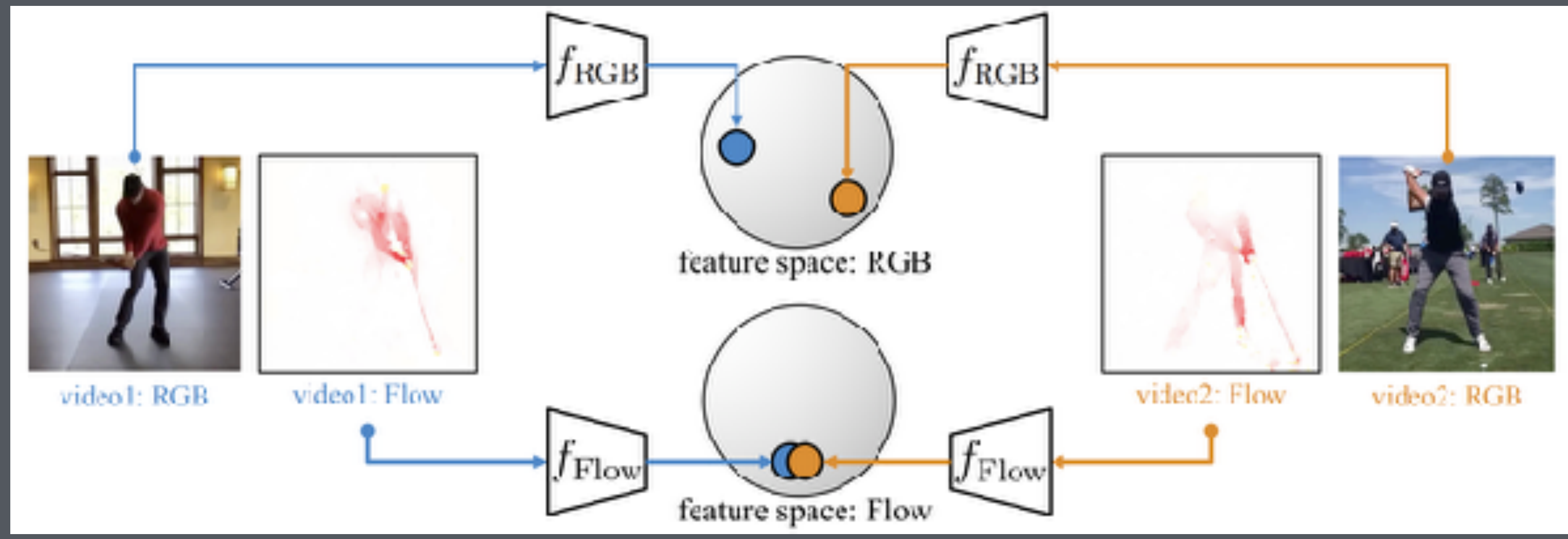
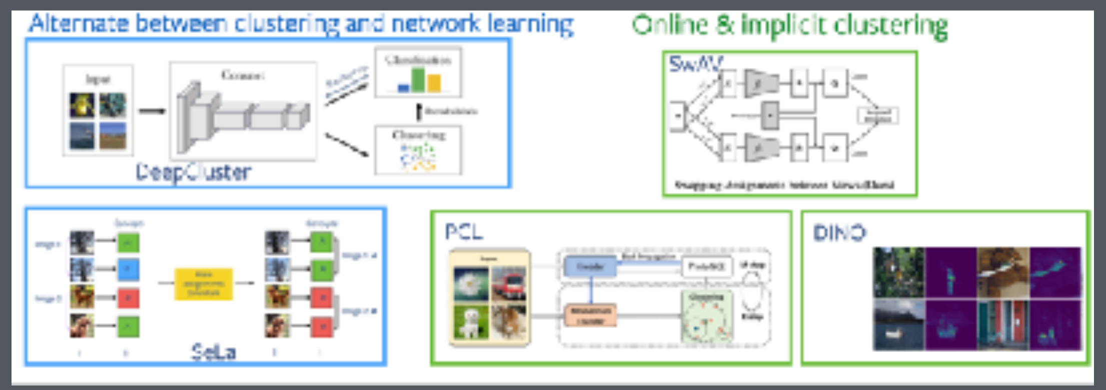
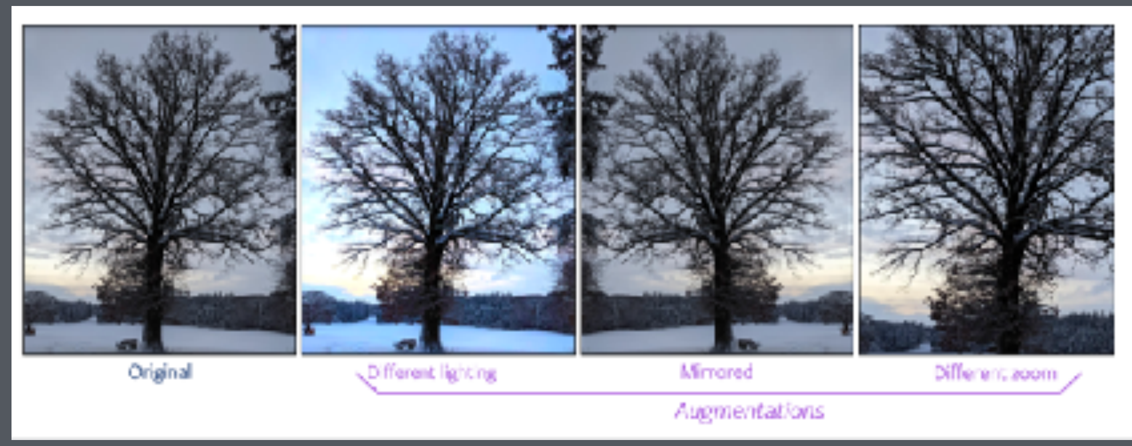
Self-supervised learning for computer vision from images, video and audio. Part 2: Multi-modal learning







@ 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

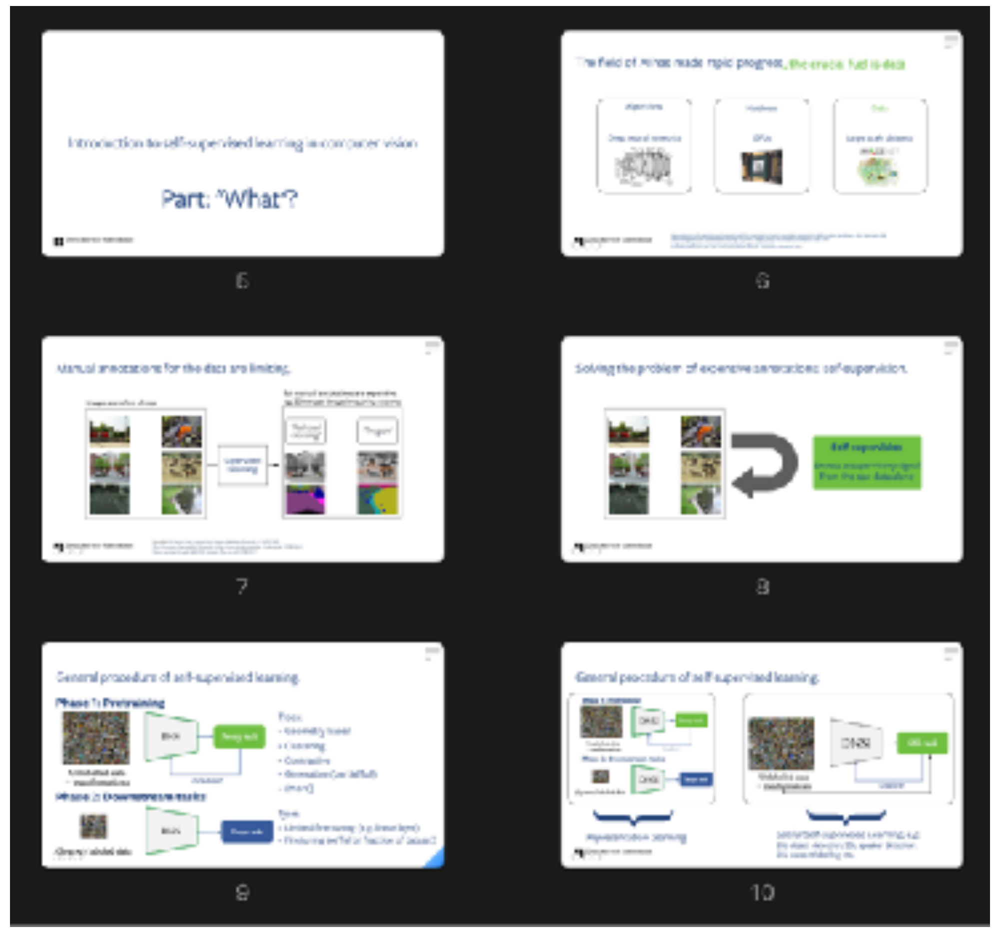
A cheat sheet with formulas will be uploaded to Canvas today.

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

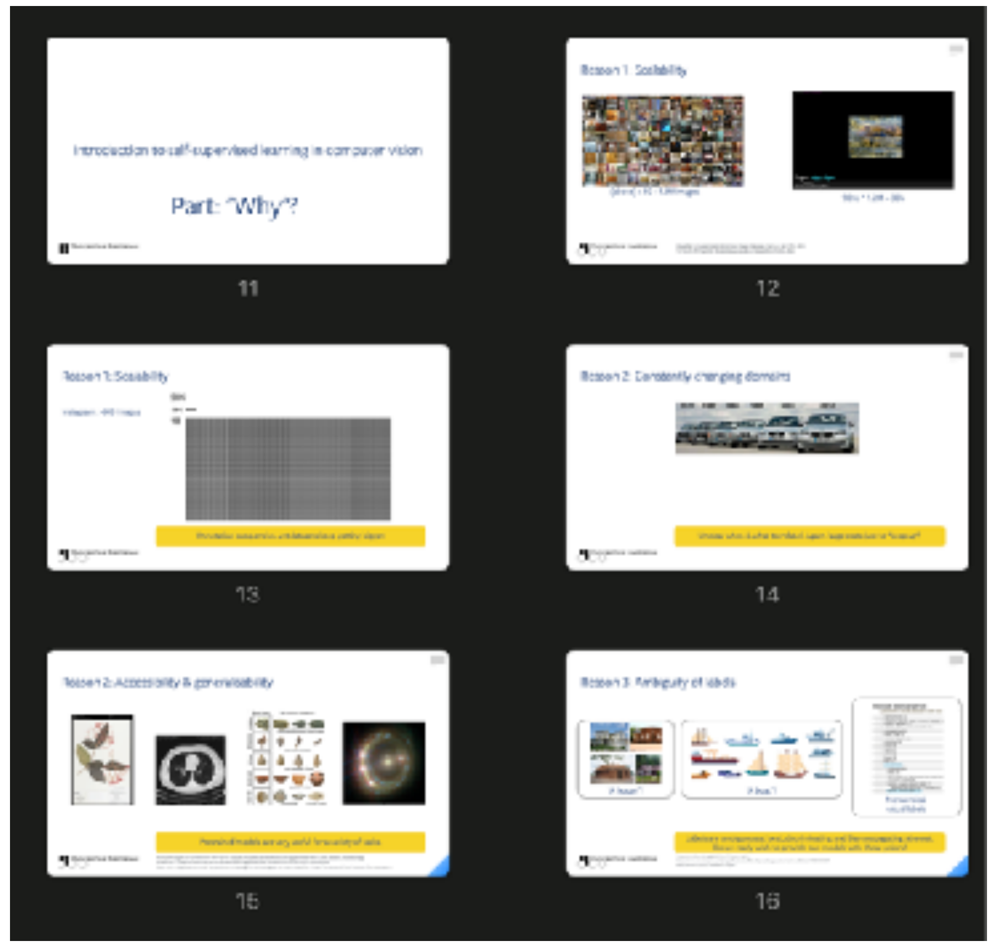
Not all of those will be relevant for the exam, of course.

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.

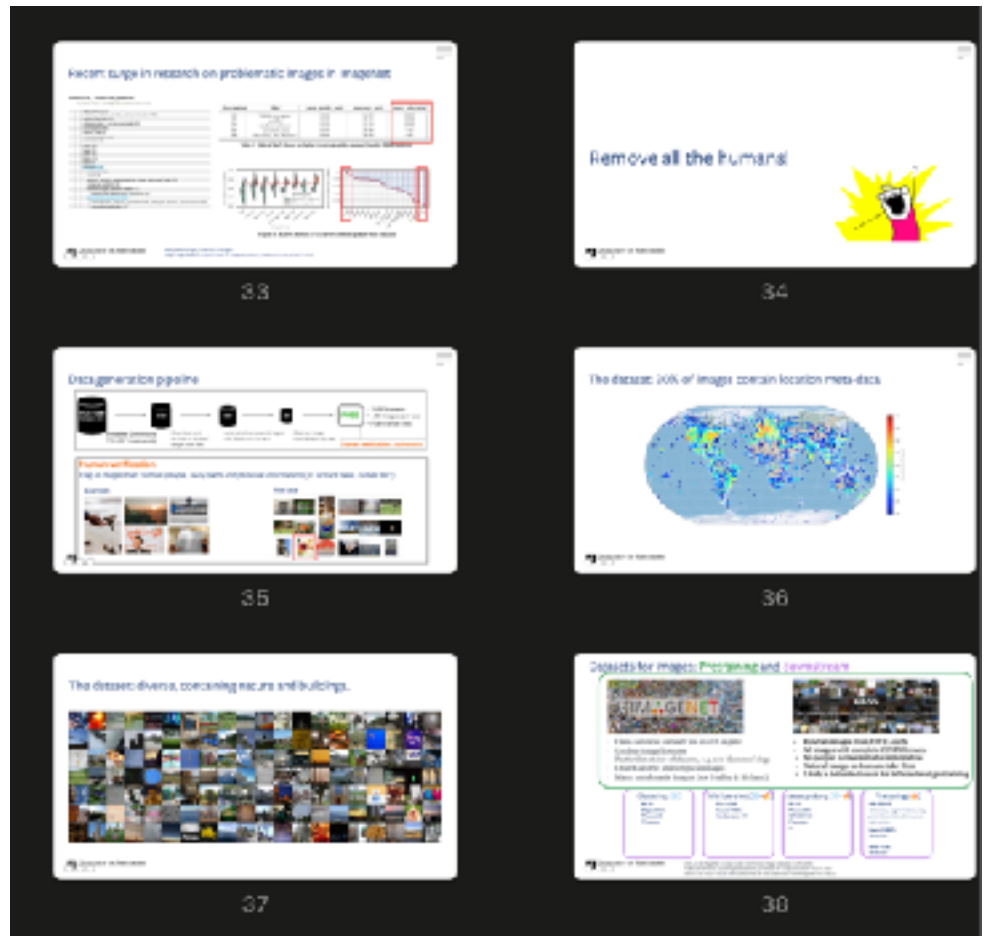
Summary last time



Why SSL?

