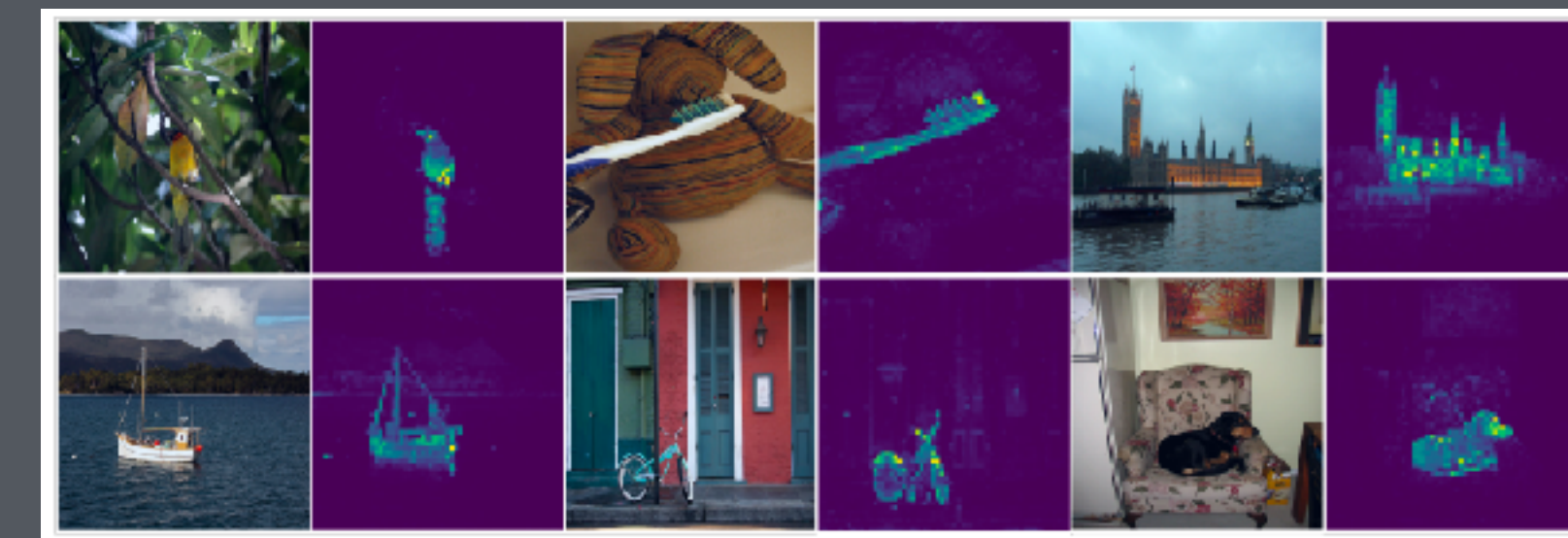
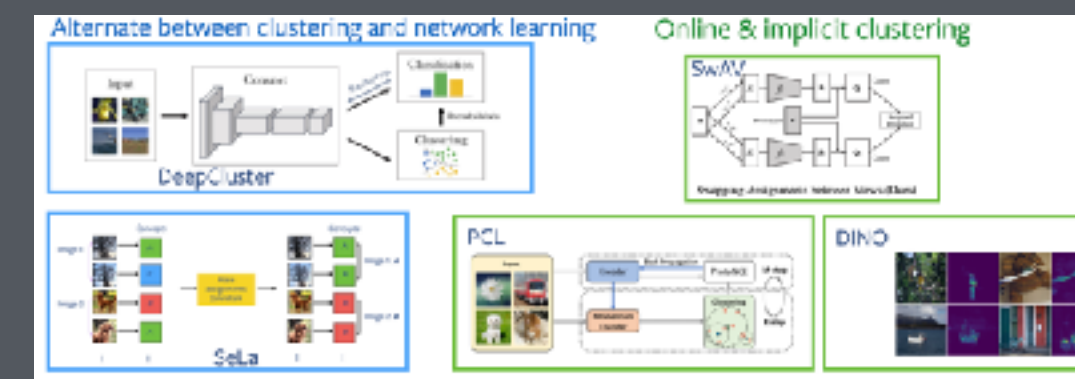
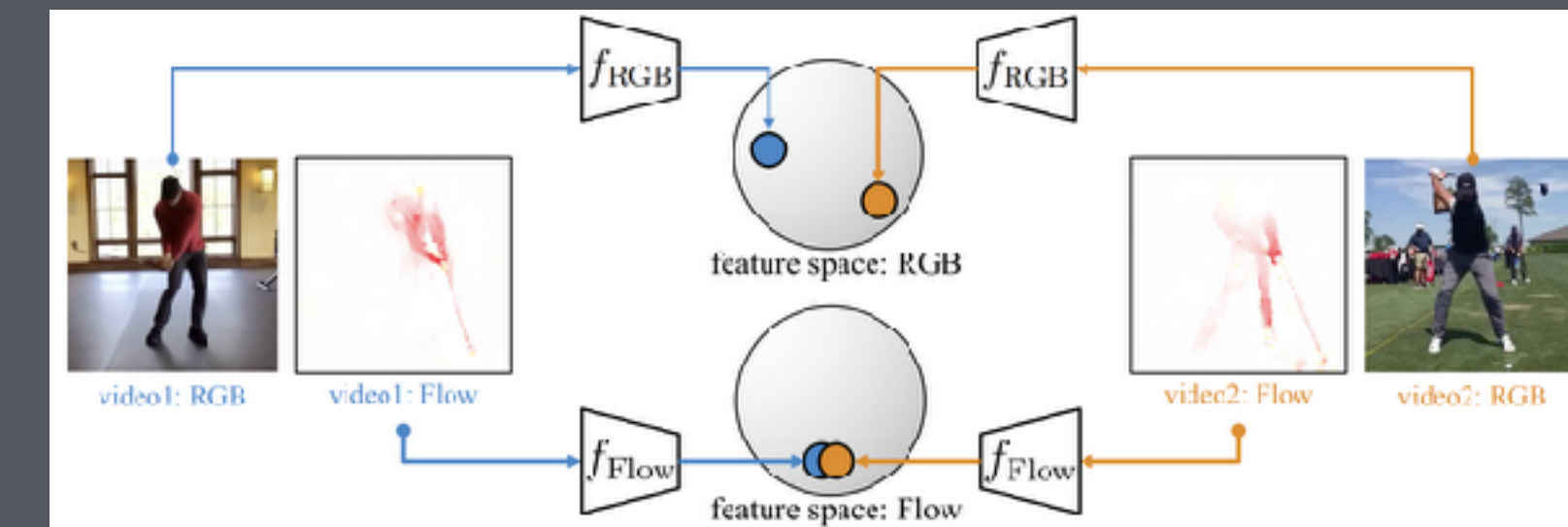
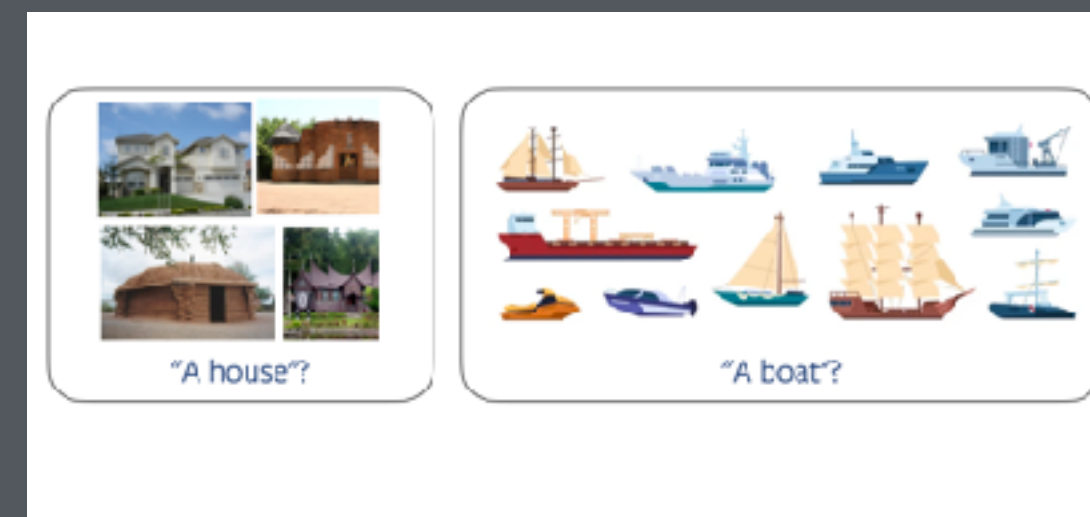


Self-supervised learning for computer vision from images, video and audio. Part 2: Multi-modal learning



@ DEEP LEARNING 1

YUKI M. ASANO

LECTURE 14

Organisation

Practicals this afternoon: merged for better synergies.
Please find below the time slots with the rooms:

11-13: D1.115

13-15: D1.114

15-17: G0.18A

Thanks for your feedback for the last (BKO) lecture. I'm very thankful for the kind words and improvement suggestions!

There's just one last feedback to be filled out after the exam (either there or online) 🙏 .

Organisation

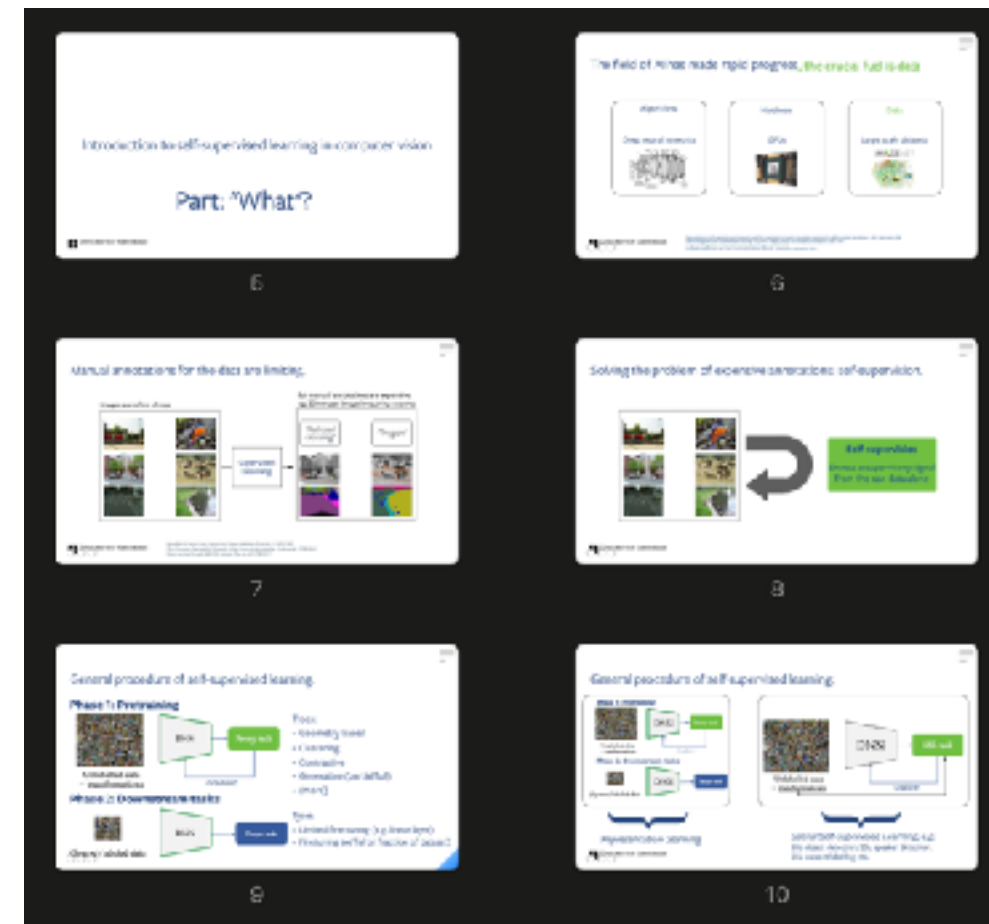
A cheat sheet with formulas will be uploaded to Canvas today.

It will contain equations and formulas such that you do not need to memorize them.

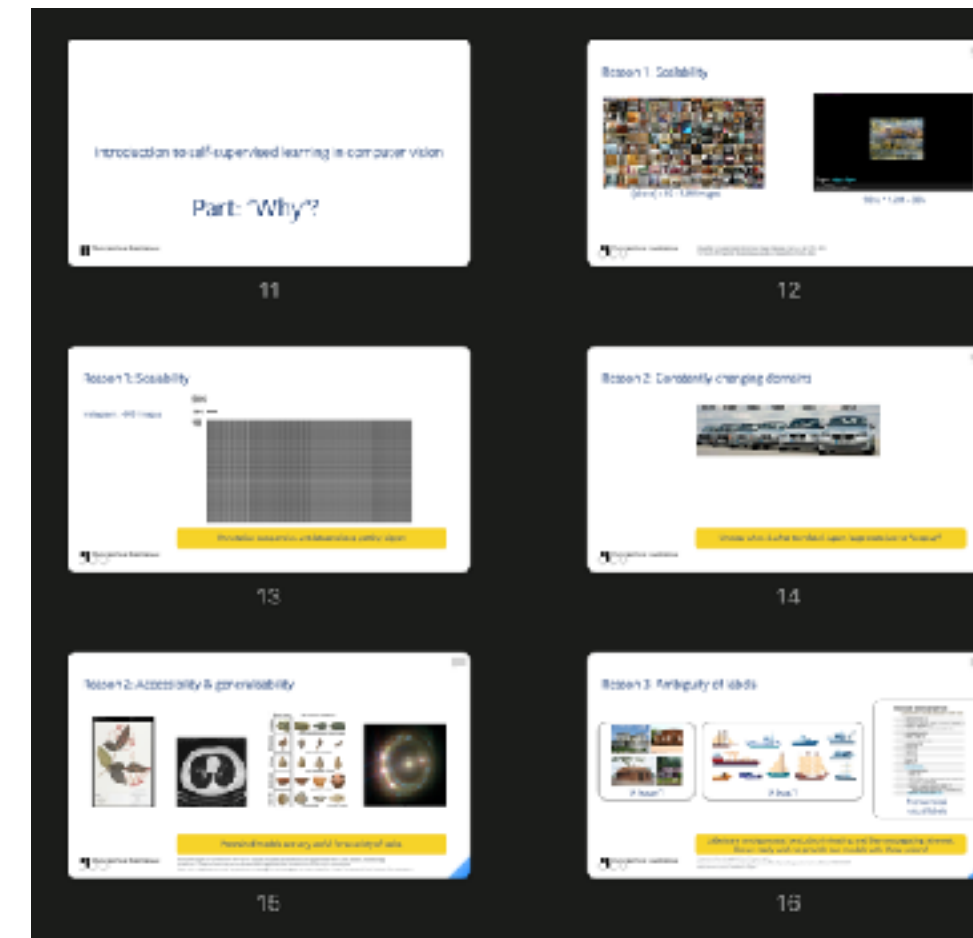
Not all of those will be relevant for the exam, of course.

The goal of the exam is to test your understanding, your ability to find solutions to deep learning questions in a quantitative and qualitative manner, to transfer your learnings to slightly different settings and to apply your critical evaluation skills.

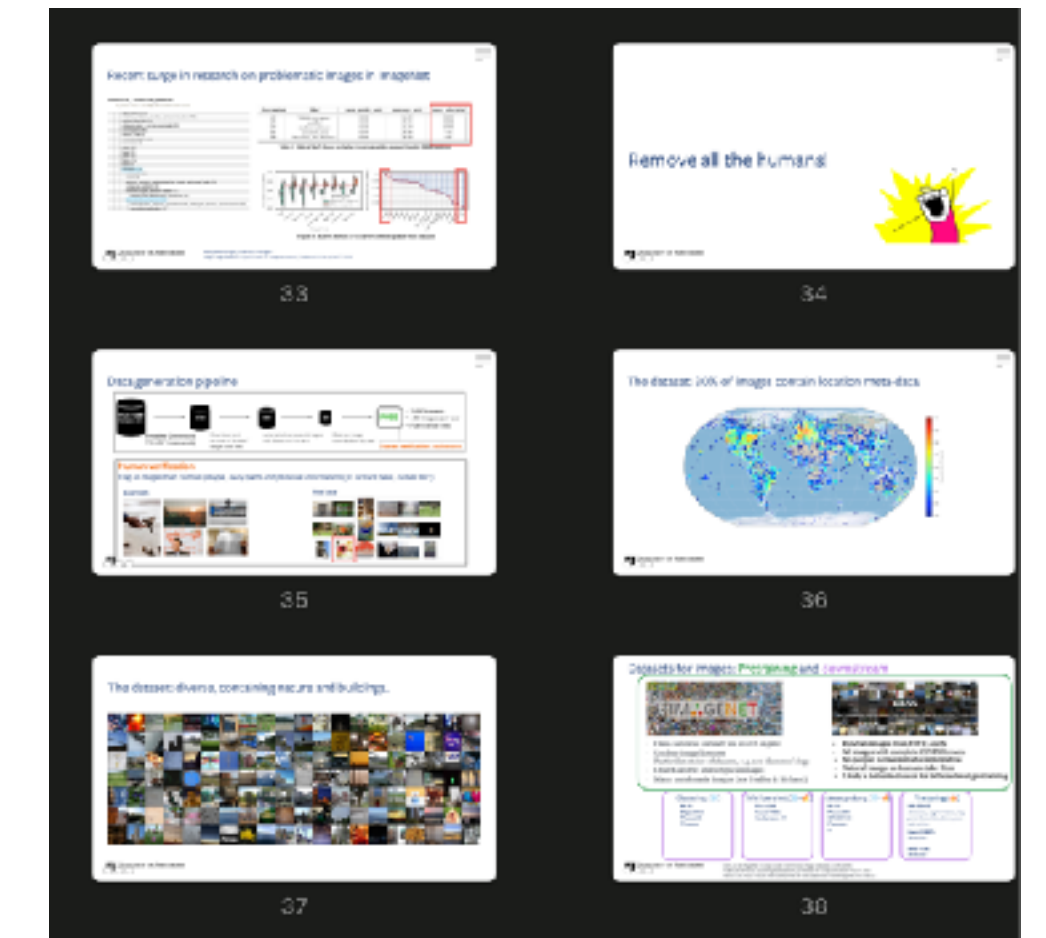
Summary last time



Why SSL?



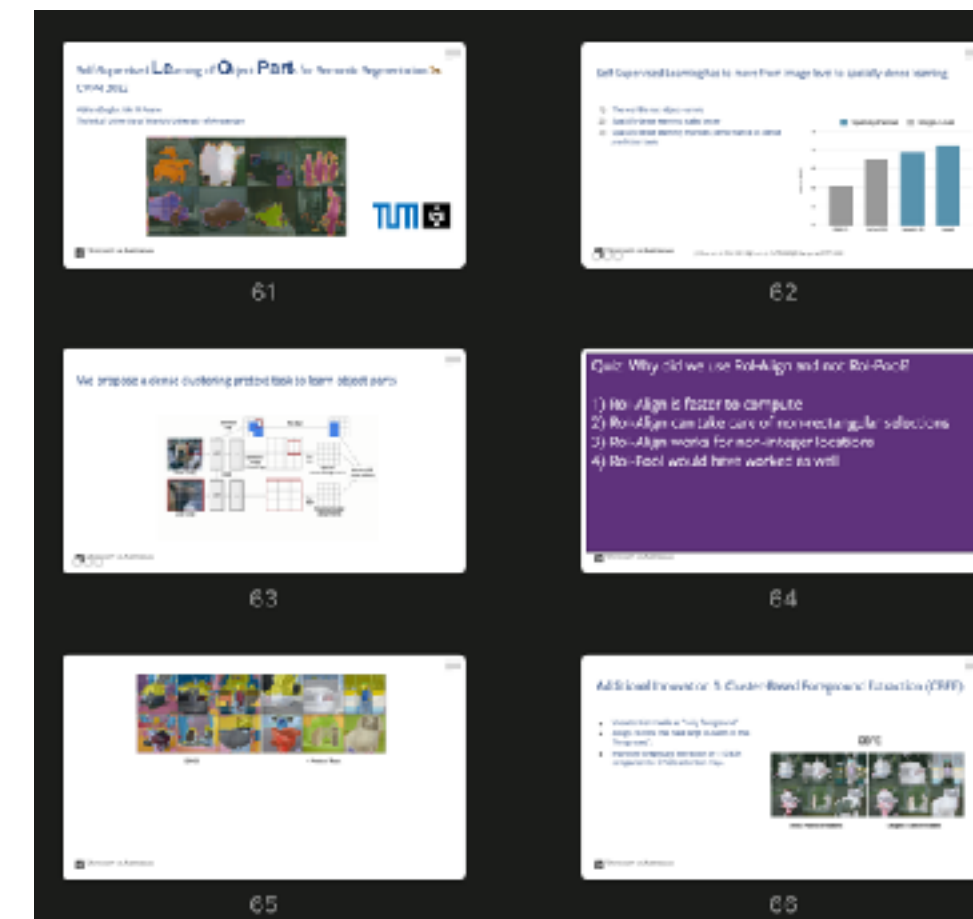
What is SSL?



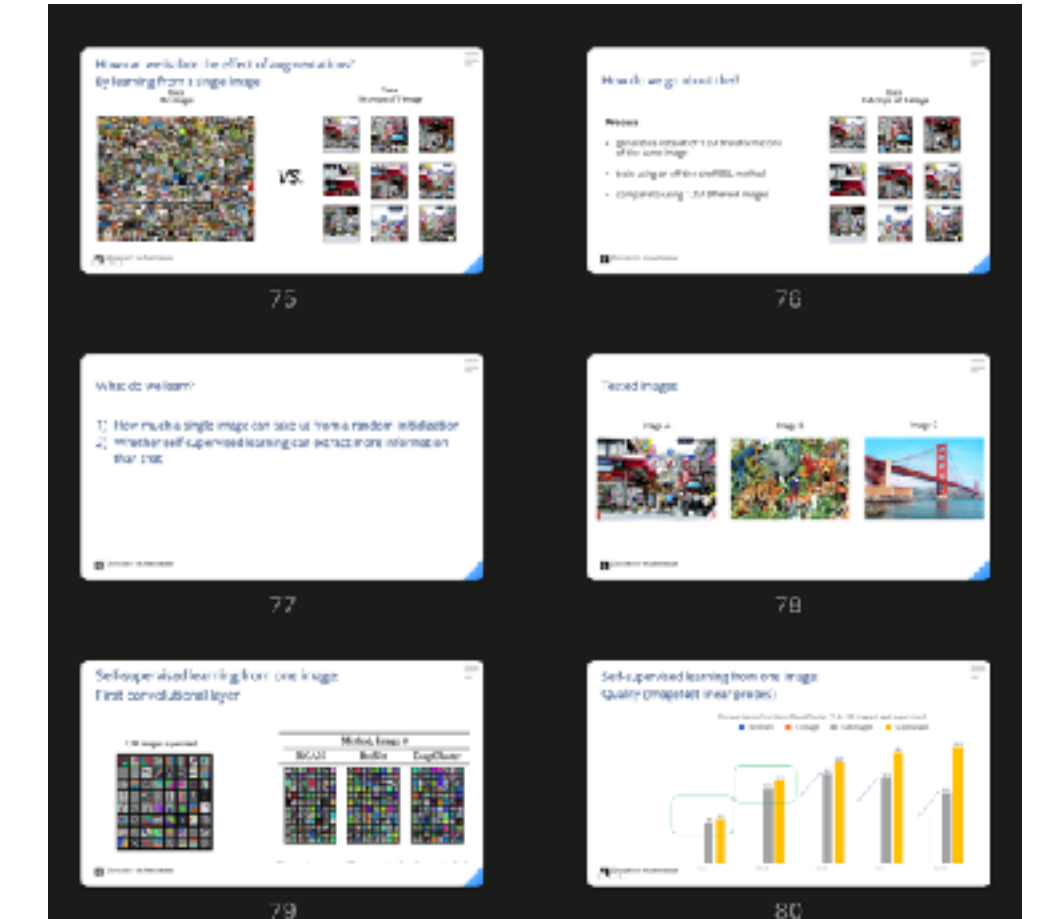
What kind of data? [1]



How SSL? (e.g. clustering [2])



SSL for segmentation [3]



Role of augmentations [4]

- [1] PASS: Pictures without humAns for Self-Supervised Pretraining. Asano et al. NeurIPS-Data'21.
- [2] Self-labelling via simultaneous clustering and representation learning. Asano et al. ICLR 2020.
- [3] Leopart: Self-Supervised Learning of Object Parts for Semantic Segmentation. Ziegler and Asano. CVPR 2022
- [4] A critical analysis of self-supervision, or what we can learn from a single image. Asano et al. ICLR 2020.

Today: Multi-modal learning

What is multi-modal data?

Why is it useful?

How is it done?

- “From” SimCLR to CLIP and GDT
- Audio-visual clustering of video-datasets
- Self-supervised object detection and classification

Present & Future

- Multi-modal Versatile Networks
- VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text
- Socratic Models
- Flamingo

What is a modality

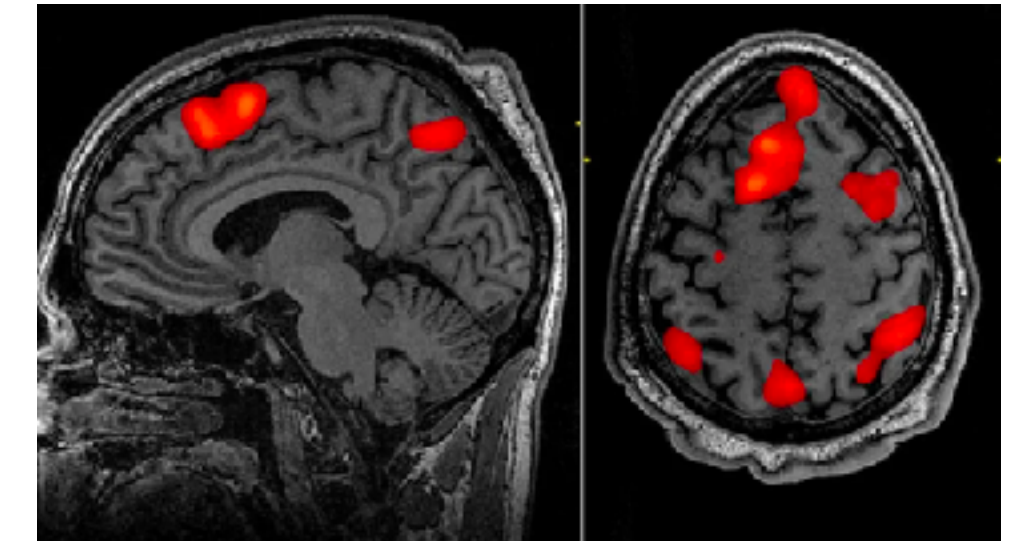
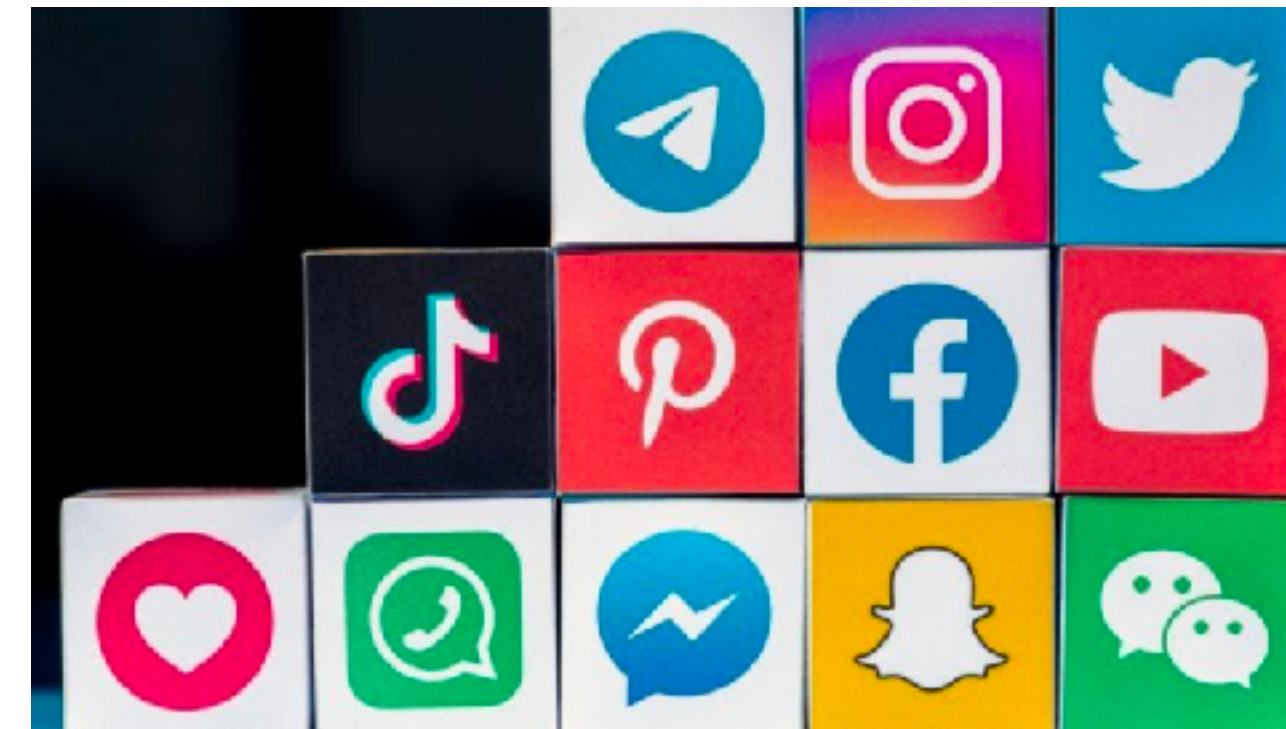
Modality:

The way in which something happens or is experienced.

- Representation format in which information is stored.
- Sensory modality: vision or touch; channel of communication.

Examples of Modalities:

- Natural language (both spoken or written)
- Visual (from images or videos)
- Auditory (including voice, sounds and music)
- Haptics / touch
- Smell, taste and self-motion
- ... Electrocardiogram (ECG), skin conductance
- ... Infrared images, depth images, fMRI



What is multi-modal learning

In general, learning that involves multiple modalities

This can manifest itself in different ways:

- Input is one modality, output is another
- Multiple modalities are learned jointly
- One modality assists in the learning of another

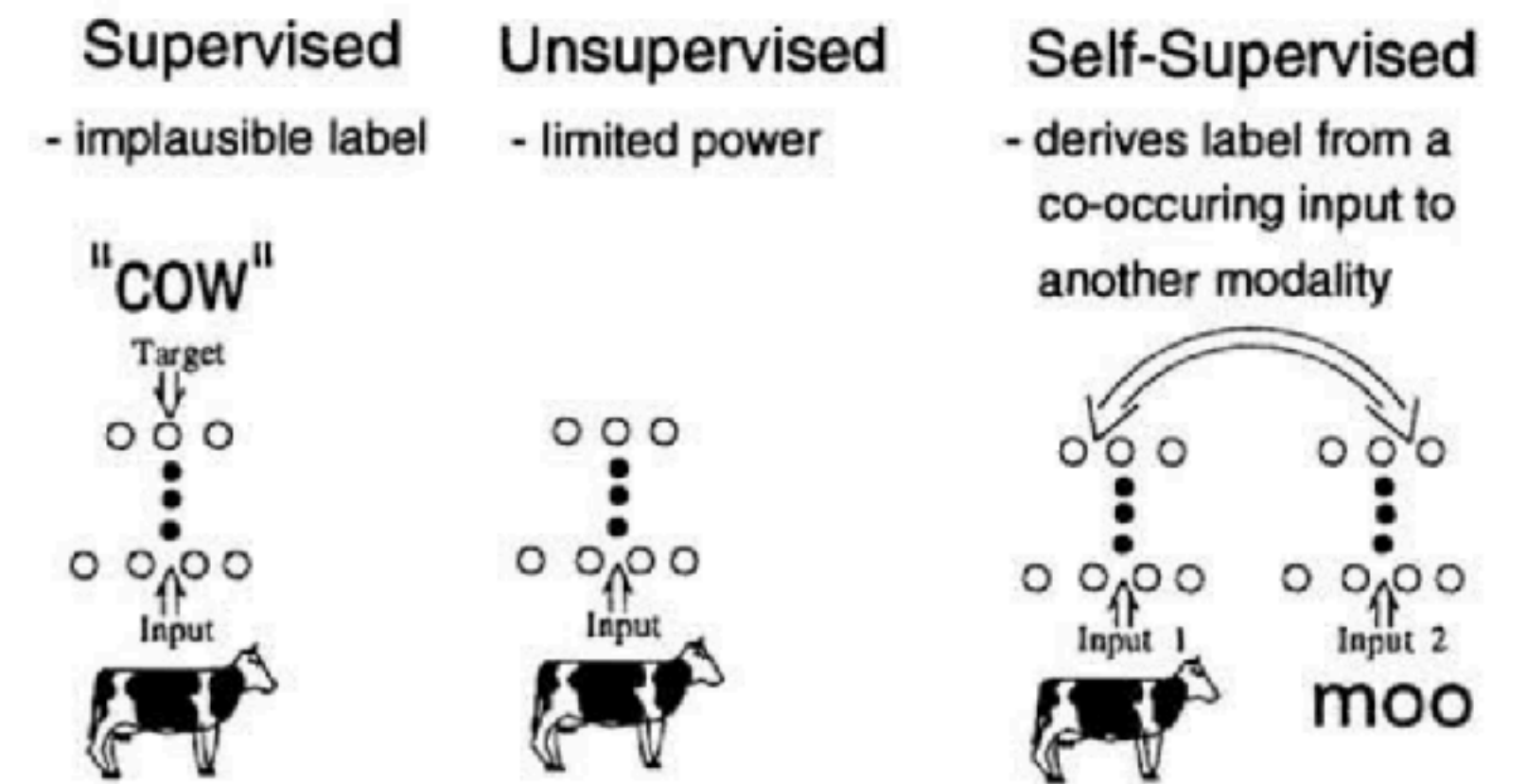


Figure 2: The idea behind the algorithm

Why multi-modal learning?

- Obviously contains more data, so it should clearly help (?)
- Noisy and missing data of a single modality
- Meaning often captured not by a single modality
- But in practice it's not so easy
 - The representation spaces vary widely
continuous (eg sound) vs ordinal (eg rankings) or discrete (eg text)

Meaning often captured not by a single modality: McGurk effect



Speech perception is not a purely auditory process.

Quiz: Multi-modal learning is a great avenue for scalable deep learning. When designing a future robot/self-driving car/generic intelligence, why would one perhaps not opt for including more and more modalities?

