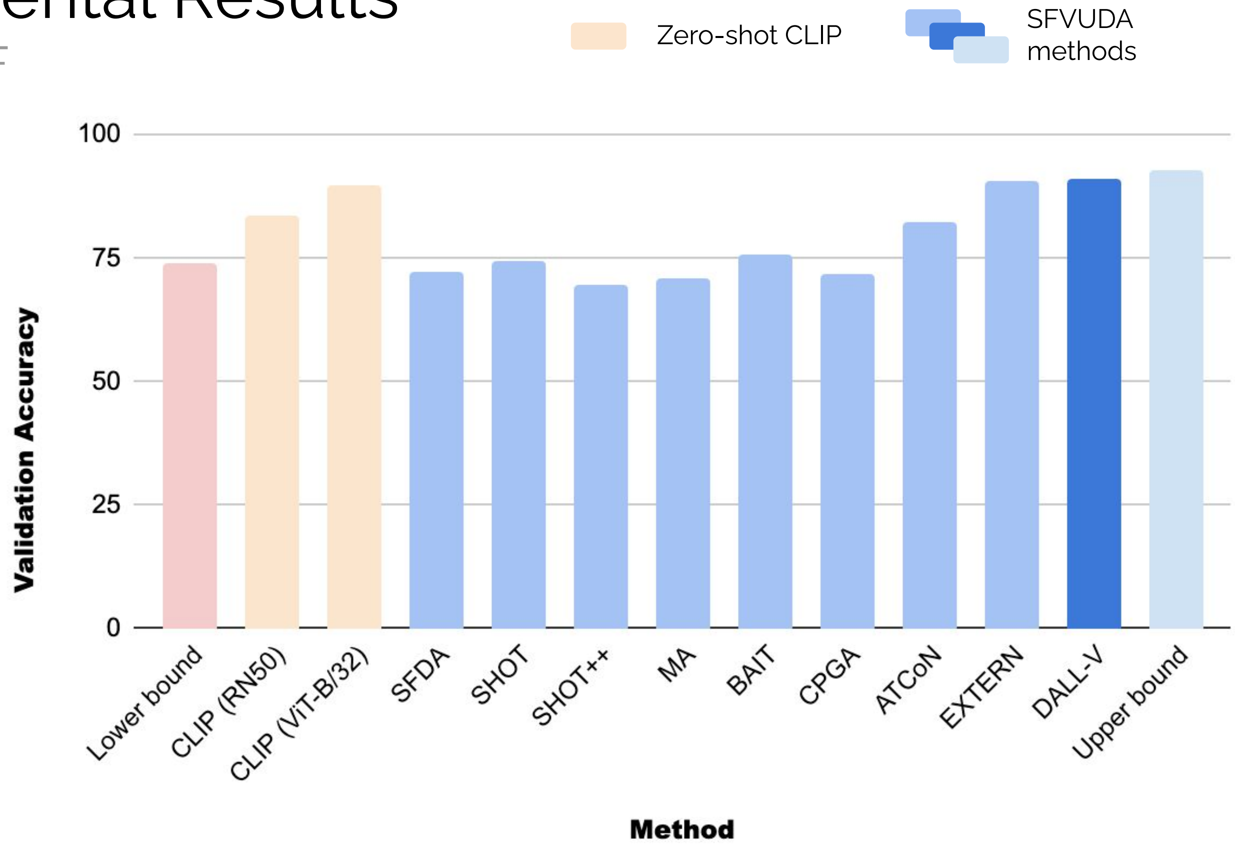
Ensemble distillation



We use the learned adapters and CLIP (ViT/B32) as **teachers**, and train a **student** adapter on top of a CLIP (RN50) encoder