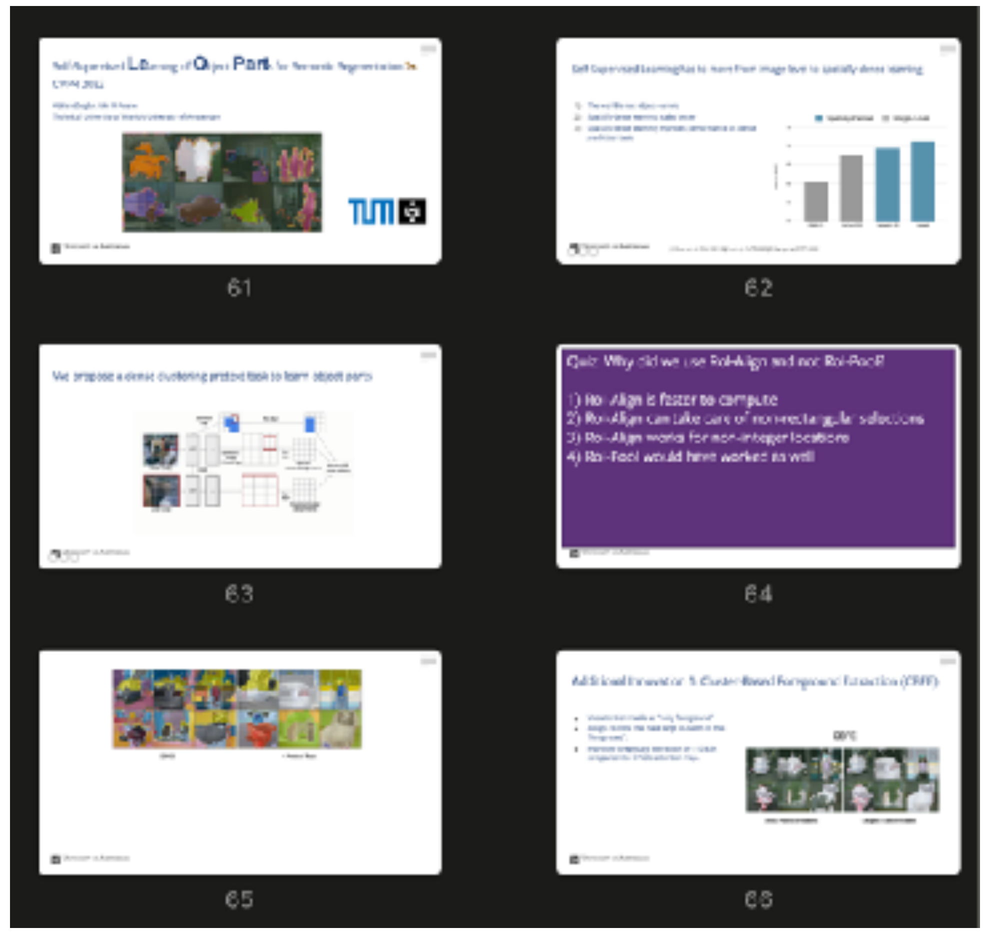


What is SSL?

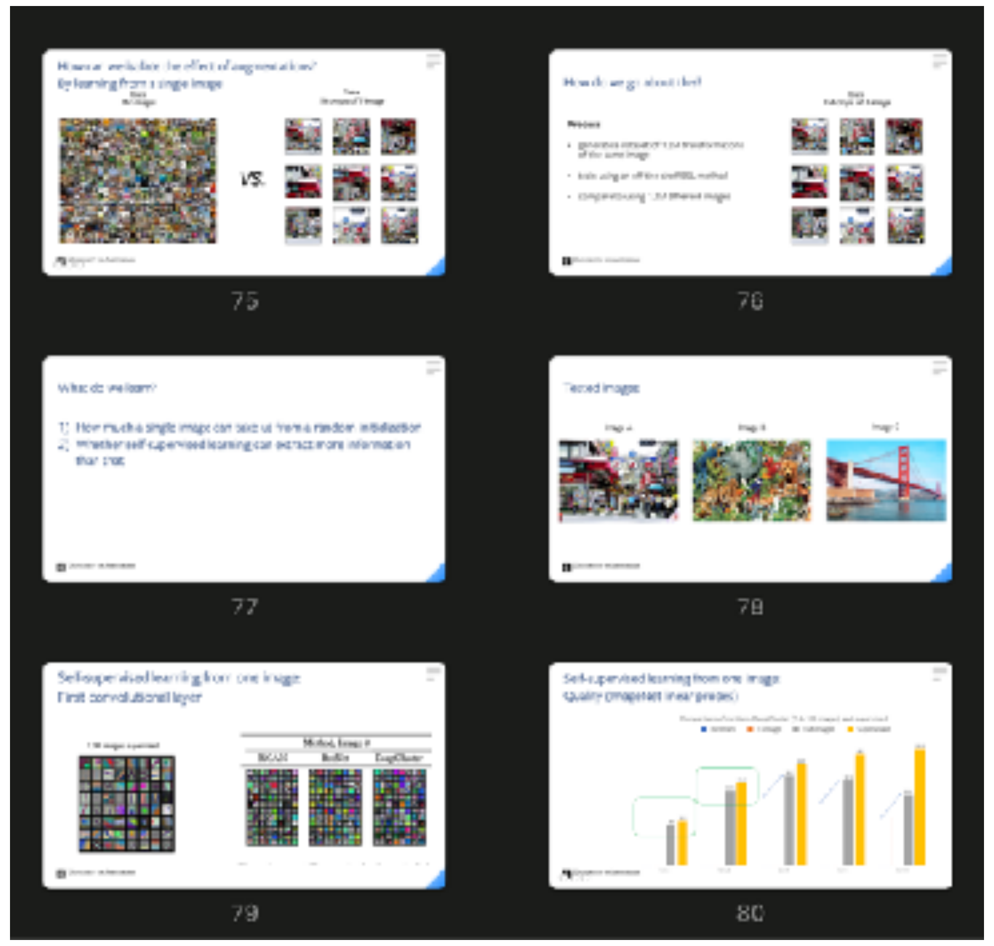


What kind of data? [1]

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

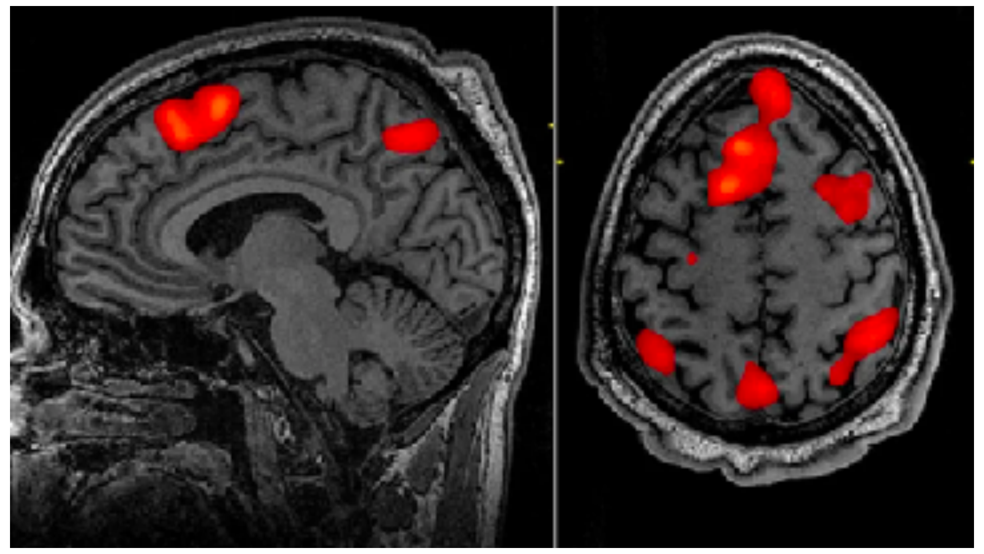
- Multi-modal Versatile Networks
- VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text
- Socratic Models
- Flamingo

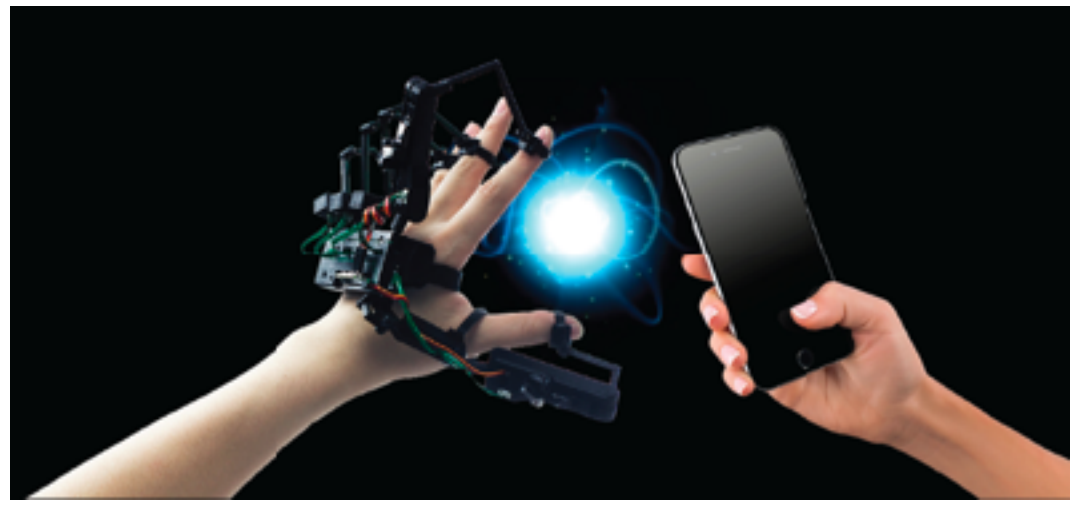
What is a modality

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

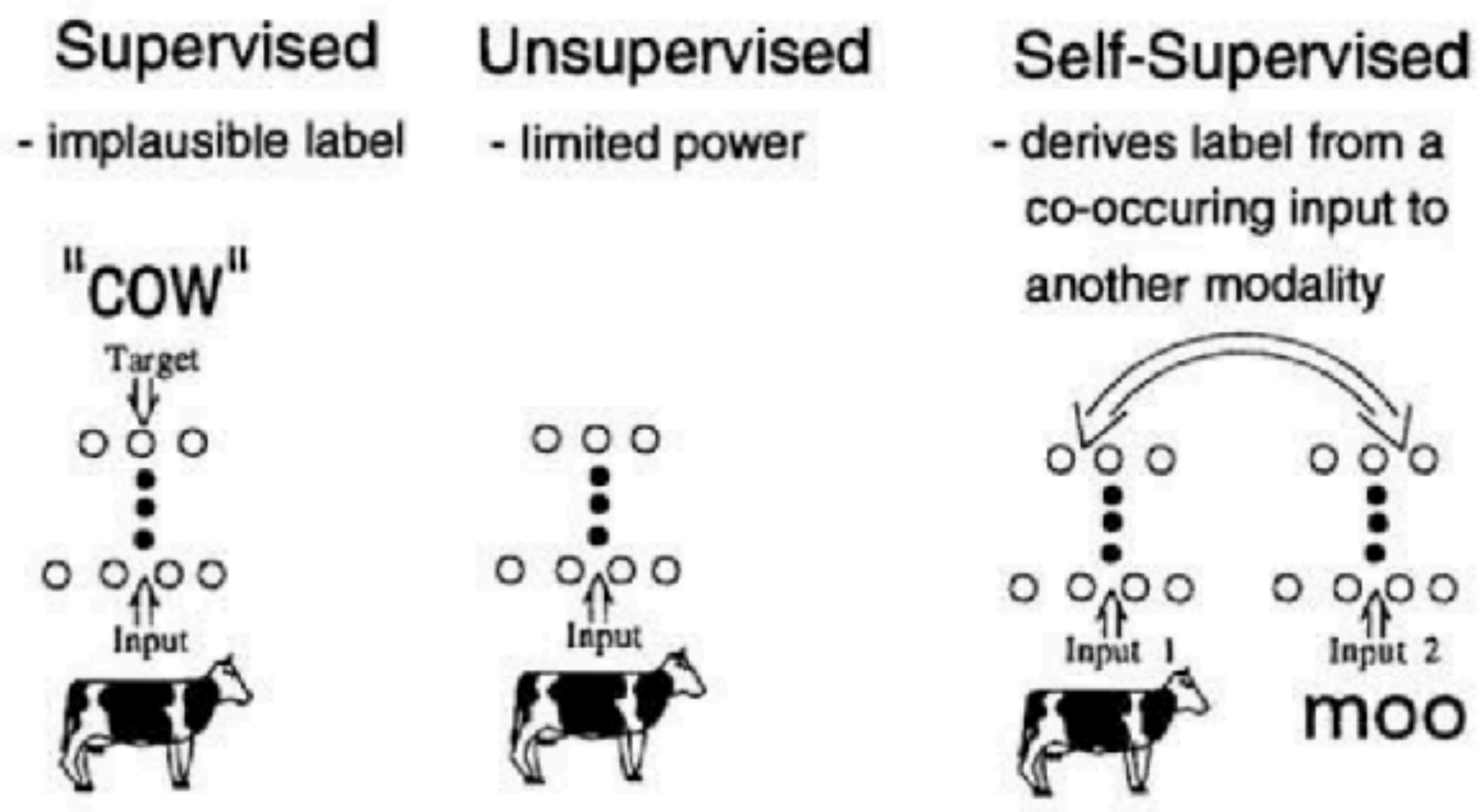


Figure 2: The idea behind the algorithm

Learning Classification with Unlabeled Data. *V. De Sa*. NeurIPS 1993

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
    continous (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