Turn to your neighbour and come up with as many reasons as you can imagine for why using *less* modalities might be sensible

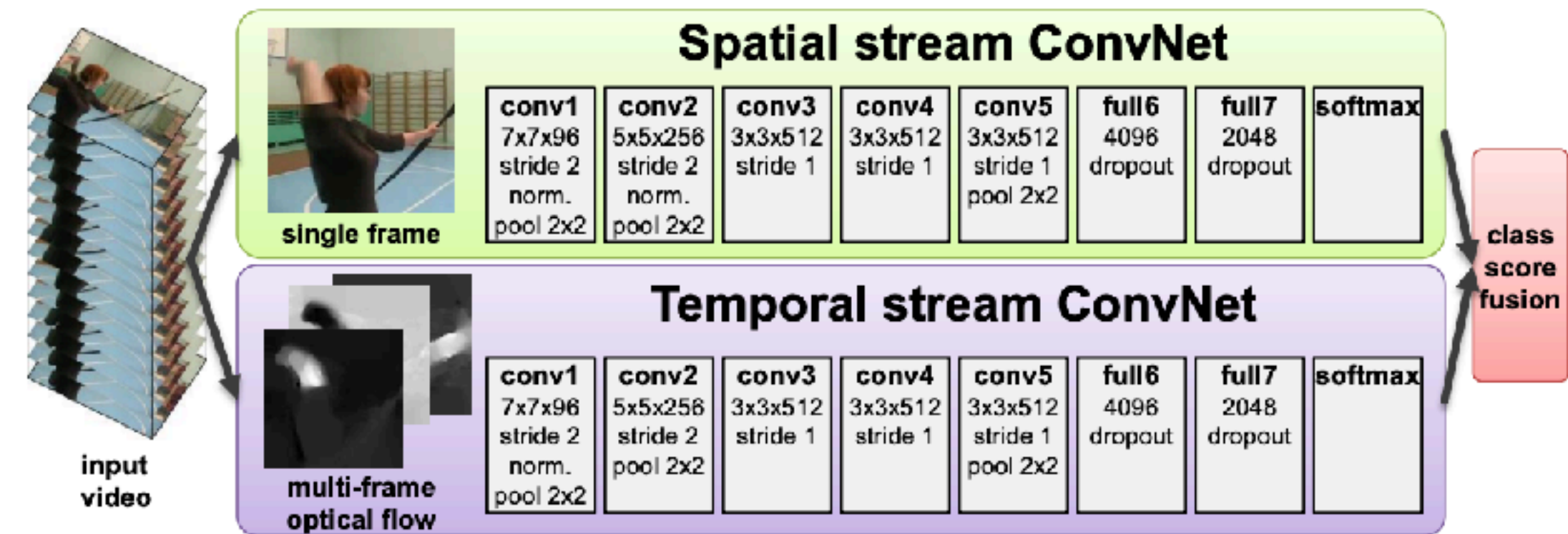
Multi-modal learning approaches

Representation learning

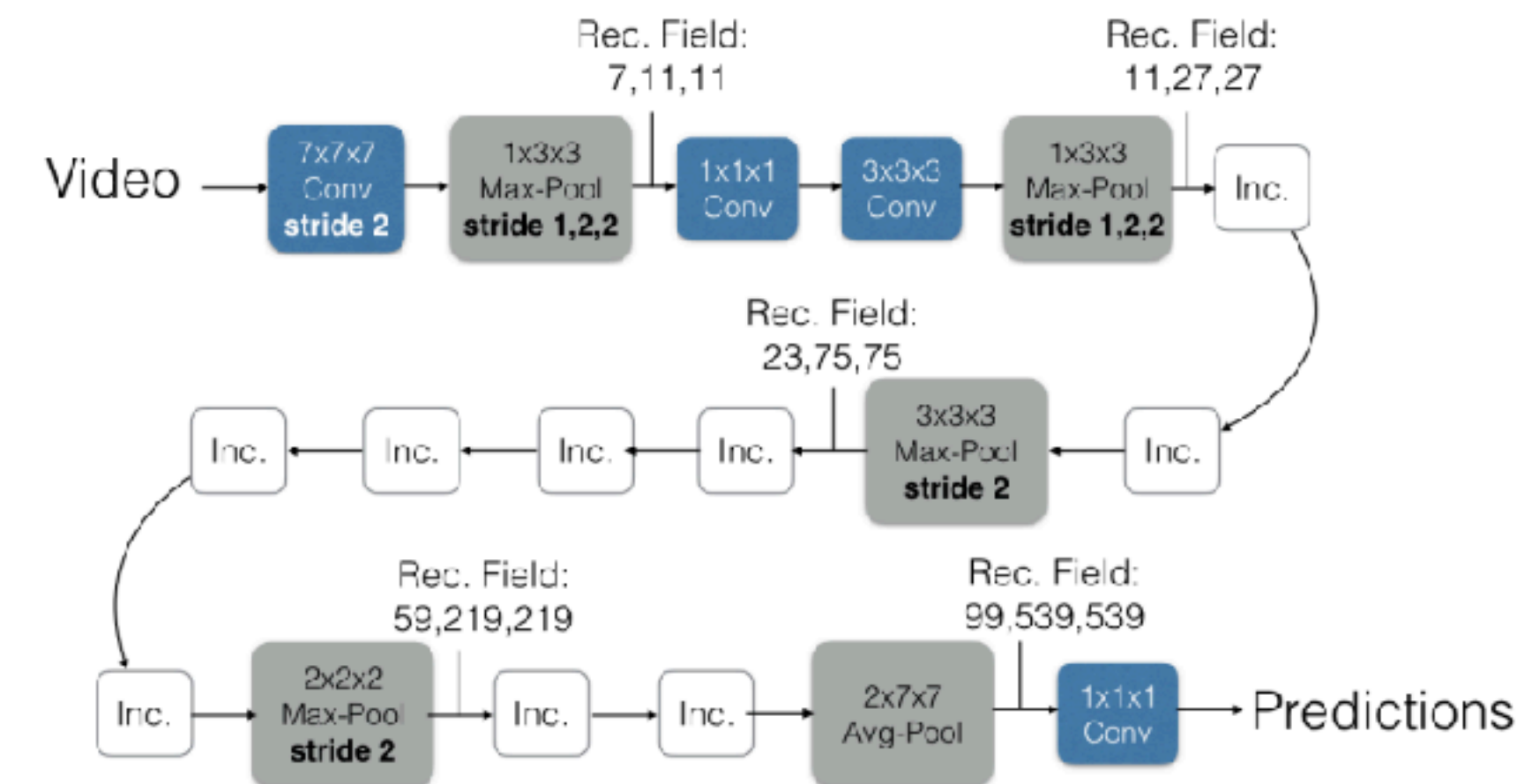
- Cross-modal
- Joint

Task learning

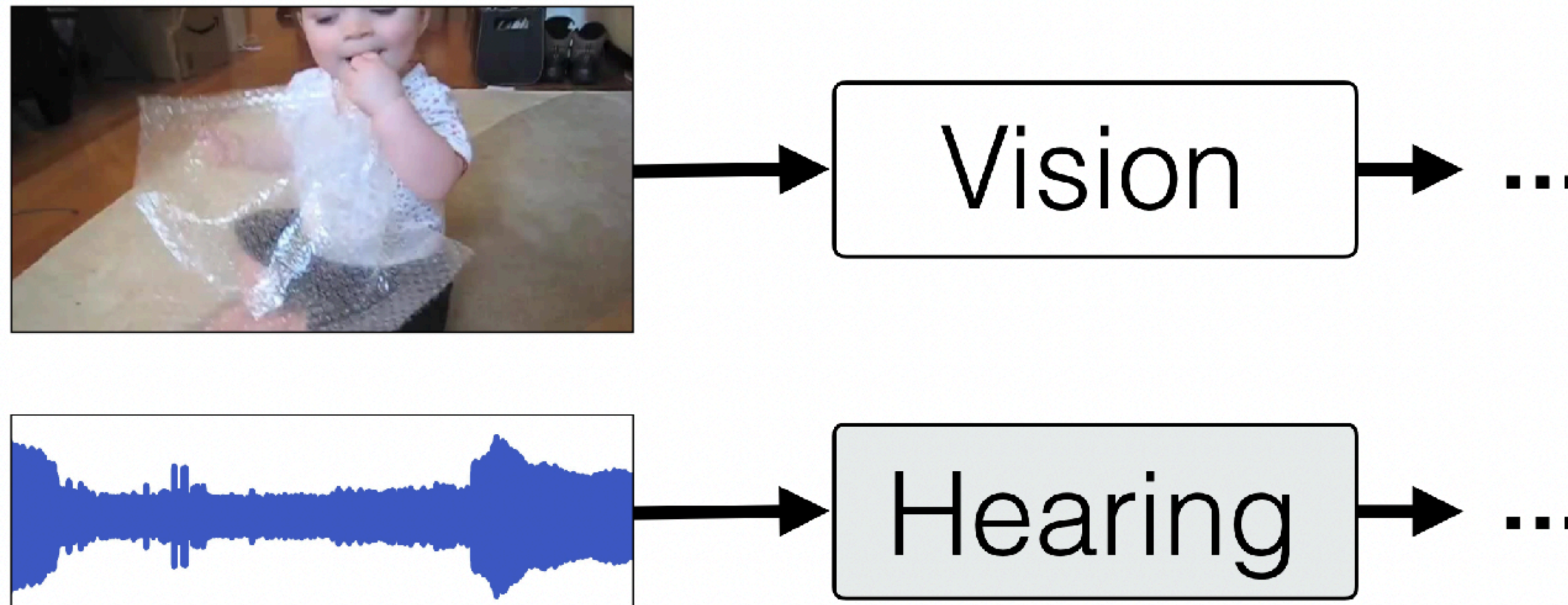
- Predict modality A from B:
 - Image/video captioning
 - Text-to-image/video generation
 - Speech(Audio)-to-text
 - Text(lyrics)-to-audio
 -



Inflated Inception-V1

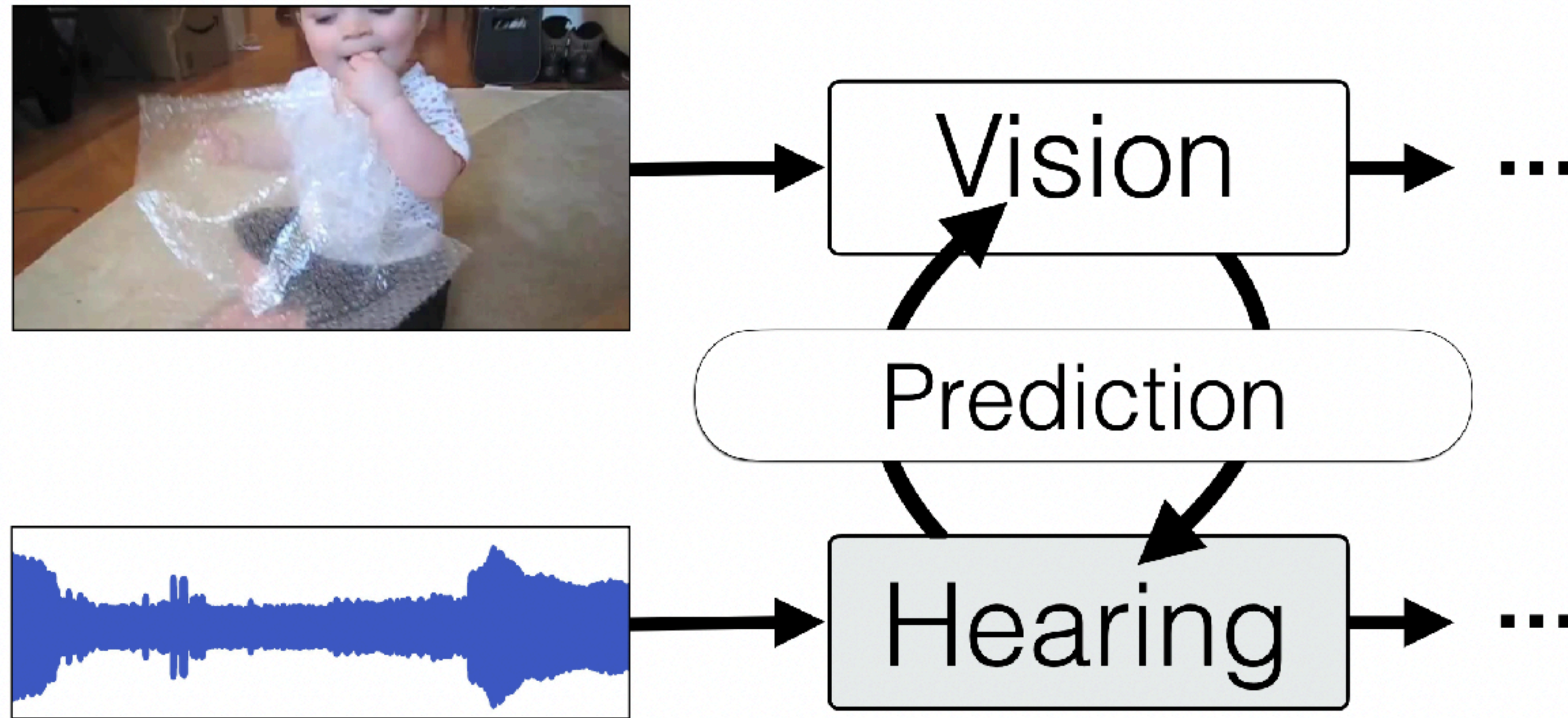


Self-supervision is the natural way to learn from paired data...



(de Sa 1994, Smith 2005)

...whereby one modality is used as supervision signal for the other



(de Sa 1994, Smith 2005)

Remember this slide?

The key to image understanding is separating meaning from appearance.



Original



Different lighting



Mirrored




Different zoom

Augmentations

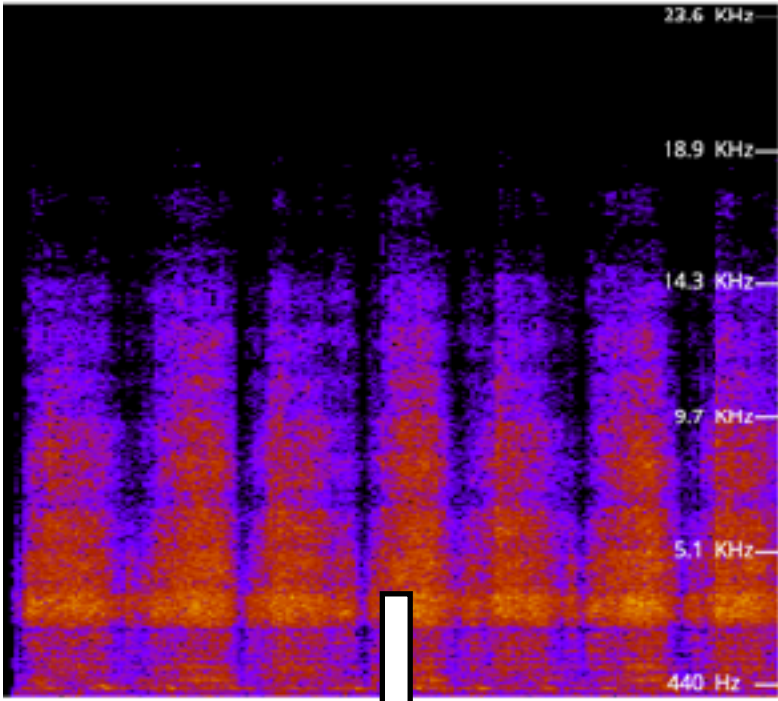


Multiple modalities can also yield such useful semantic information.


Video



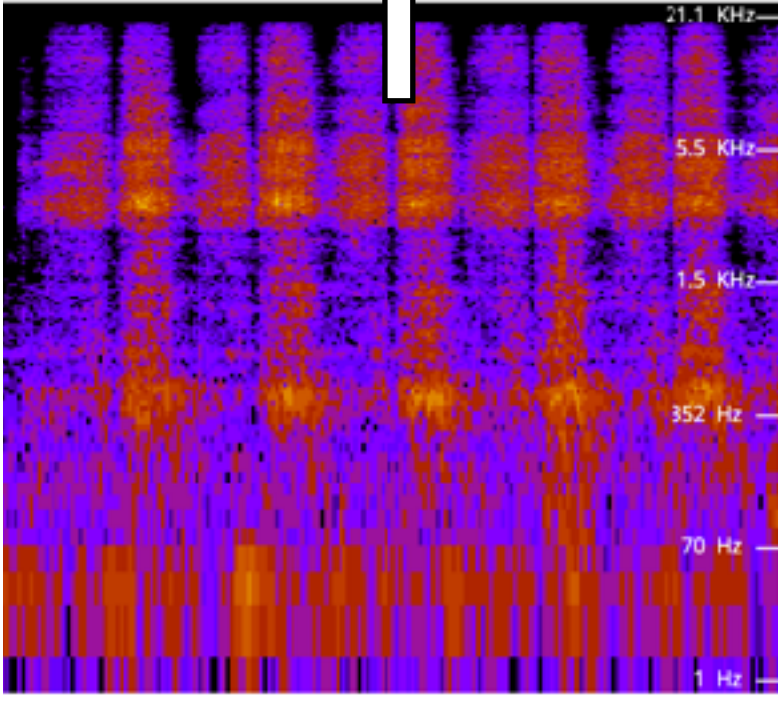
Audio



Video




Audio

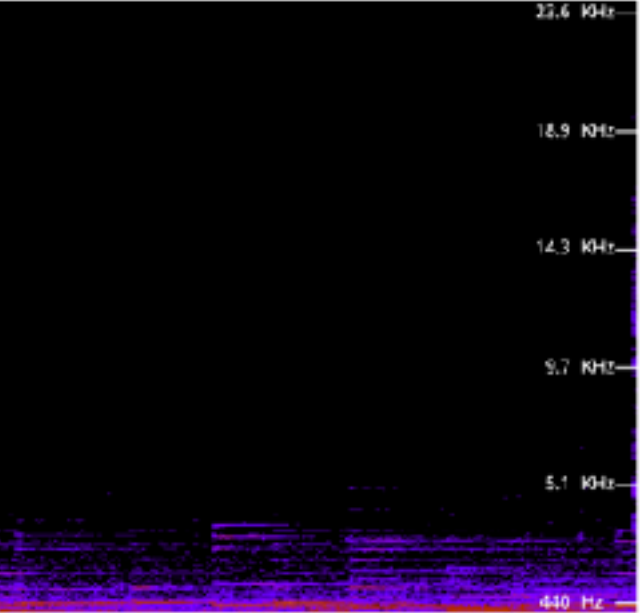


Two large white arrows point downwards from the top audio spectrogram to the bottom audio spectrogram, indicating a relationship or comparison between the two sounds.


Video



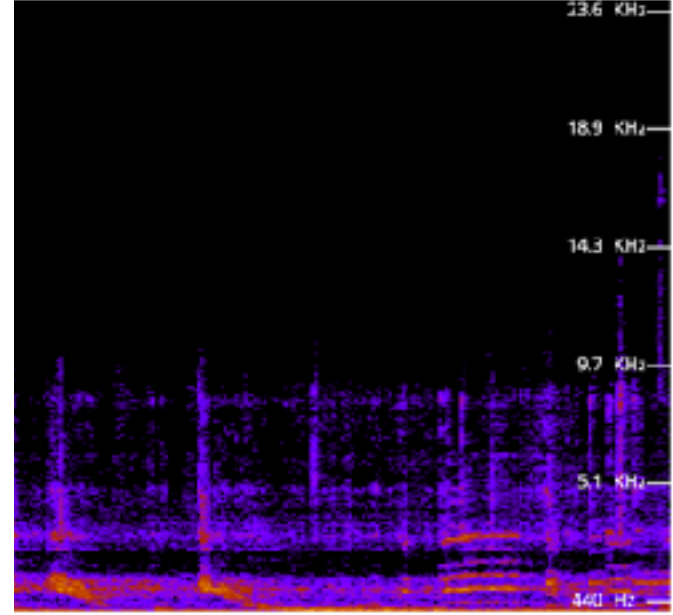
Audio



Video



Audio



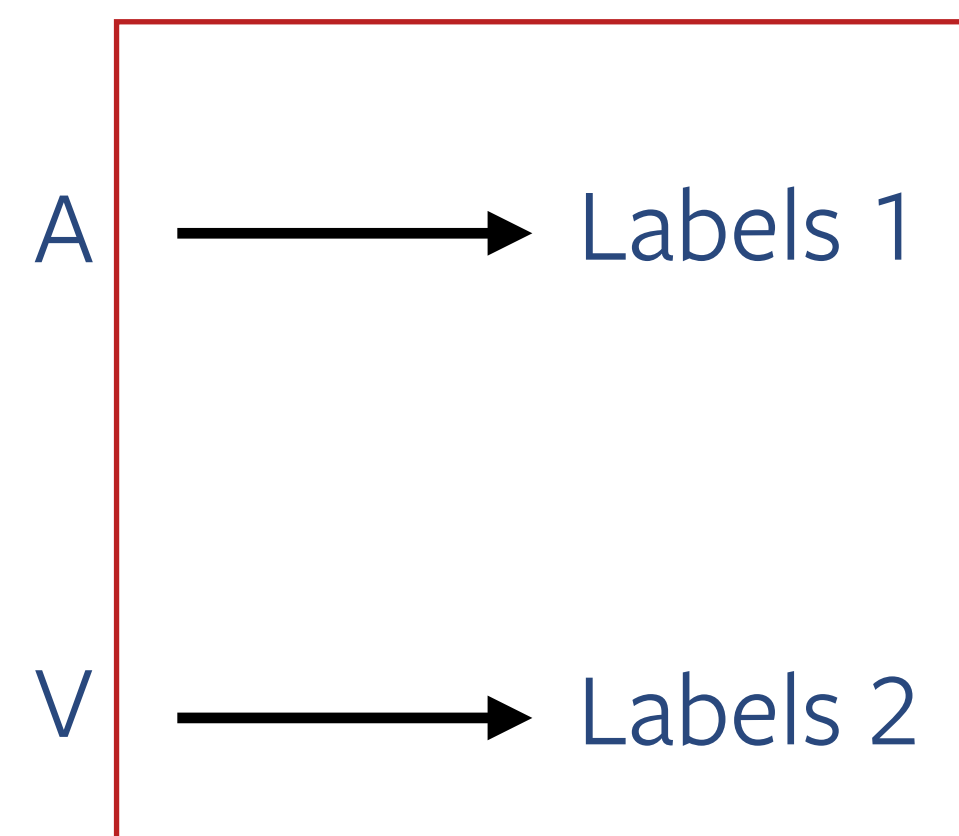
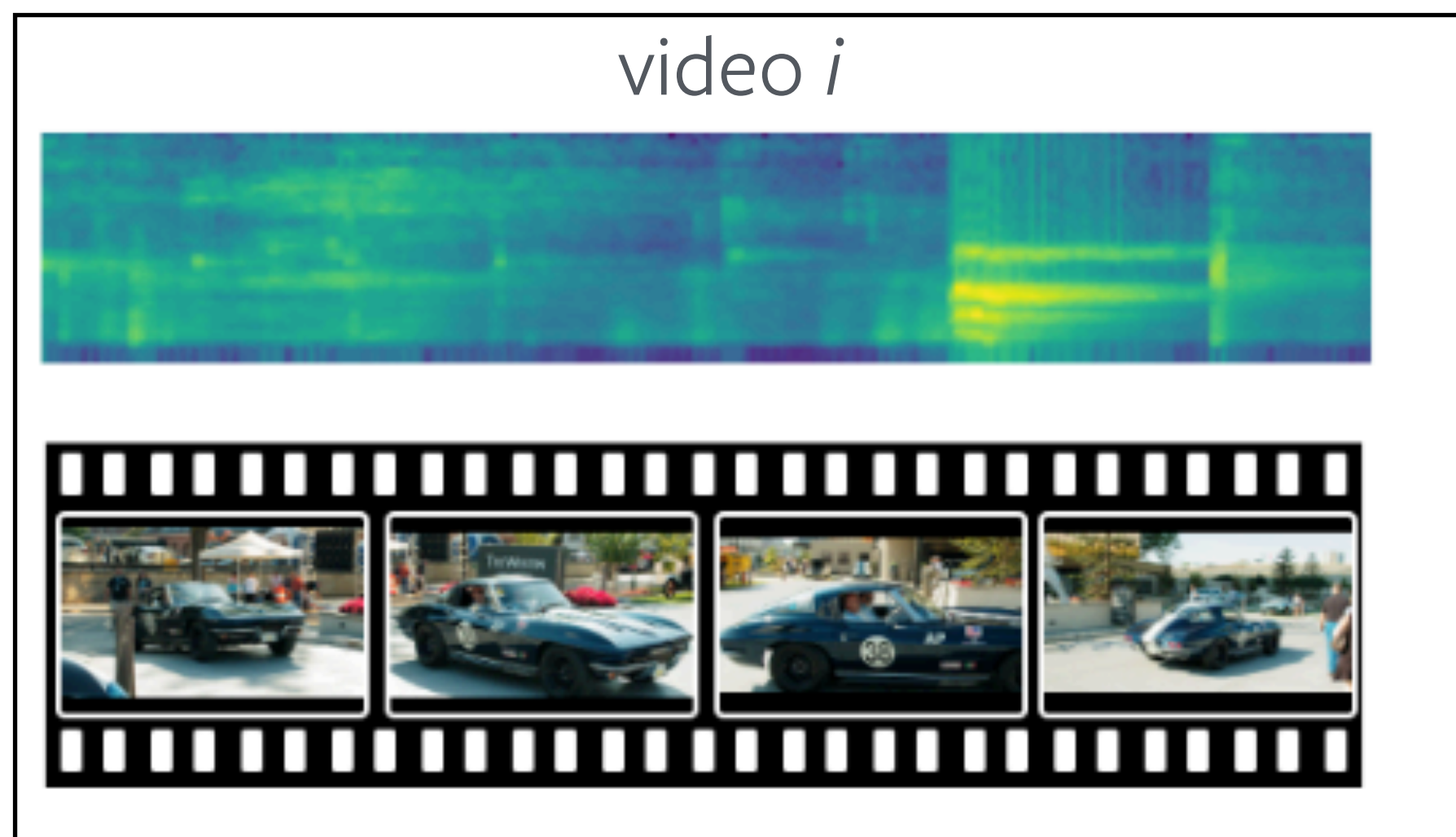
Two large white arrows point downwards from the top audio spectrogram to the bottom audio spectrogram, indicating a relationship or comparison between the two sounds.

Why clustering for videos?

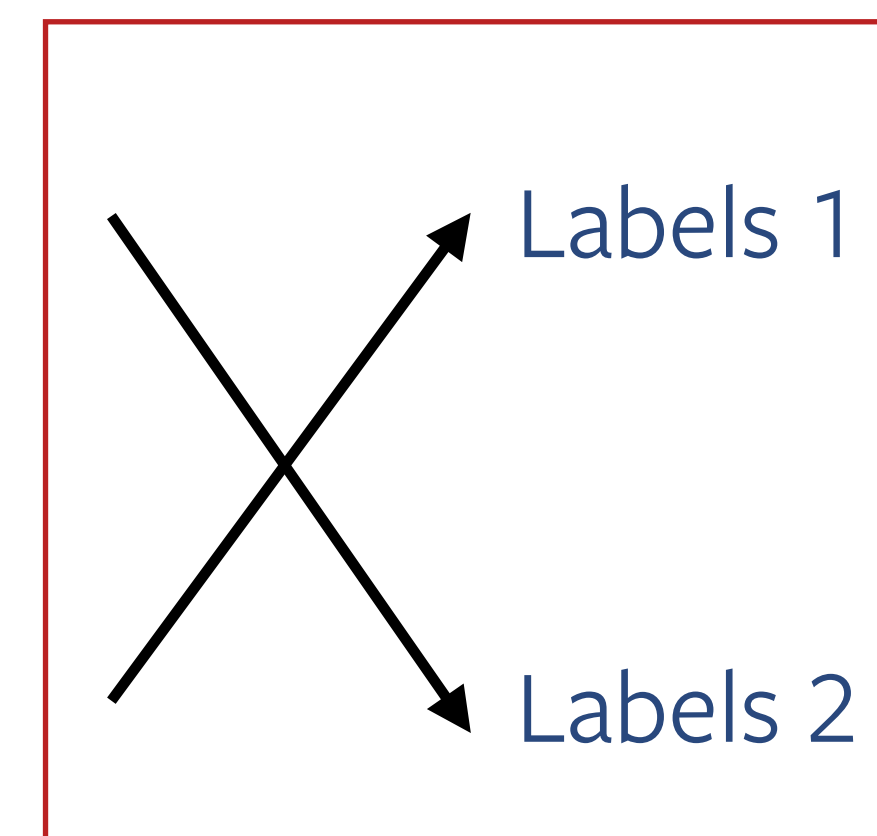
- 🤔 Clustering works well for images
- 💰 Videos are expensive to annotate.
- 📈 Video content is rapidly increasing.



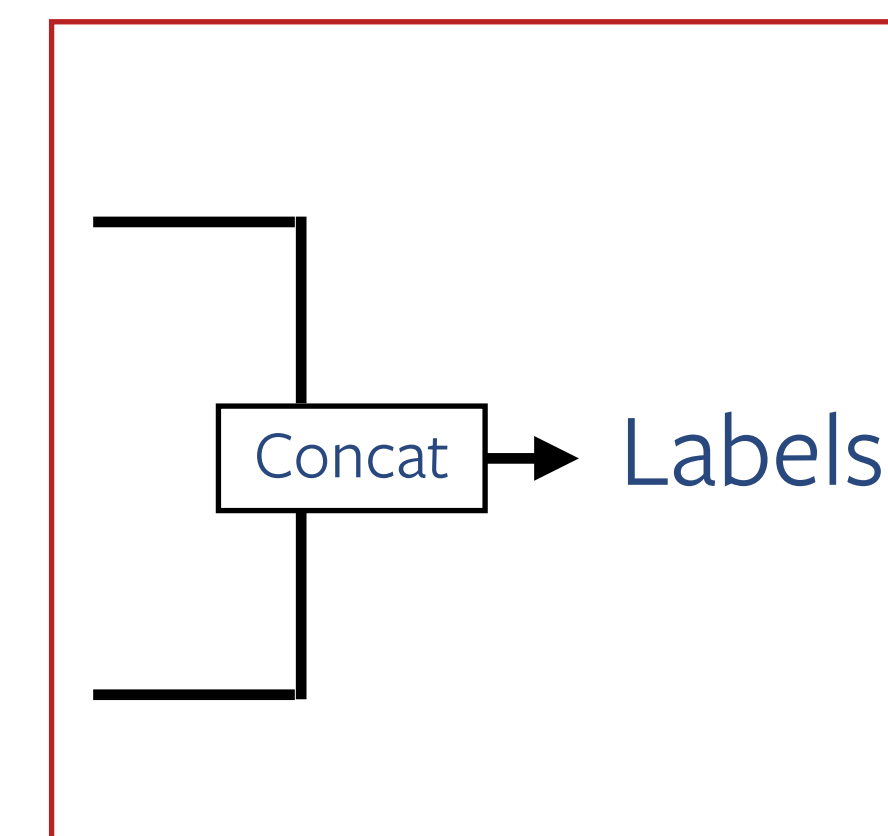
Clustering multi-modal data



- × does not use same-source information
- × two different sets of clusters

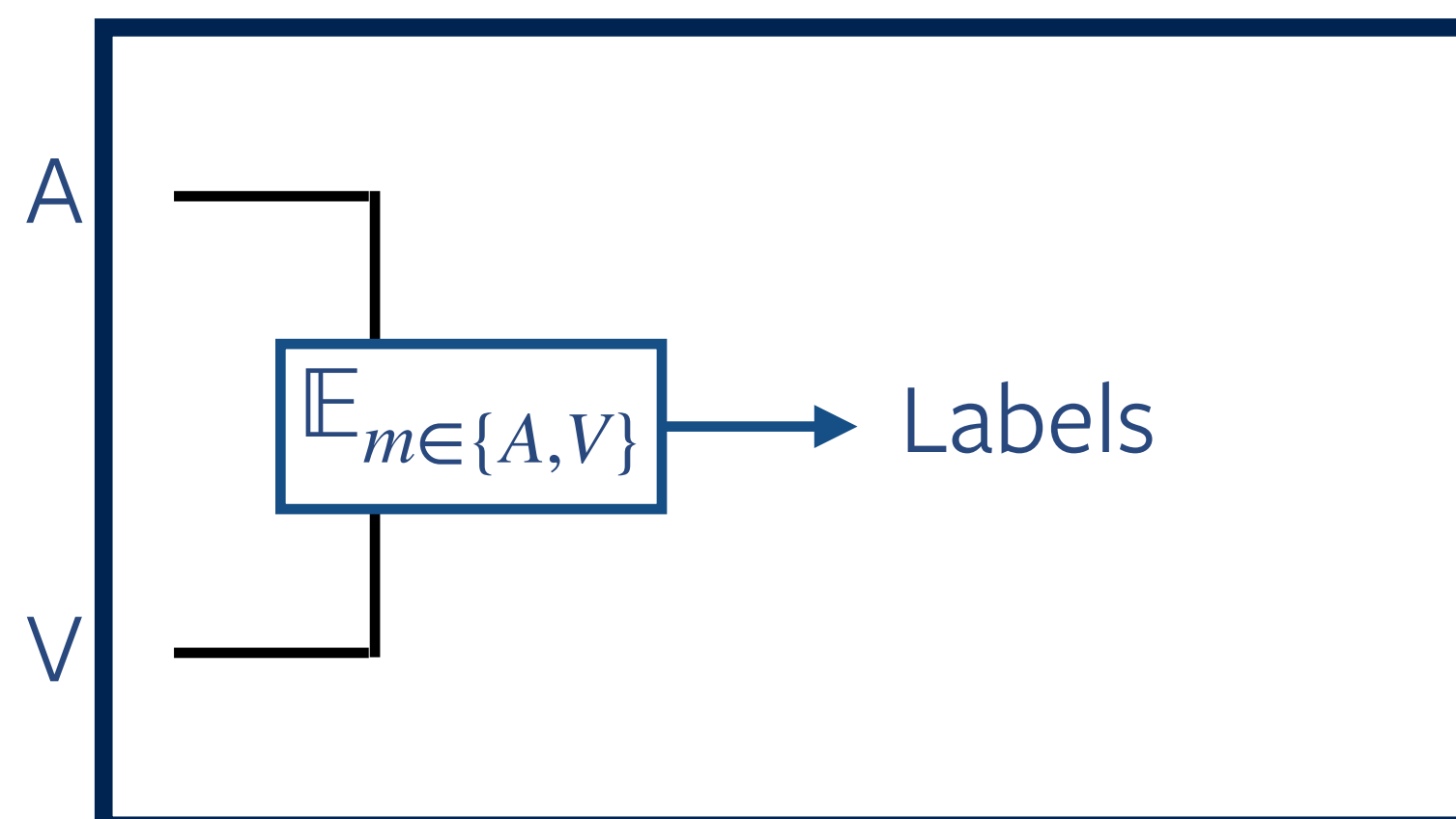


- × two different sets of clusters
- × hard to interpret



- × concatenation can just rely on stronger modality and ignore the other

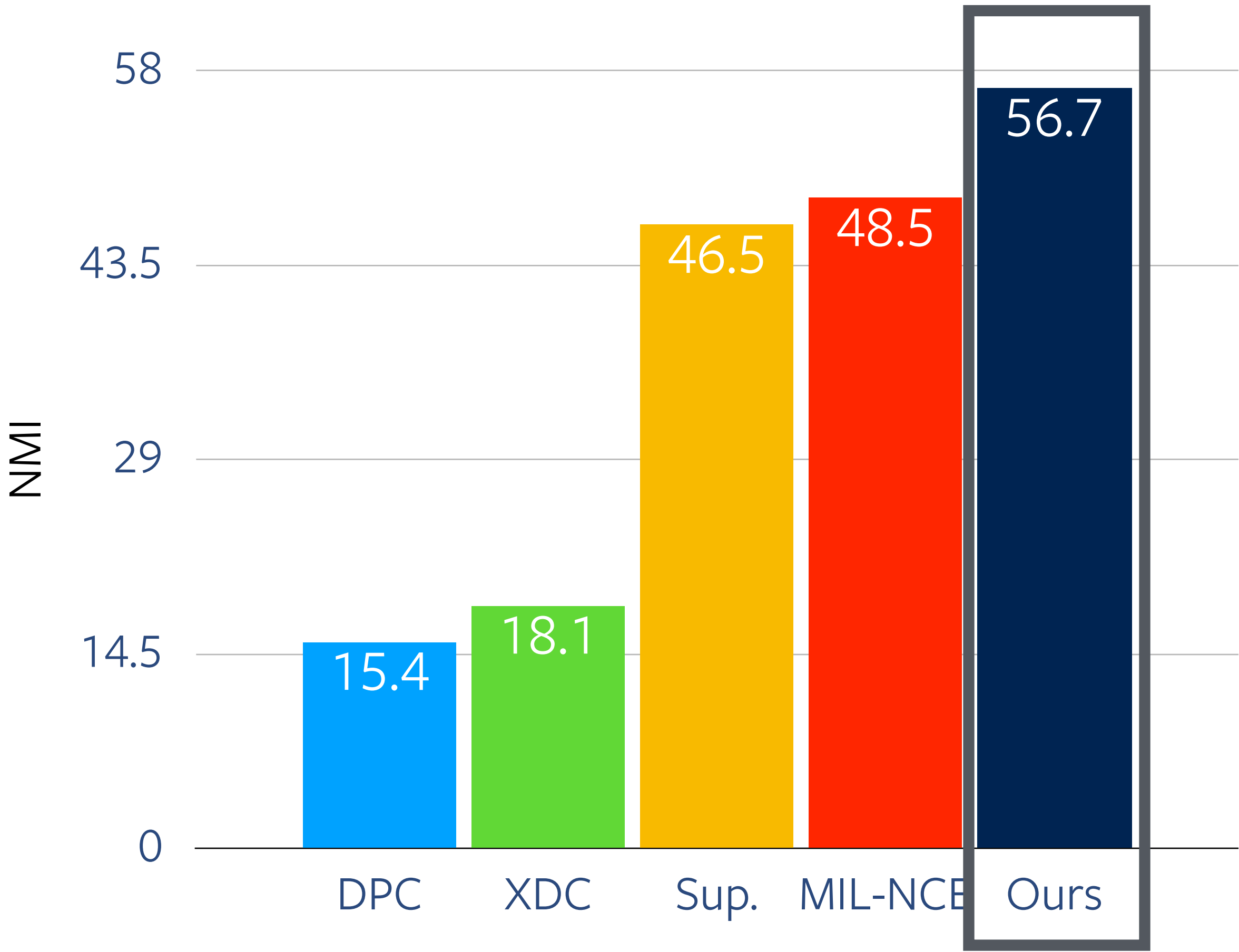
Audio



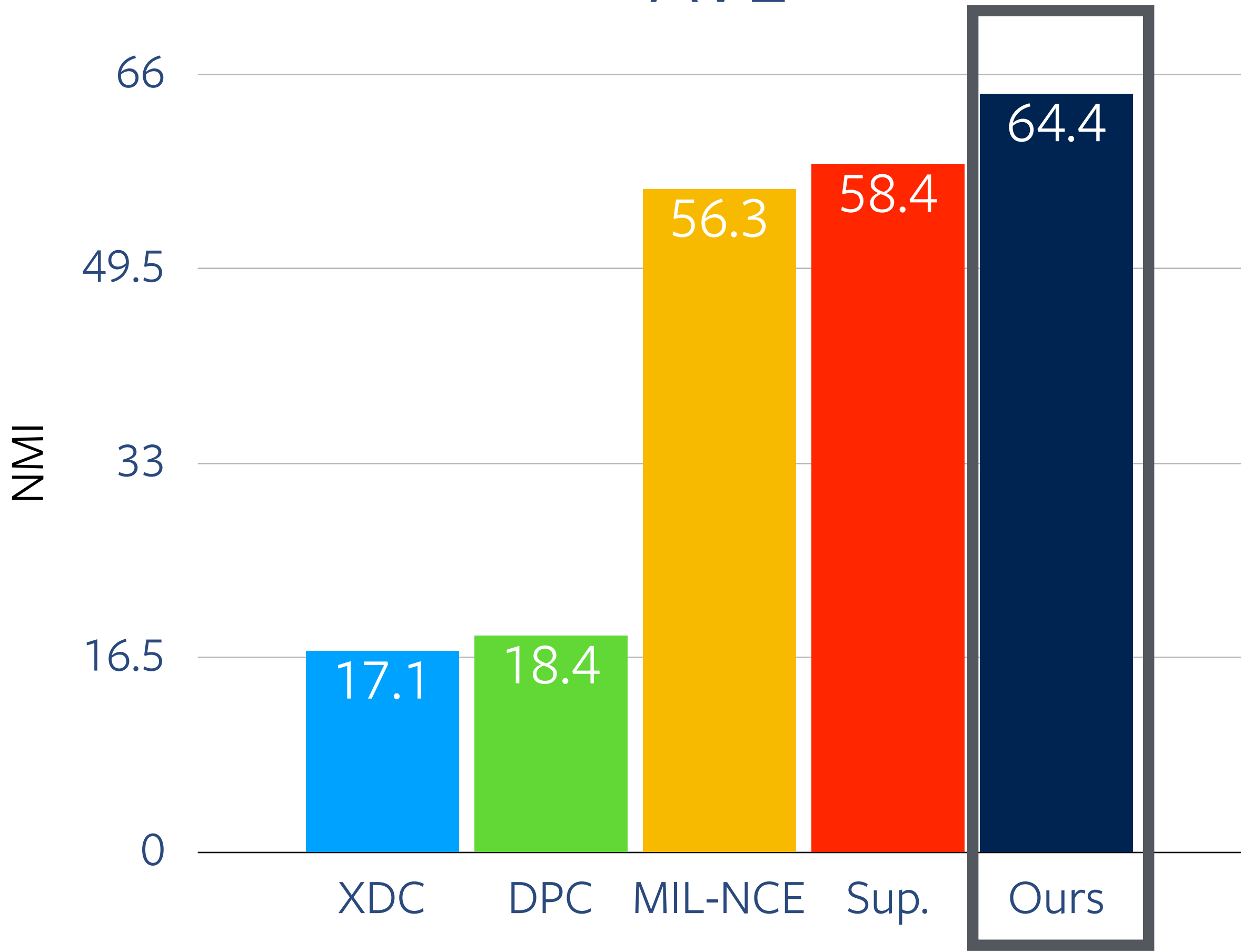
Our idea:
view each modality
as an *augmentation*