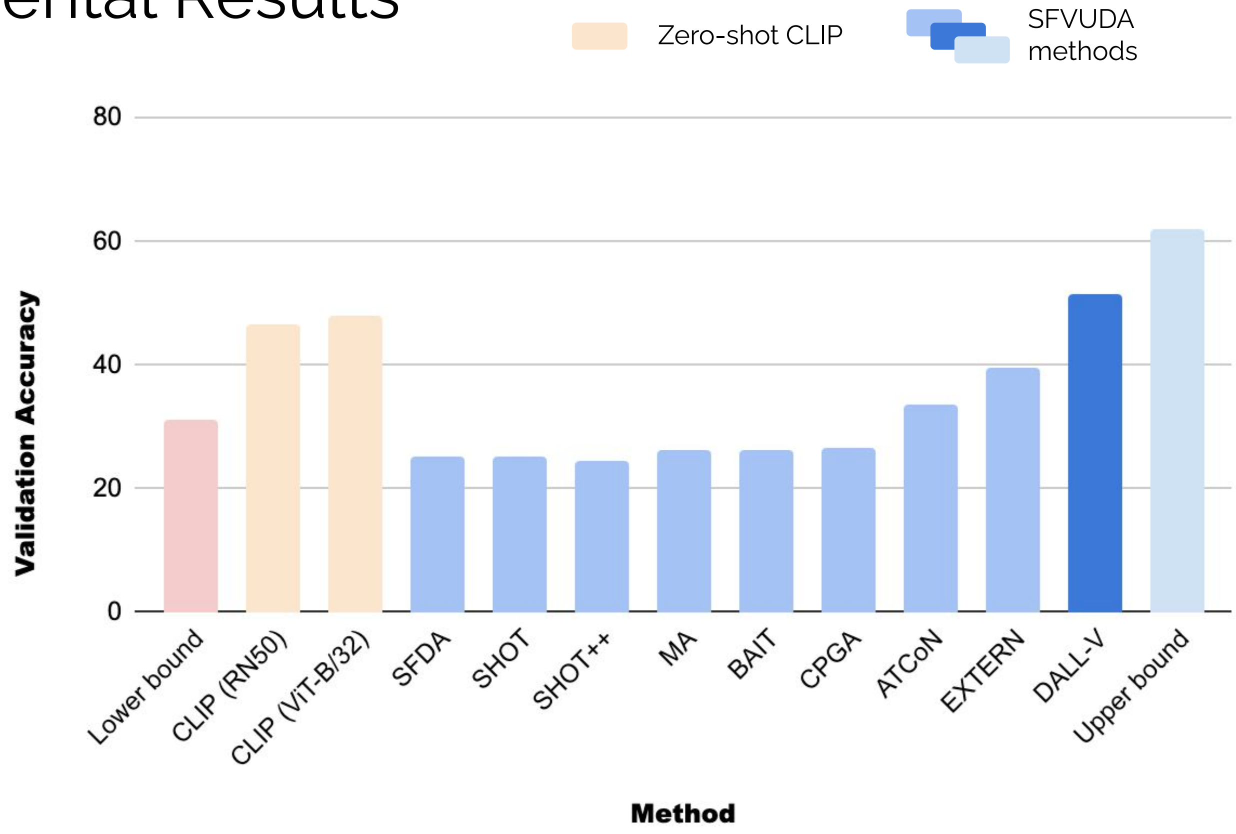
Experimental Results

HMDB-UCF



Experimental Results

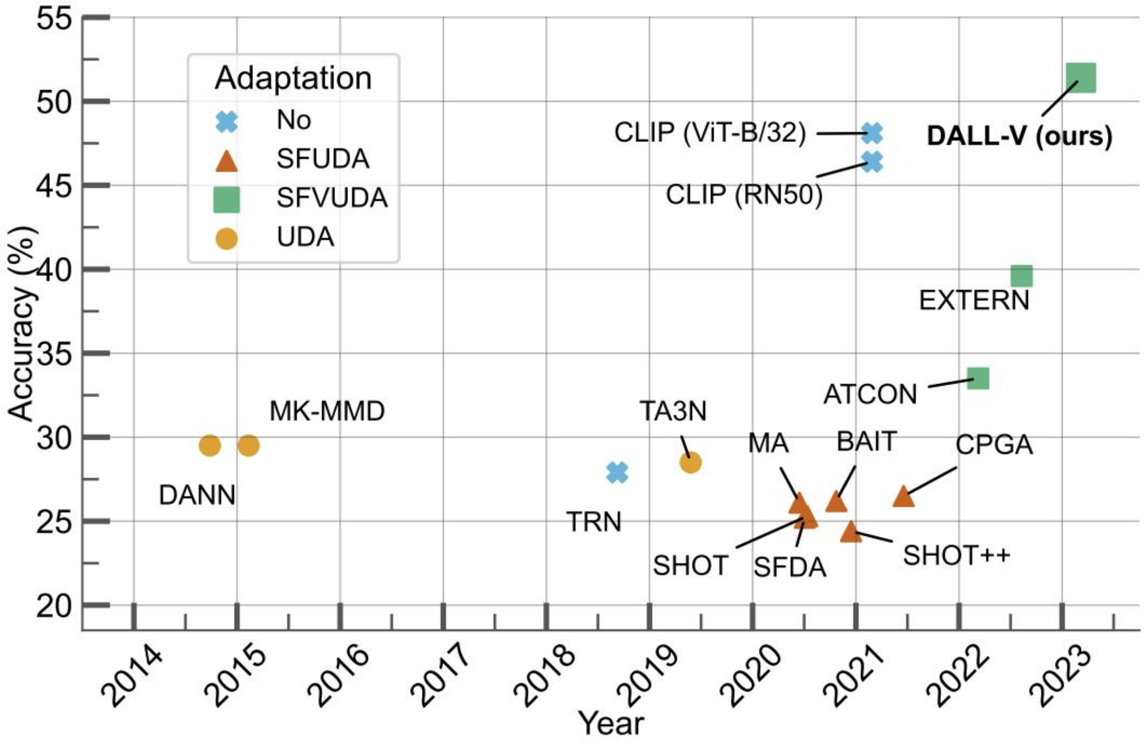
Daily-DA



Take home messages

- Simple but novel approach for SFVUDA
- Combination of complementary information derived from domain-specific models and the powerful CLIP-based LLVMs
- Extensive evaluation on two standard benchmarks for VUDA repurposed for the source-free scenario
- Comparison with existing methods and a selection of CLIP-based baseline, showing state-of-the-art results

Did we close the gap?
(thanks to Large Language & Vision Models)



Upper bound (target model) is 61%

Future research directions

- Can we describe the test domain with language?
- How can we further exploit language (e.g. other captioning models)?
- Domain Generalization
- Mandatory to reduce computational burden

Thank you for the attention!

UNIVERSITÀ DI TRENTO

FONDAZIONE BRUNO KESSLER

People

G. Zara

A.Conti

V. G.T. Da Costa

T. O. Dos Santos

P. Rota

S. Roy

S. Lathuillere

N. Sebe

V.Murino