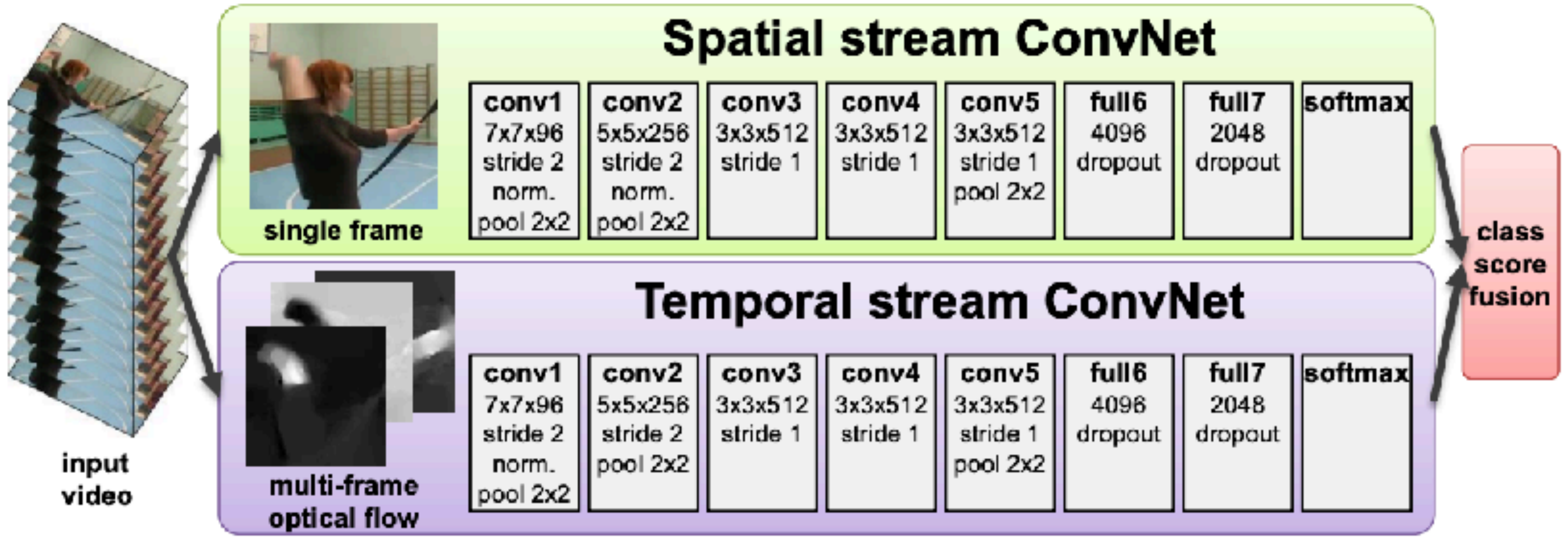
Multi-modal learning approaches

Representation learning

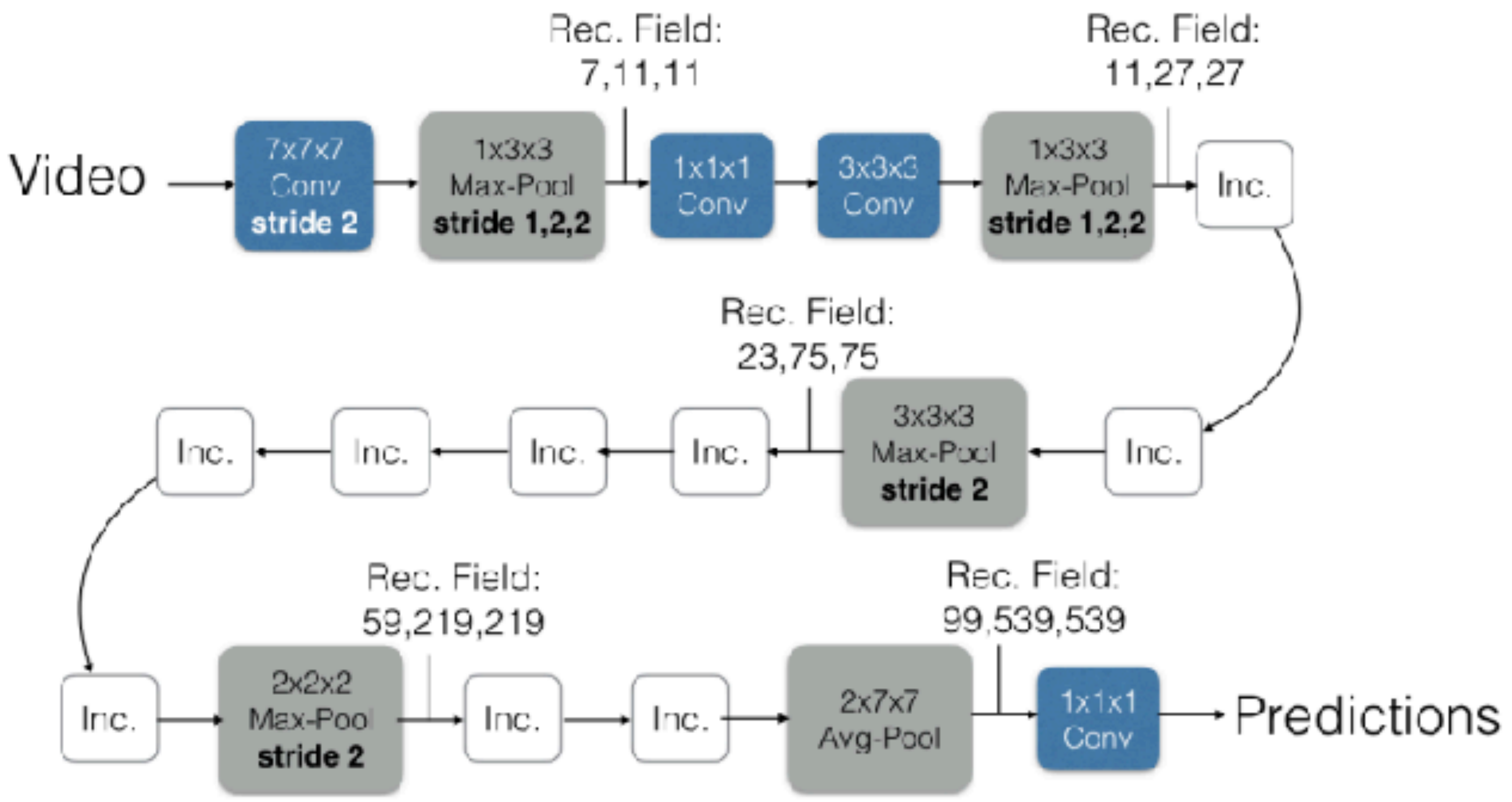
- Cross-modal
- Joint

Task learning

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

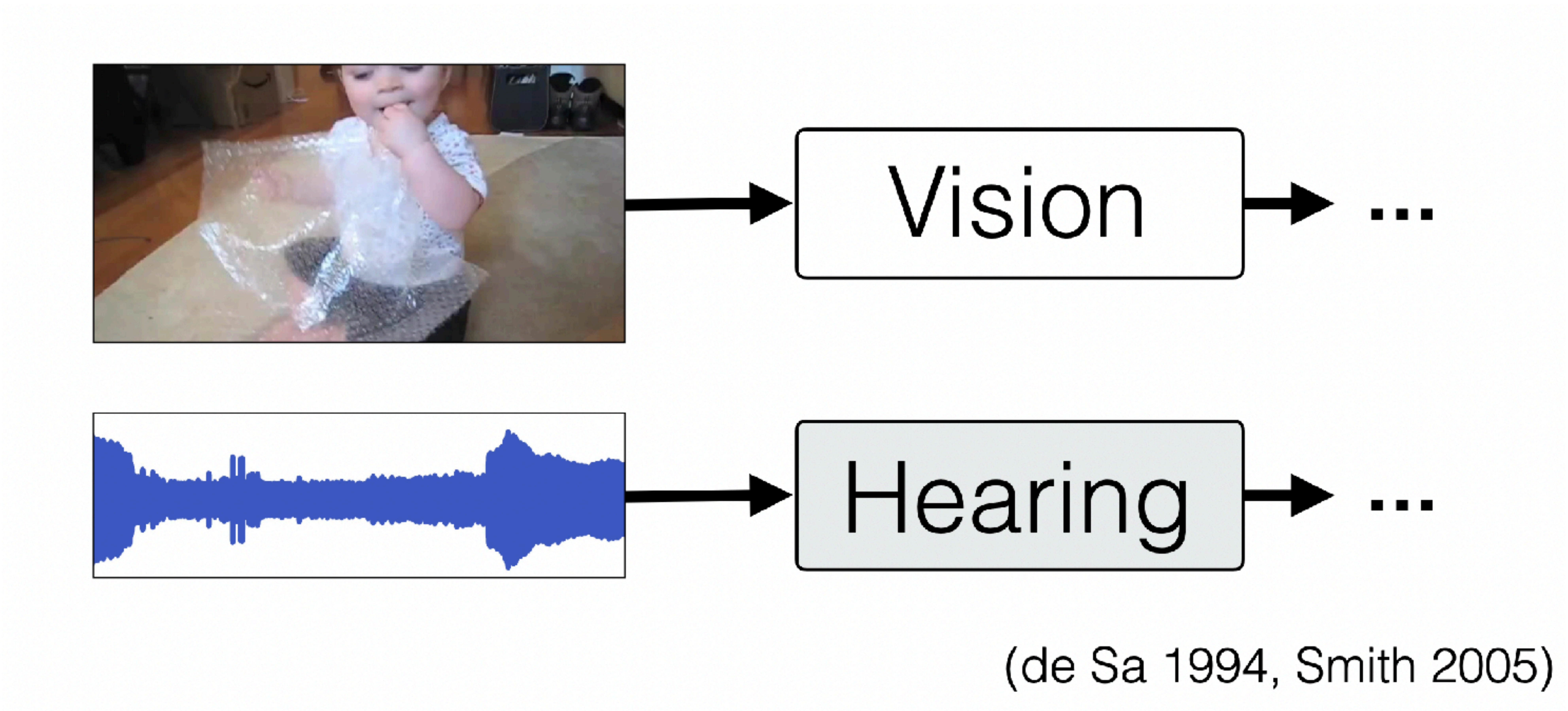


Inflated Inception-V1

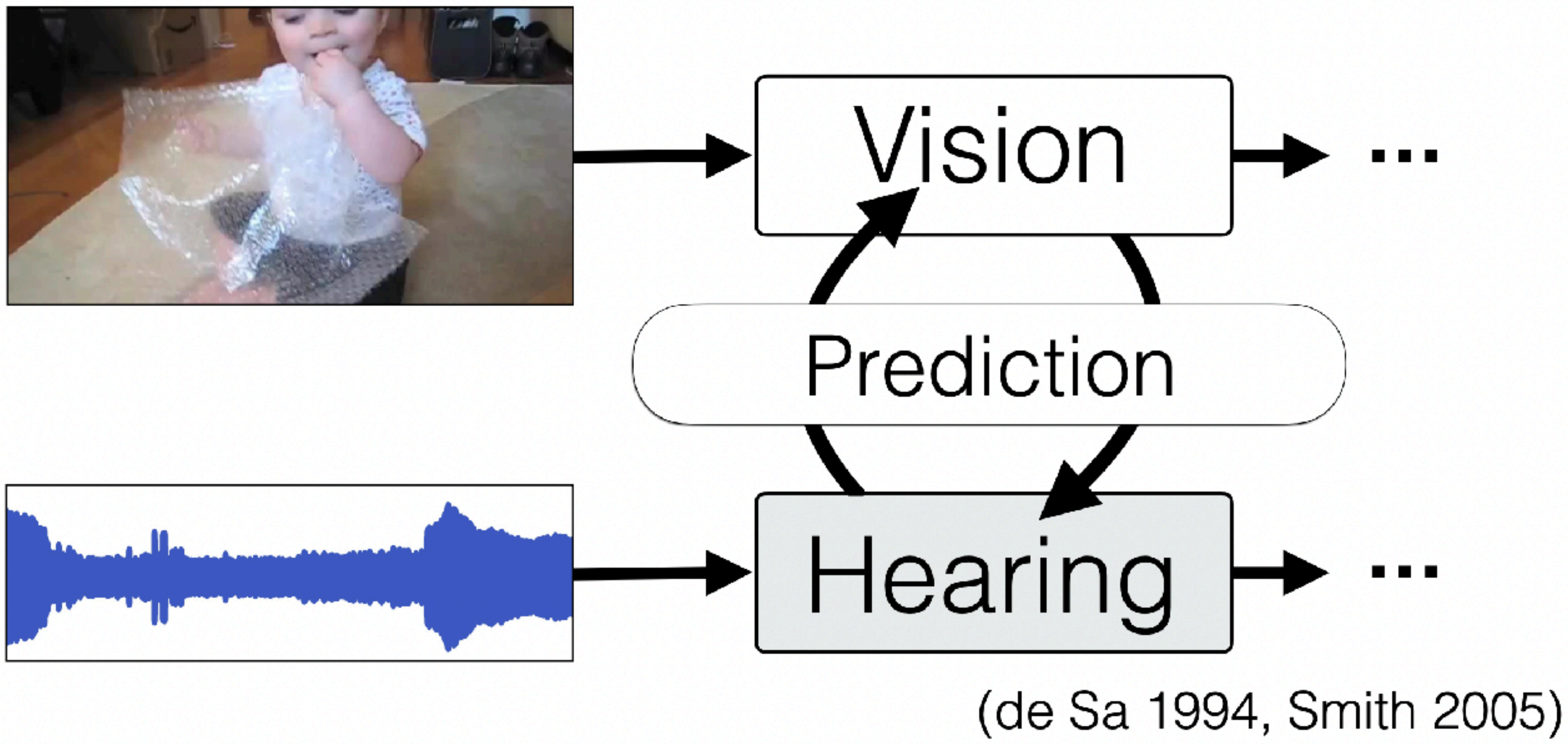


Two-Stream Convolutional Networks for Action Recognition in Videos. Simonyan and Zisserman. NeurIPS 2014
Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Carreira and Zisserman. CVPR 2017