Good feature representations \rightarrow good clustering

VGG-Sound

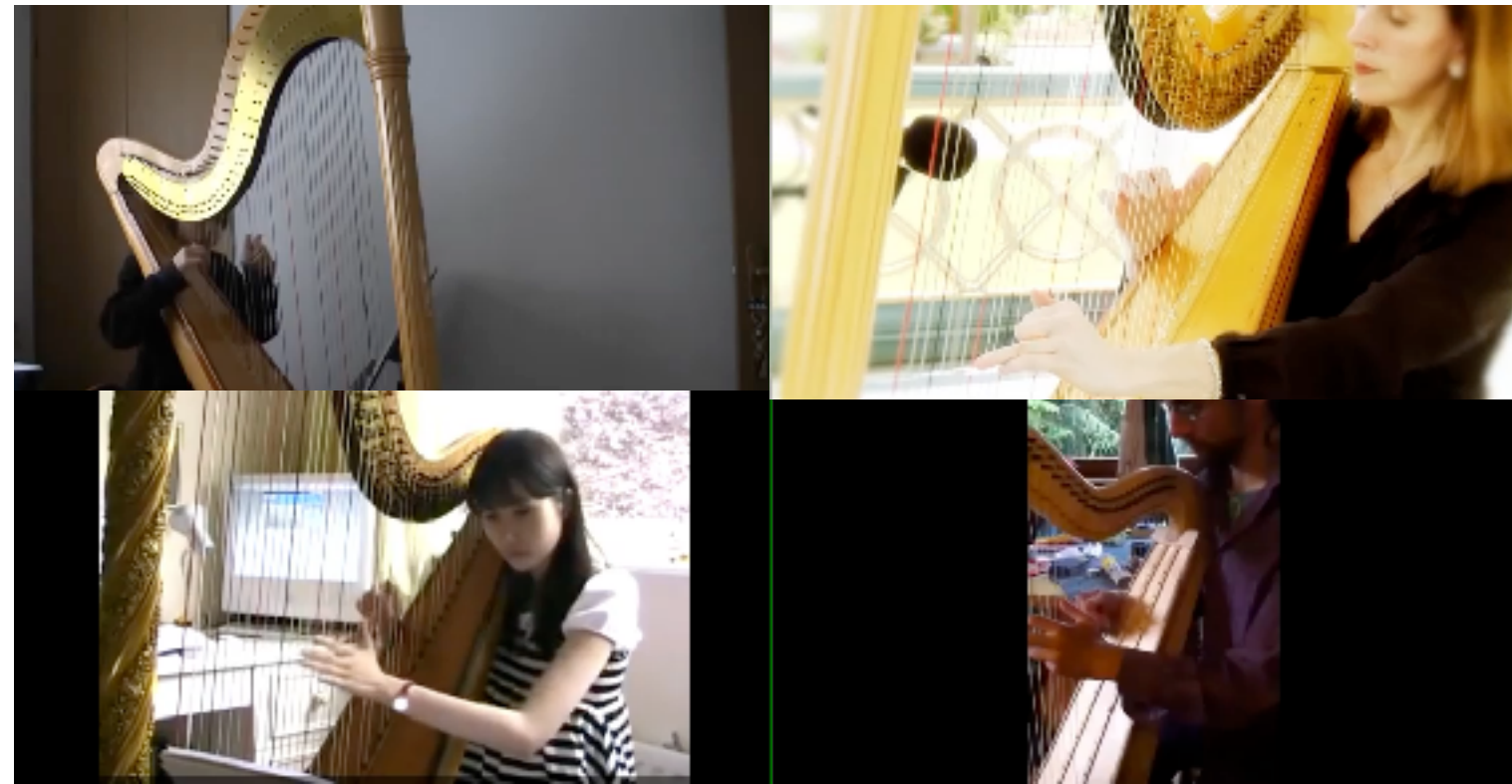


AVE



[Han et al., arXiv; Alwassel et al., NeurIPS 2020; Miech et al., CVPR 2020]

Discovering concepts without manual annotations from 230K videos



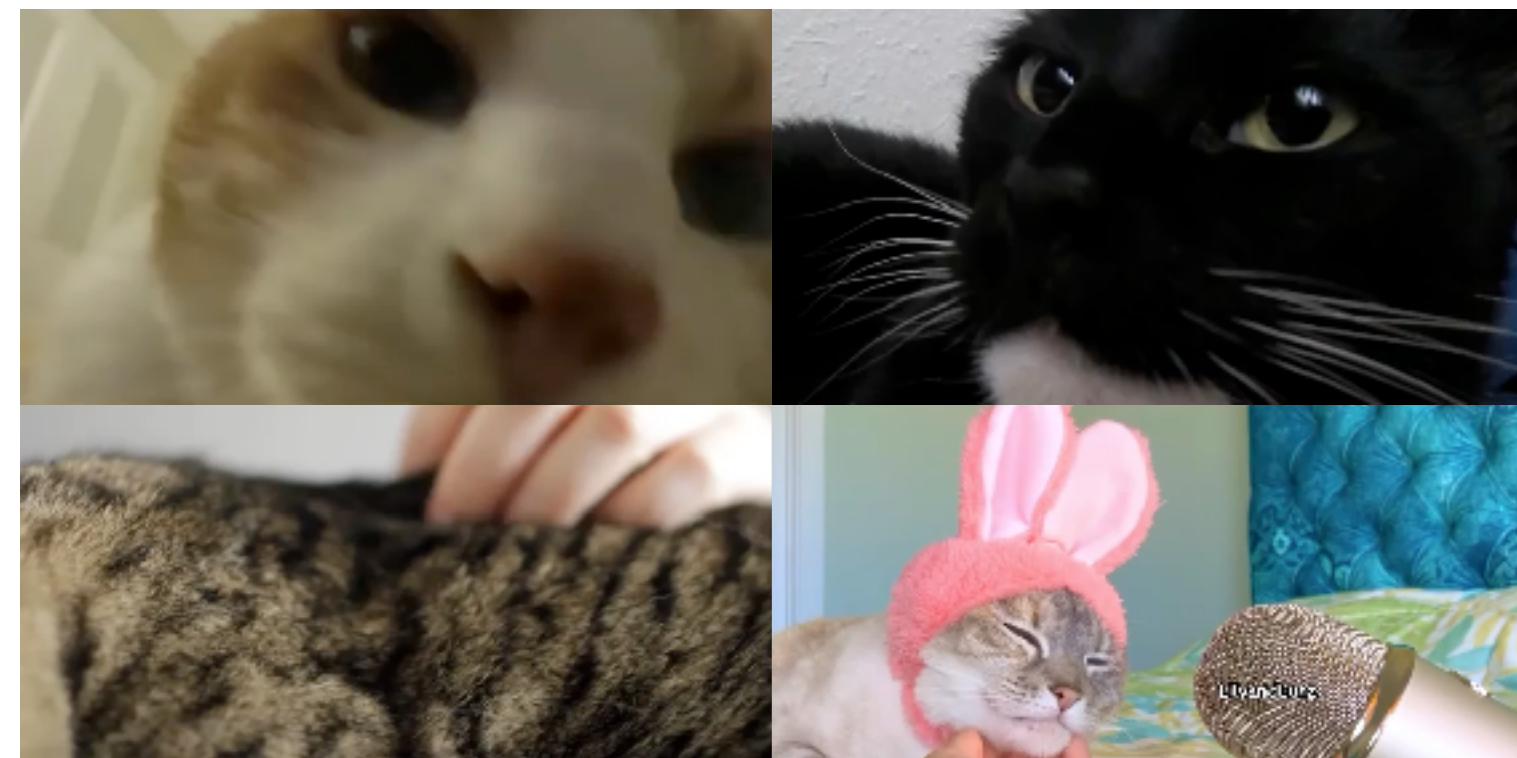
"Playing harp"



"Fireworks"



"Hockey game"



"Cat growling"



"Electric guitar"



"Vehicle driving"

“Self-Labeling” videos: three main findings

- Clustering framework of SeLa is well-suited
- Generalizeable to model any cluster distribution (Zipf/exponential etc.)
- Good feature representations do not imply good clustering

For completeness: here's a reference of common video datasets

Pretraining:

Kinetics-400 (600/700)

miniKinetics

VGGSound

HowTo100M

AudioSet

Youtube8M

Explore some datasets:

<https://www.robots.ox.ac.uk/~vgg/research/selavi/#demo>

Evaluation:

Kinetics/HMDB-51/UCF-101

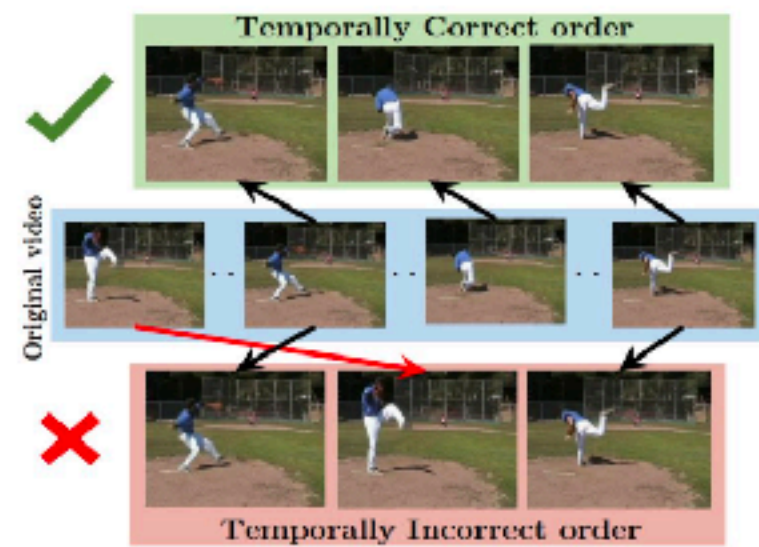
SomethingSomething-v2

Oops

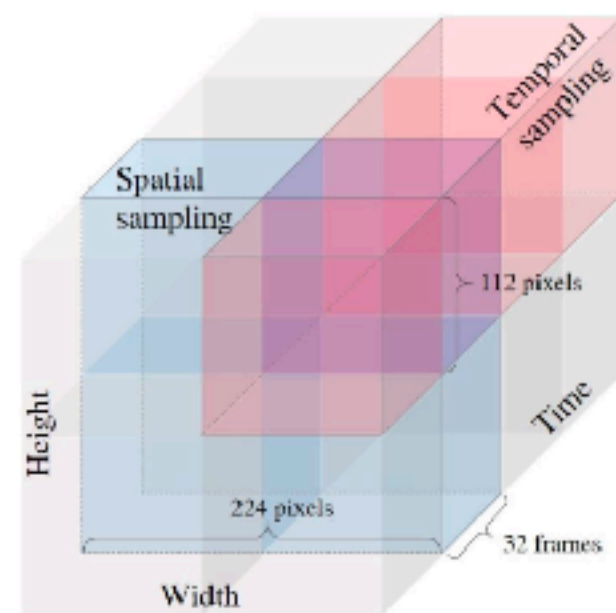
MSRVTT/VATEX/DiDeMo/ActivityNet

Youtube-VOS/VIS

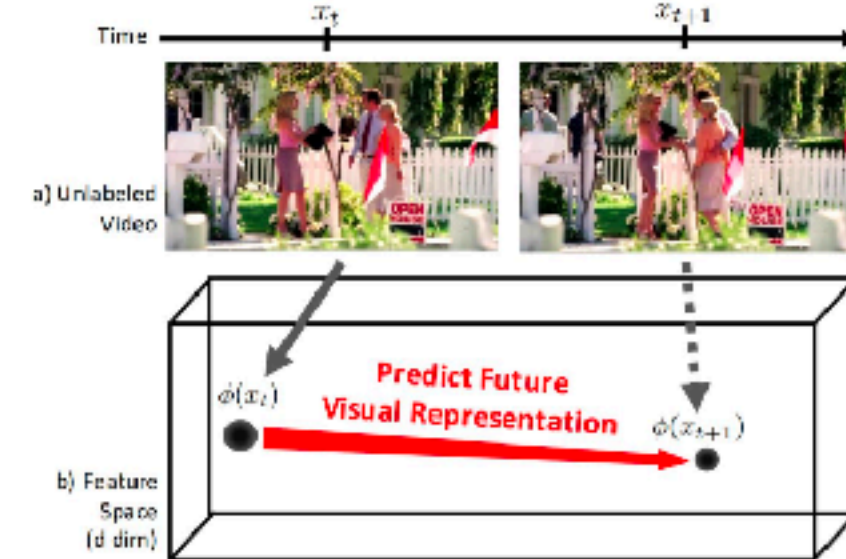
Other popular self-supervised learning approaches



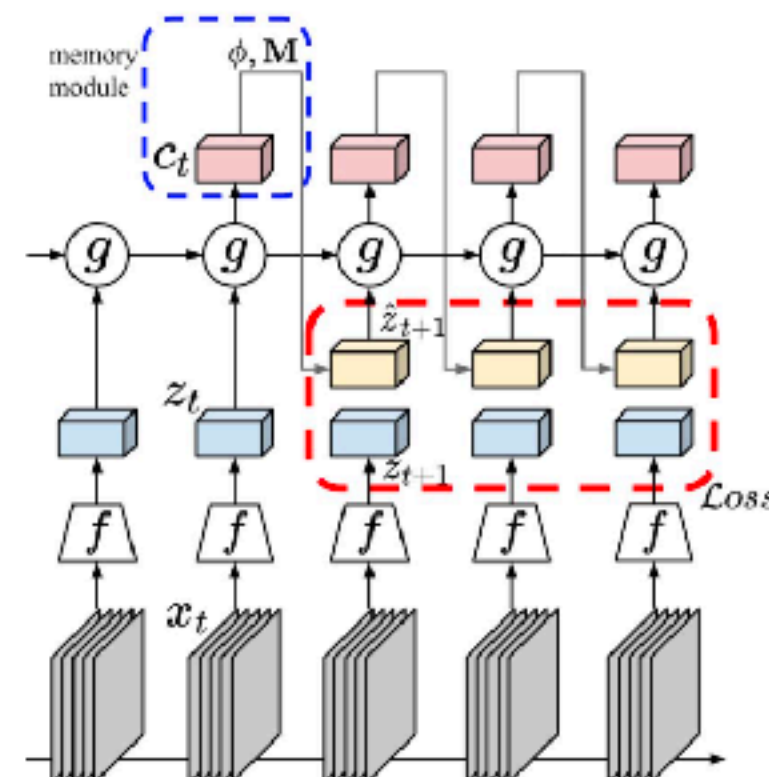
Shuffle & Learn. Misra et al.



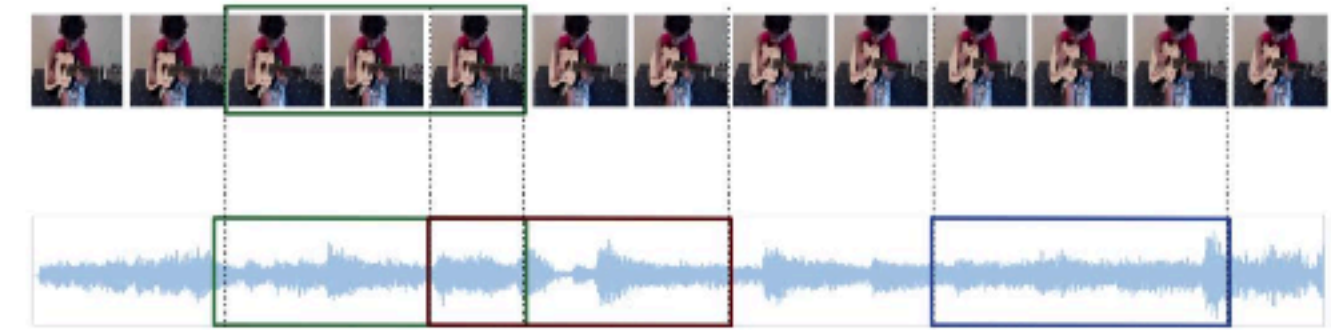
3D ST-Puzzle. Kim et al.



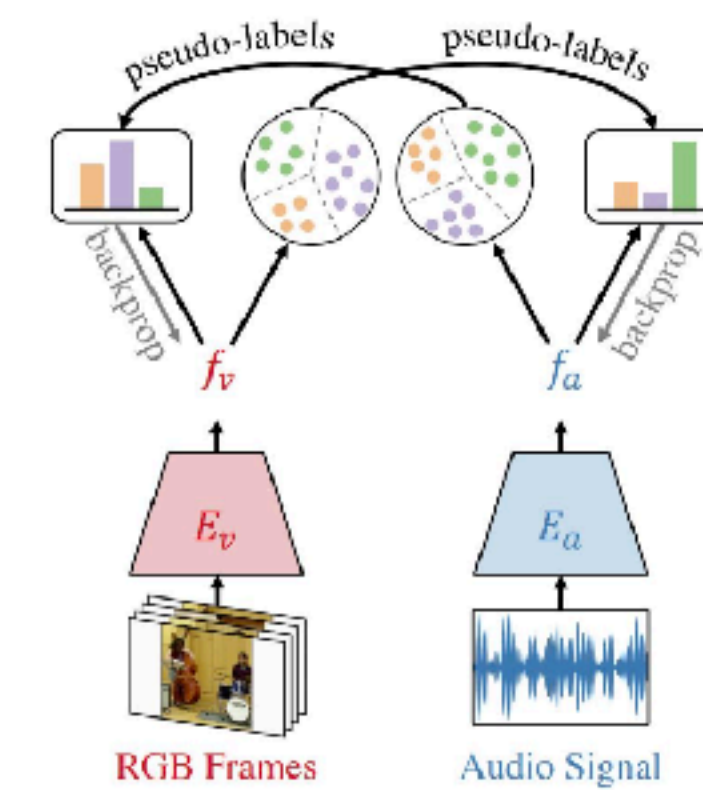
Anticipating Representation. Vondrick et al.



DPC/MemDPC. Han et al.



Audio-Video Synchronization. Korbar et al.



Cross-Modal Audio-Video Clustering. Alwassel et al.



MIL-NCE. Miech et al.

On Compositions of Transformations in Contrastive Self-Supervised Learning

Mandela Patrick*, Yuki M. Asano*, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, Andrea Vedaldi
ICCV'21

Invariance vs distinctiveness

In contrastive learning, we define positives and negatives.

Should the representations enforce invariance or distinctiveness?



?
=



Learning hypotheses we test:

1. Sample Distinctiveness



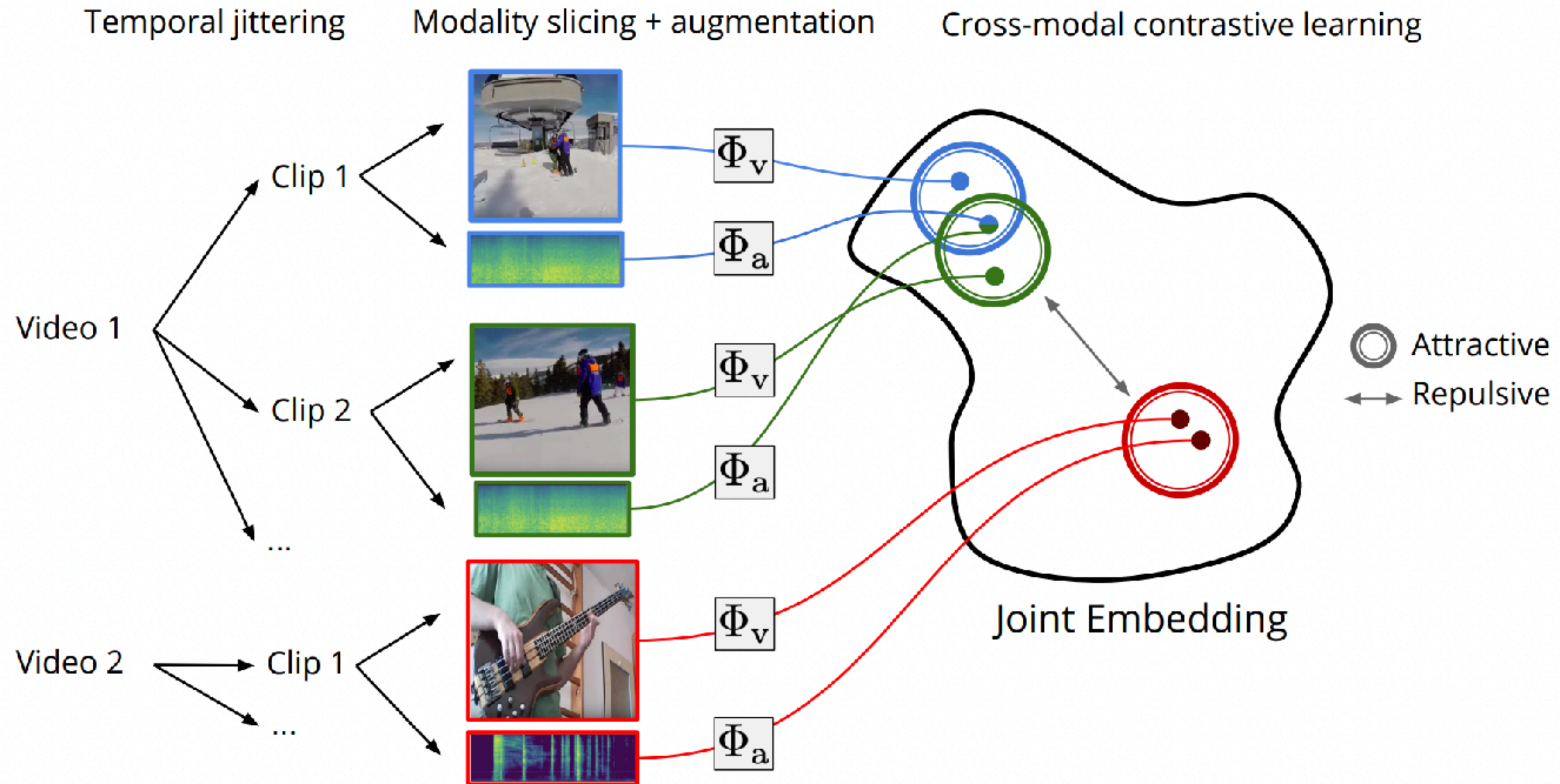
2. Time Reversal



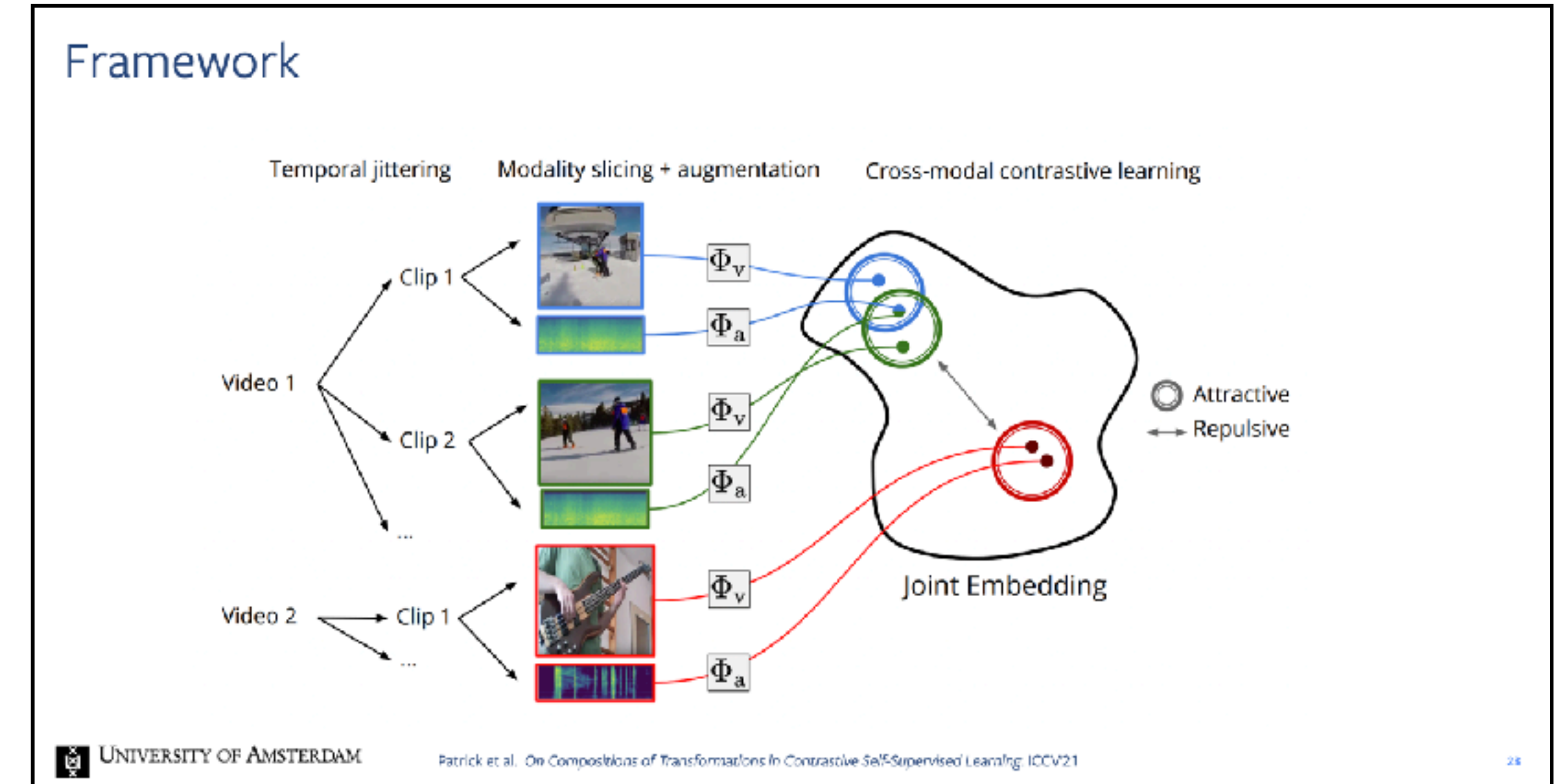
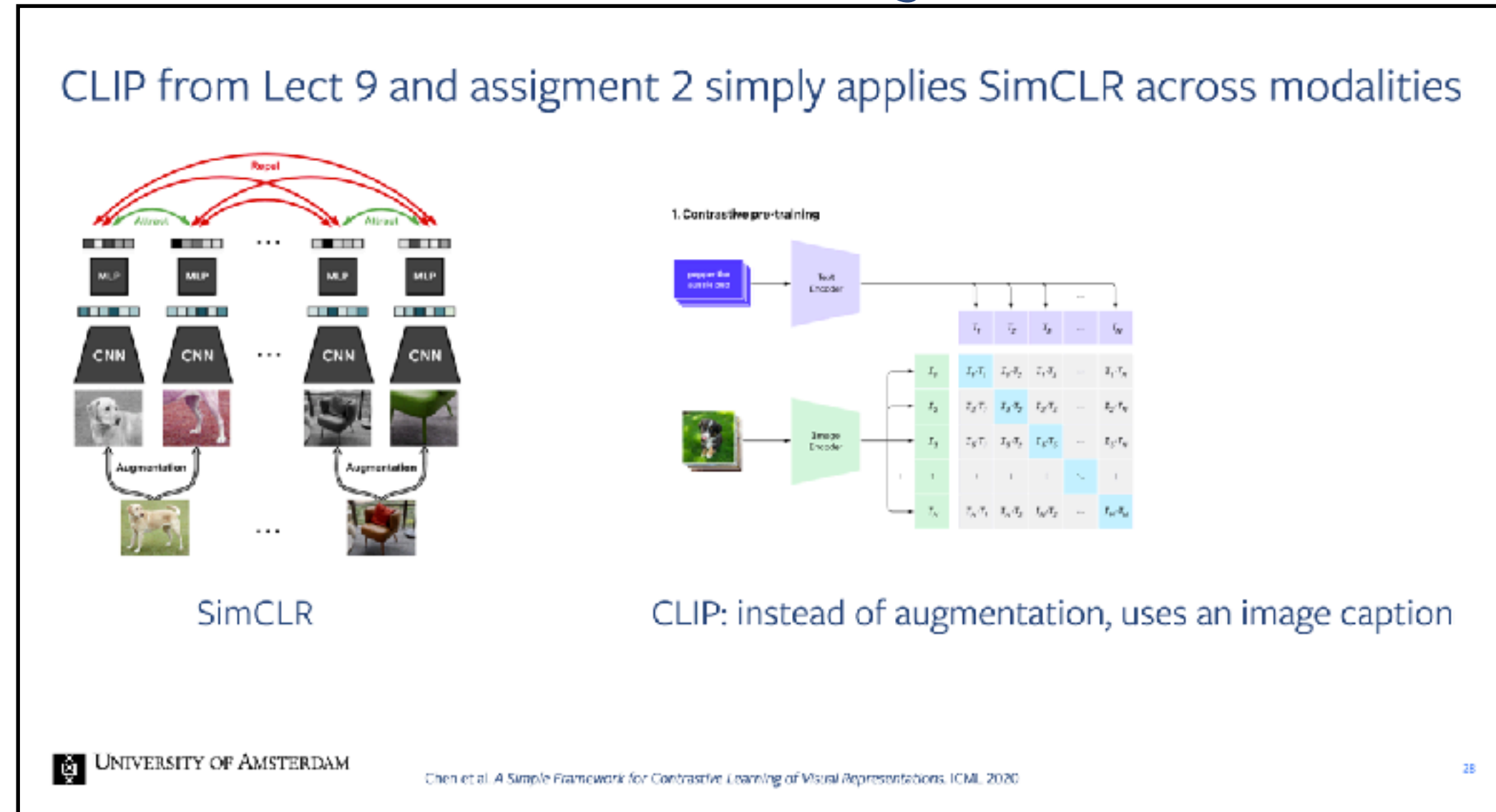
3. Time Shift



Framework



Notice the similarity to SimCLR:

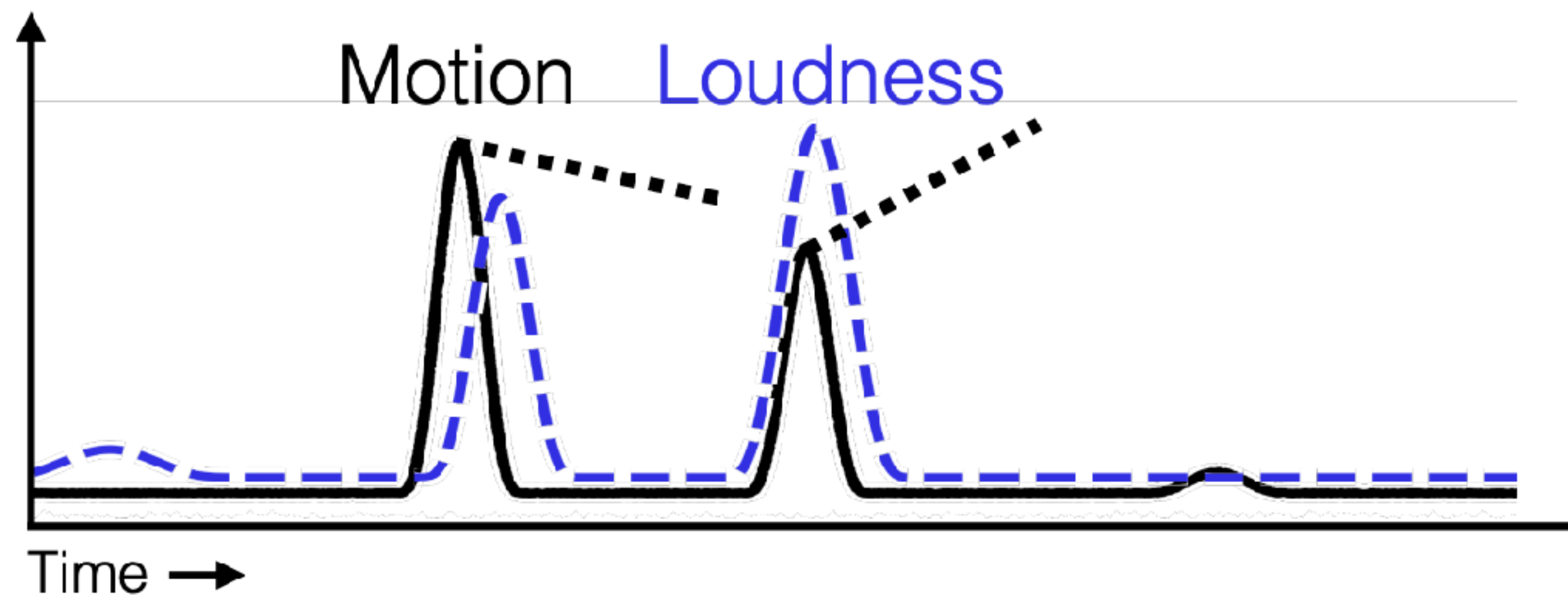


Quiz: this work resembles “simply applying SimCLR to audio-video”, what does one need to be careful about?