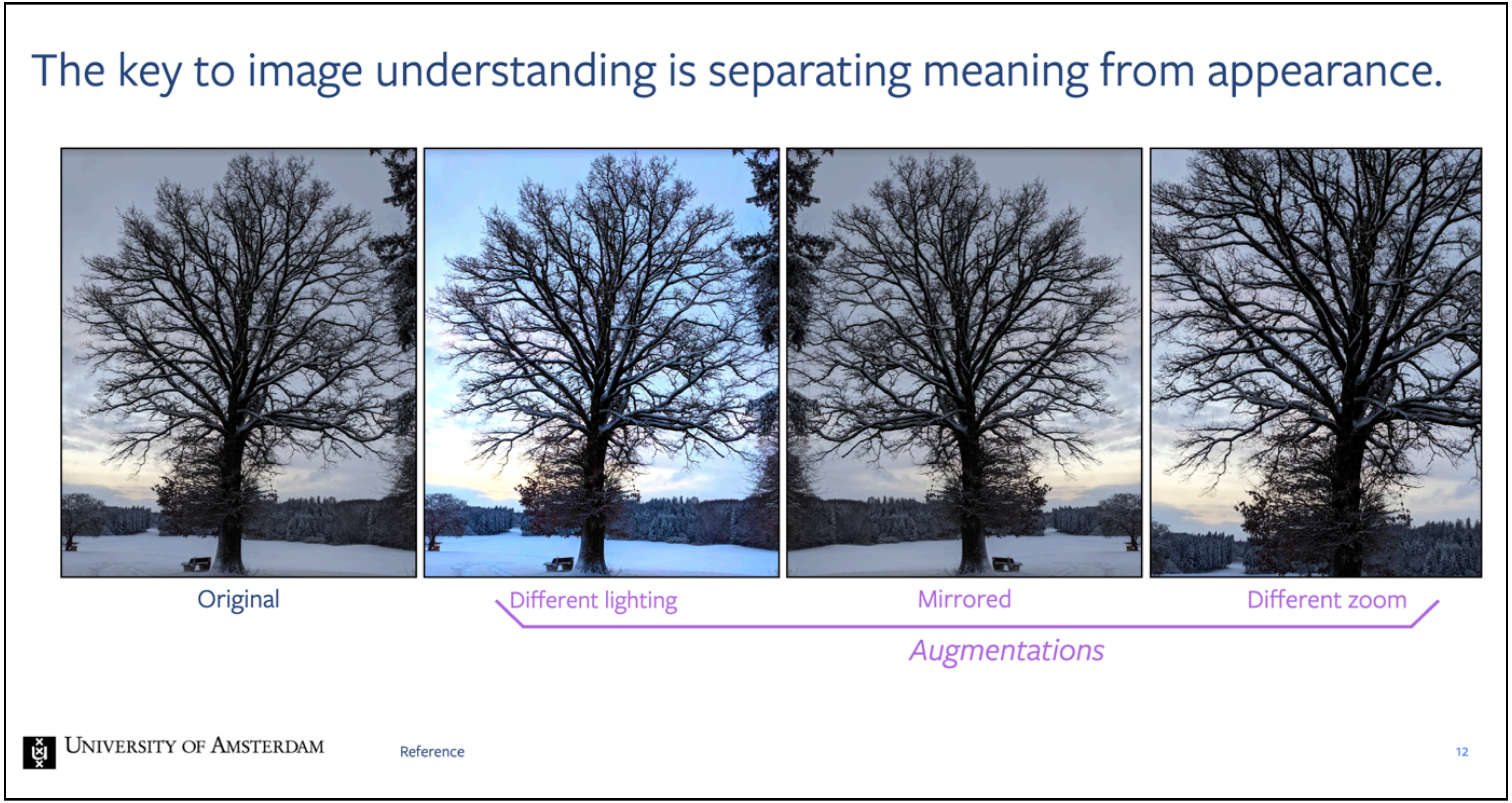
Self-supervision is the natural way to learn from paired data….



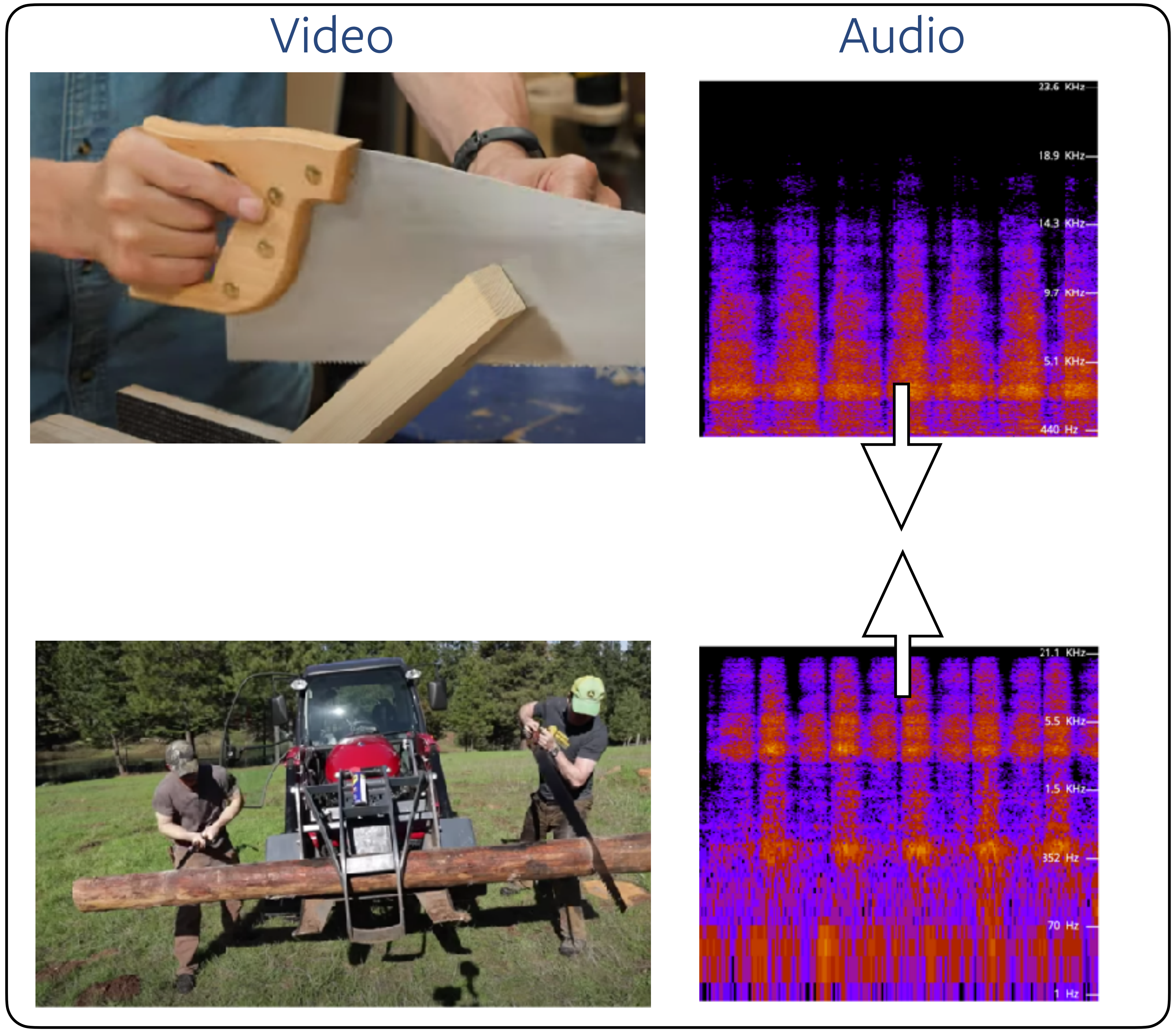
…whereby one modality is used as supervision signal for the other

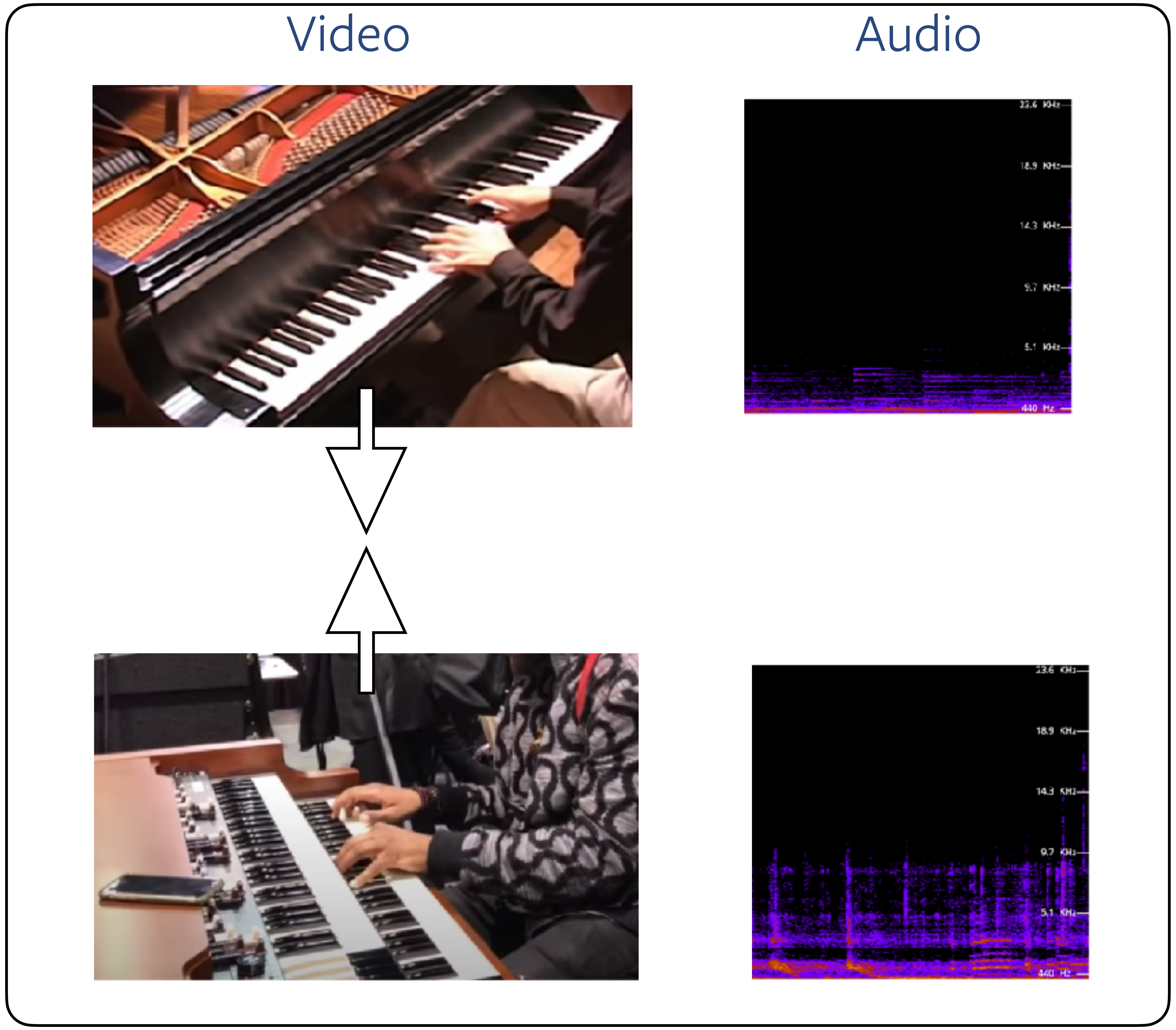


Remember this slide?



Multiple modalities can also yield such useful semantic information.





https://www.youtube.com/watch?v={ZlS8t8_nAHo, tvVYdaykaZM, Kt8_u_i-anQ, J4X6tc1HYts}

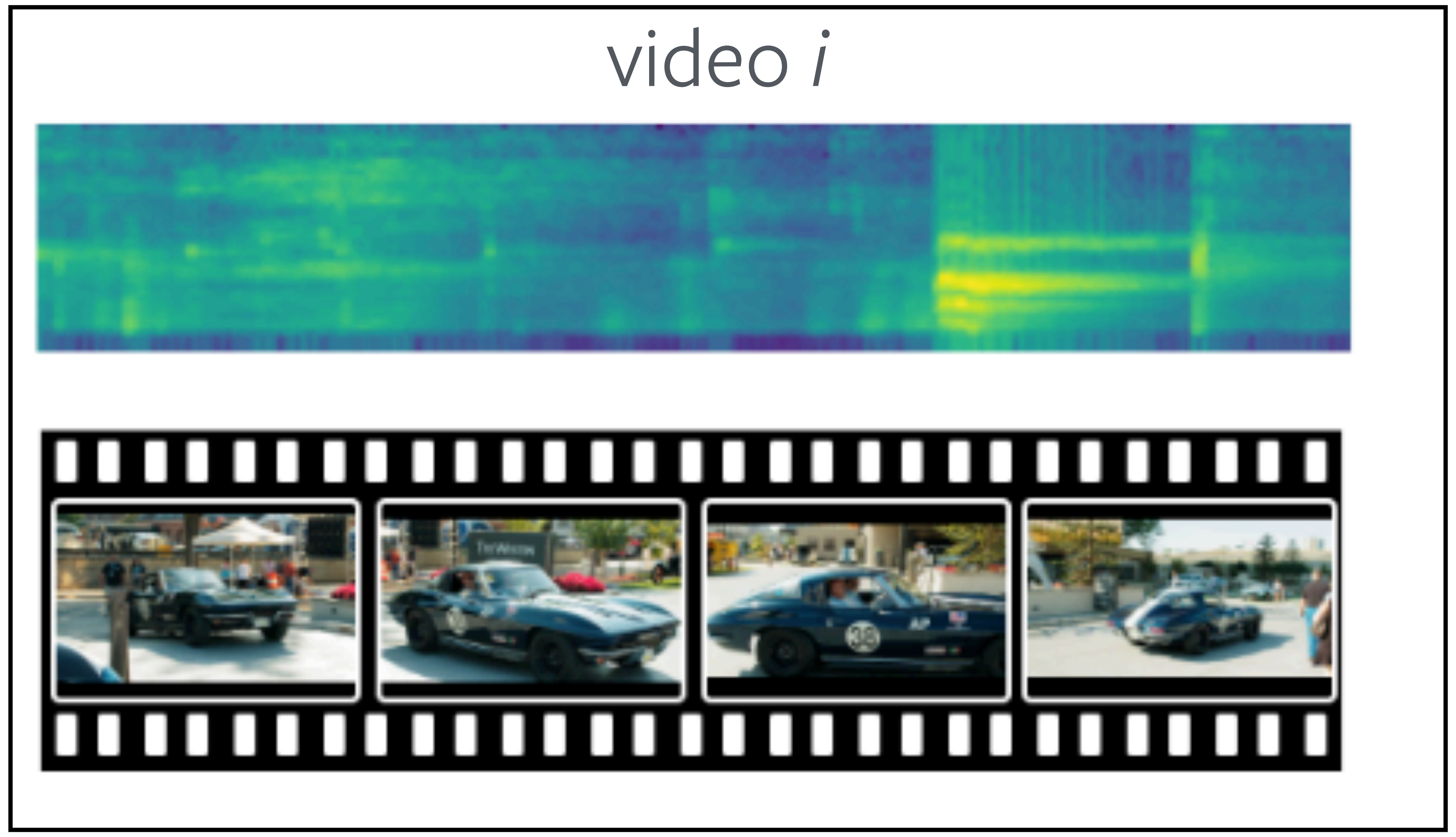
Why clustering for videos?