- 1) Video augmentations like time-shift can change the semantics
- 2) The audio inputs cannot handle augmentations
- 3) Constructing a large enough batch (for contrastive learning) is more difficult
- 4) The contrastive loss should not be applied symmetrically

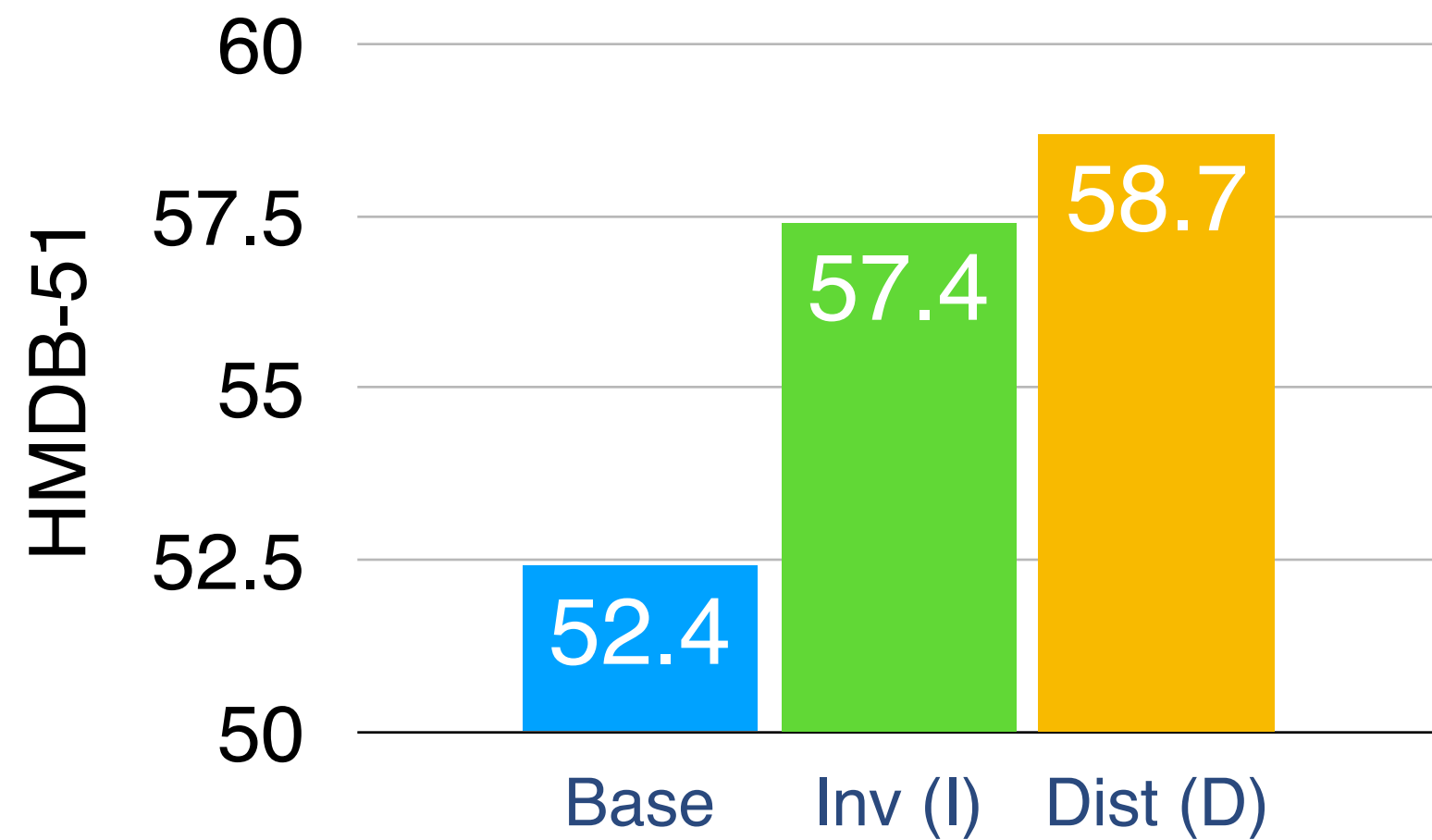
Detecting time-shifted pairs as negatives is not easy

- The scene-level semantics match, but the timing doesn't
- This requires fine-grained recognition
- However, there's work* that shows time can also be used as an augmentation to some degree

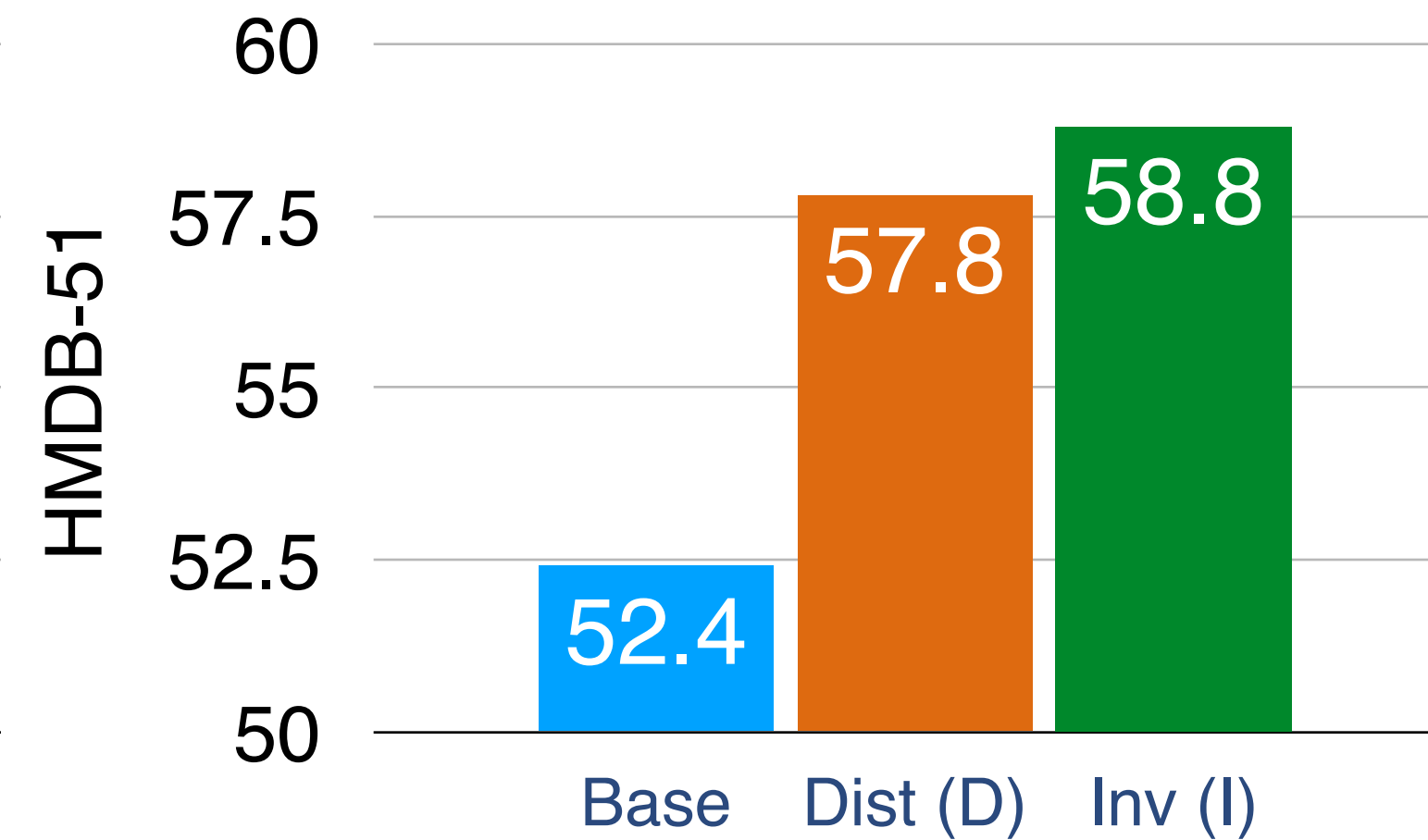


Gains from hypotheses

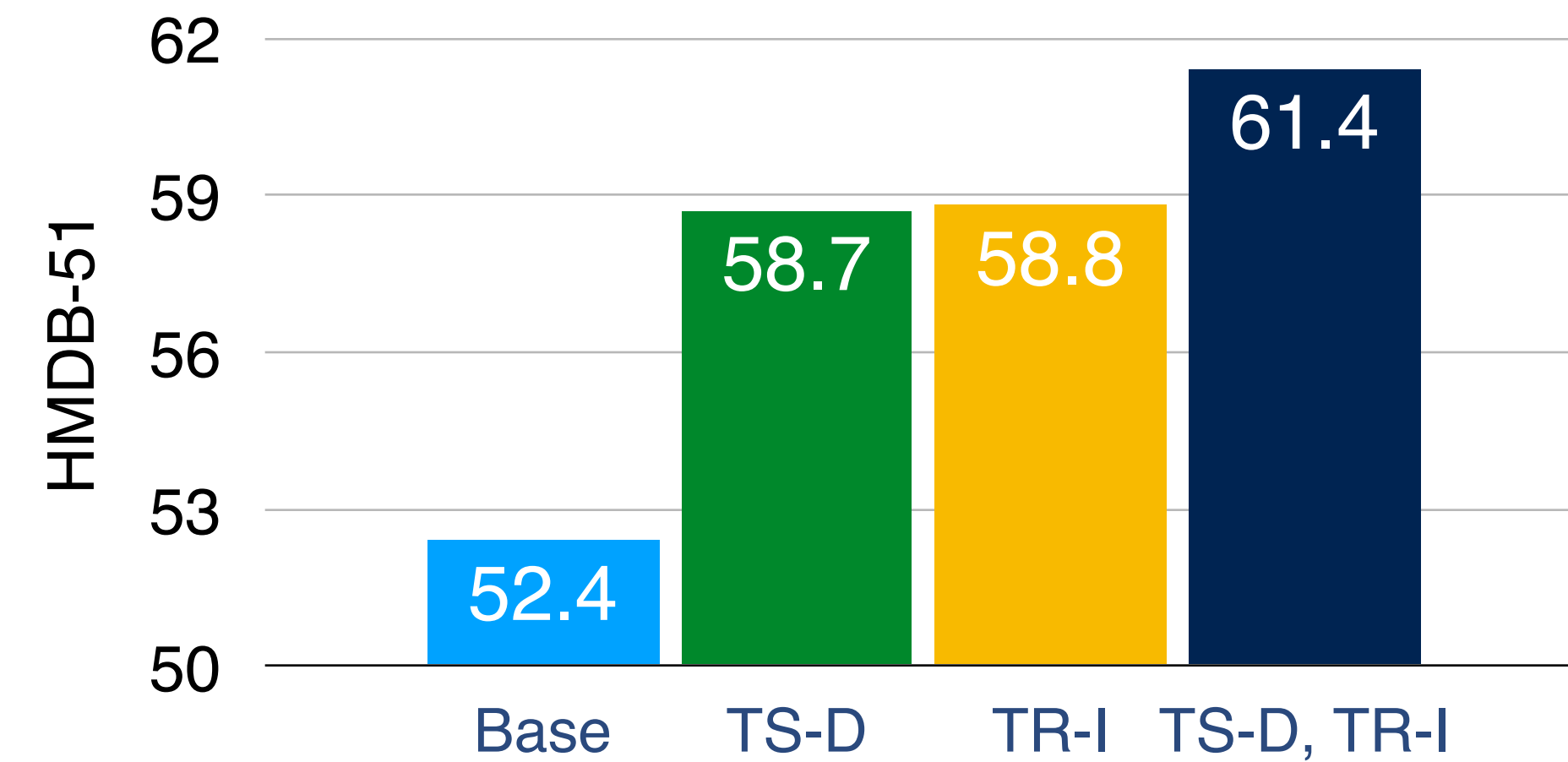
Timeshift (TS)



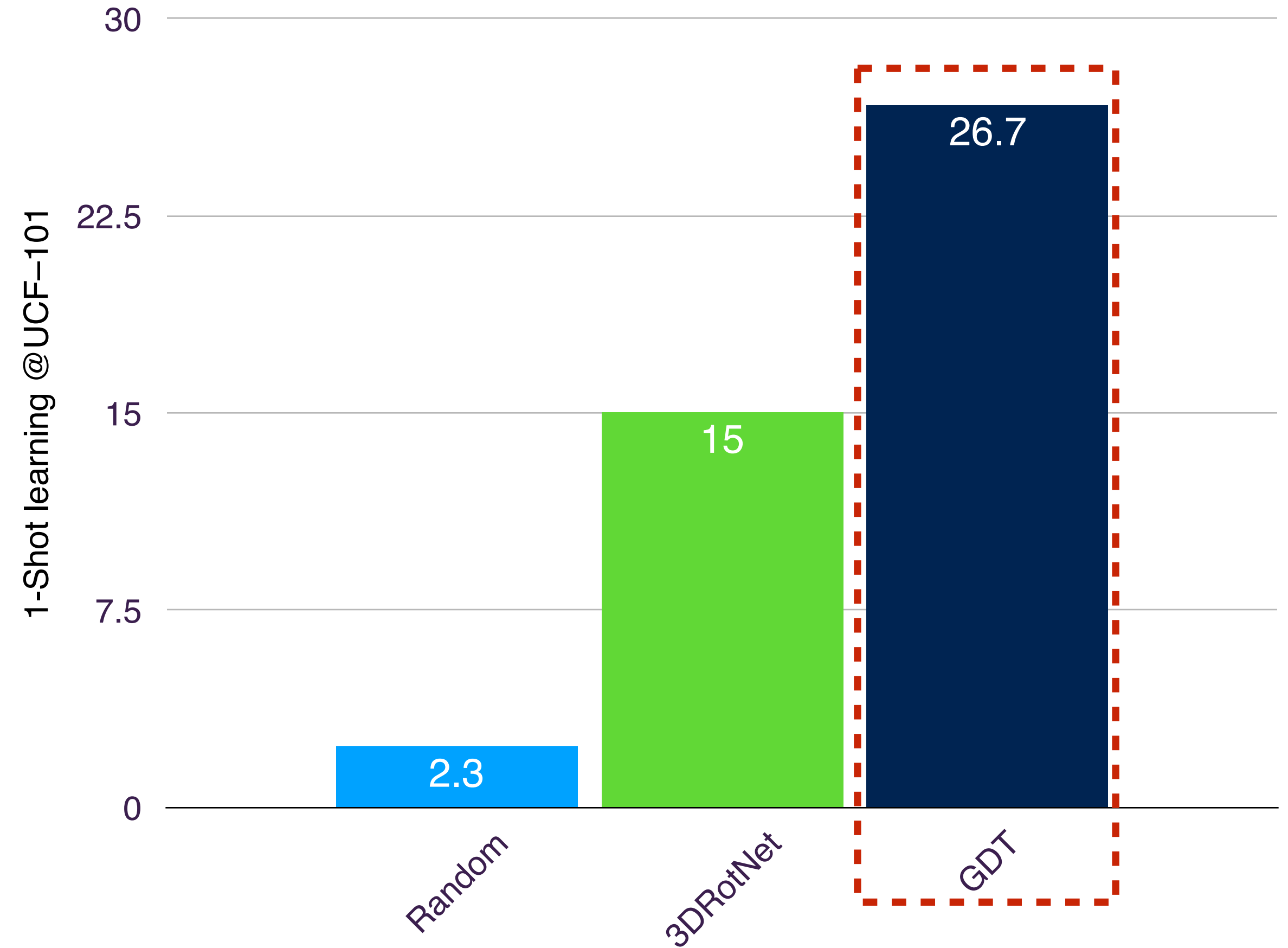
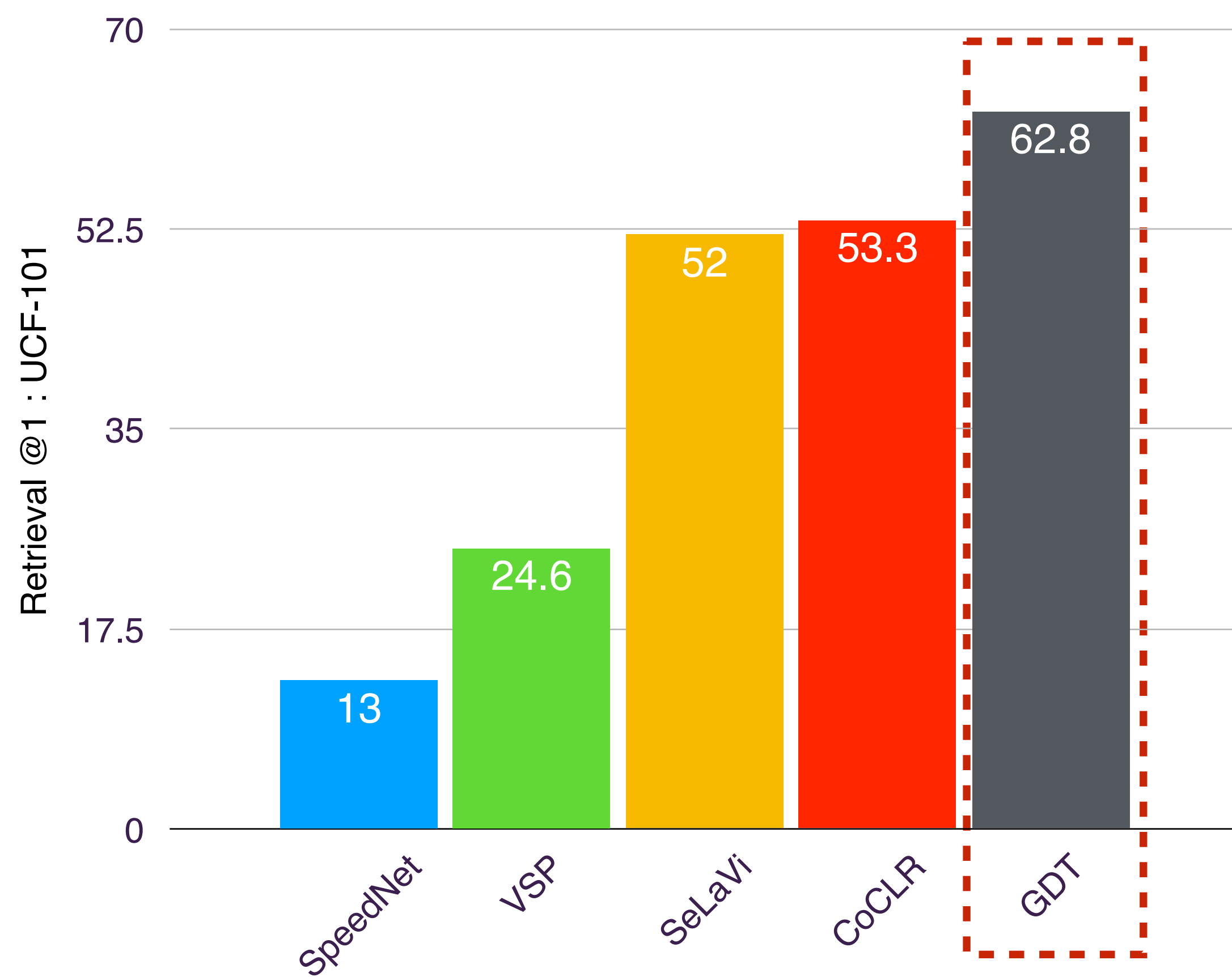
Time reversal (TR)



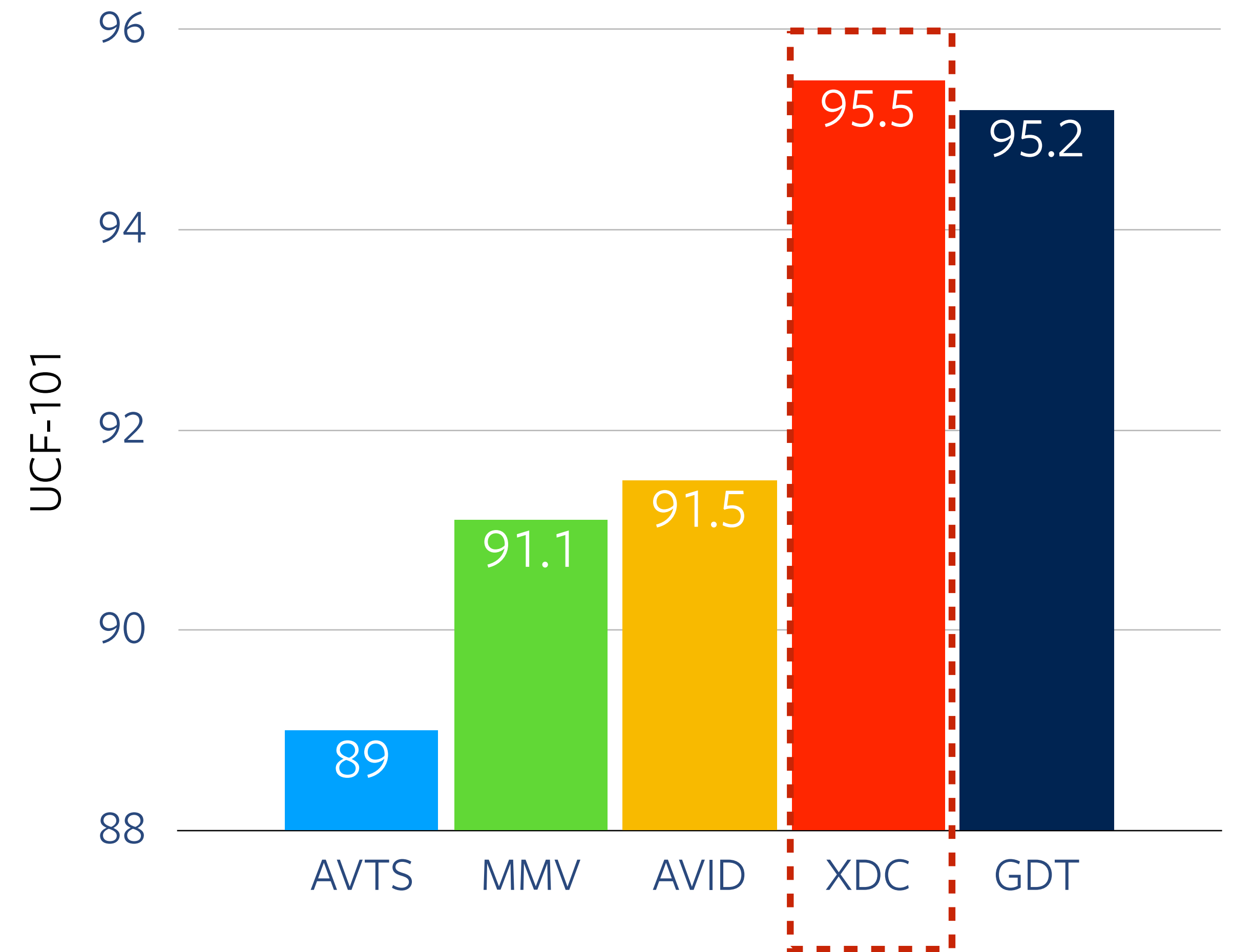
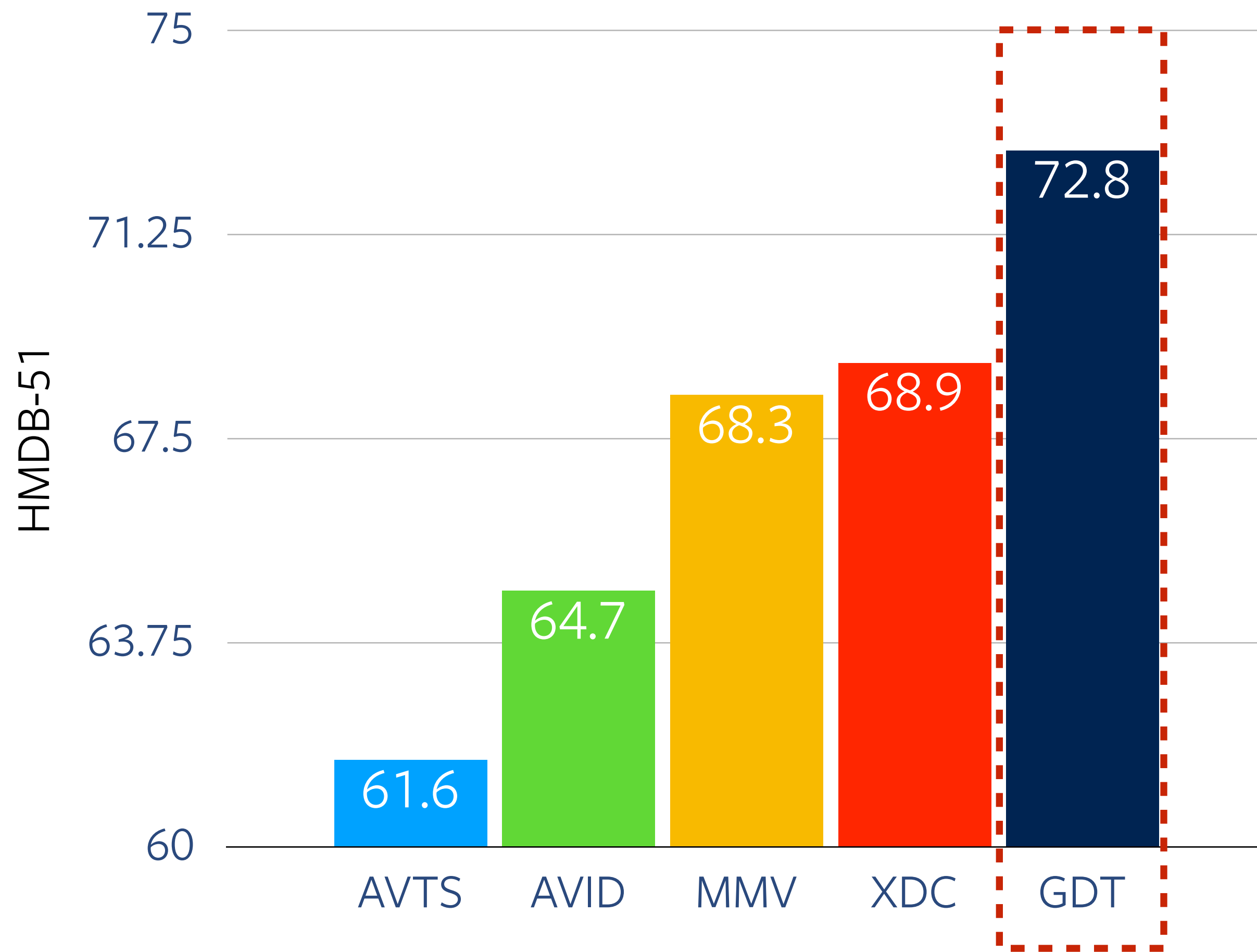
Combinations



SOTA video action retrieval and few-shot learning results



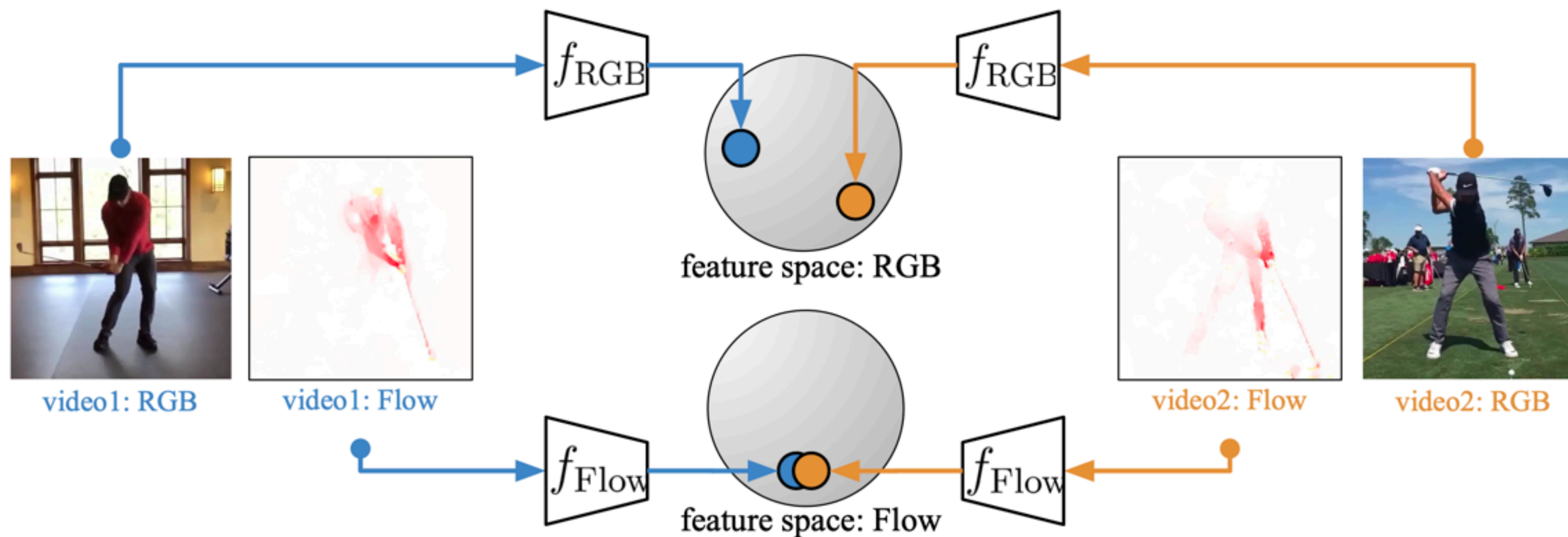
SOTA finetuning video-action recognition results



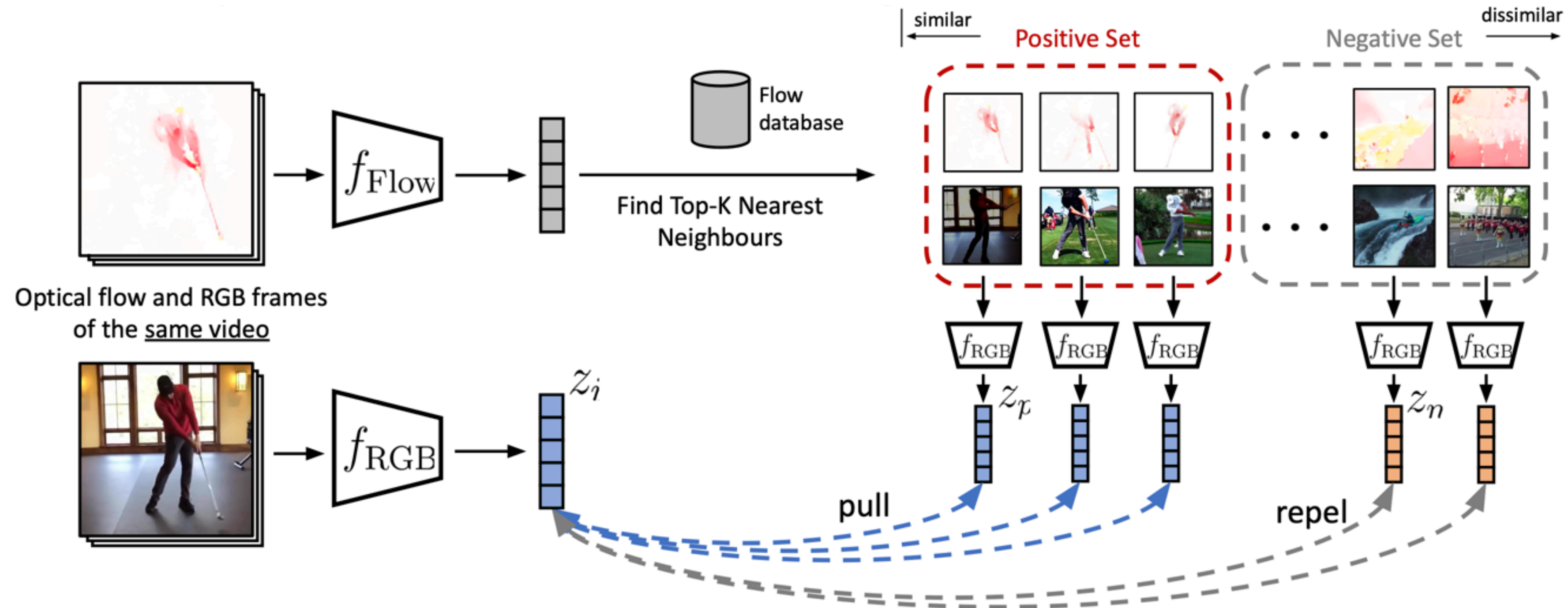
Another work that generalises SimCLR to more modalities: CoCLR

- Improving the sampling mechanism:

- Video involves multiple modalities, e.g. appearance, motion, audio, narrations.
- Dissimilar instances in RGB stream might be **naturally** similar in other modalities.
- Simultaneously co-train two networks, e.g. RGB, Flow.