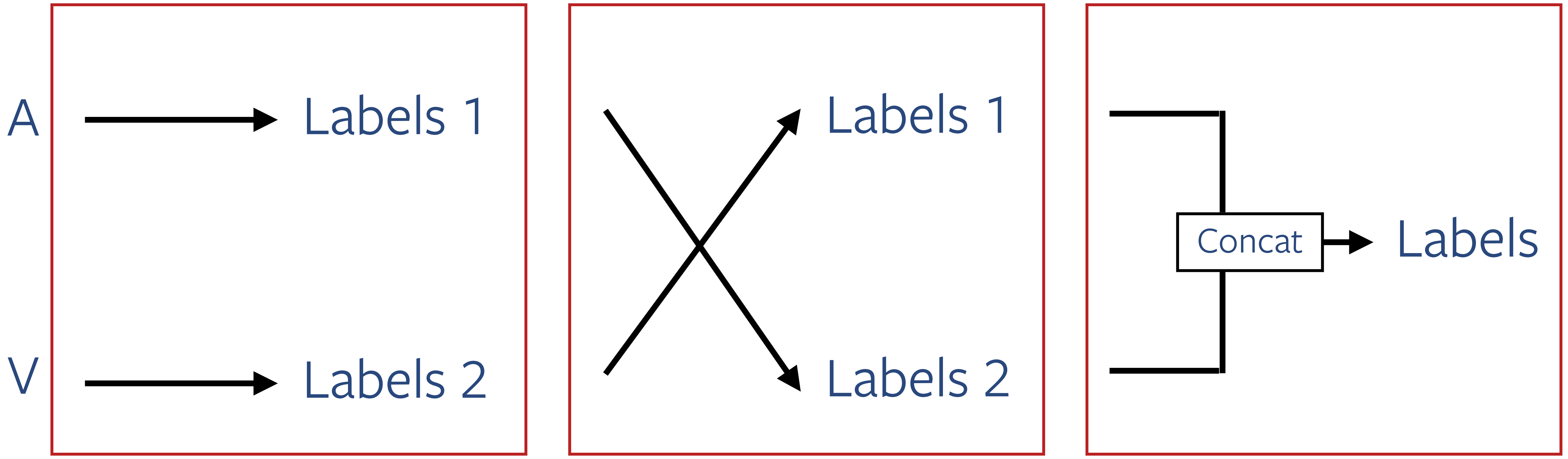
🤔 Clustering works well for images

🤑 Videos are expensive to annotate.

📈Video content is rapidly increasingly.

Clustering multi-modal data



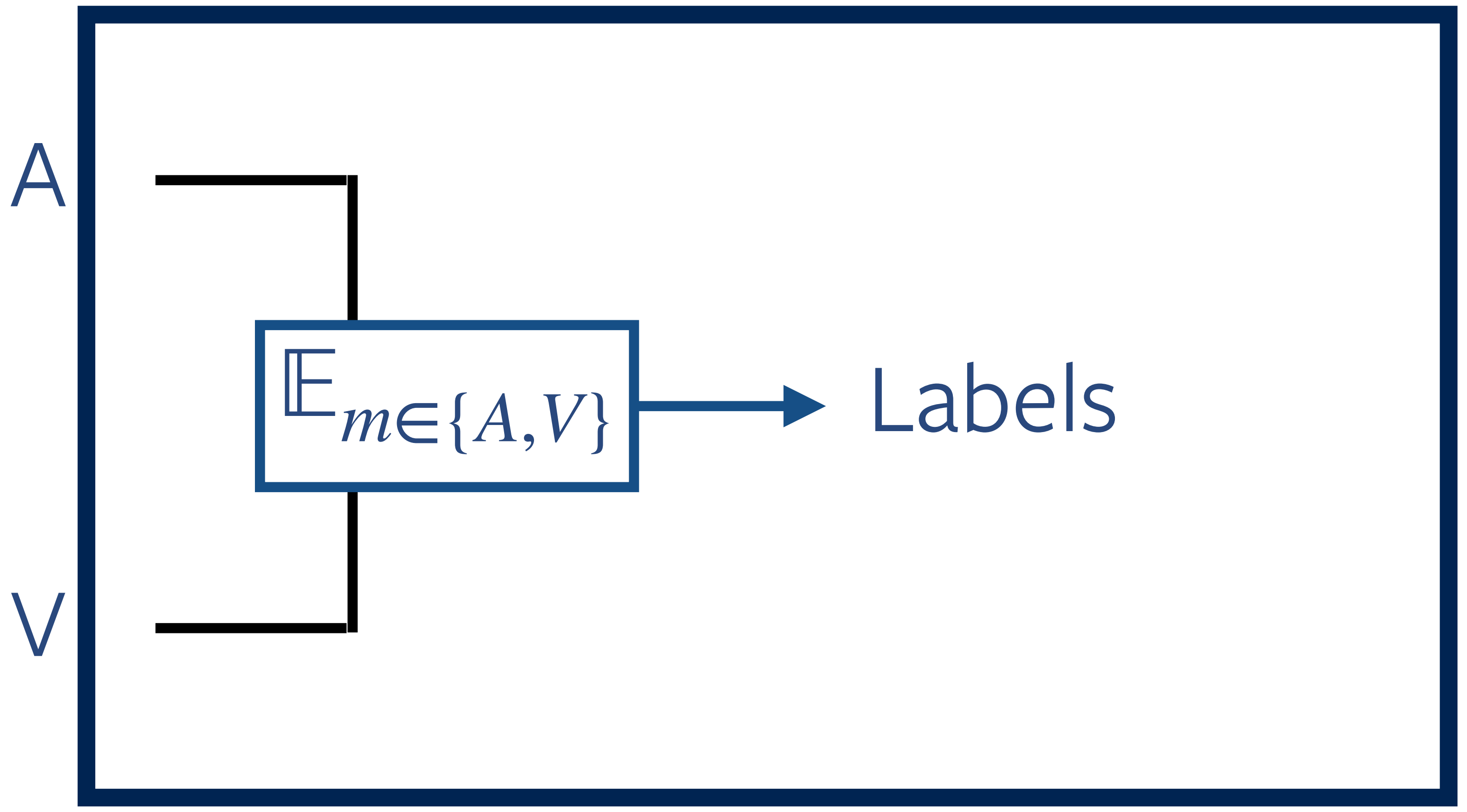


- × does not use same-source information
- × two different sets of clusters

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

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

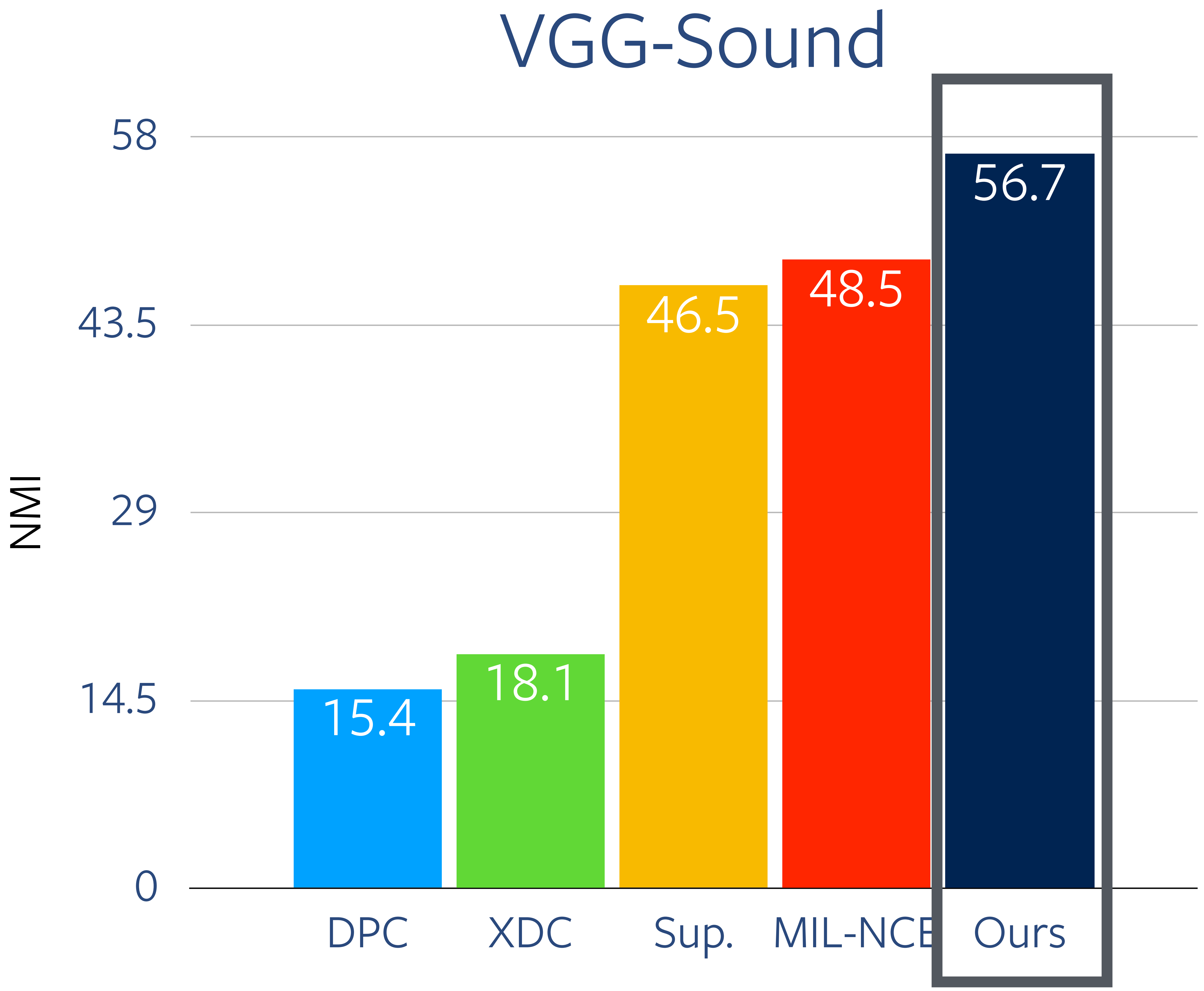
Audio

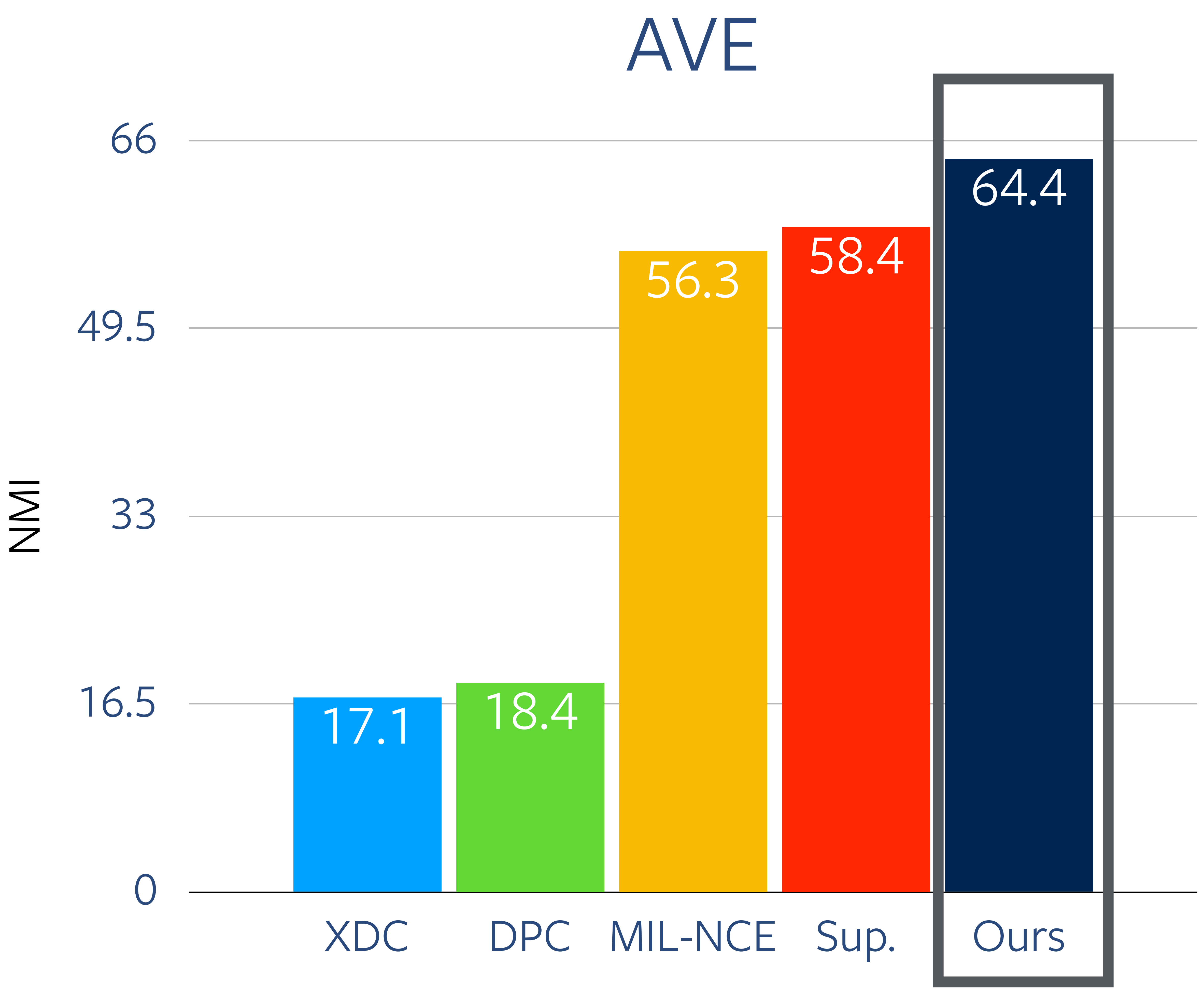


Our idea:
view each modality as an *augmentation*

Asano et al. *Labelling unlabelled videos from scratch with multi-modal self-supervision.* NeurIPS 2020