CoCLR defined positive pairs by finding nearest neighbors in a different modality



Multi-Instance Contrastive Loss (MIL-NCE):

- Features from the positive set are pulled together
- Features NOT from the positive set are pushed apart
- Optical flow helps RGB frames to go beyond instance discrimination

$$\mathcal{L}_{\text{CoCLR}} = -\mathbb{E} \left[\log \frac{\sum_{p \in \mathcal{P}_i} \exp(z_i \cdot z_p)}{\sum_{p \in \mathcal{P}_i} \exp(z_i \cdot z_p) + \sum_{n \in \mathcal{N}_i} \exp(z_i \cdot z_n)} \right]$$

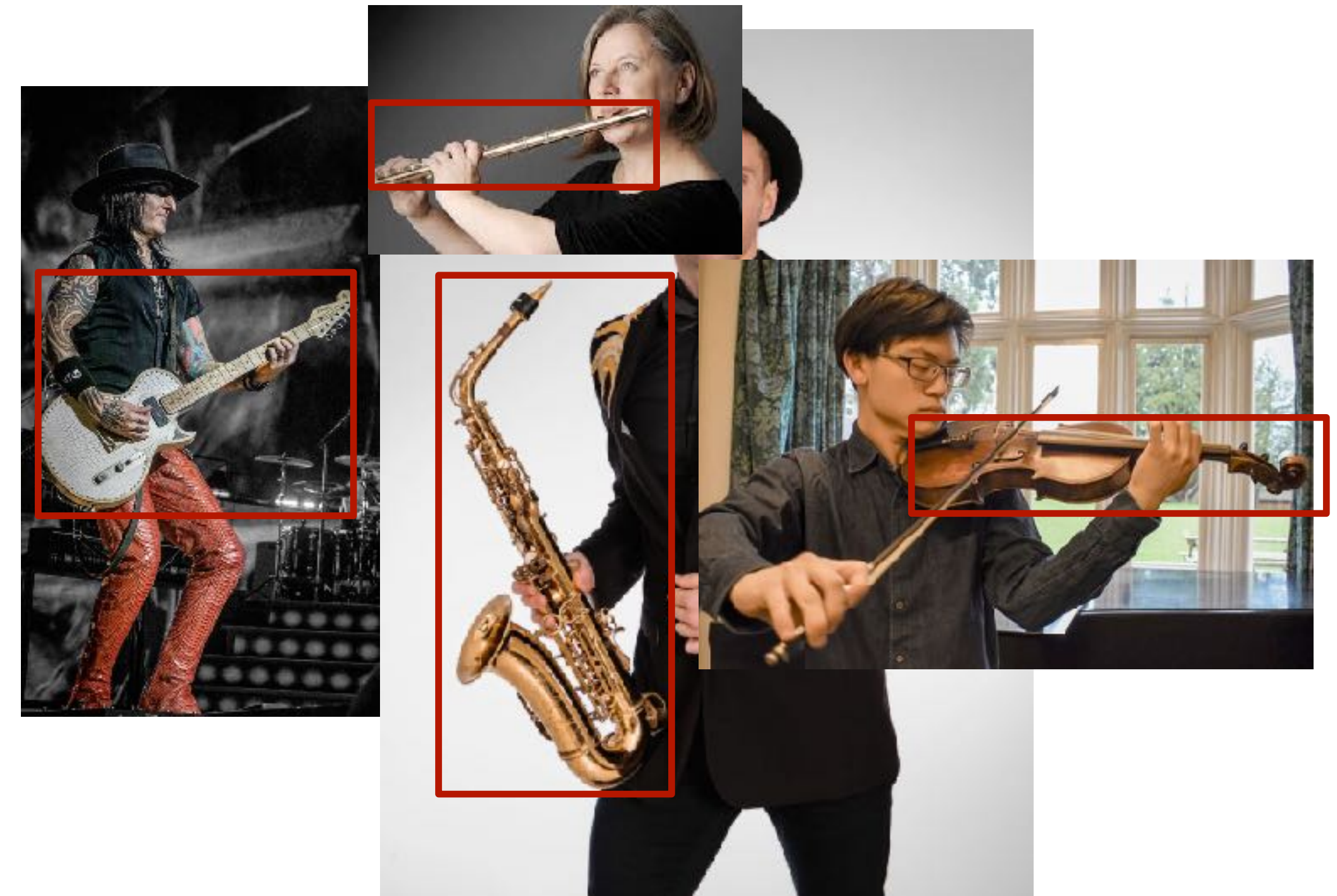
Self-supervised object detection from audio visual correspondence CVPR'22

TRANTAFYLLOS AFOURAS*, YUKI M. ASANO*,
FRANCOIS FAGAN, ANDREA VEDALDI, FLORIAN METZE

Object detection - supervised training

Detector

Supervised
training

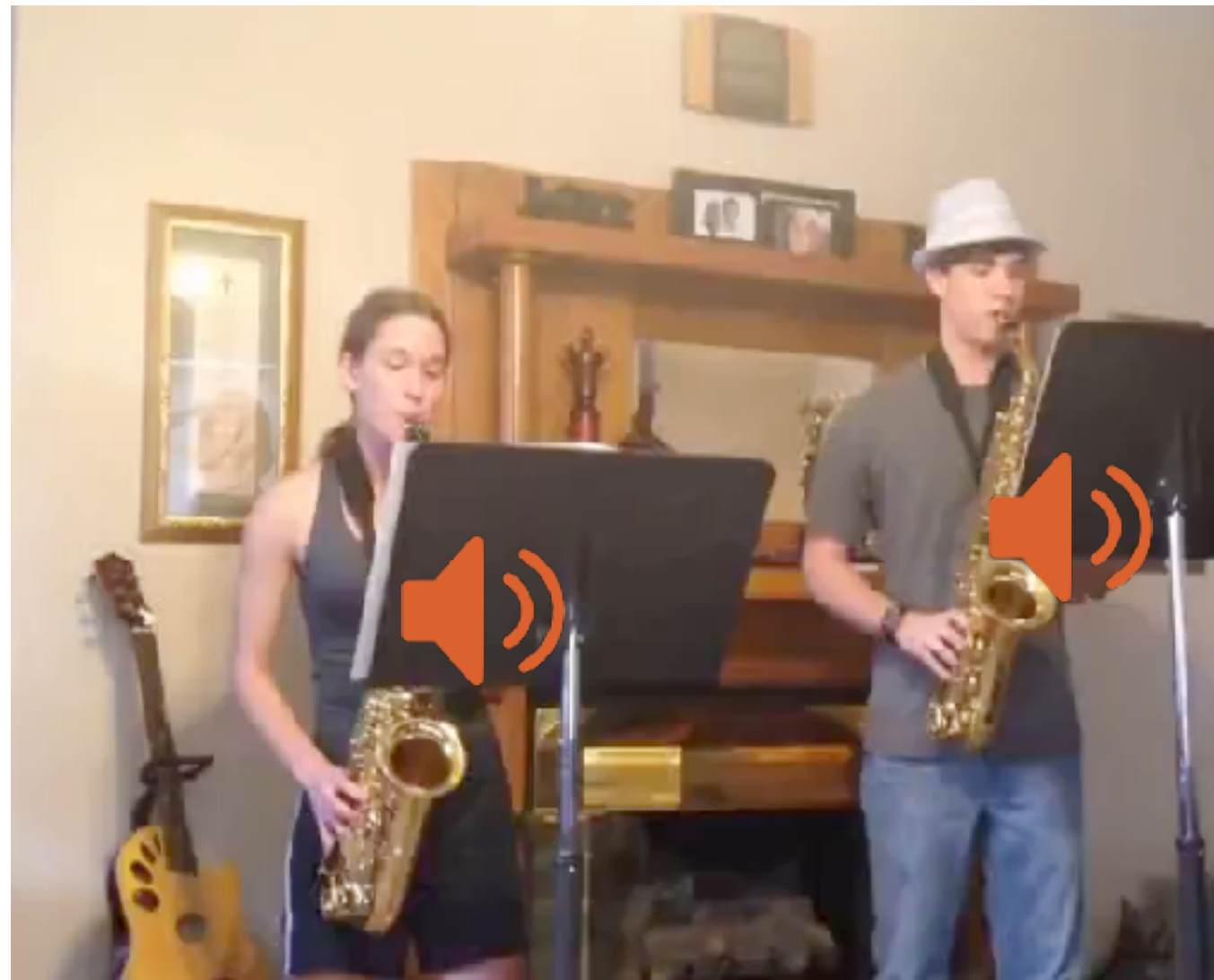


Data annotation expensive

Process hard to generalise

Lots of
annotated samples

What we propose instead:



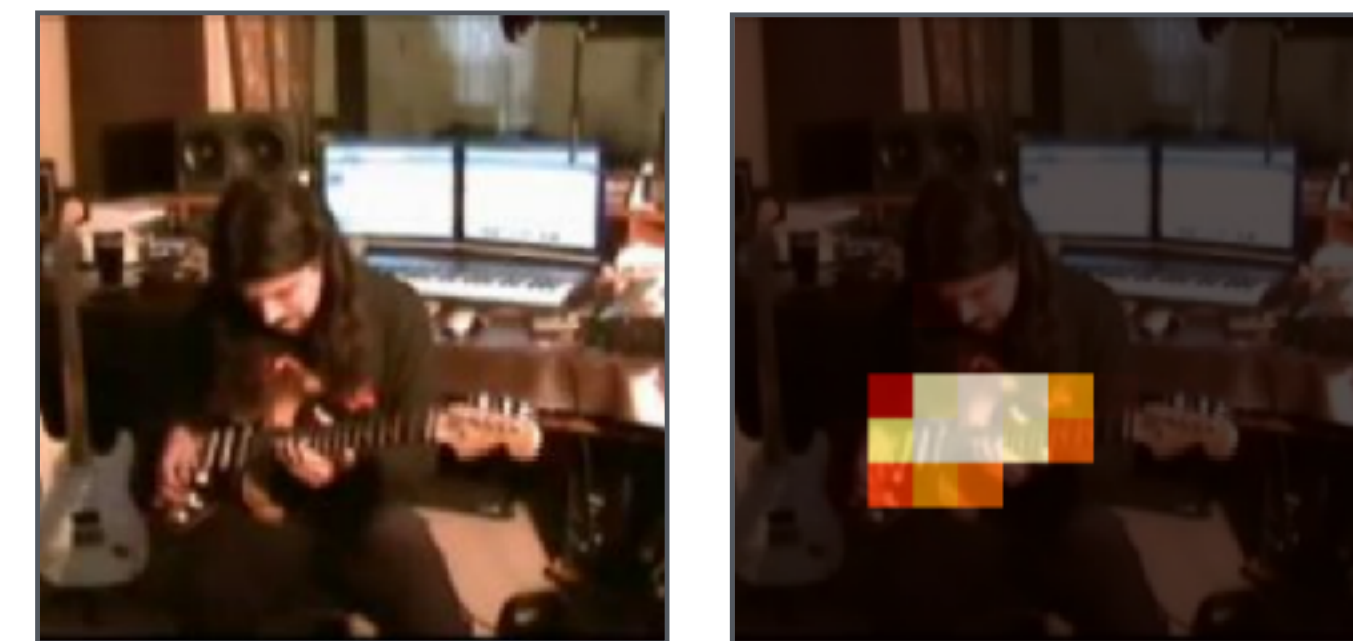
Use sound as supervision for detection

e.g. 1 second window
around frame

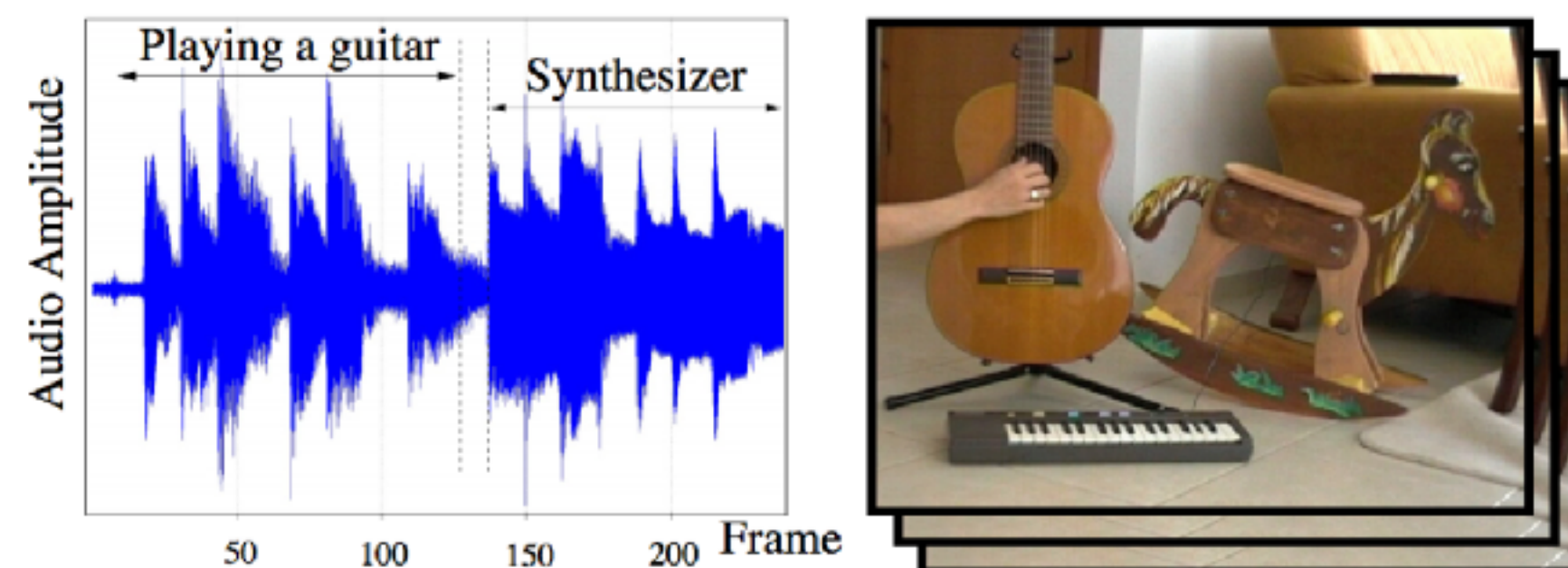
Related work: sound source localisation



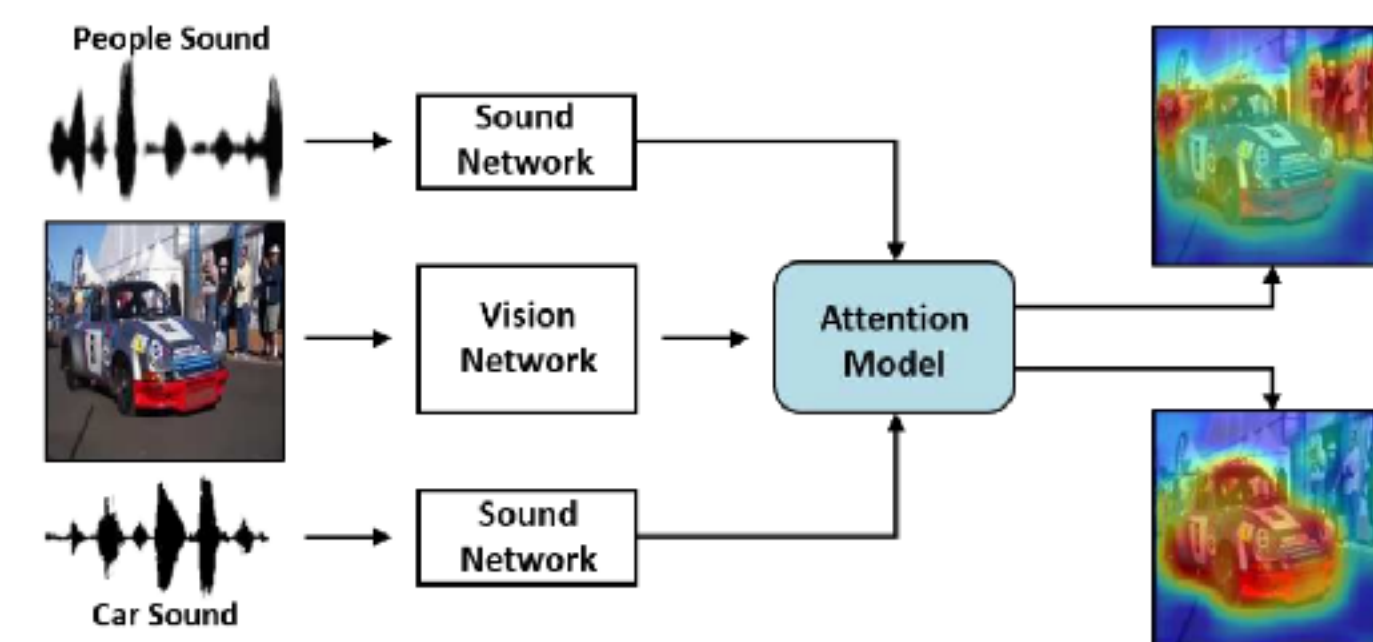
Audio vision: Using audio-visual synchrony to locate sounds. Hershey and Movellan, NeurIPS 2000.



Objects that Sound. Arandjelović and Zisserman, ECCV 2018.



Pixels that Sound. Kidron et al., CVPR 2005.



Learning to Localize Sound Source in Visual Scenes. Senocak et al., CVPR 2018.

Related work: Limitations



Owens et al., *Audio-Visual Scene Analysis with Self-Supervised Multisensory Features*, ECCV 2018

Outputs are heatmaps



Arandjelović et al., *Objects that Sound*, ECCV 2018.

No class labels

Inference requires audio

Goal of this paper:

Combining self-labelling and multi-modal learning to detect multiple objects

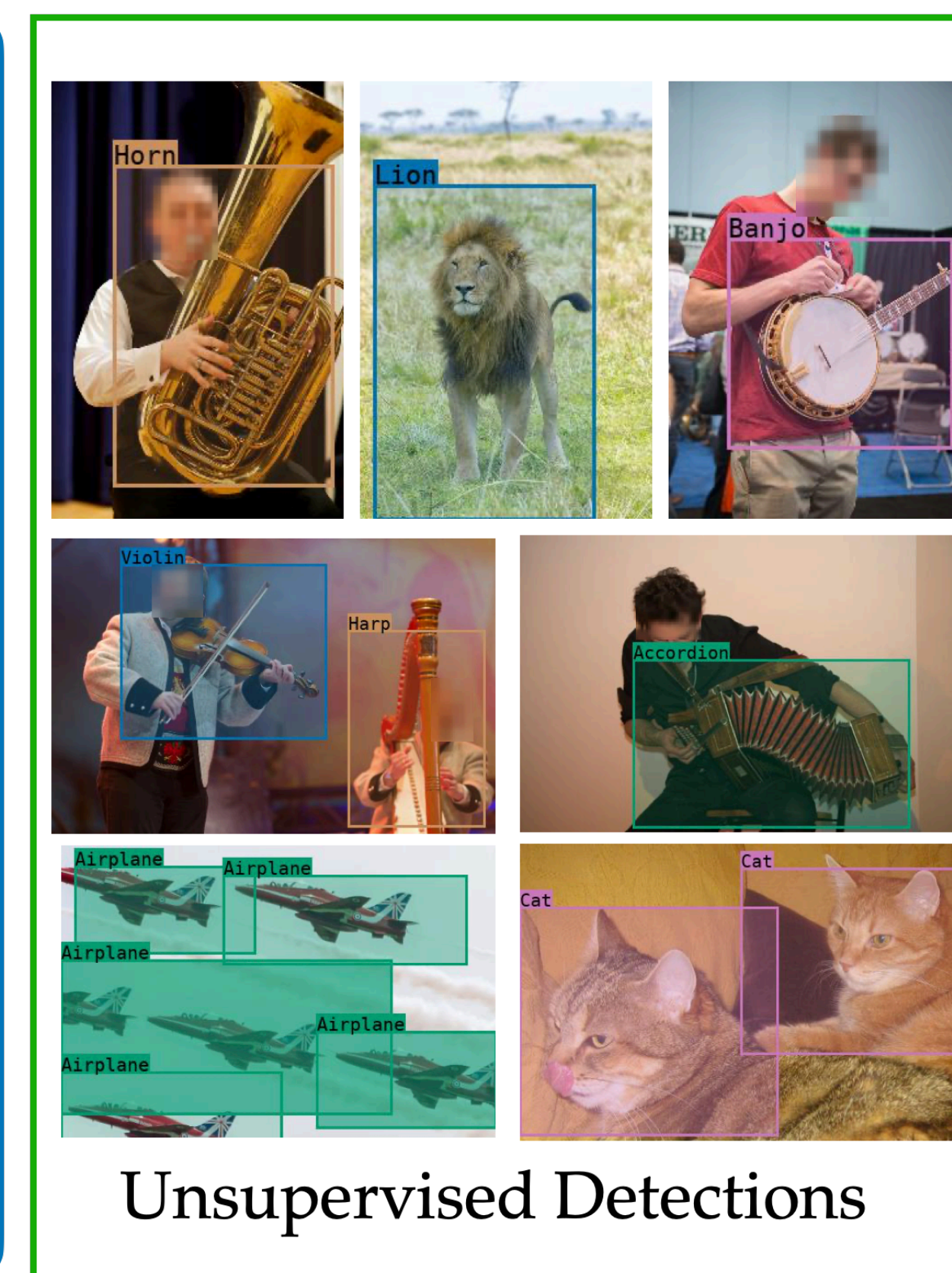
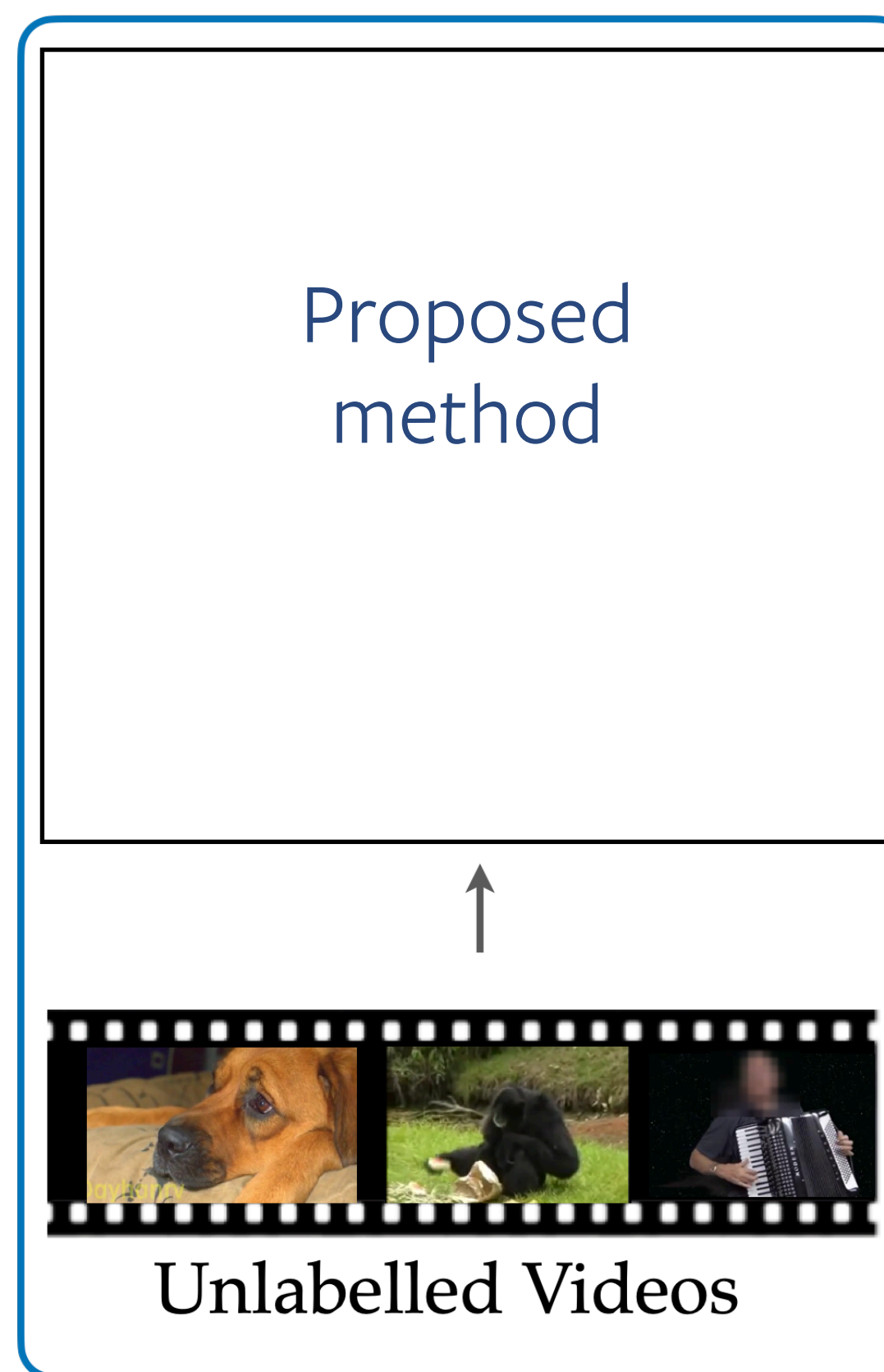
Training object detectors without labels:

- ✓ Use only free audio as “supervision”
- ✓ Output bounding boxes & class labels — not just heatmaps
- ✓ No audio required during inference
- ✓ Retrieve all visible instances, not just the actively sounding ones

Multi-modal Training

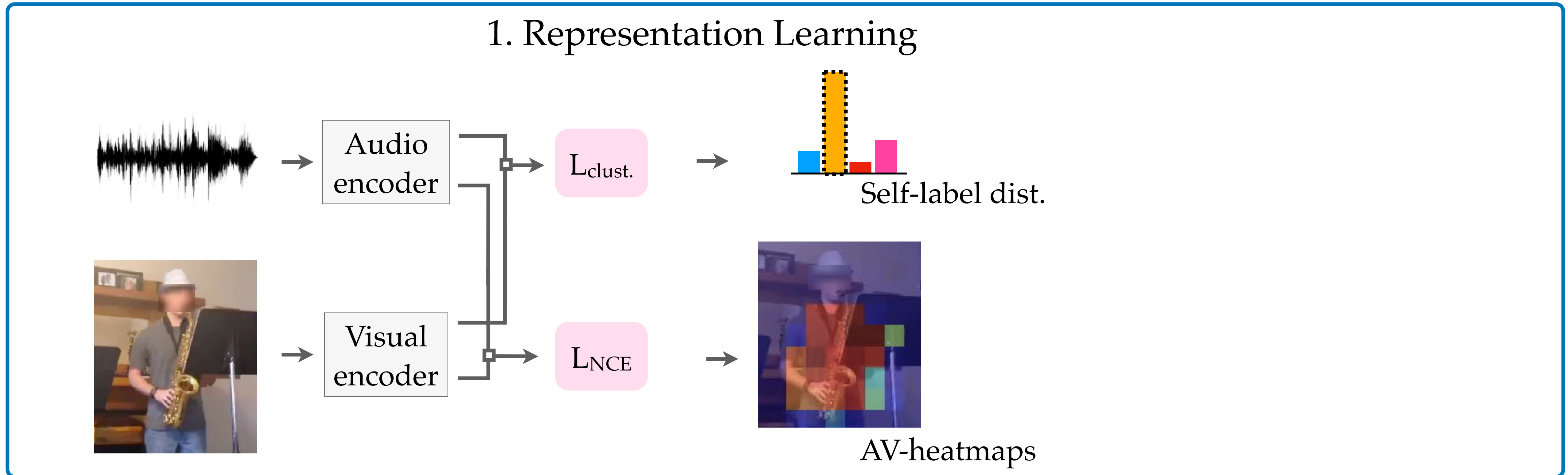


Inference on Images

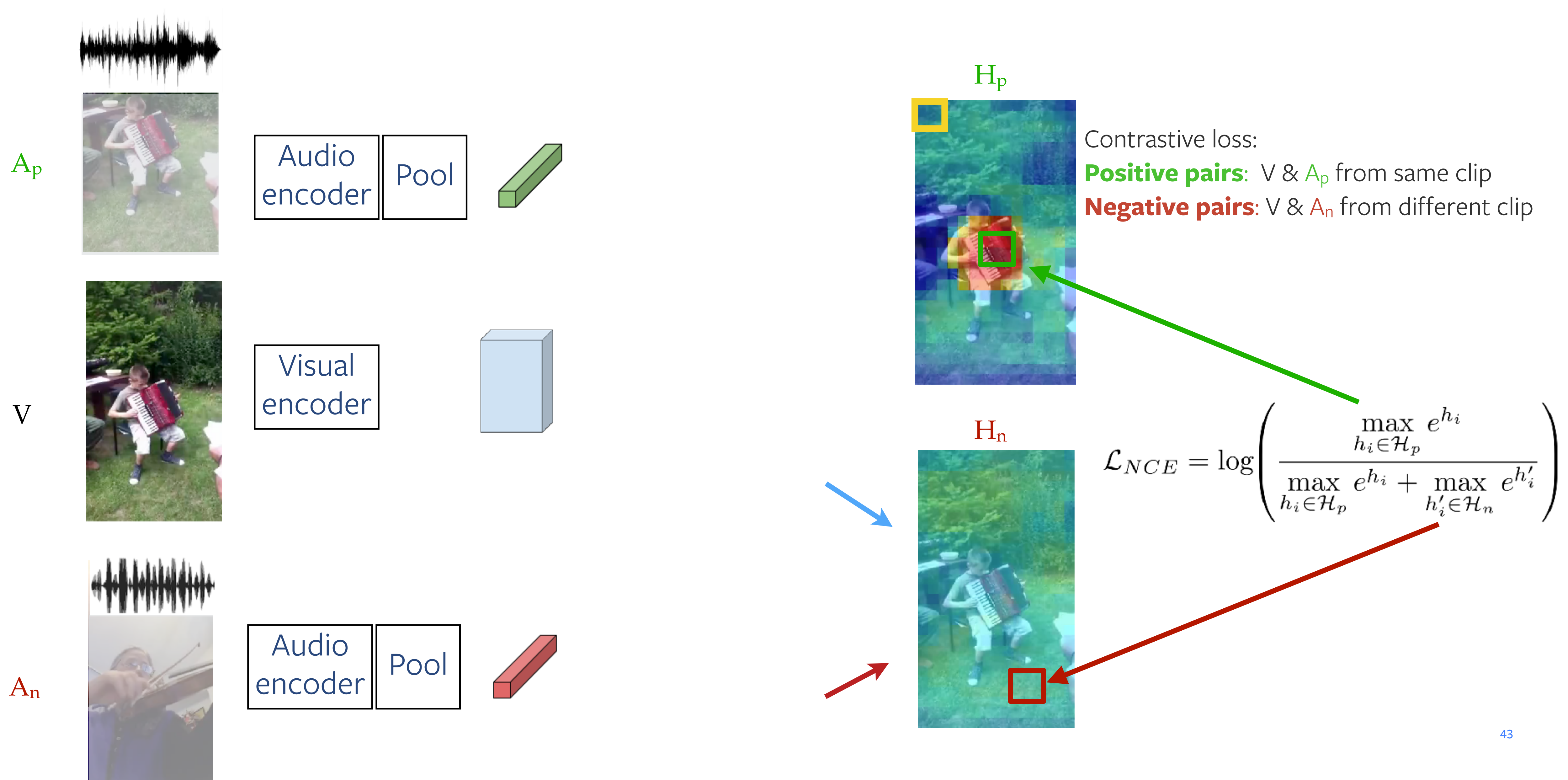


“Useful”
Self-supervised Learning

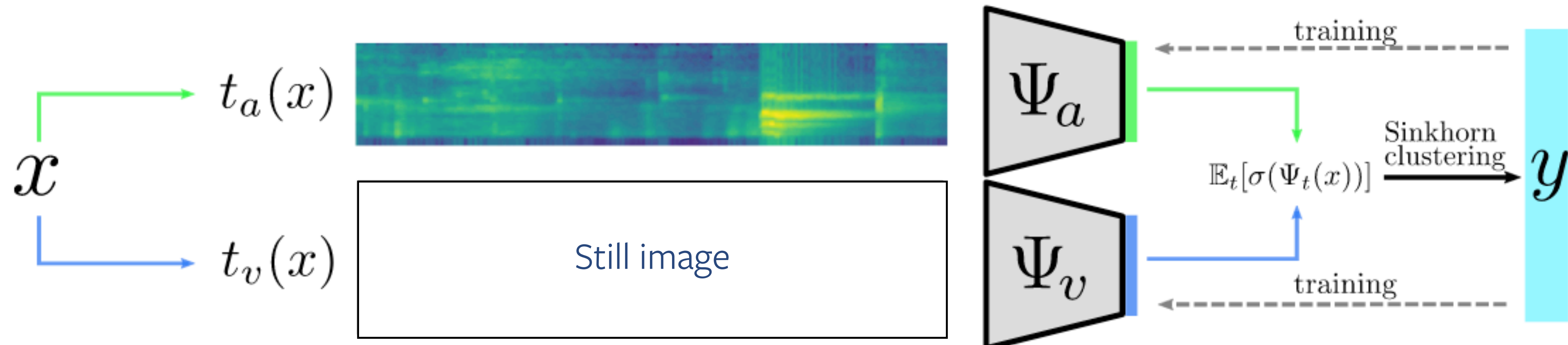
Framework overview



Ingredient 1: training heat maps



Ingredient 2: Self-labels training ($\mathcal{L}_{\text{clust.}}$)

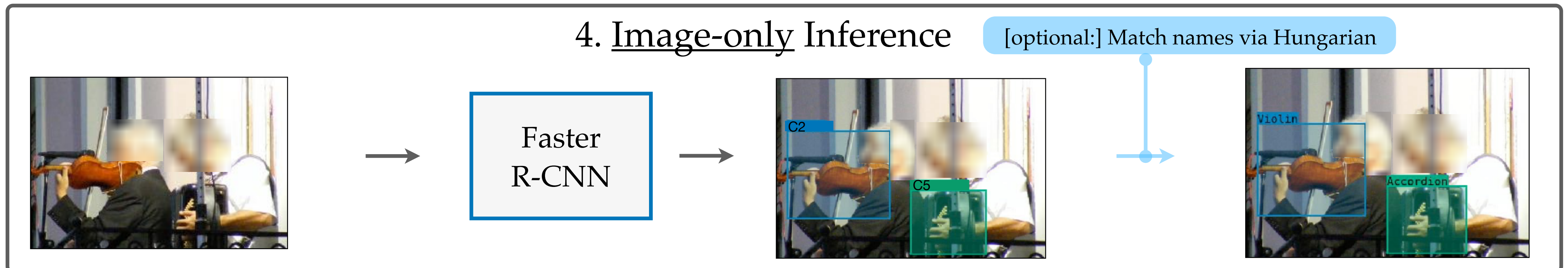
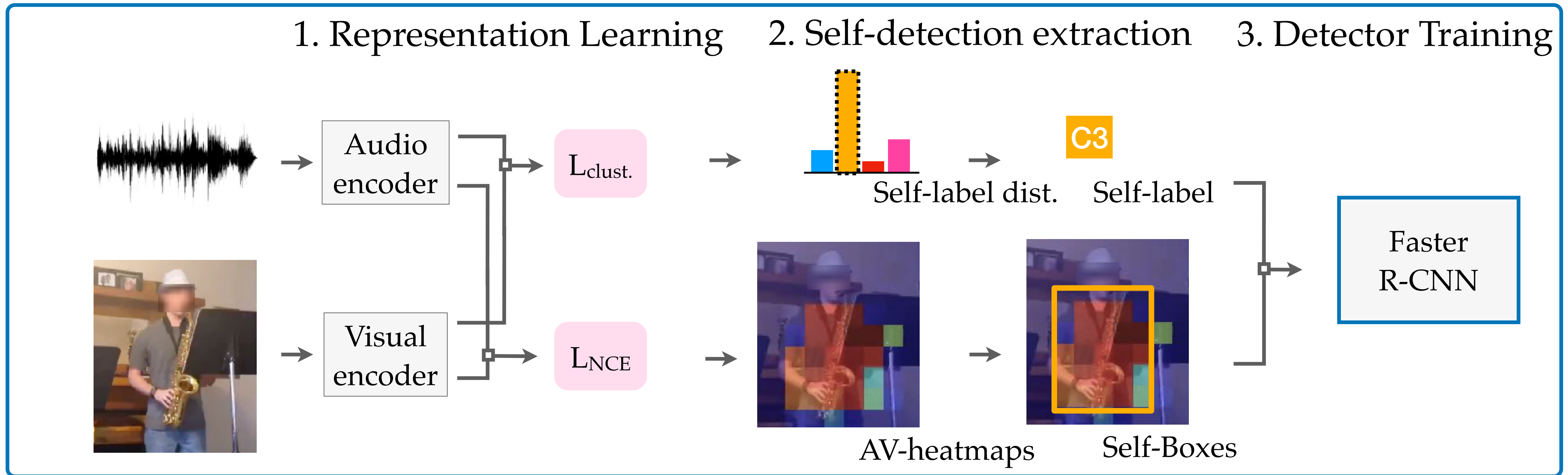


$$\mathcal{L}_v(\mathcal{B}|y) = -\frac{1}{|\mathcal{B}|} \sum_{(v,a) \in \mathcal{B}} \log \text{softmax}(y(v, a) | \Psi_v(v))$$

$$\mathcal{L}_a(\mathcal{B}|y) = -\frac{1}{|\mathcal{B}|} \sum_{(v,a) \in \mathcal{B}} \log \text{softmax}(y(v, a) | \Psi_a(a))$$

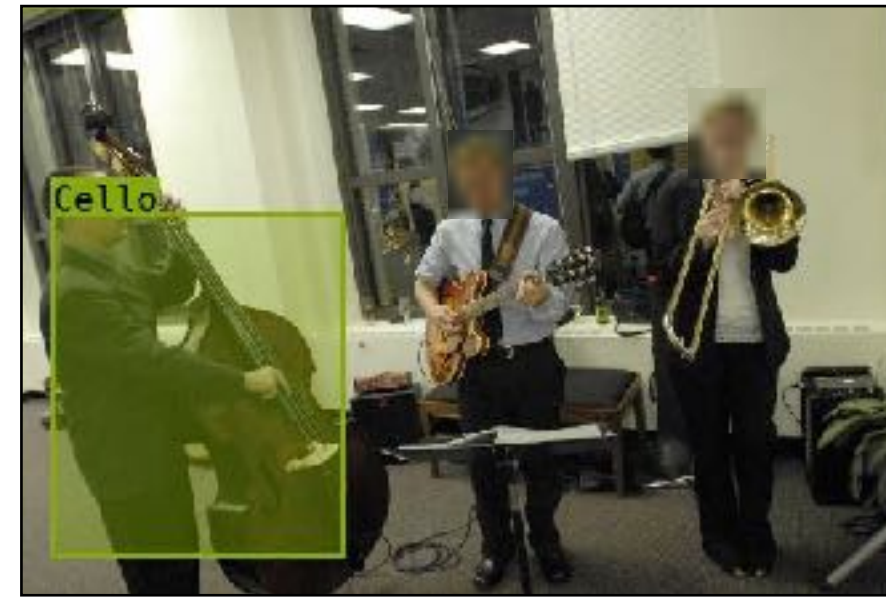
$$\mathcal{L}_{\text{clust}}(\mathcal{B}|y) = (\mathcal{L}_v(\mathcal{B}|y) + \mathcal{L}_a(\mathcal{B}|y))/2$$

Framework overview

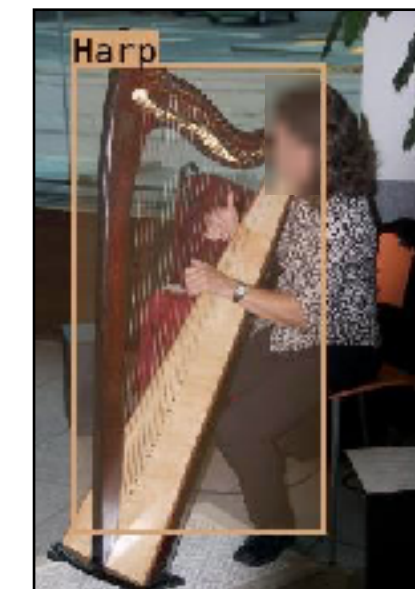
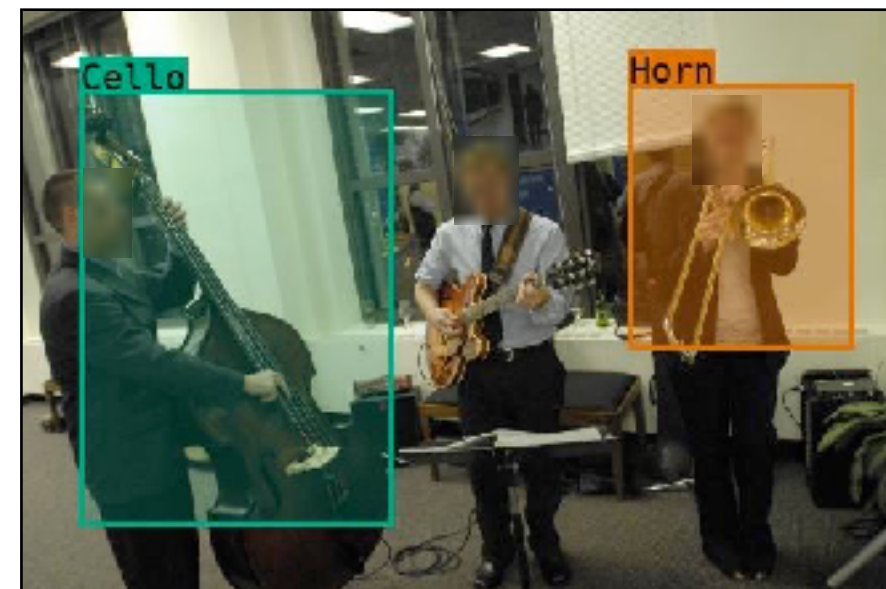
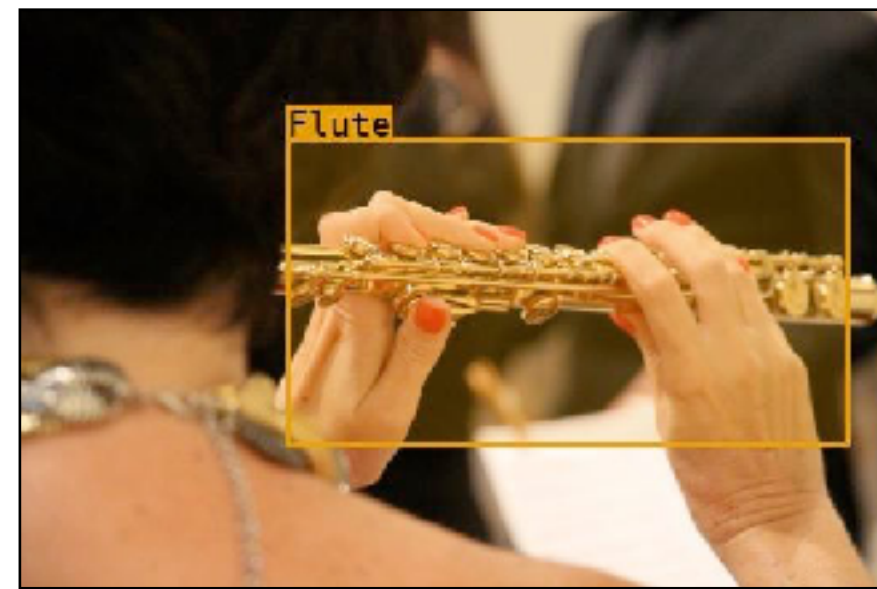
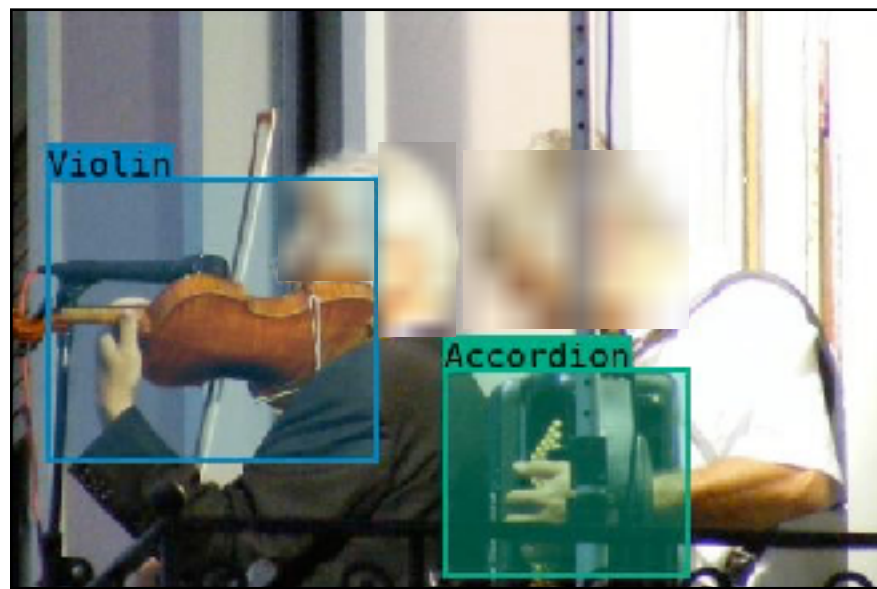


Qualitative results compared to weakly-supervised

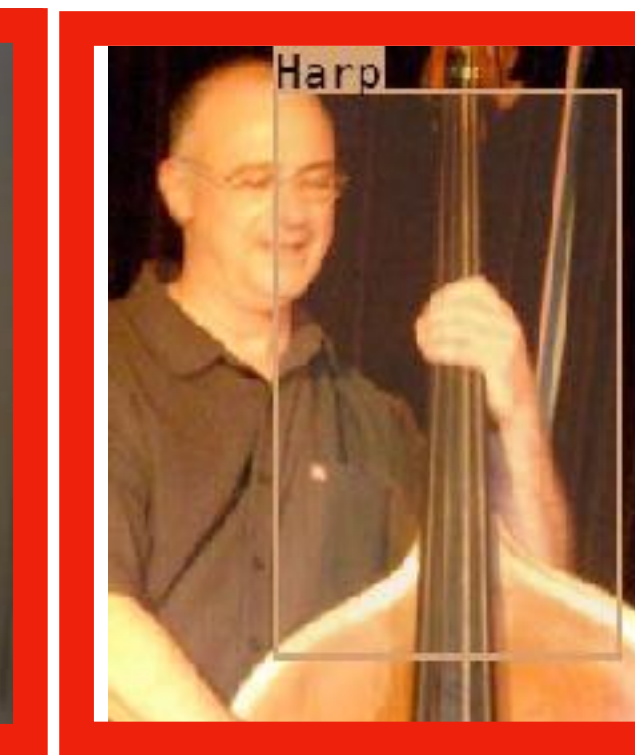
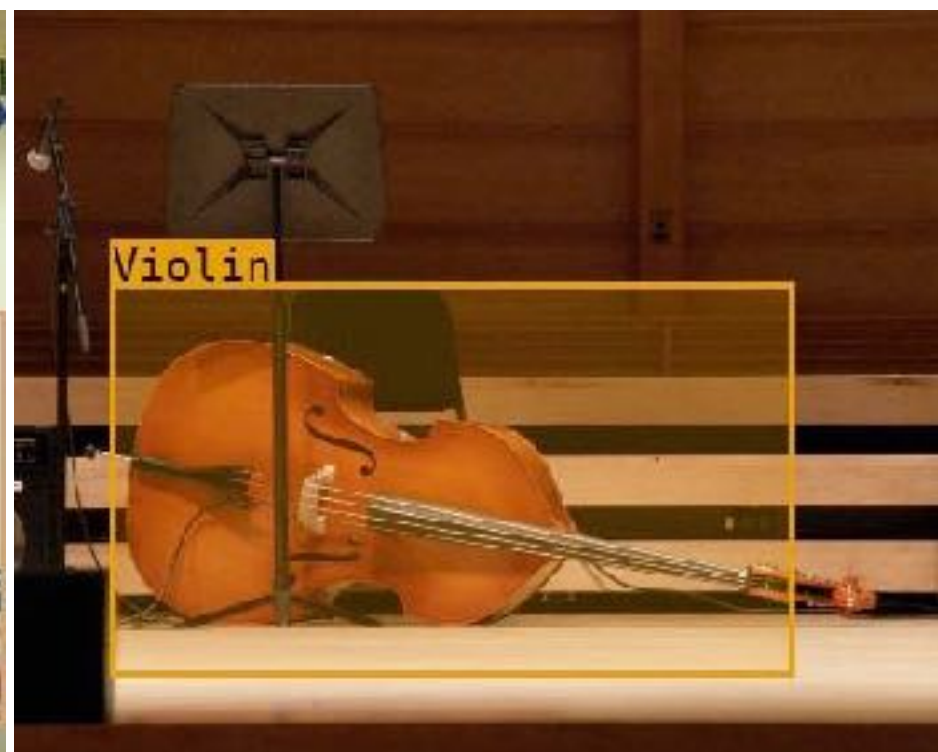
PCL



Ours



Detection examples and failure cases



What about more general objects, beyond instruments?

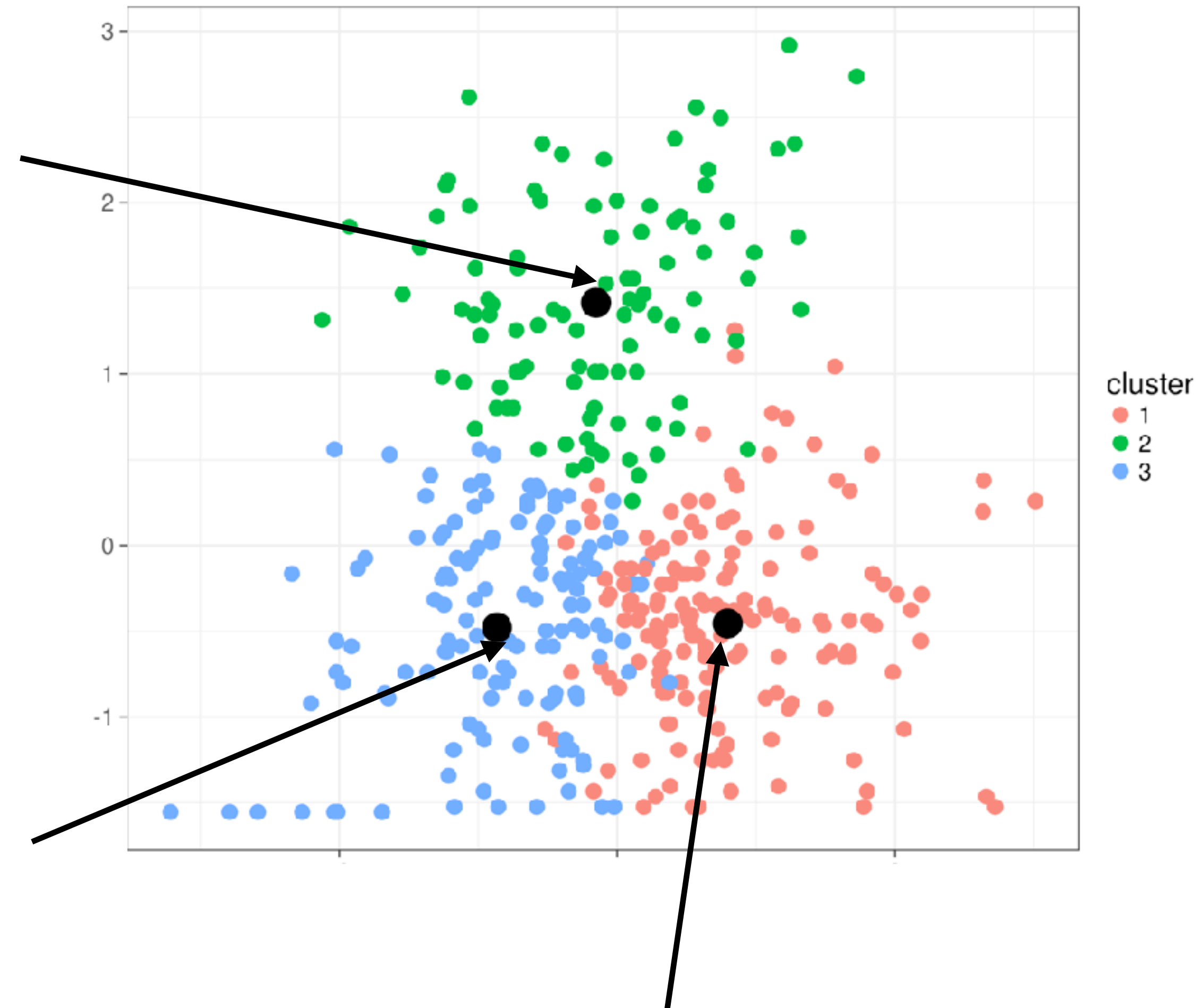
We train on all ~300 VGGSound classes, learning 300 clusters.

- Same method — no changes at all
- Only match class labels *after* the detector is trained
- Match class labels with as few as 1 sample per cluster (ie 300 “labels”)

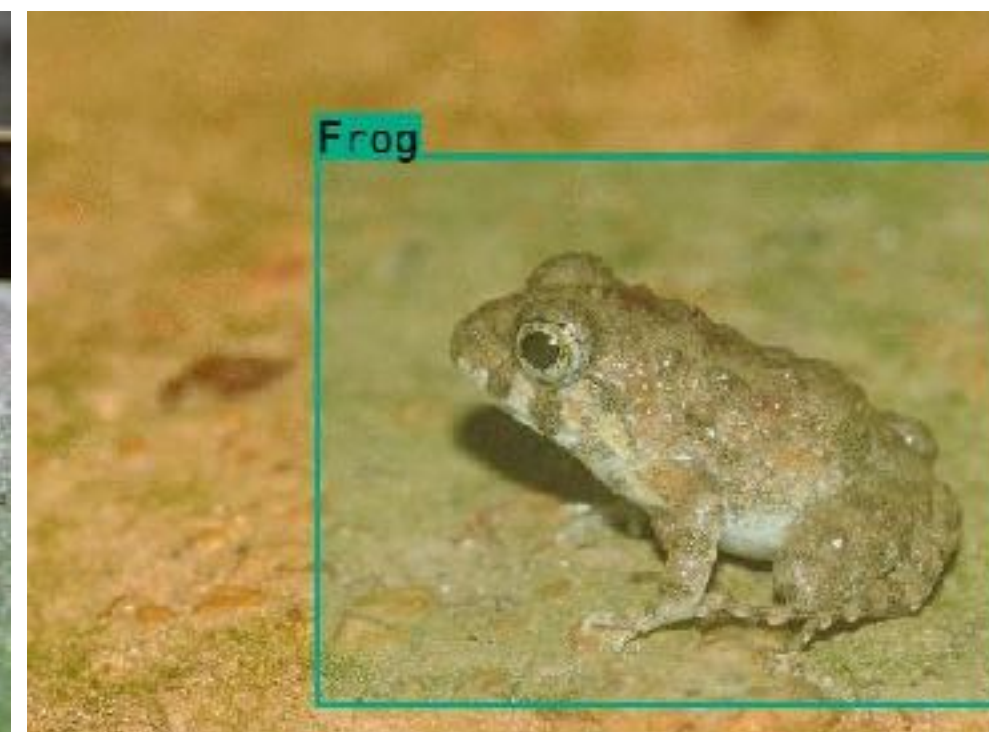
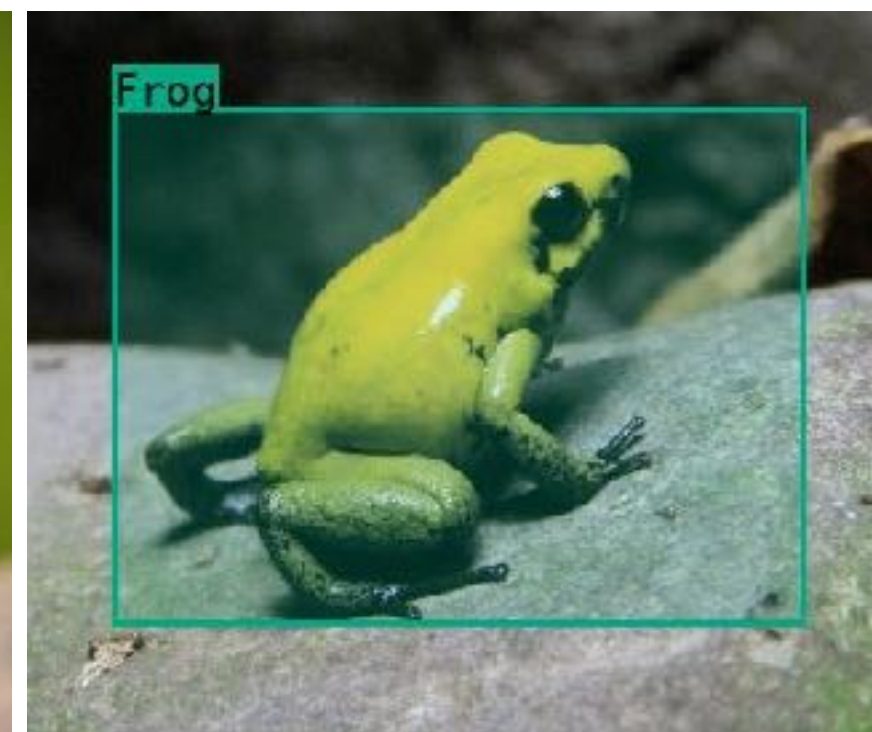
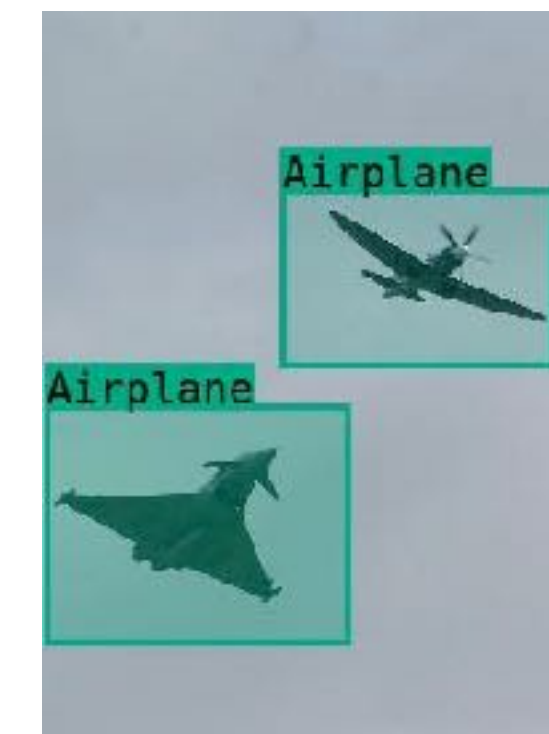
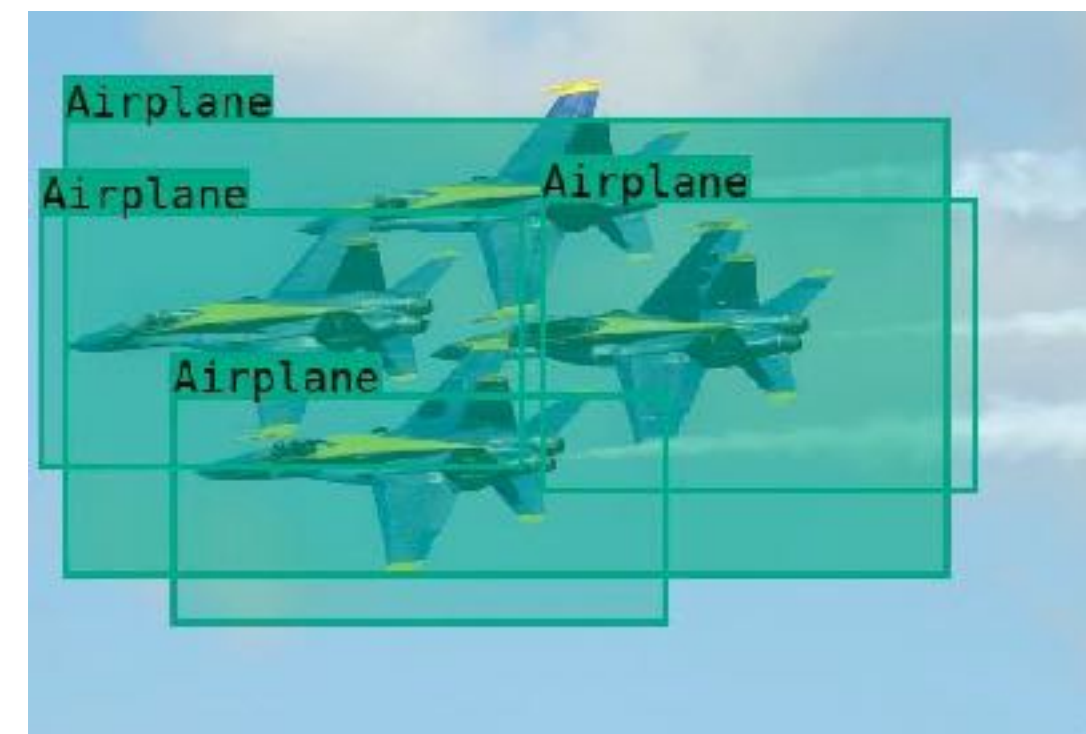
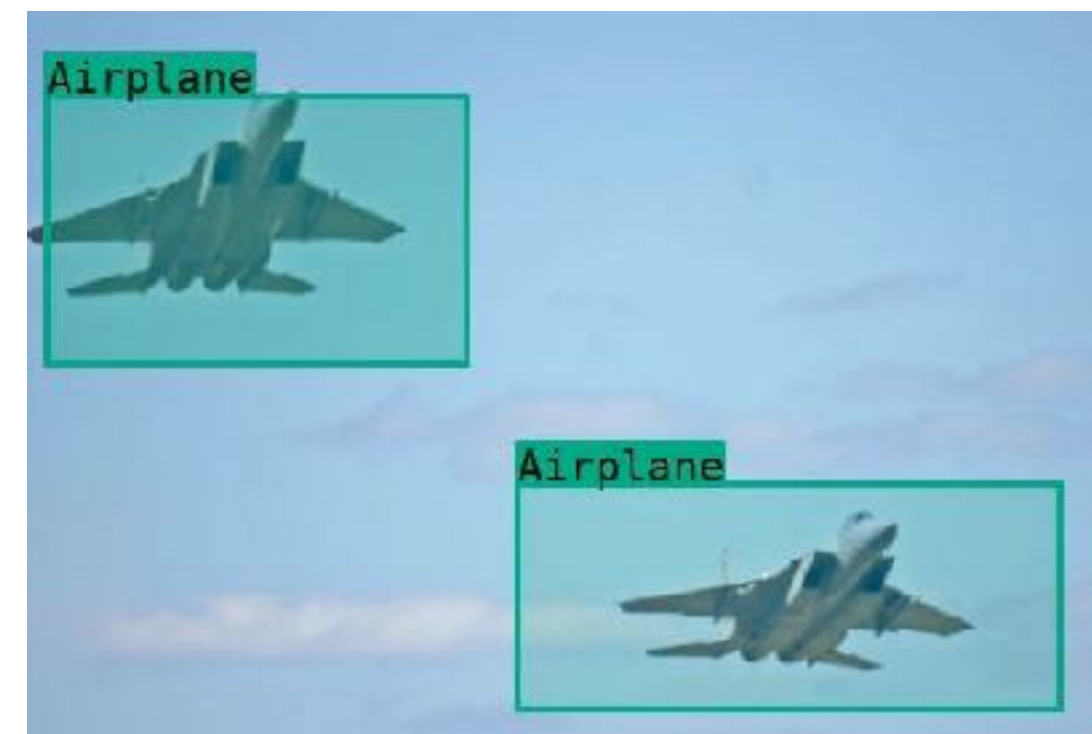
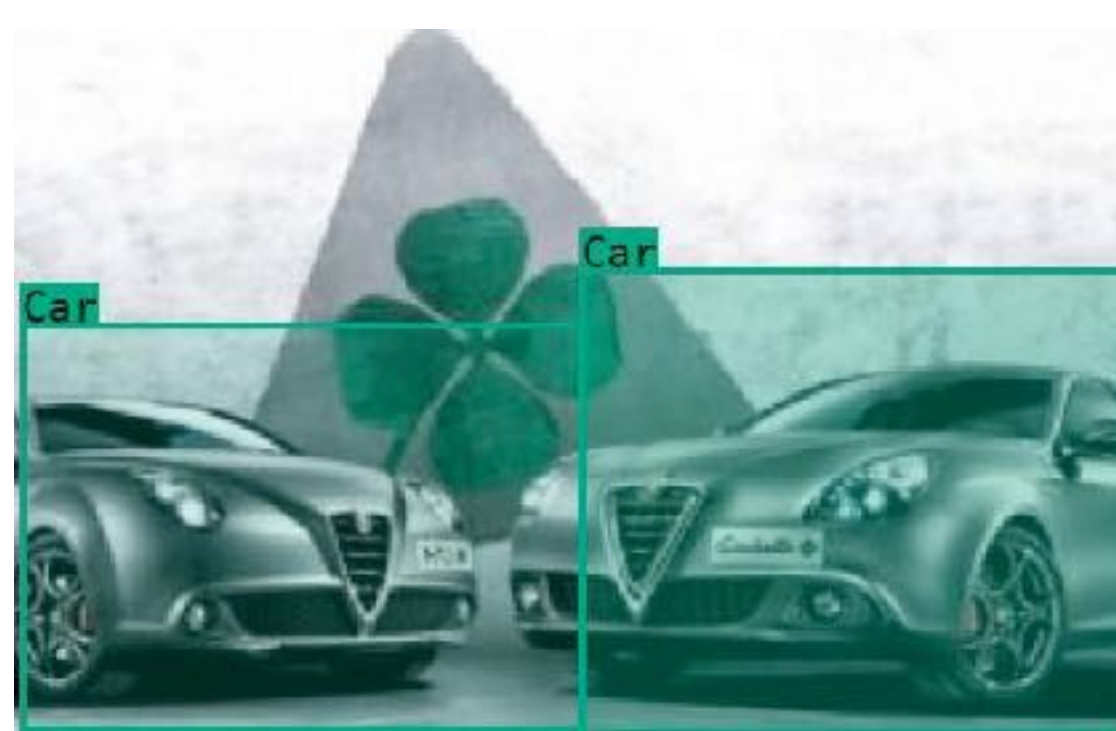
Prototype efficient matching: obtain labels of samples closest to centroids

Matching	mAP ₅₀	
	VGGS	O.Images
Hung.	39.4	28.5
Argmax	39.6	30.1
Manual	41.0	29.5
1-shot	36.4	25.1
10-shot	37.1	25.8