Good feature representations ↛ good clustering



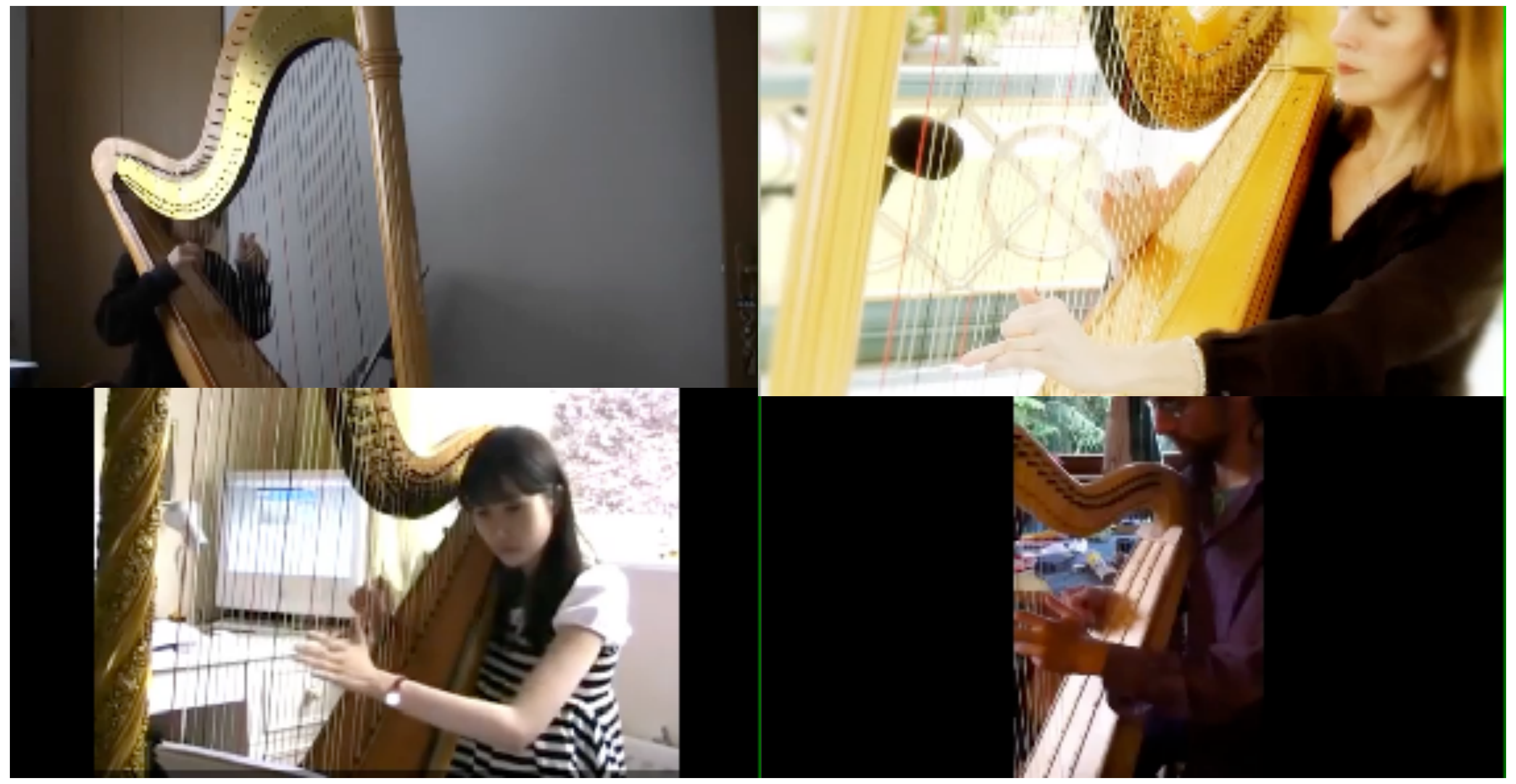


[Han et al., arXiv; Alwassel et al., NeurIPS 2020; Miech et al., CVPR 2020]

UNIVERSITY OF AMSTERDAM

Asano et al. *Labelling unlabelled videos from scratch with multi-modal self-supervision.* NeurIPS 2020

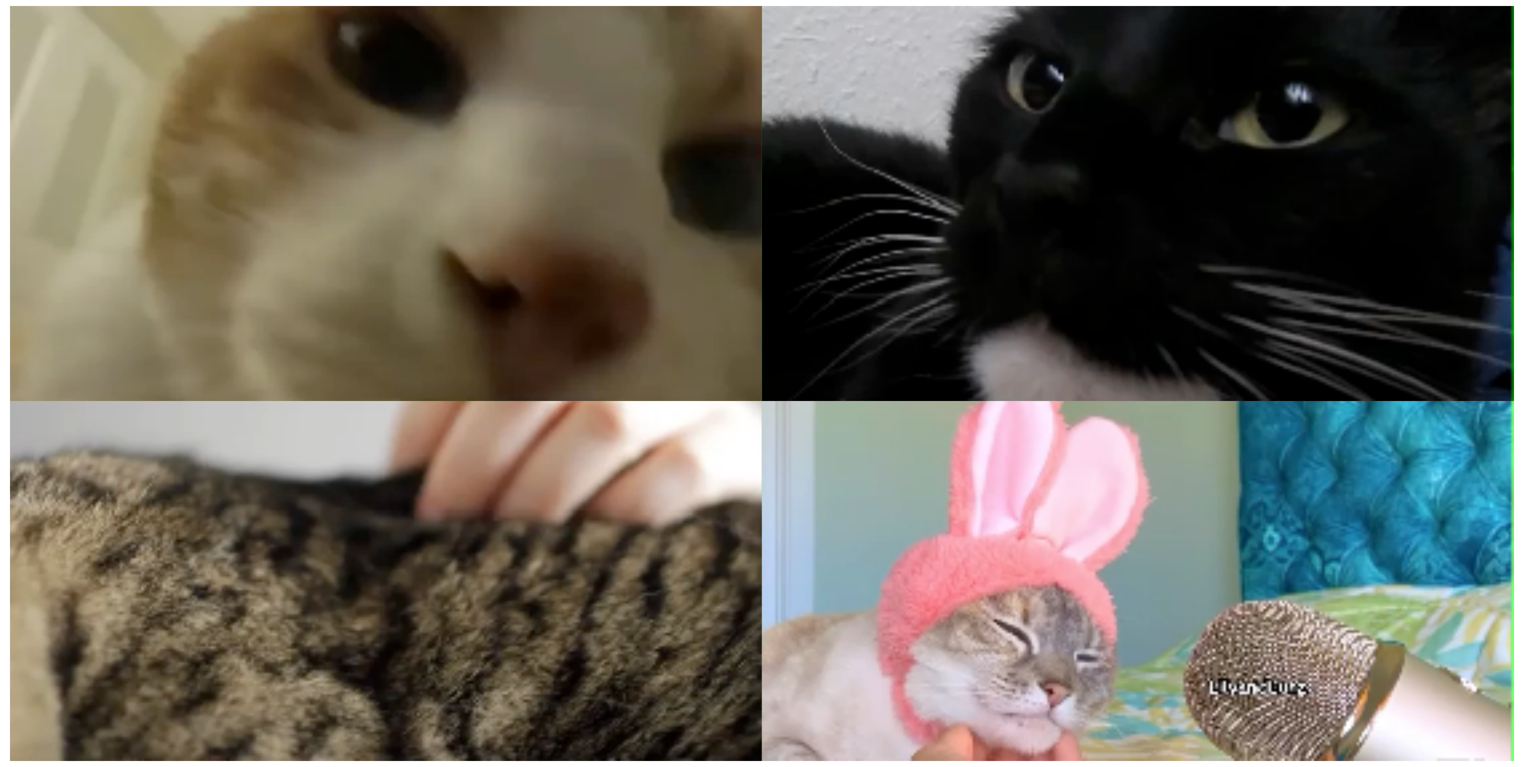
Discovering concepts without manual annotations from 230K videos



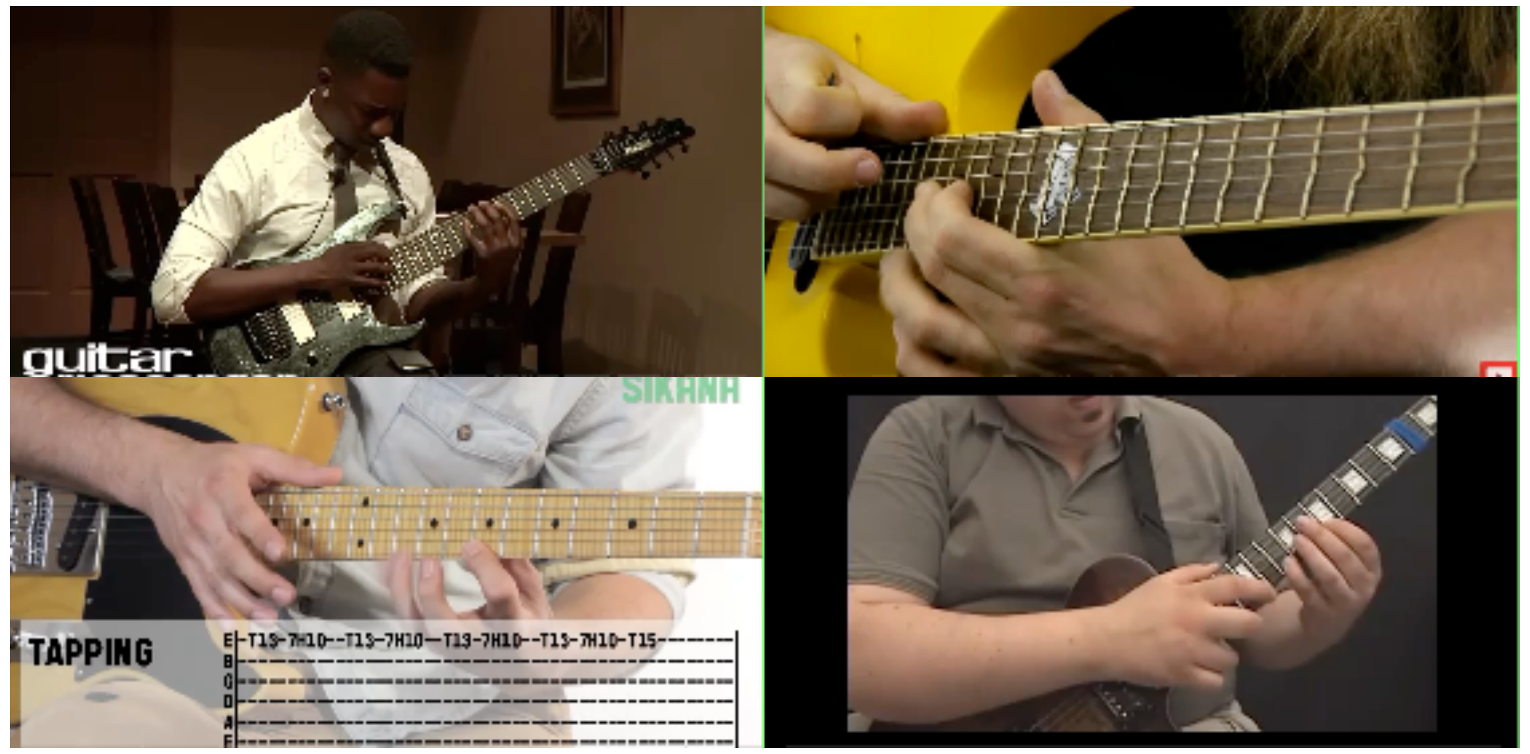
"Playing harp"

"Fireworks"

"Hockey game"



"Cat growling"



"Electric guitar"

"Vehicle driving"

Asano et al. *Labelling unlabelled videos from scratch with multi-modal self-supervision.* NeurIPS 2020

"Self-Labelling" 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

Asano et al. Labelling unlabelled videos from scratch with multi-modal self-supervision. NeurIPS 2020

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

Pretraining:

Kinetics-400 (600/700)
miniKinetics
VGGSound
HowTo100M
AudioSet
Youtube8M

Evaluation:

Kinetics/HMDB-51/UCF-101
SomethingSomething-v2
Oops
MSRVTT/VATEX/DiDeMo/ActivityNet
Youtube-VOS/VIS

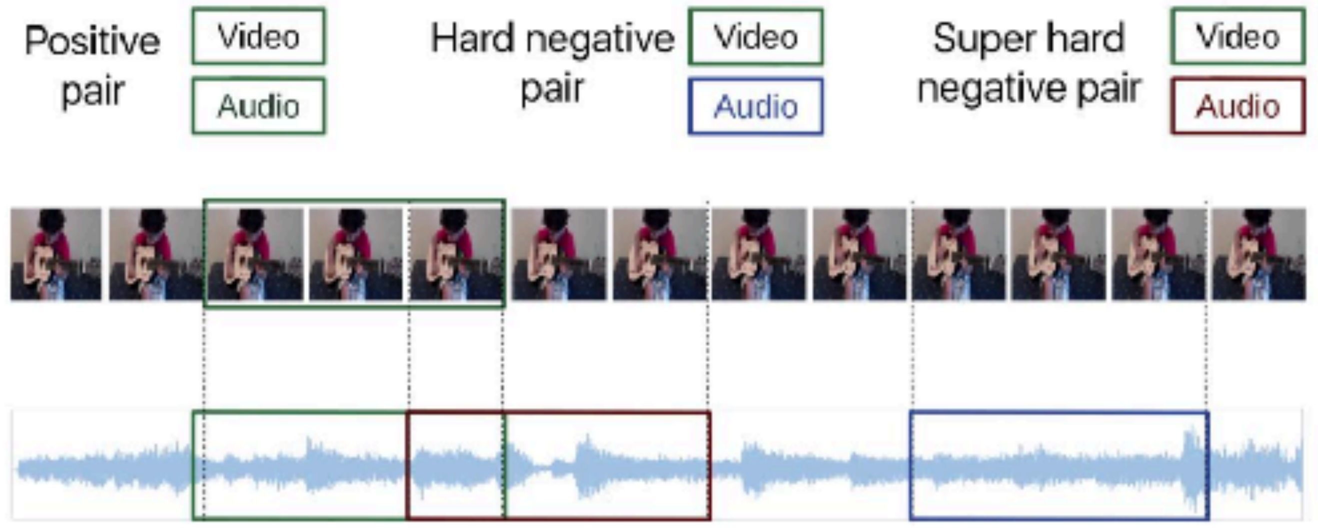
Explore some datasets:
https://www.robots.ox.ac.uk/~vgg/research/selavi/#demo