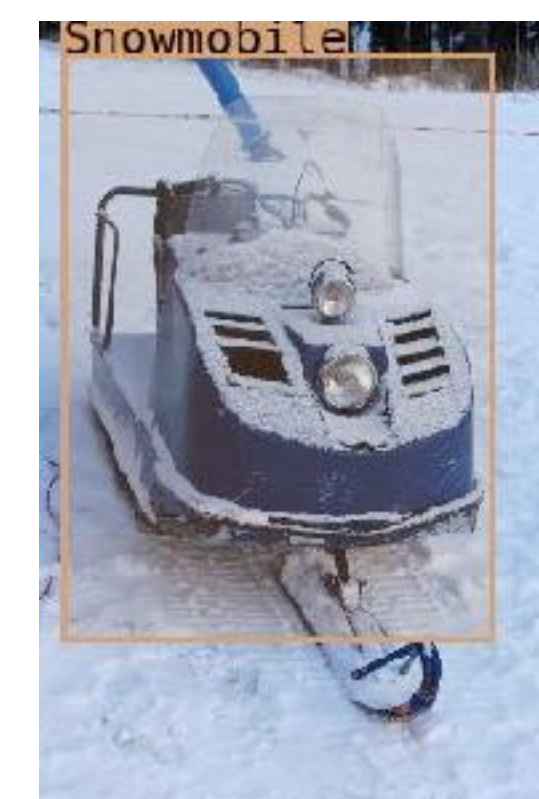
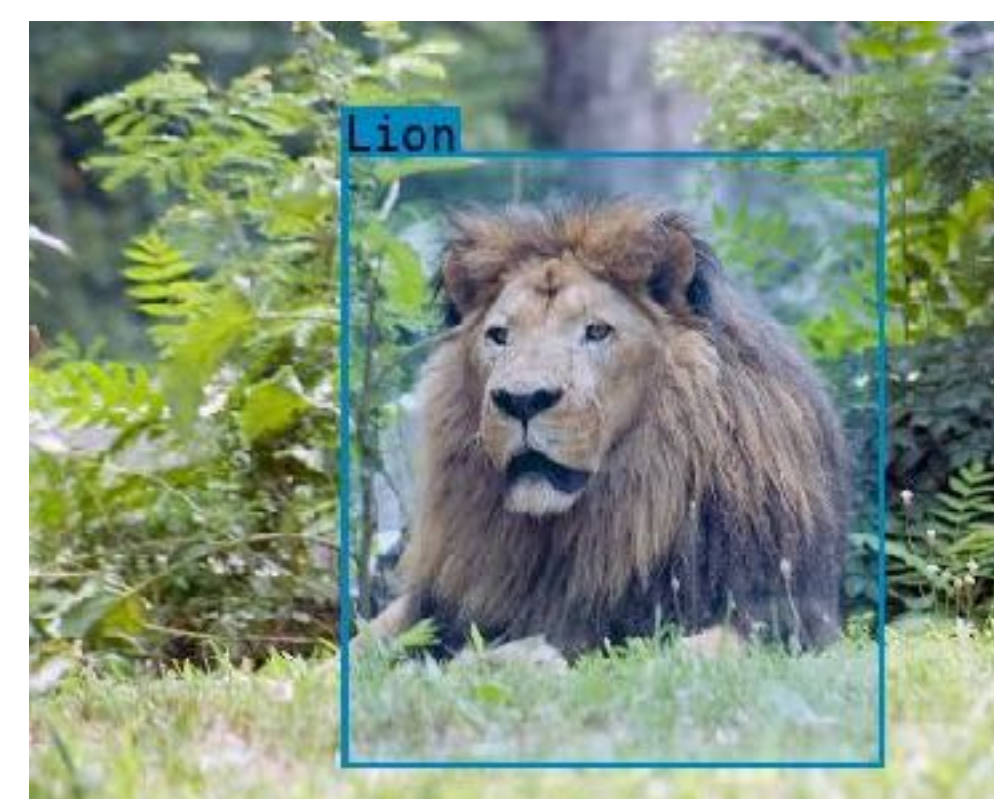
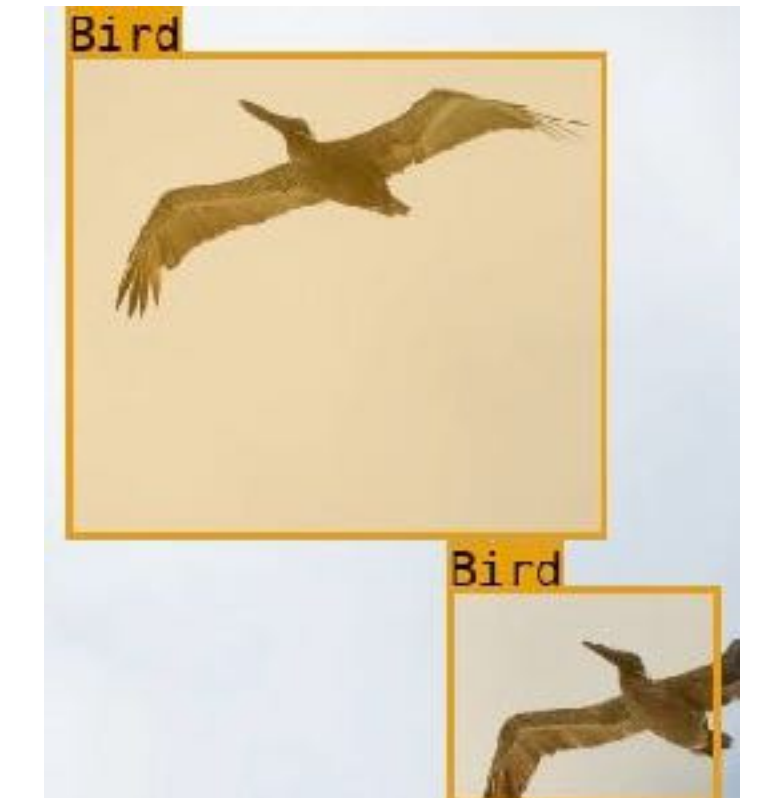
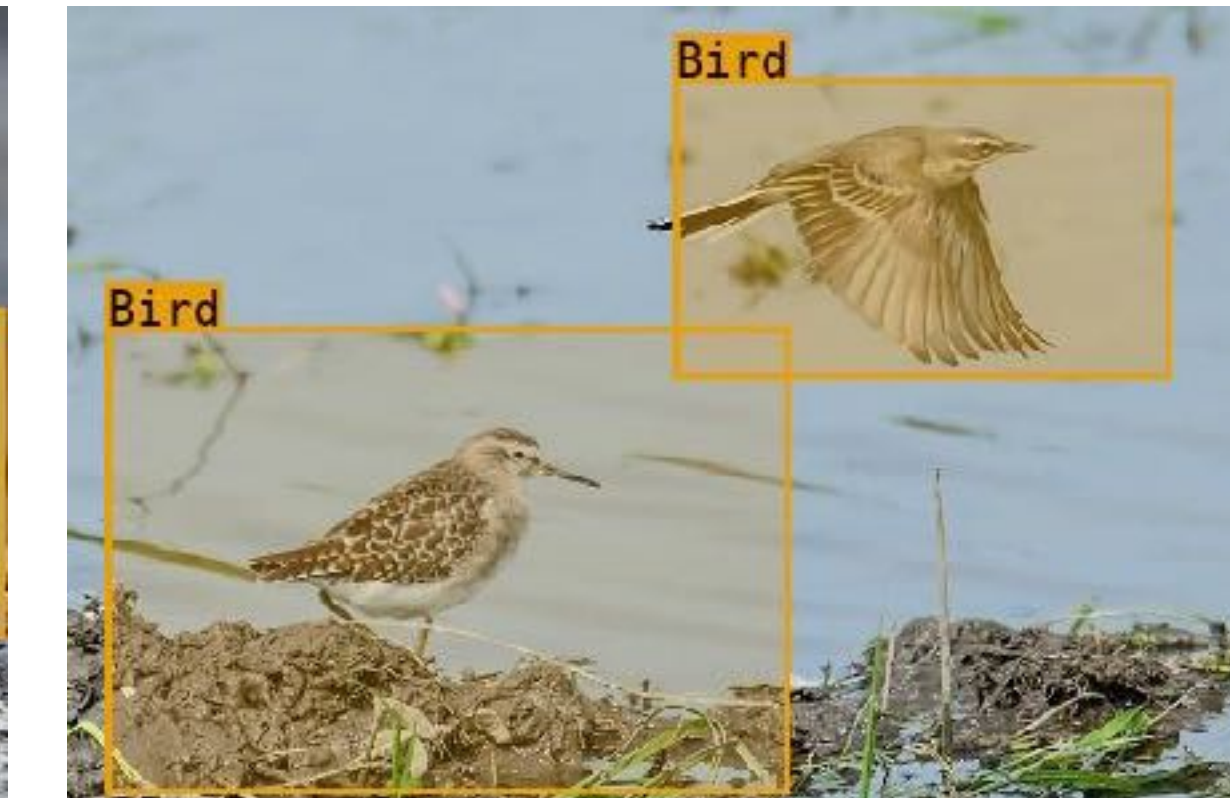
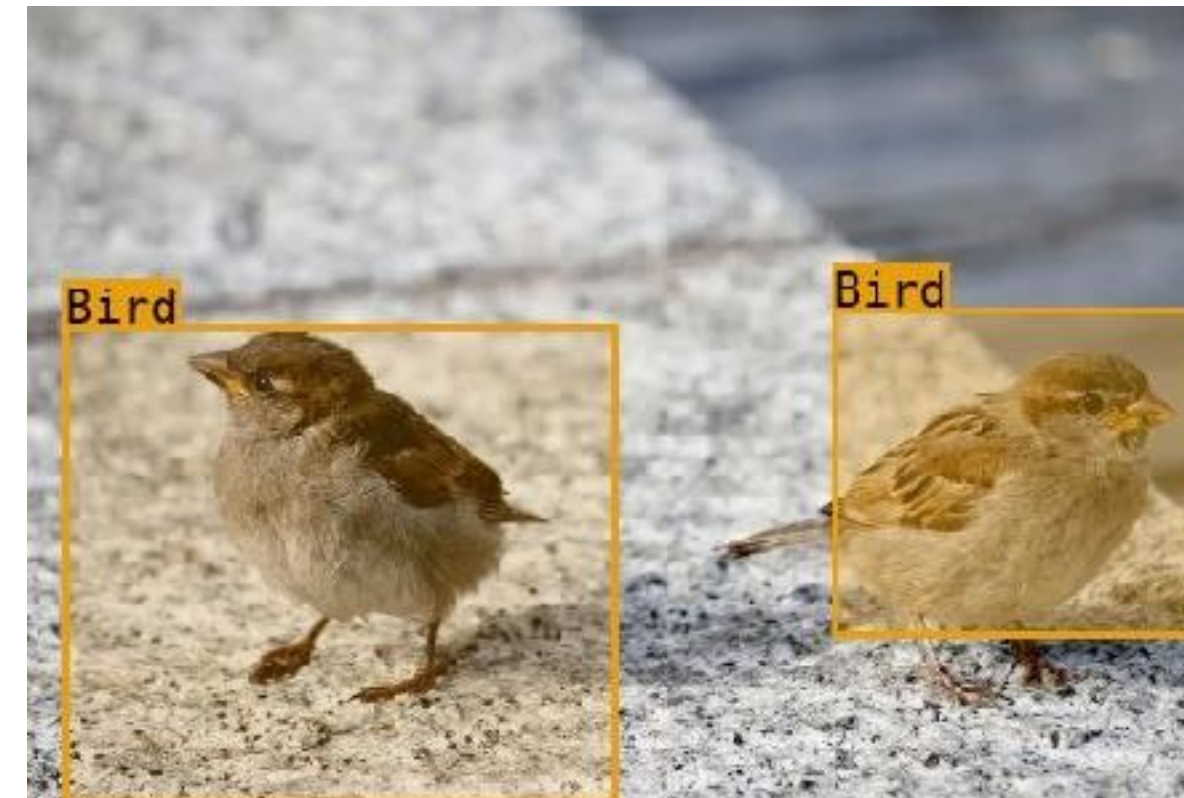
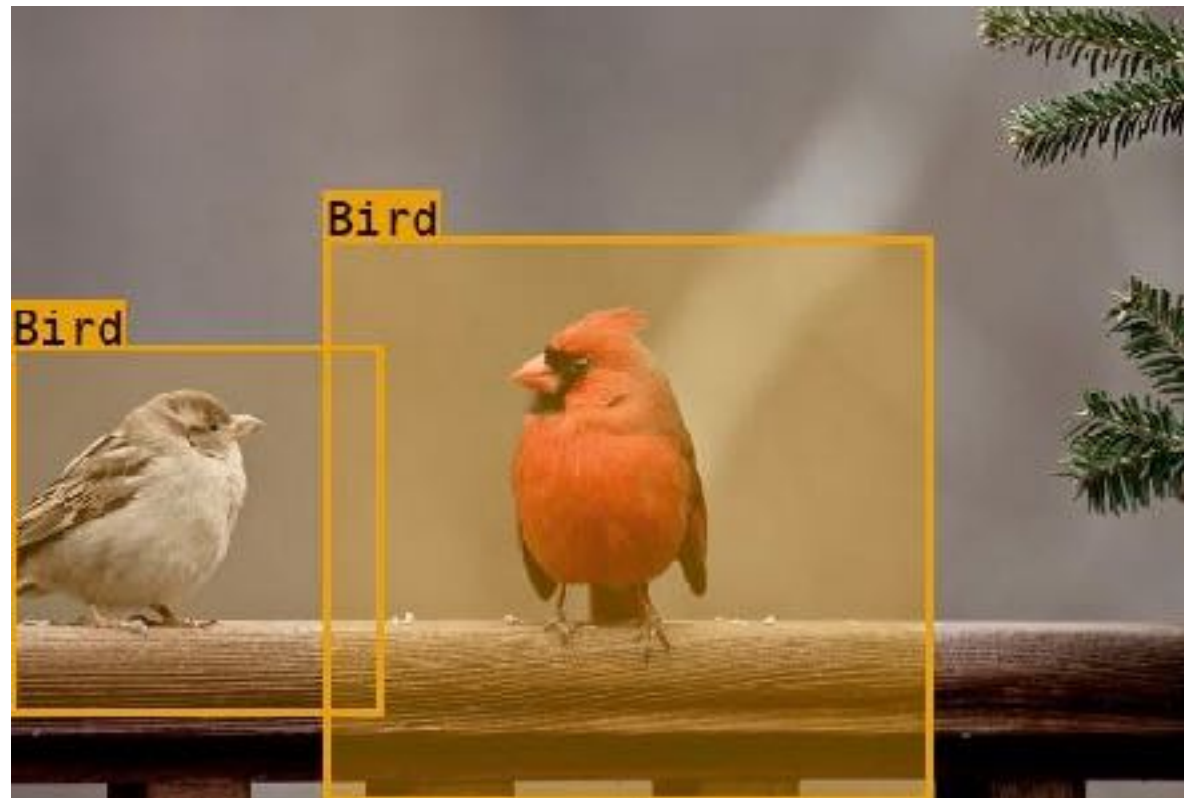
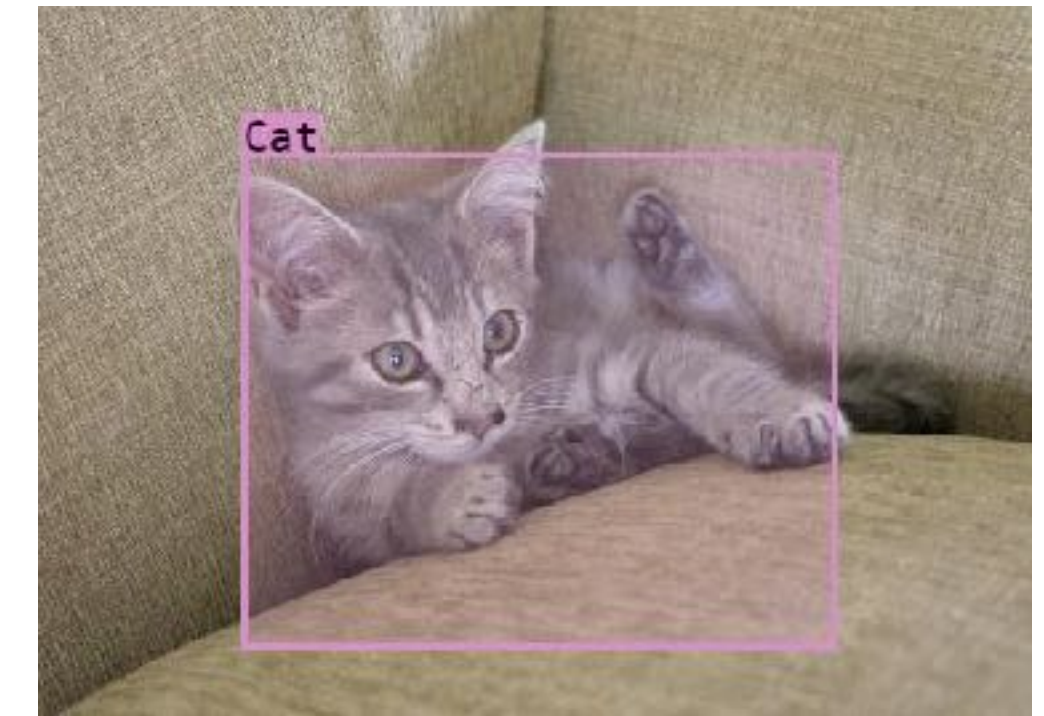
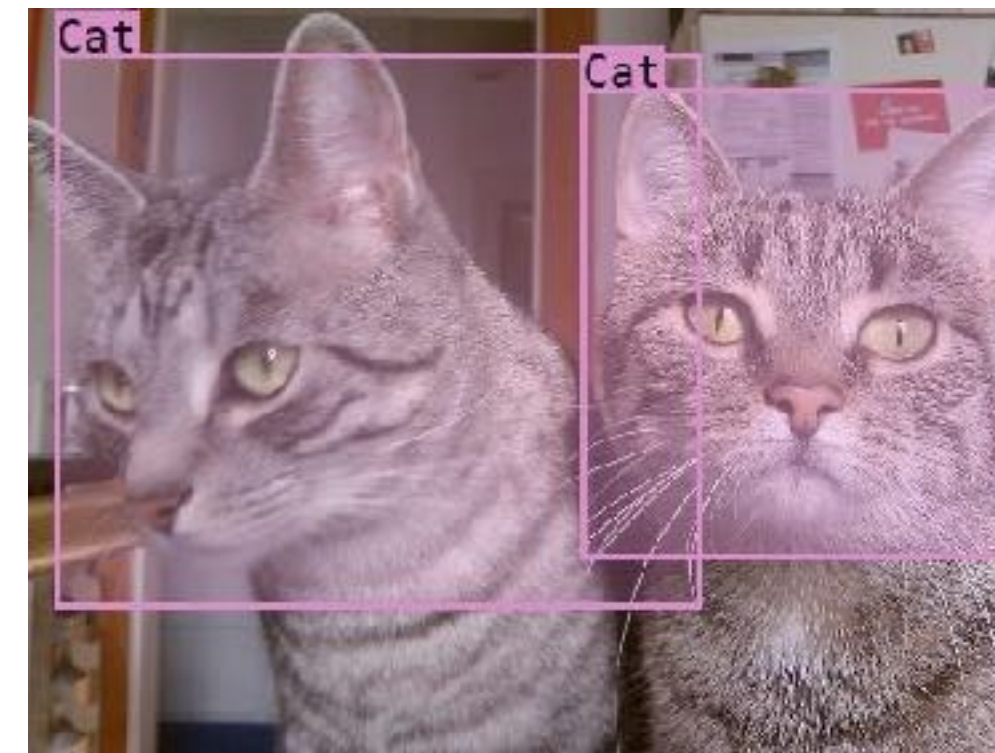
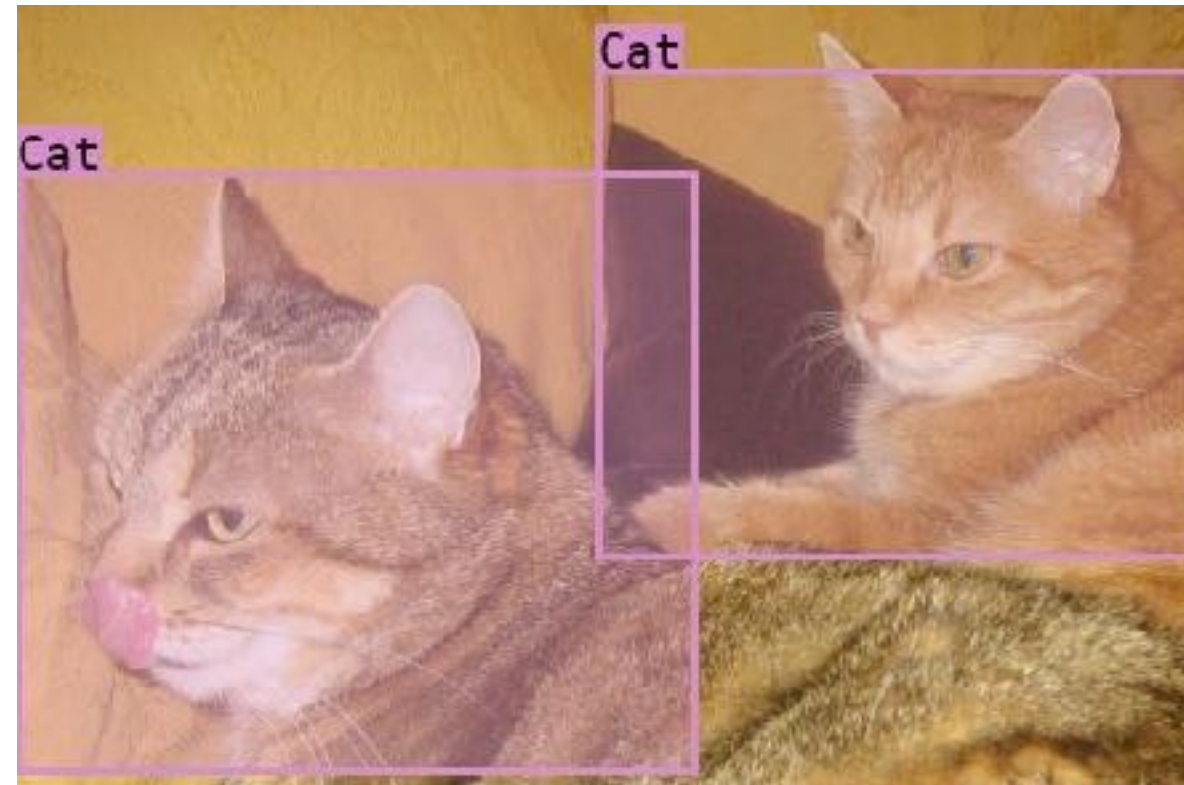
Table 6: **Matching strategies.** Even with as little as 39 labels, our method can detect and classify objects accurately.



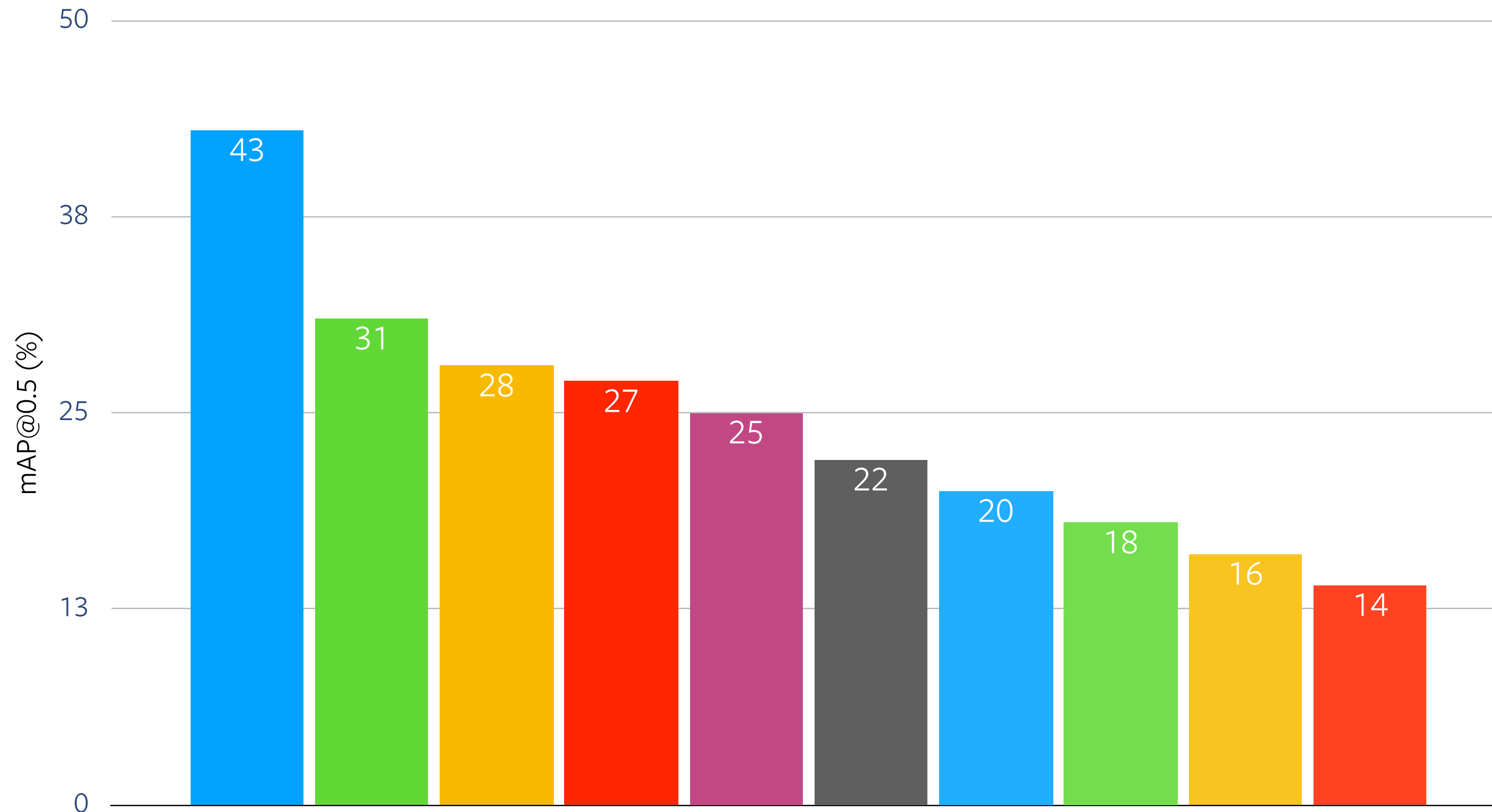
VGGSound object detections



VGGSound object detections



Per class performances



- Still long way to go for most objects
- mAP @0.5 is a hard measure

How research gets done: part 10



Previous parts:

[fundamental understanding/read papers, how-to-read-papers, implement & tinker with code, realise and seek *funny* moments, MVP/principles/benchmarks/baselines, when to (not) give up/impact-vs-work, importance of Ablations, storyline]

- So you have made great research and submitted your paper to some venue
- Congratulations. Please celebrate this.
 - Acceptance of a paper is sometimes stochastic, but finishing a piece of work needs to be celebrated on its own
- Take a break (it will allow you to see things from a new perspective)
- Alongside an ML paper, we
 - Publish the code on github, please follow reproducibility guidelines: <https://github.com/paperswithcode/releasing-research-code>
 - Make a website for a paper:
examples: <https://richzhang.github.io/colorization/> <https://single-image-distill.github.io/> <https://www.di.ens.fr/willow/research/mil-nce/> <https://www.matthewtancik.com/nerf>
 - Sometimes make a twitter thread about it, or write a layperson-blogpost about it
- All in order to increase the accessibility and reach of our research, because
 - The field moves so quickly it's hard to keep track but,
 - Research should be accessible, available and understandable

Multi-modal research now-a-days



MMV

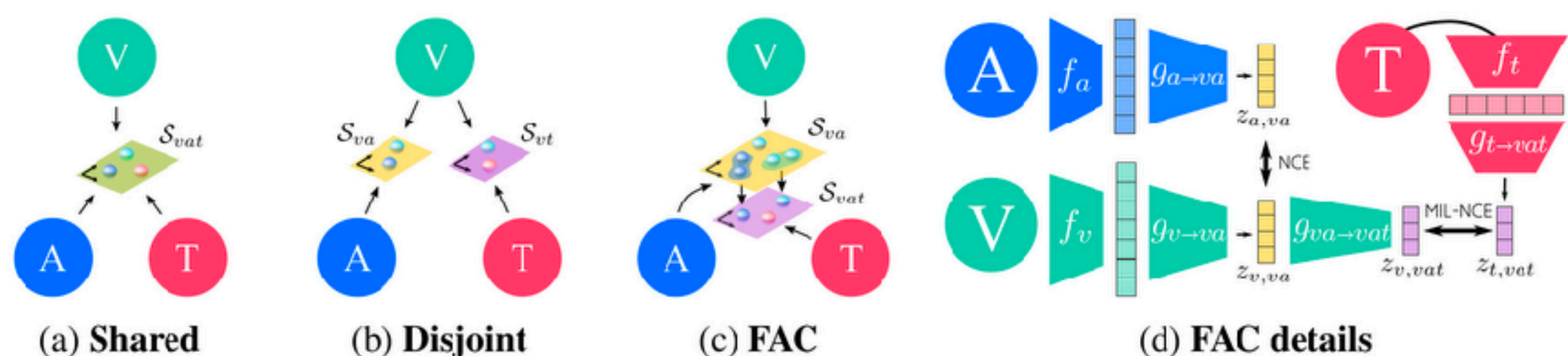


Figure 1: (a)-(c) Modality Embedding Graphs, (d) Projection heads and losses for the FAC graph. V=Vision, A=Audio, T=Text.

Extend contrastive learning to three modalities:

- Requires being careful about what features to compare
- Text modality is more granular

(a) Benefits of multiple modalities on HT

Modalities	UCF	HMDB	YC2	MSRVTT	ESC-50
VT	82.7	55.9	33.6	27.5	/
VA	75.5	51.6	/	/	79.0
VAT (FAC)	84.7	57.3	32.2	28.6	78.7

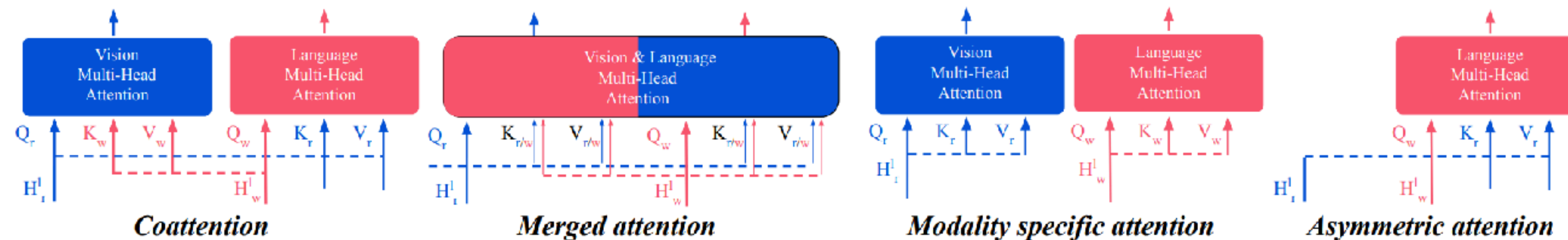
Table 2: Comparison of learnt representations versus the state-of-the-art. Results are averaged over all splits. The “Mod.” column shows which combinations of modalities are used by the methods, possibilities: Vision, Audio, Text, Flow. Dataset abbreviations: AudioSet, HowTo100M, Instagram65M [23], SoundNet [7], 2M videos from YouTube8M [2], Kinetics600; their length in years is given in the “years” column. †[71] uses a non-linear classifier. We report top-1 accuracy for UCF101, HMDB51, ESC-50, Kinetics600 and mAP for AudioSet.

Method	f_v (#params)	Train data	years	Mod.	UCF101		HMDB51		ESC-50	AS	K600
					Linear	FT	Linear	FT	Linear	MLP	Linear
MIL-NCE [49]	I3D (12.1M)	HT	15	VT	83.4	89.1	54.8	59.2	/	/	
MIL-NCE [49]	S3D-G (9.1M)	HT	15	VT	82.7	91.3	53.1	61.0	/	/	
AVTS [41]	MC3 (11.7M)	AS	1	VA		89.0		61.6	80.6		
AVTS [41]	MC3 (11.7M)	SNet	1	VA					82.3		
AA+AV CC [32]	RN-50 (23.5M)	AS	1	VA						28.5	
CVRL [67]	R3D50 (33.3M)	K600	0.1	V							64.1
XDC [4]	R(2+1)D-18 (33.3M)	AS	1	VA		91.2		61.0	84.8		
XDC [4]	R(2+1)D-18 (33.3M)	IG65M	21	VA		94.2		67.4			
ELo [64]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [55]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7	89.2		
GDT [62]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1	88.5		
GDT [62]	R(2+1)D-18 (33.3M)	IG65M	21	VA		95.2		72.8			
VA only (ours)	R(2+1)D-18 (33.3M)	AS	1	VA	83.9	91.5	60.0	70.1	85.6	29.7	55.5
VA only (ours)	S3D-G (9.1M)	AS	1	VA	84.7	90.1	60.4	68.2	86.1	29.7	59.8
VA only (ours)	S3D-G (9.1M)	AS+HT	16	VA	86.2	91.1	61.5	68.3	87.2	30.6	59.8
MMV FAC (ours)	S3D-G (9.1M)	AS+HT	16	VAT	89.6	92.5	62.6	69.6	87.7	30.3	68.0
MMV FAC (ours)	TSM-50 (23.5M)	AS+HT	16	VAT	91.5	94.9	66.7	73.2	86.4	30.6	67.8
MMV FAC (ours)	TSM-50x2 (93.9M)	AS+HT	16	VAT	91.8	95.2	67.1	75.0	88.9	30.9	70.5
Supervised [19, 40, 64, 71, 87]						96.8	71.5	75.9	86.5 [†]	43.9	81.8

Multimodal Transformers

Multimodal Transformers

- **Coattention**: given queries from one modality (image), keys and values can be taken only from the other modality (language)
- **Merged** attention: For a given query (from one modality), consider keys and values from all input tokens regardless of the modality type.
- **Modality-specific** attention: single modality attention where queries, keys, and values all come from either the image or text modality.
- **Asymmetric** attention: queries are either from language or image, while keys and values are from image or language, respectively.

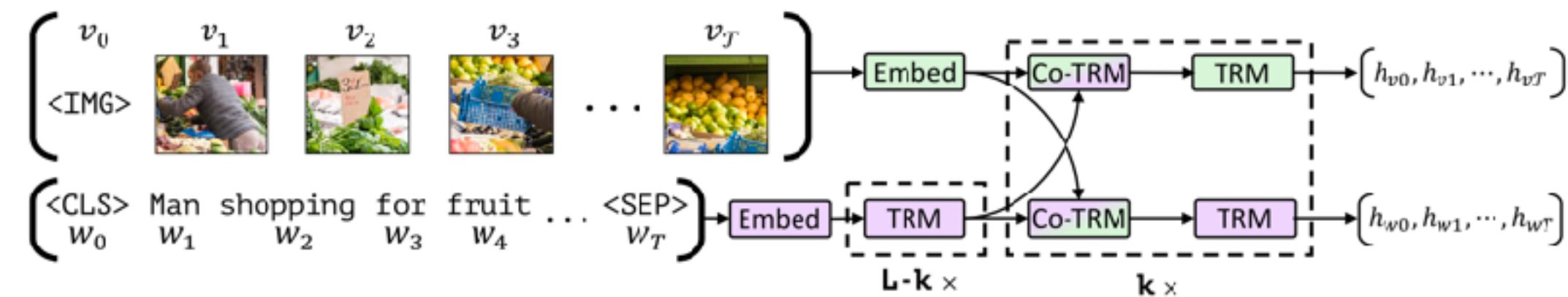


¹ Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers, Hendricks et al. (2018)

Multimodal Transformers

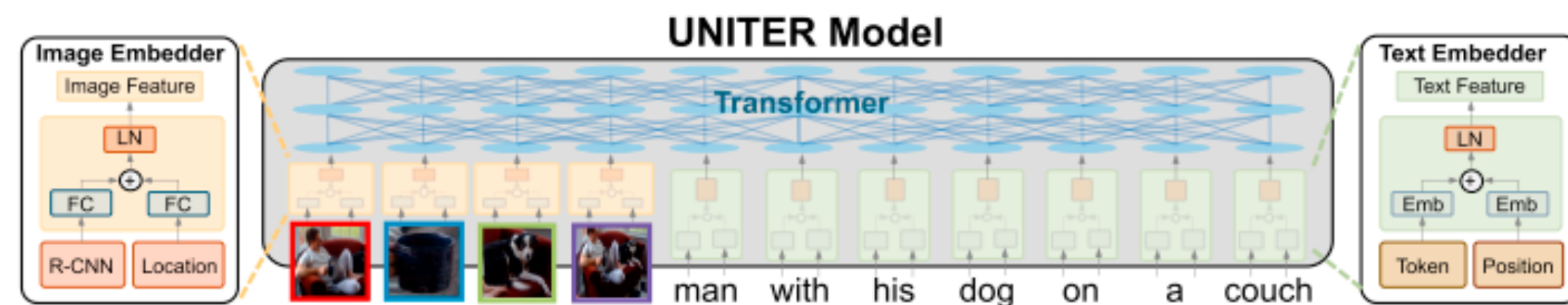
Two-stream Multimodal Transformers: two Transformers are applied to images and text independently, which is fused by a third Transformer in a later stage.

Vision & Language BERT (ViLBERT)¹



One-stream Multimodal Transformers: a single Transformer is applied to both images and text.

UNiversal Image-Text Representation (UNITER)²



¹ ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, Lu et al. (2019)

² UNITER: UNiversal Image-Text Representation Learning, Chen et al. (2020)

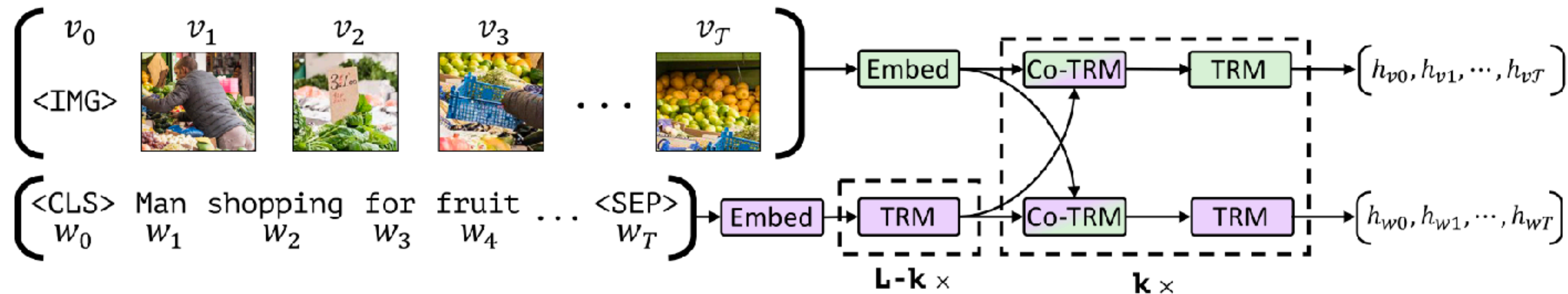
Multimodal Transformers: ViLBERT



ViLBERT¹: two-stream multimodal Transformer for vision-language

It consists of two parallel streams for visual (green) and linguistic (purple) with co-attentional layers.

This structure allows for variable depths for each modality and enables sparse interaction through co-attention.



¹ ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, Lu et al. (2019)

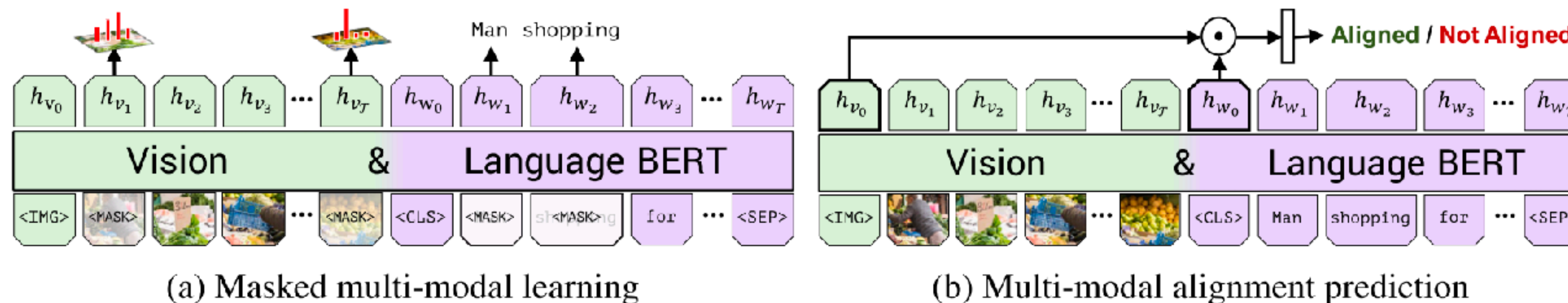
Multimodal Transformers: ViLBERT



Pre-training: ViLBERT is trained on the image-captions dataset with two tasks:

Masked multi-modal learning: reconstructing image region categories or words from masked inputs

Multi-modal alignment prediction: predicting whether the caption describes the image content.



Multimodal Transformers: ViLBERT



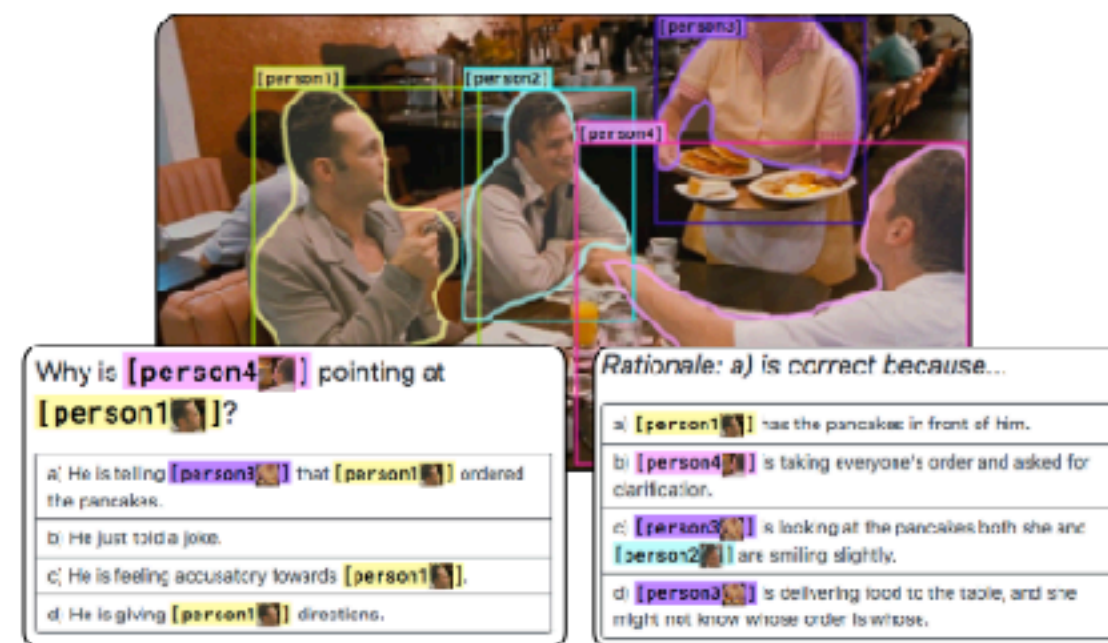
For the downstream tasks: learning a classification layer.

Wide range of tasks



VQA

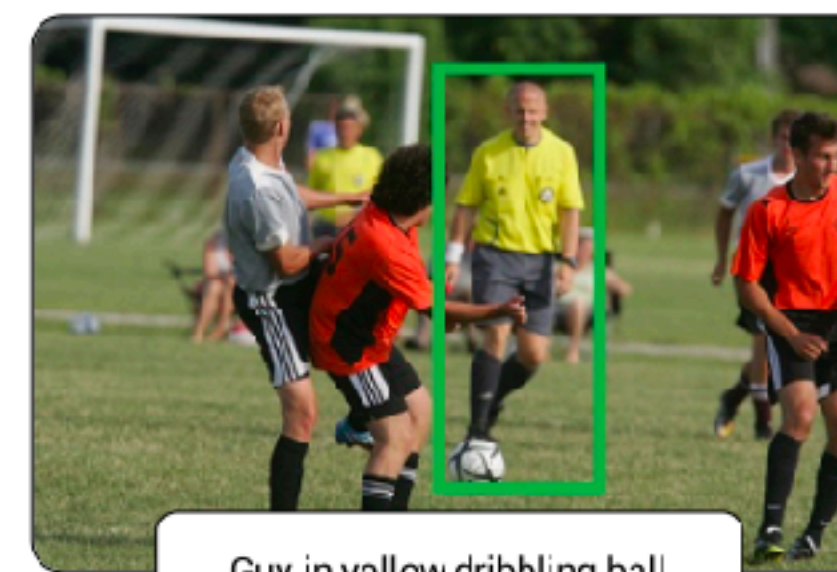
answering natural language questions about images



VCR Q→A

VCR QA→R

posed as multiple-choice problems



Referring Expressions

localize an image region given a natural language reference



Caption-Based Image Retrieval

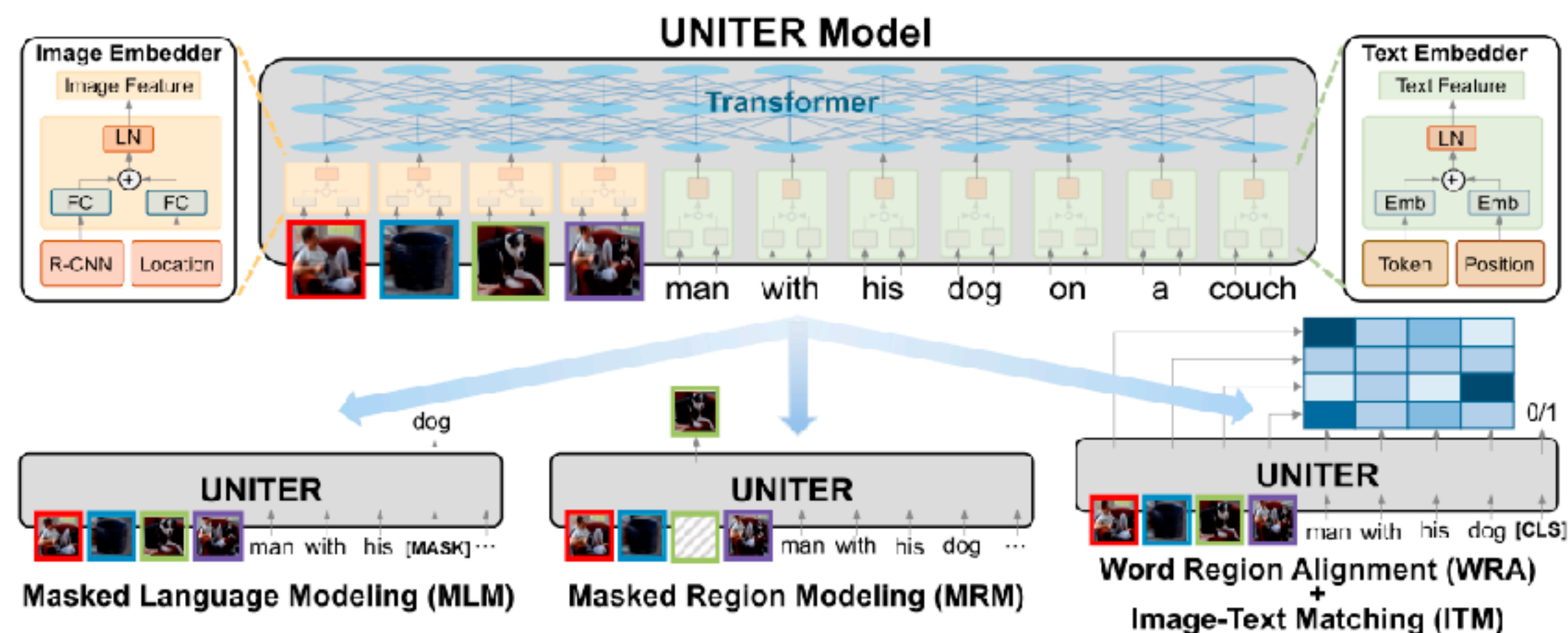
identifying an image from a pool given a caption describing its content

Multimodal Transformers: UNITER

UNITER¹ is one-stream pre-trained multimodal Transformer.

Uses visual features and bounding box features and word tokens and positions.

Various pretraining losses are applied



- (i) Masked Language Modeling conditioned on image;
- (ii) Masked Region Modeling conditioned on text;
- (iii) Image-Text Matching;
- (iv) Word-Region Alignment.

¹ UNITER: UNiversal Image-TExt Representation Learning, Chen et al. (2020)

VATT: MMV but with audio-inputs and *one* model instead

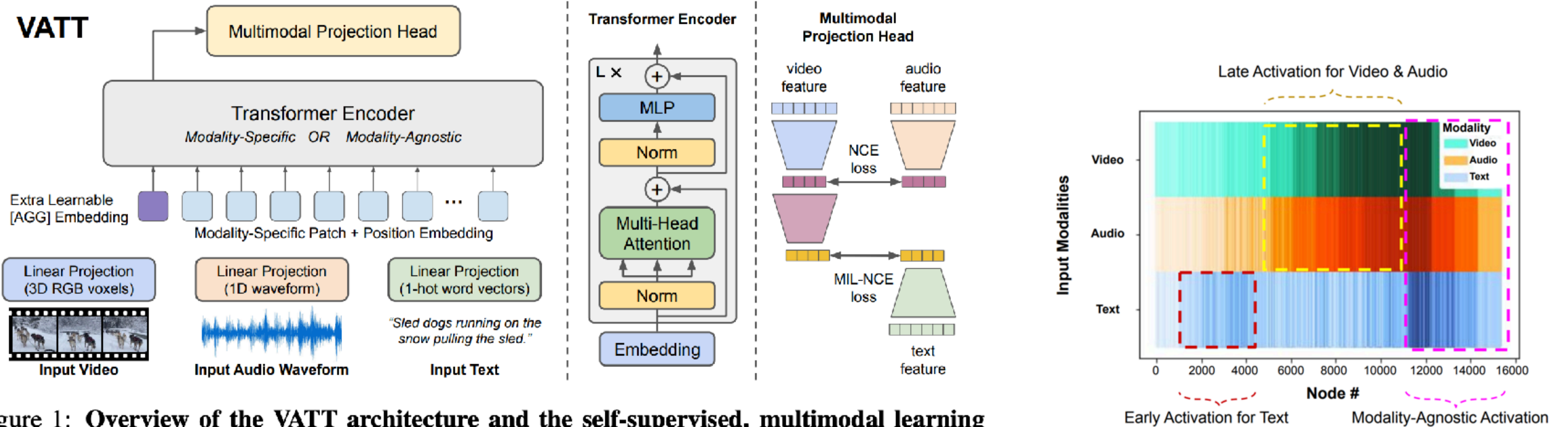
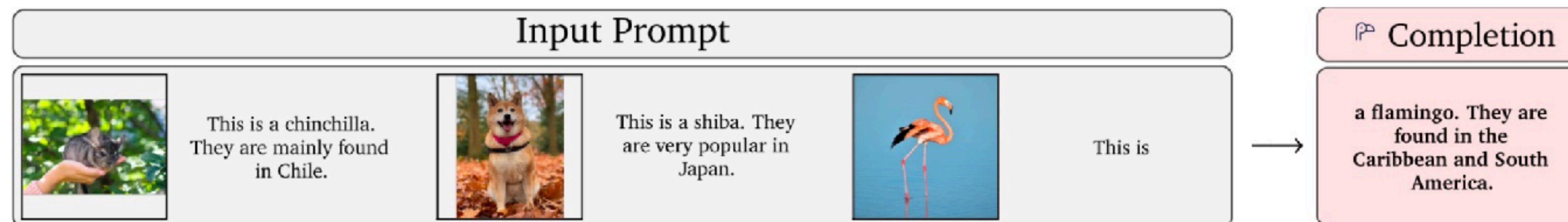


Figure 1: **Overview of the VATT architecture and the self-supervised, multimodal learning strategy.** VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. We define a semantically hierarchical common space to account for the granularity of different modalities and employ the Noise Contrastive Estimation (NCE) to train the model.

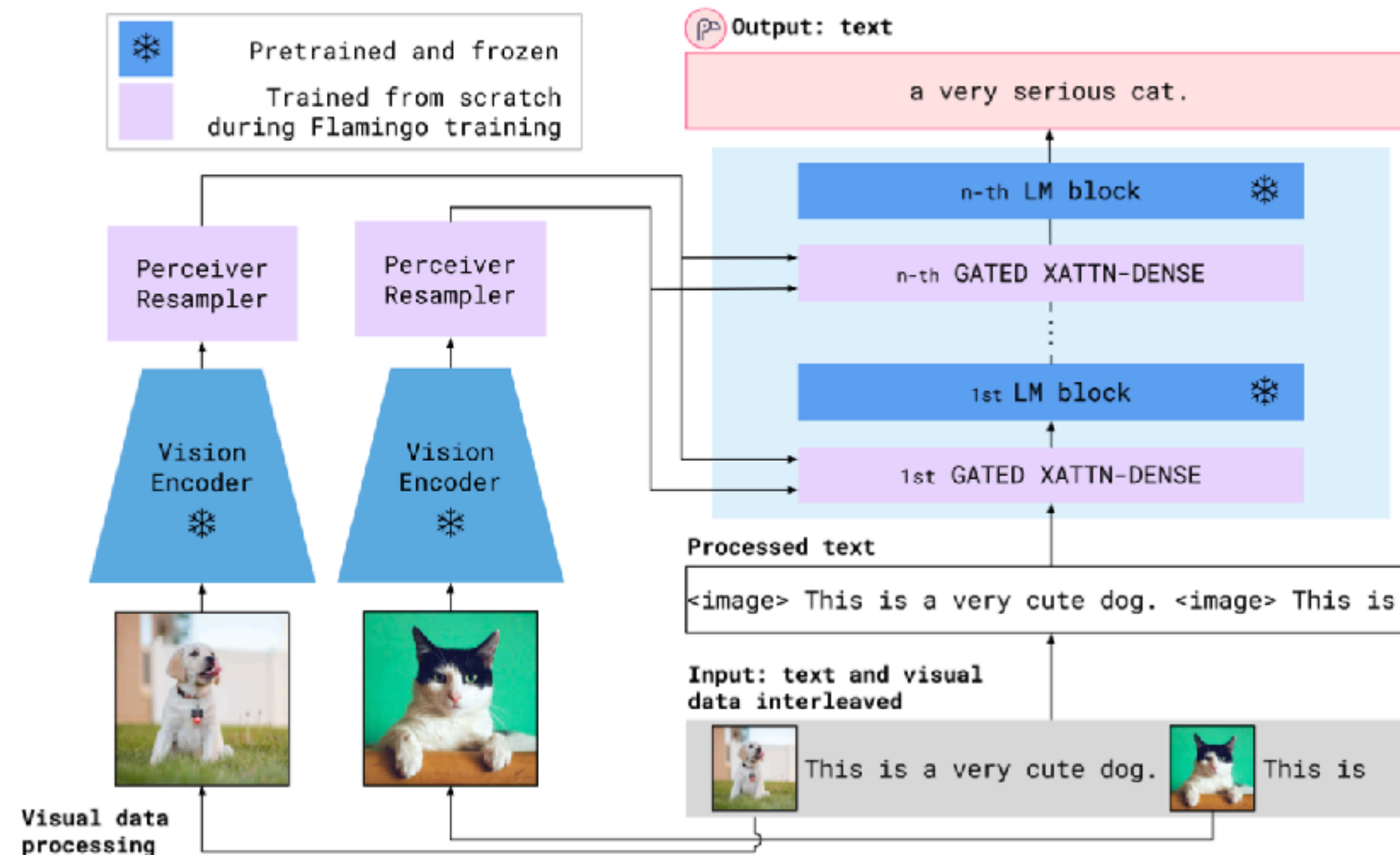
Flamingo

- Flamingo is a Transformer-based architecture for multimodal few-shot tasks (image captioning, visual dialogue or visual question answering)
- Able to learn from only a few input/output examples i.e., *in few-shot settings*.
 - It processes arbitrarily interleaved images and text as prompt;
 - And it generates output text in an open-ended manner.
- Basically: it performs in-context learning (like GPT) but with images and text as context (prompt).



Flamingo

- On the vision side: a vision encoder with a contrastive text-image approach, à la CLIP
- On the language side: existing autoregressive LM trained on a large text corpus
- Linked via a learnable attention component (the Perceiver)
 - It outputs a fixed-size set of visual tokens.
 - Which are used to condition the frozen LM, trained to generate text.

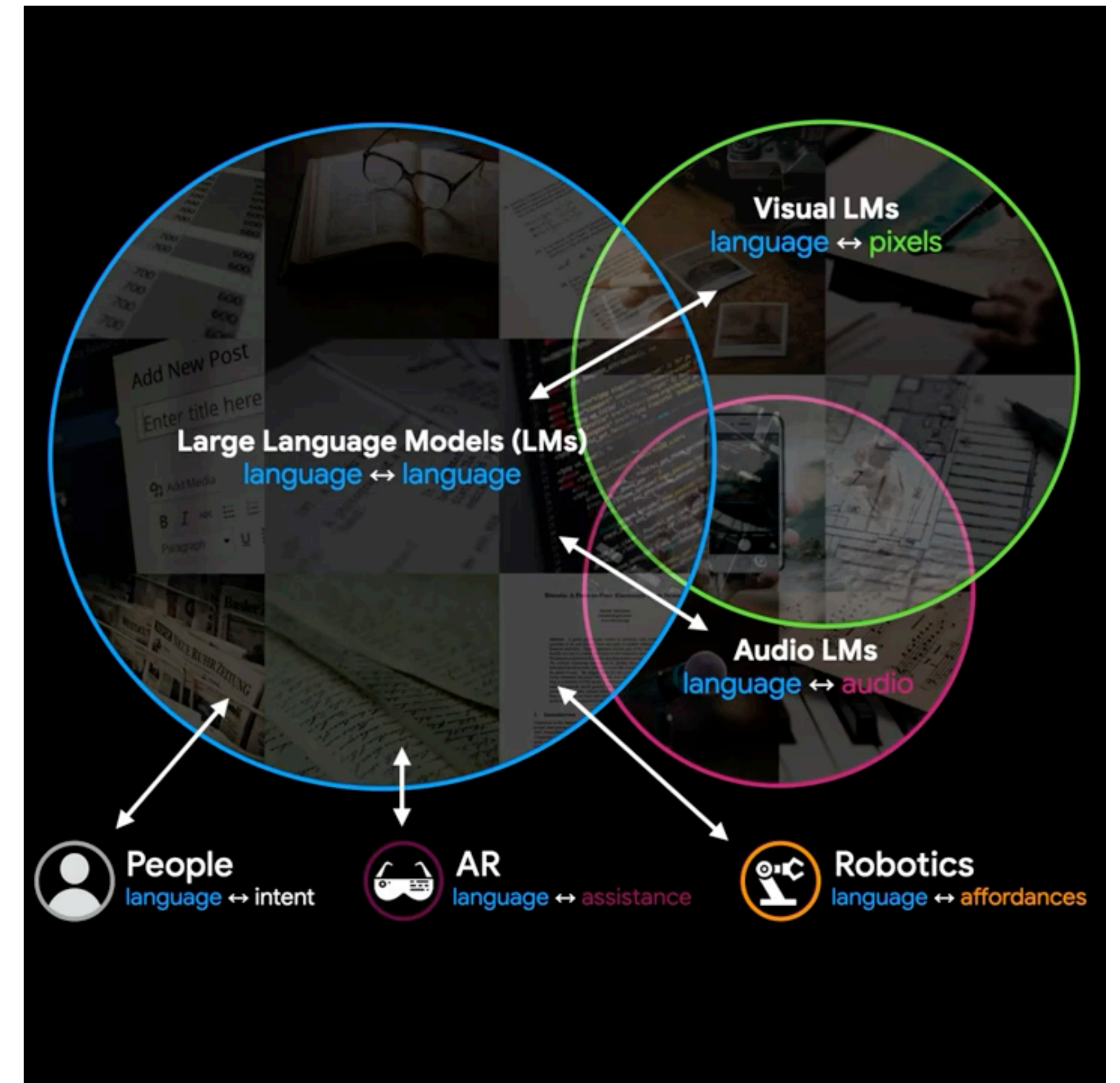


Socratic Models

$$\text{activity} = f_{\text{LM}}(f_{\text{VLM}}(f_{\text{LM}}(f_{\text{ALM}}(f_{\text{LM}}(f_{\text{VLM}}(\text{video}))))))$$

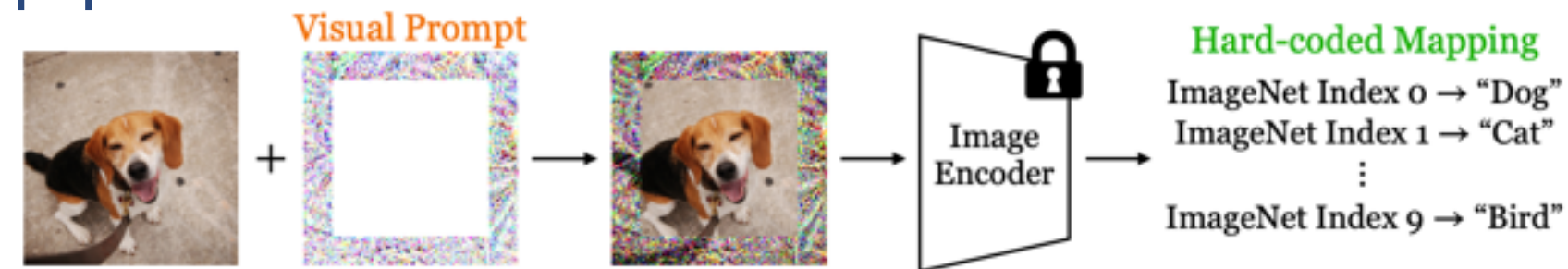
- (i) the VLM detects visual entities,
- (ii) the LM suggests sounds that may be heard,
- (iii) the ALM chooses the most likely sound,
- (iv) the LM suggests possible activities,
- (v) the VLM ranks the most likely activity,
- (vi) the LM generates a summary of the Socratic interaction.

Essentially language as the lingua franca

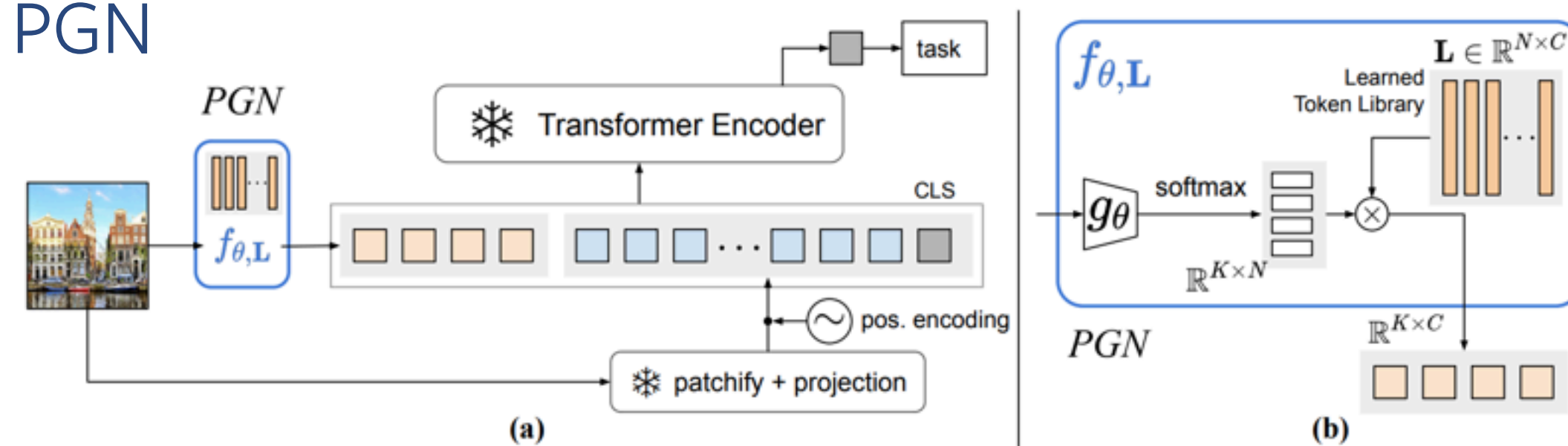


Simpler models that utilize pretrained models

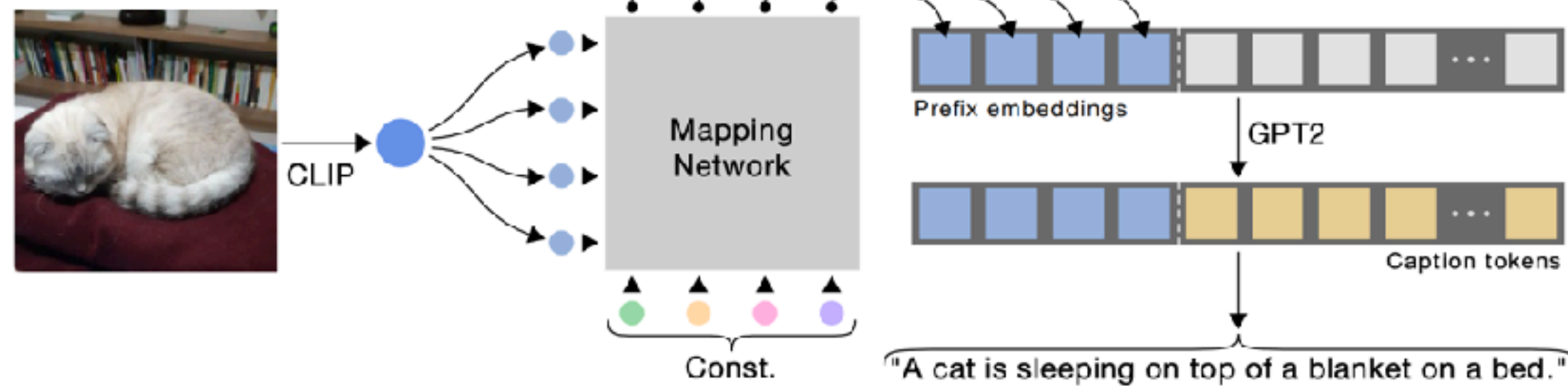
VPT



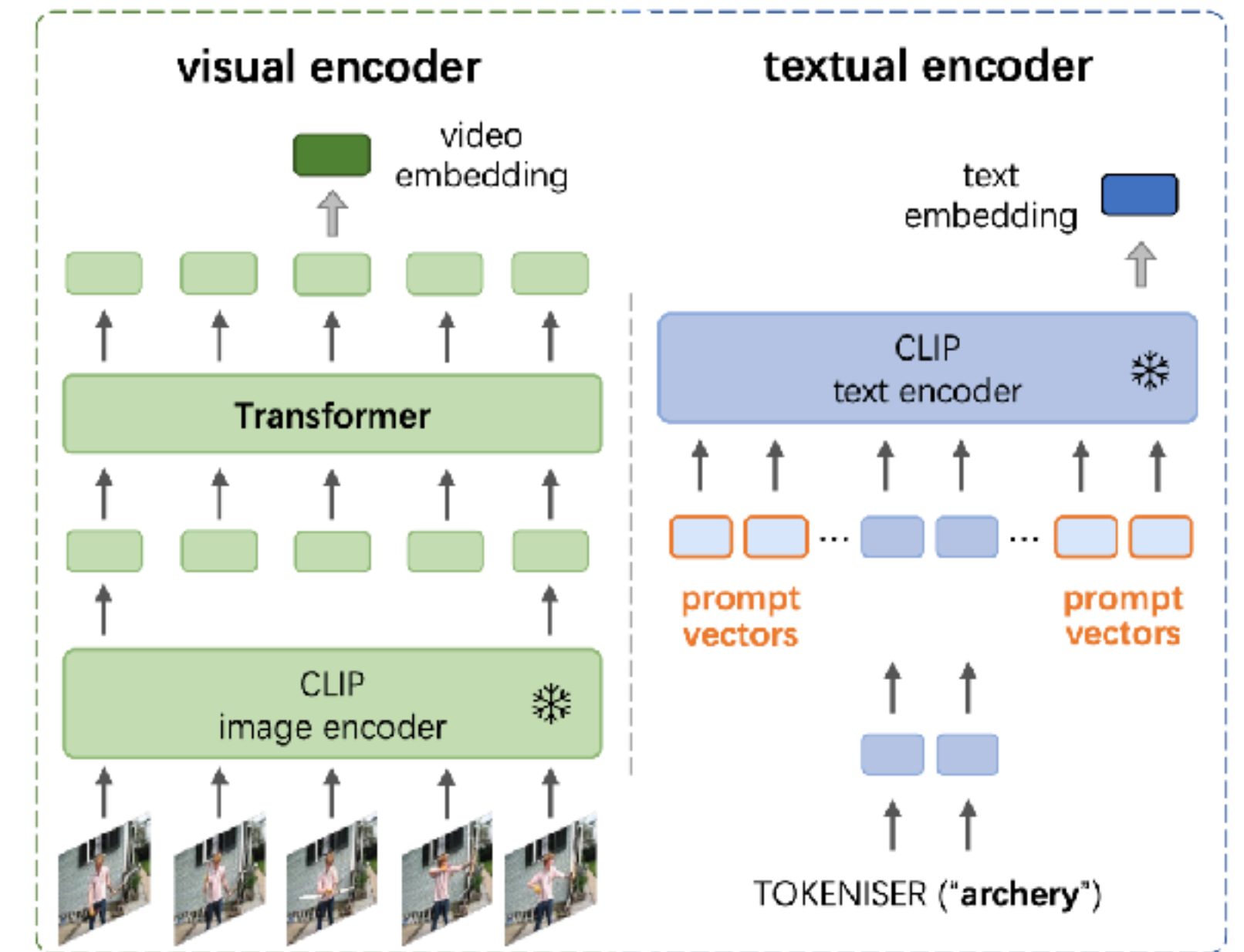
PGN



ClipCap



Frozen CLIP tuning for videos



Exploring Visual Prompts for Adapting Large-Scale Models. Bahng et al. ArXiv 2022

Prompt Generation Networks for Efficient Adaptation of Frozen Vision Transformers. Loedeman et al. arXiv 2022

Prompting Visual-Language Models for Efficient Video Understanding. Ju et al. ECCV 2022

ClipCap: CLIP Prefix for Image Captioning. Mokady et al. ArXiv 2022

My bets for future “big” research directions

Combination of more and more modalities

Useful self-supervised learning for *most* tasks (and a merger of 2D and 3D learning)

Combination of very large-scale “foundation” models

Move into more difficult domains like robotics

AI for science, self-coding language models

Renewed focus and work on privacy and bias (soon)



Summary

Today: Multi-modal learning

What is multi-modal data?
Why is it useful?
How to use it?

- Visual Question Answering (VQA) and CAPTCHA
- Multimodal clustering of video-datasets
- Self-supervised object detection and classification

Practical Pitfalls

- Multi-modal variational networks
- WATT: Transformers for Multimodal Self-Supervised Learning on Raw Video, Audio and Text
- Socratic Models
- Flamingo

Exercises in Attention

5

What is a modality?

Modality: The set of sensor streams (inputs or outputs) that are used to interact with the environment.

Examples of modalities

- Visual (eyes) (with glasses or contact lenses)
- Auditory (ears) (with hearing aids)
- Tactile (skin) (with sensors)
- Olfactory (nose) (with sensors)
- Vestibular (inner ear) (with sensors)
- Proprioceptive (body position) (with sensors)
- ... (Hand images, Brain images, MRI)

Exercises in Attention

6

Why multi-modal learning?

- Chosen to combine modalities, as it doesn't clearly help (?)
- Noisy and missing data of a single modality
- Meaning often captured not by a single modality
- But in practice it's not so easy
- The representation spaces vary widely: continuous (eg sound) vs ordinal (eg rankings) or discrete (eg text)

Exercises in Attention

7

Meaning often captured not by a single modality: McGurk effect

Speech perception is not a purely auditory process.

Exercises in Attention

9

Quiz: Multi-modal learning is a great avenue for scalable deep learning. When designing a future robot, what things, on top of its eyes, why modality, perhaps not opt for including more and more modalities?

Turn to your neighbour and come up with as many reasons as you can imagine for why using less modalities might be sensible.

Exercises in Attention

10

What and why multi-modal

...whereby one modality is used as supervision signal for the other

Vision → Prediction → Hearing

Exercises in Attention

13

Remember this slide!

The key to understanding is repeating meaning from association

Exercises in Attention

14

Multiple modalities can also yield such useful semantic information.

Exercises in Attention

15

Clustering multi-modal data

Exercises in Attention

17

Good feature representations ⇒ good clustering

VGG-Sound

AVE

Exercises in Attention

18

Audio-visual clustering

invariance vs distinctiveness

Exercises in Attention

25

Learning hypotheses we test:

1. Sample Distinctness

2. Time Invariant

3. Time Shift

Exercises in Attention

26

Framework

Exercises in Attention

27

Notice the similarity to SimCLR

Exercises in Attention

28

Variance and invariance: insensitive to sample & time shift, invariant to modality

Exercises in Attention

29

Detecting time-shifted pairs as regulars is not easy

Exercises in Attention

30

Multi-modal variances learning

Goal of this paper: Combining self-supervised and multi-modal learning to detect multiple objects

Exercises in Attention

41

FRAMEWORK OVERVIEW

Exercises in Attention

42

Ingredient 1: training heat maps

Exercises in Attention

43

Ingredient 2: Self-supervised training (L_{SSL})

Exercises in Attention

44

Framework overview

Exercises in Attention

45

Qualitative results compared to weakly-supervised

Exercises in Attention

46

SSL object detection

Multimodal Transformers: VIT-RFT

For the downstream task: learning a classification layer

Exercises in Attention

61

Multimodal Transformers: LITR

Exercises in Attention

62

WATT: NMMV but with audio inputs and one model instead

Exercises in Attention

63

Flamingo

Exercises in Attention

65

Socratic Models

Exercises in Attention

66

Large-scale multi-modal learning

One Final Note

This. And much more. You've learned a tremendous amount. A huge well done to all of you. And thank you for your curiosity, patience, critiques and motivation to learn. Please take time to rest over the holidays.

Depictions of AI

<https://beta.imagesofai.org/about>

- Compare this to <https://www.youtube.com/watch?v=...>
- With new hype about AI, important to stay critical.

ImageNet: side notes

- Most commonly used services: ImageNet-42: 1K categories, 1.2M images, 154GB
- Explore them here: <https://www.image-net.org/>
- Important to also "use" the data, don't just throw a neural network at it!

Also check out: On the quality of machine learning datasets: A critical history of ImageNet. Cohen et al. 2011

ImageNet 2012 winner: AlexNet

More weights than samples in the dataset!

Momentum

2022 Cheese Rolling

Hubel and Wiesel: Nobel Prize for Physiology or Medicine in 1981

WE NEED TO GO DEEPER

Interesting results with randomly initialized networks

- Hidden in a randomly weighted fully connected (FC) network (with random weights) that is smaller than, but matches the performance of a ResNet-50 trained on ImageNet! (randomly do these "natural shortcuts" exist, but we provide an algorithm to effectively find them.)
- Random matrices are essentially full-rank (cf. Johnson-Lindenstrauss Lemma)
- Currently one of my MSc students working on generalizing this!

What do the different layers in a deep neural network learn

Layer 1, Layer 2, Layer 3

Fibers, Image patches for activation, Feature maps

<https://osslandscape.com/explorer>

- Explore: play around with momentum, it etc.

Hubel and Wiesel: Nobel Prize for Physiology or Medicine in 1981

1x1 Convolution: a computationally cheap method

Eg 28x28x192 -> 28x28x32

Number of Operations: (28*28*192) * (28*28*32) = 226,422,720 Ops

Number of Operations for 2x2 Conv Step: (28*28*192) * (28*28*16) = 226,422,720 Ops

Number of Operations for 3x3 Conv Step: (28*28*192) * (28*28*16) = 226,422,720 Ops

Multimodal Transformer architecture: CLIP

- CLIP solves an easier proxy pre-training task of predicting which text and image is paired with which image.
- Mediate the cosine similarity of the image and text embeddings of true pairs (I_i * T_i)
- Minimize the cosine similarity of the embeddings of incorrect pairs (I_i * T_j)
- Formally said: CLIP optimizes a contrastive loss.

Combining this with text as coord. inputs: DALL-E v2 / "imCLIP"

Don't forget about the RLHF data and the 3B parameters

Graphs! They're everywhere

Convolution theorem: $x * w = \Phi^{-1}(\Lambda(w) \cdot \Phi(x))$

spatial domain, frequency domain

Graph Convolutional Networks (GCN)

- Main idea: keep this as simple as possible, get larger neighborhood by stacking. Choose polynomial with just order of 1
- Each node has a feature vector (row-wise)
- Left multiplying normalized Laplacian, we **smooth** features in neighborhood
- Right-multiplying with a weight matrix, we "can" $y = \text{ReLU}(L_{\text{norm}} * X * W)$

The last few lectures

CONVOLUTIONS FOR EQUIVARIANCE, ATTENTION GENERALIZES CONVOLUTIONS, CHINESE MORE GENERAL TRANSFORMERS

Why generative modelling?

Discriminative model p is normalized by outputs, but not for inputs

Generative model: p is normalized for inputs

VQ-VAE and VQ-VAE2

Neural Discrete Representation Learning, Van den Oord et al. ICLR'19

Let's take a breath. Where are we?

- Autocoders: nice idea but does not give us probabilities
- idea 1: lets model it with latent variables $z \rightarrow x$
- if we simply fix $p(z)$ in some manner, $p(x)$ can be computed as $\sum_z p(x|z)p(z)$
- Nice
- idea 2: how do we do inference (going from observations to model parameters)?
- We do this by learning $p(z|x)$ terms which is not used for a given x
- However the "normalization constant" integral is intractable to compute

How research gets done part 8

- Ok, so you didn't give up and you're on to something new.
- Next: how do you know if what's happening is why your method is better?
- Answer: Ablation
- The key idea is to "only vary one thing at a time"
- (Same principle behind when designing experiments in the investigation phase)
- Never change how things at the same time, you won't know if it was A or B that helped
- Some examples:
- Show simple way to understand cases (sometimes try examples)
- One idea per paper!

Fitting the variational posterior

We can fit the variational posterior to the exact posterior by maximizing the ELBO w.r.t. ϕ , which minimizes the

$$KL(q_{\phi}(z|x) || p_{\theta}(z|x))$$

This is reasonable because we can not compute $KL(q_{\phi}(z|x) || p_{\theta}(z|x))$ or even $p_{\theta}(z|x)$.

$$\mathcal{L}_{\text{ELBO}}(\phi) = -\mathbb{E}_{q_{\phi}(z|x)} \log p_{\theta}(z|x) - KL(q_{\phi}(z|x) || p_{\theta}(z|x))$$

ELBO is the difference between two intractable quantities, which is tractable!!

GAN: Intuition: arms race

- Police: wants to detect fake money as reliably as possible
- Counterfeiter: wants to make as realistic fake money as possible
- At beginning, both have no clue
- The police forces the counterfeiter to get better as it compares it to real money + and vice versa
- Convergent solution - Nash equilibrium (game theory)

Quiz: what dimensions need to be considered when thinking about developing the next generative model?

- Ethics of dataset use (bias, representation, consent)
- Copyright of inputs and outputs
- Malicious uses/cases of the trained model
- Malicious uses/cases of adapting the trained model
- Fraud or surveillance/creative industry/hack/news
- All of the above and more

See also "Broader Impacts" section in papers, e.g. [this guide](#)

Not only images: All suggest and 40,000 new possible chemical structures in just six hours

Pre-trained continuous surface embeddings (CSEs)

Matching CSEs learned on 3D model

Deep learning will be transforming the natural sciences

Multiscale Convolution, Protein Folding, Planetary Control