For fun: the Oops dataset

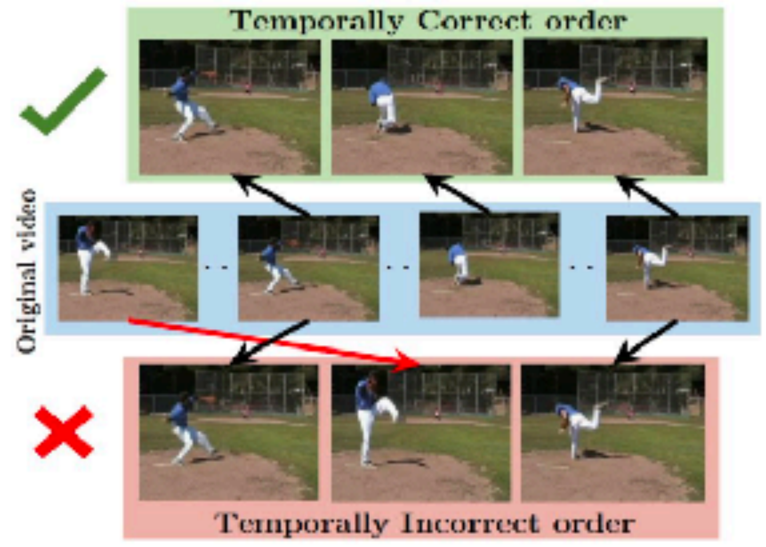
We present the Oops dataset for studying unintentional human action. The dataset consists of 20,723 videos from YouTube fail compilation videos, adding up to over 50 hours of data. These clips, filmed by amateur videographers in the real world, are diverse in action, environment, and intention. The dataset covers many causes of failure and unintentional action, including physical and social errors, errors in planning and execution, limited agent skill, knowledge, or perceptual ability, and environmental factors.

https://oops.cs.columbia.edu/data/ CVPR 2020 paper

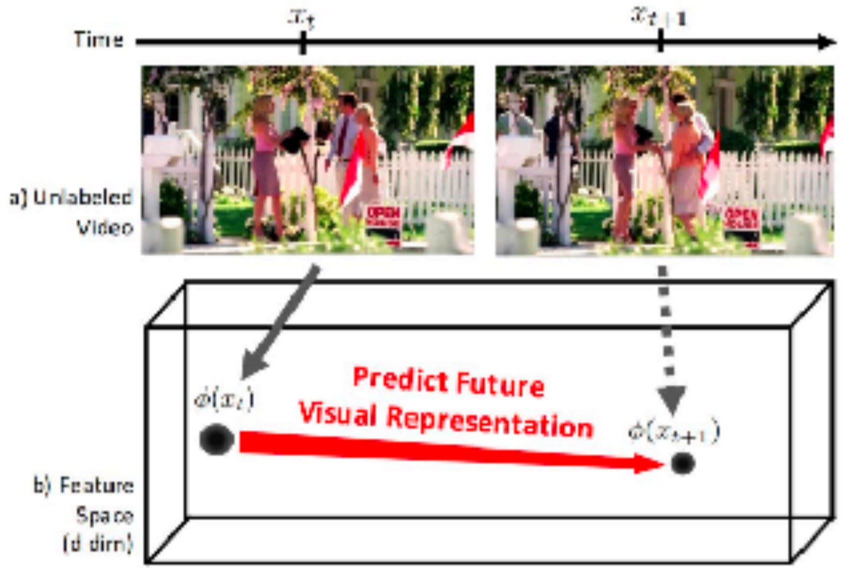
Other popular self-supervised learning approaches



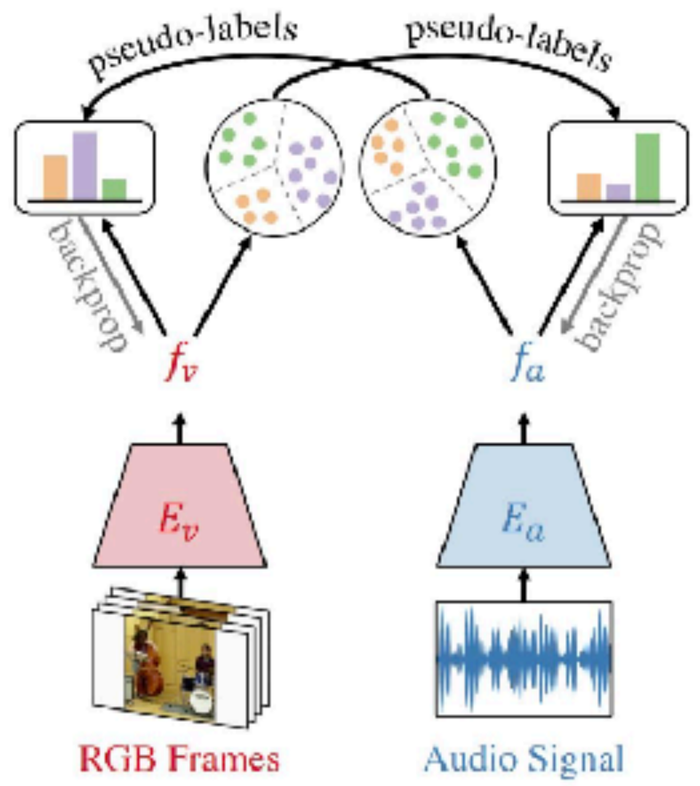
Audio-Video Synchronization. Korbar et al.



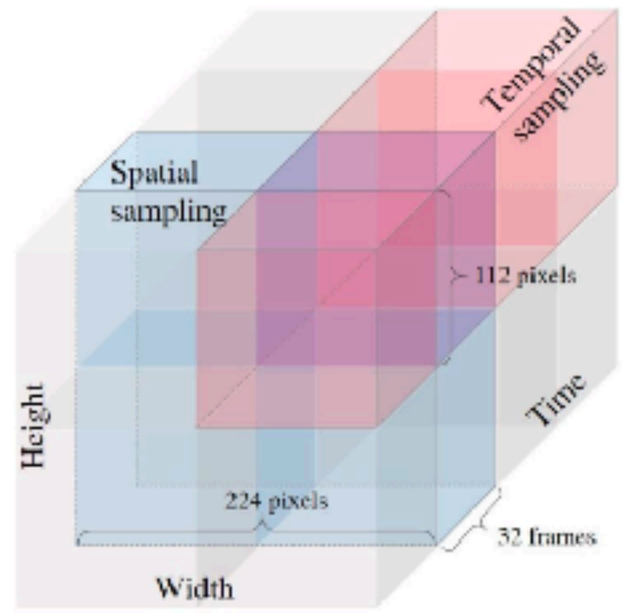
Shuffle & Learn. Misra et al.



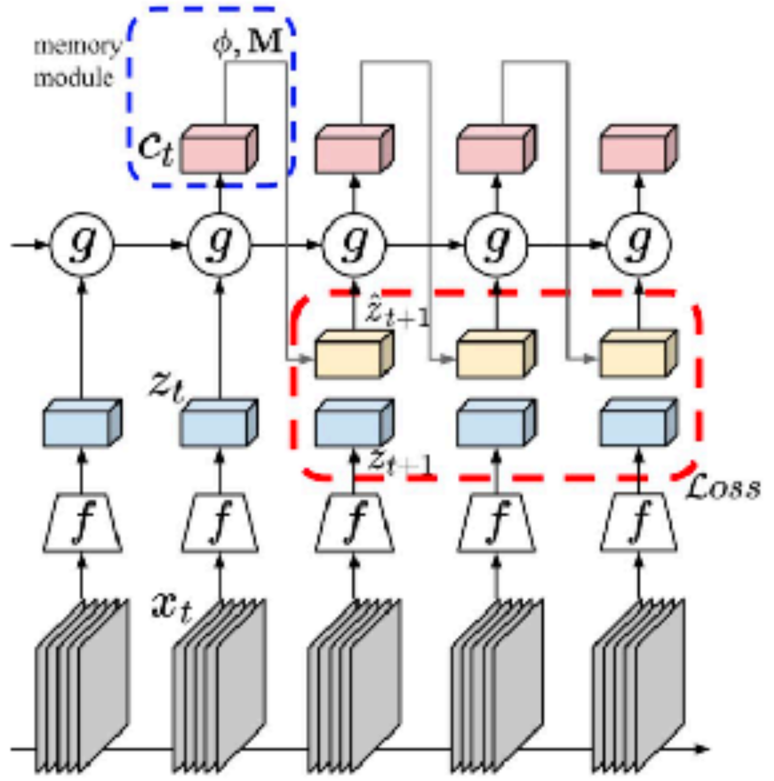
Anticipating Representation. Vondrick et al.



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



3D ST-Puzzle. Kim et al.



DPC/MemDPC. Han 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

Invaniance *vs* distinctiveness

In contrastive learning, we define positives and negatives.

Should the representations enforce invariance or distinctiveness?

$\stackrel{?}{=}$

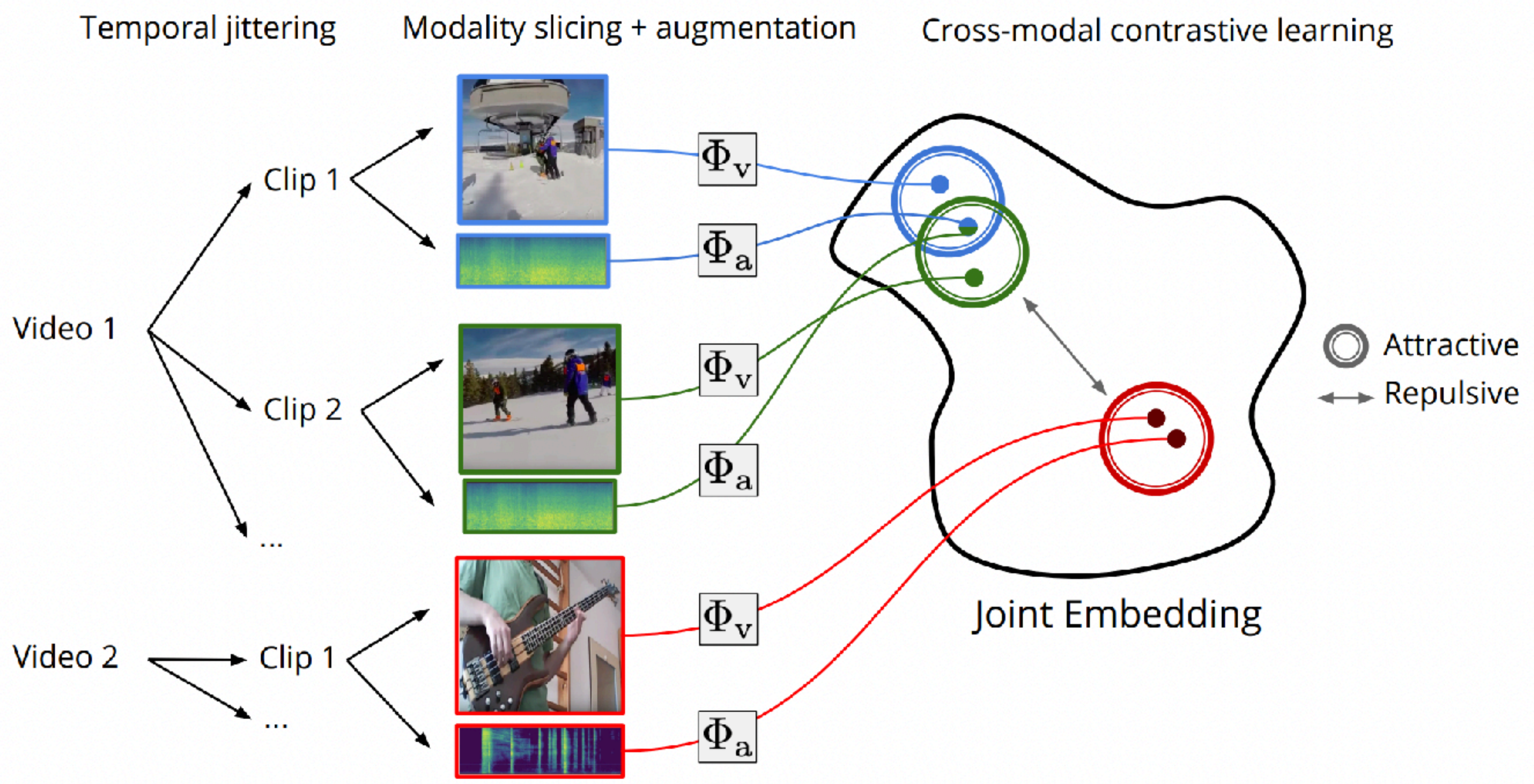
Learning hypotheses we test:

1. Sample Distinctiveness

2. Time Reversal

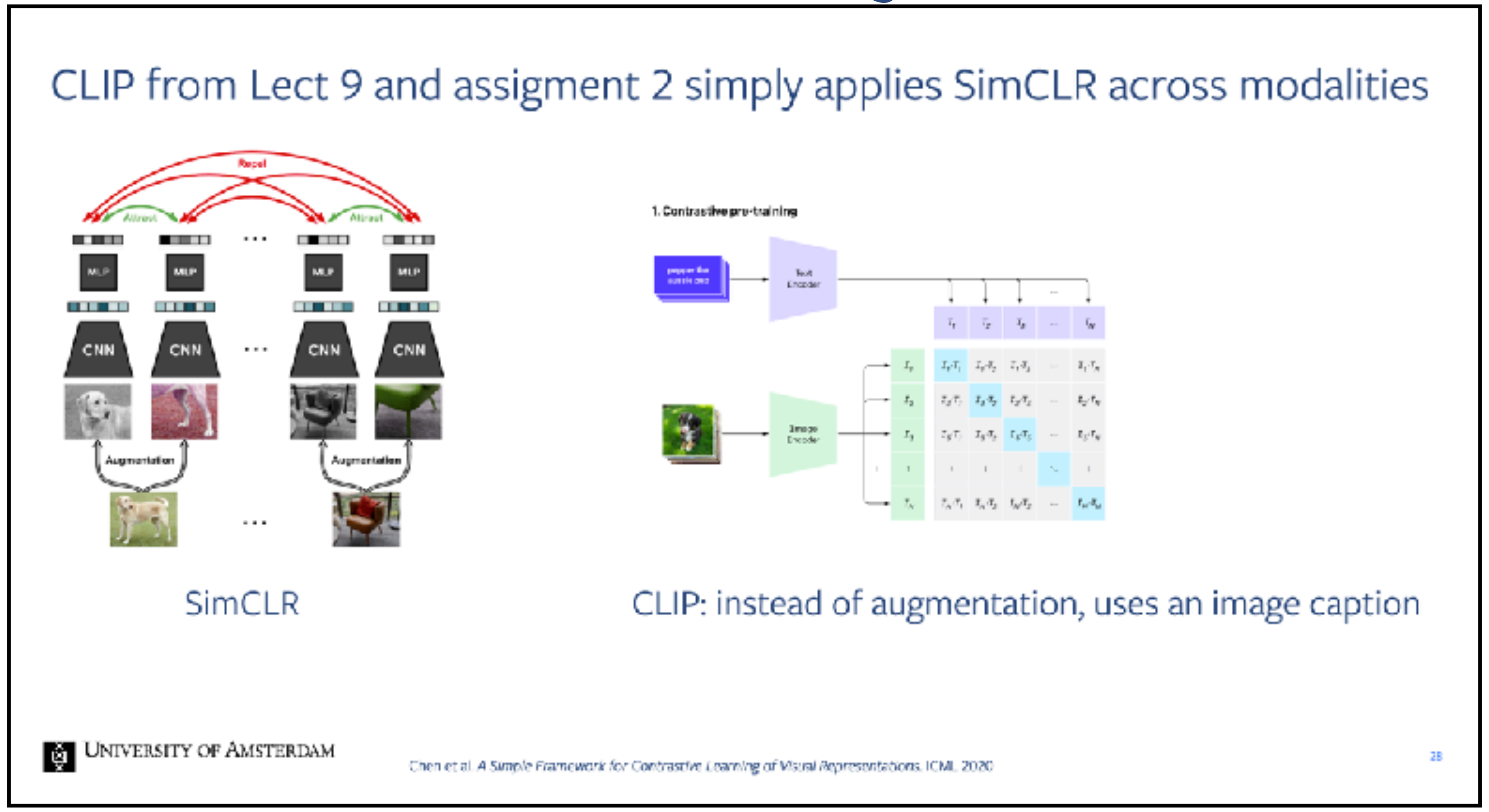
3. Time Shift

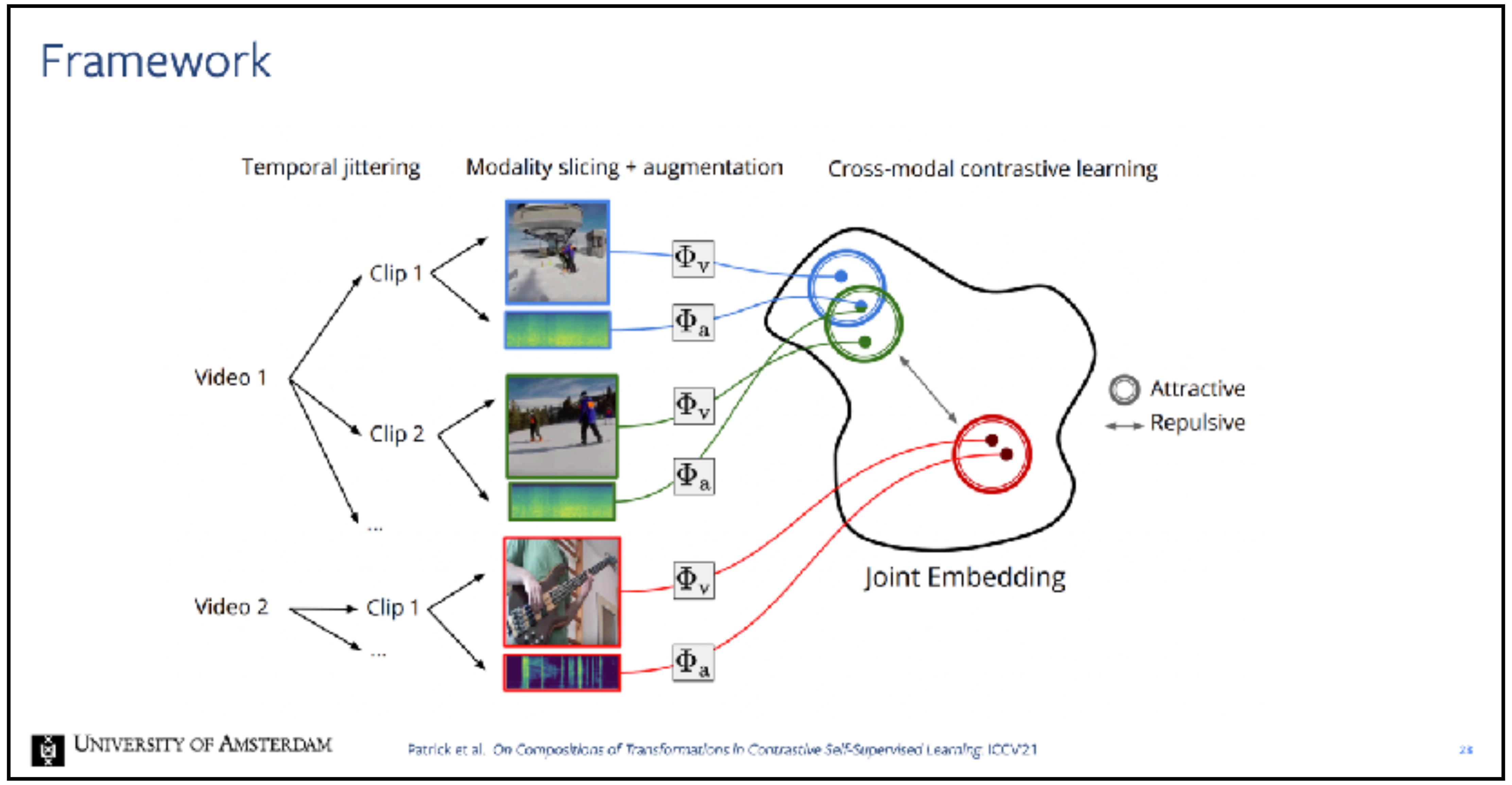
Framework



Patrick et al. *On Compositions of Transformations in Contrastive Self-Supervised Learning*. ICCV'21

Notice the similarity to SimCLR:

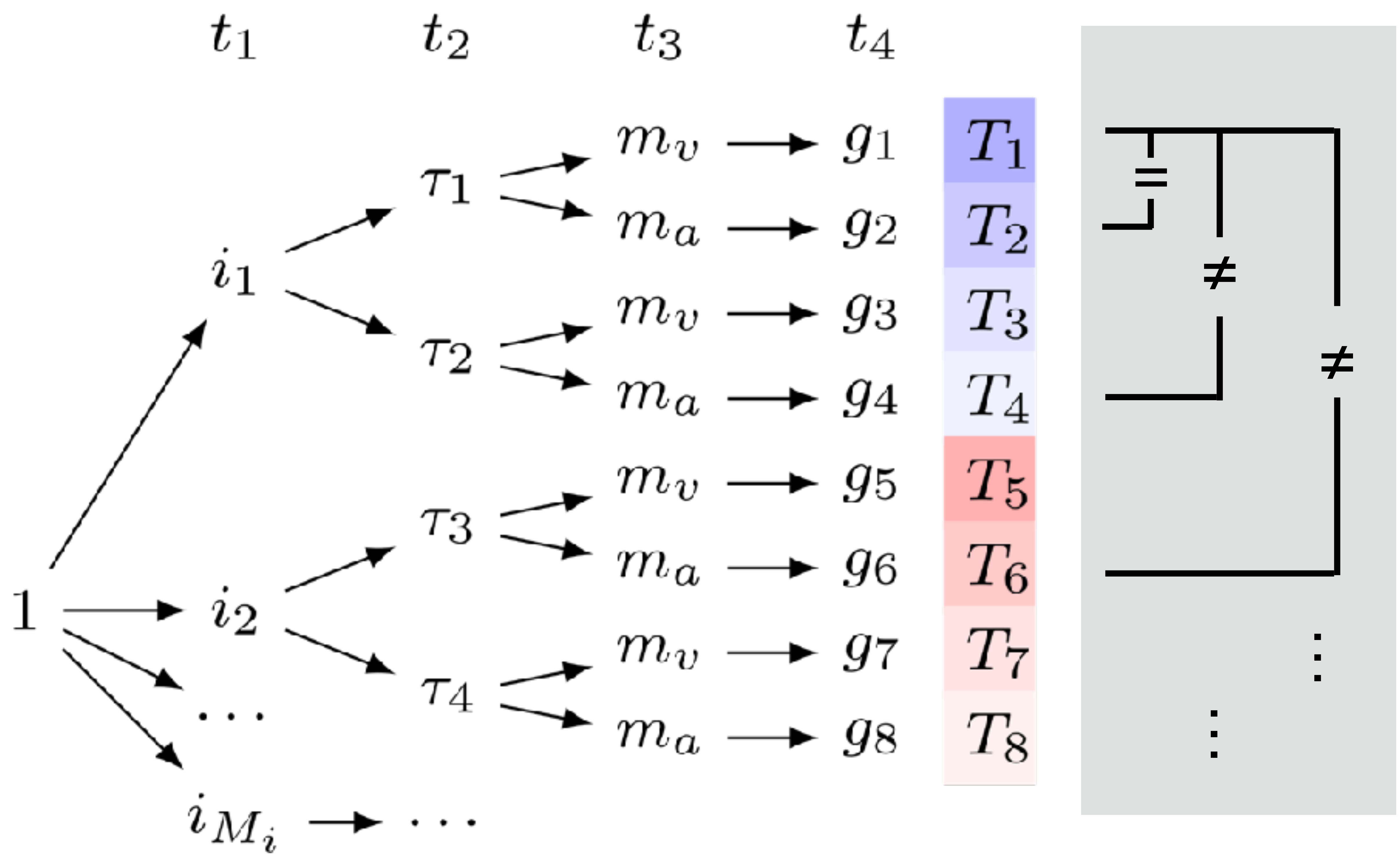




Quiz: this work resembles "simply applying SimCLR to audio-video", what does one need to be careful about?

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

Variance and invariances: distinctive to sample & time shift, invariant to modality



Same embedding for video and audio at *same* time from *same* video

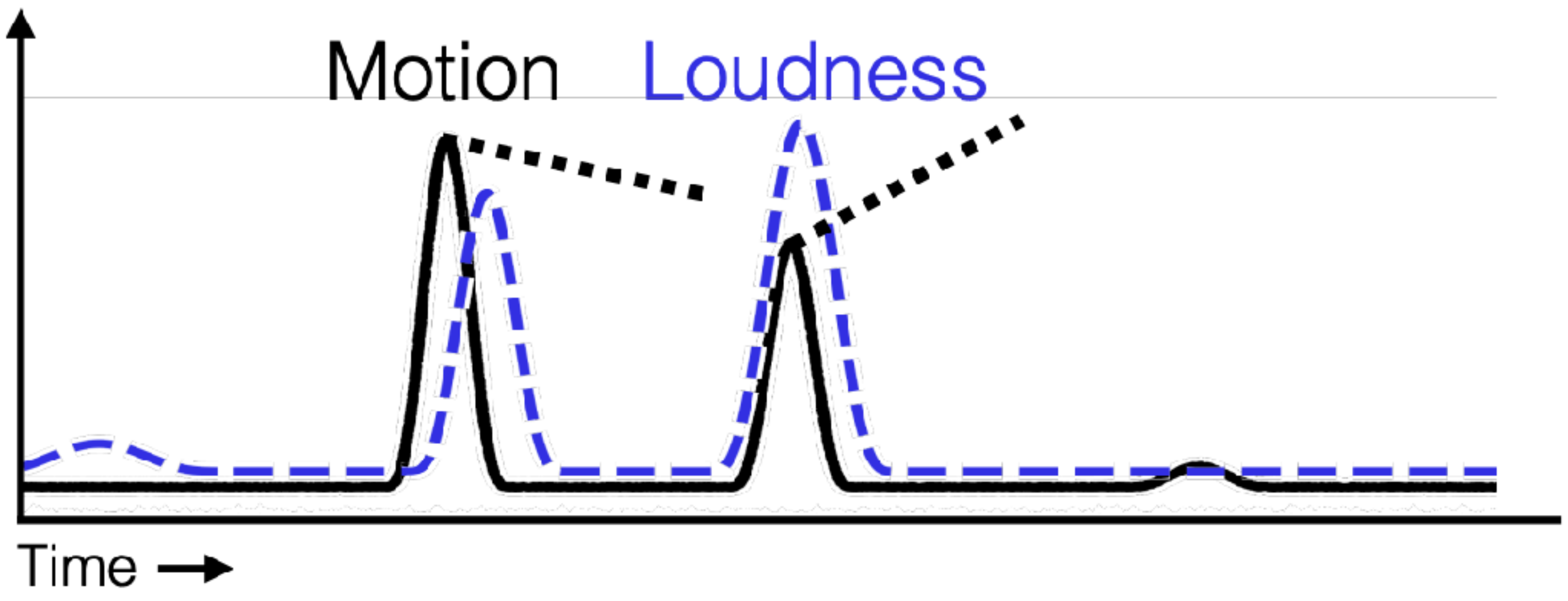
Different embedding for *different* time-indices from the *same* video

Different embedding for *different* video

Cross-modally

Patrick et al. *On Compositions of Transformations in Contrastive Self-Supervised Learning*. ICCV'21

Detecting time-shifted pairs as negatives is not easy

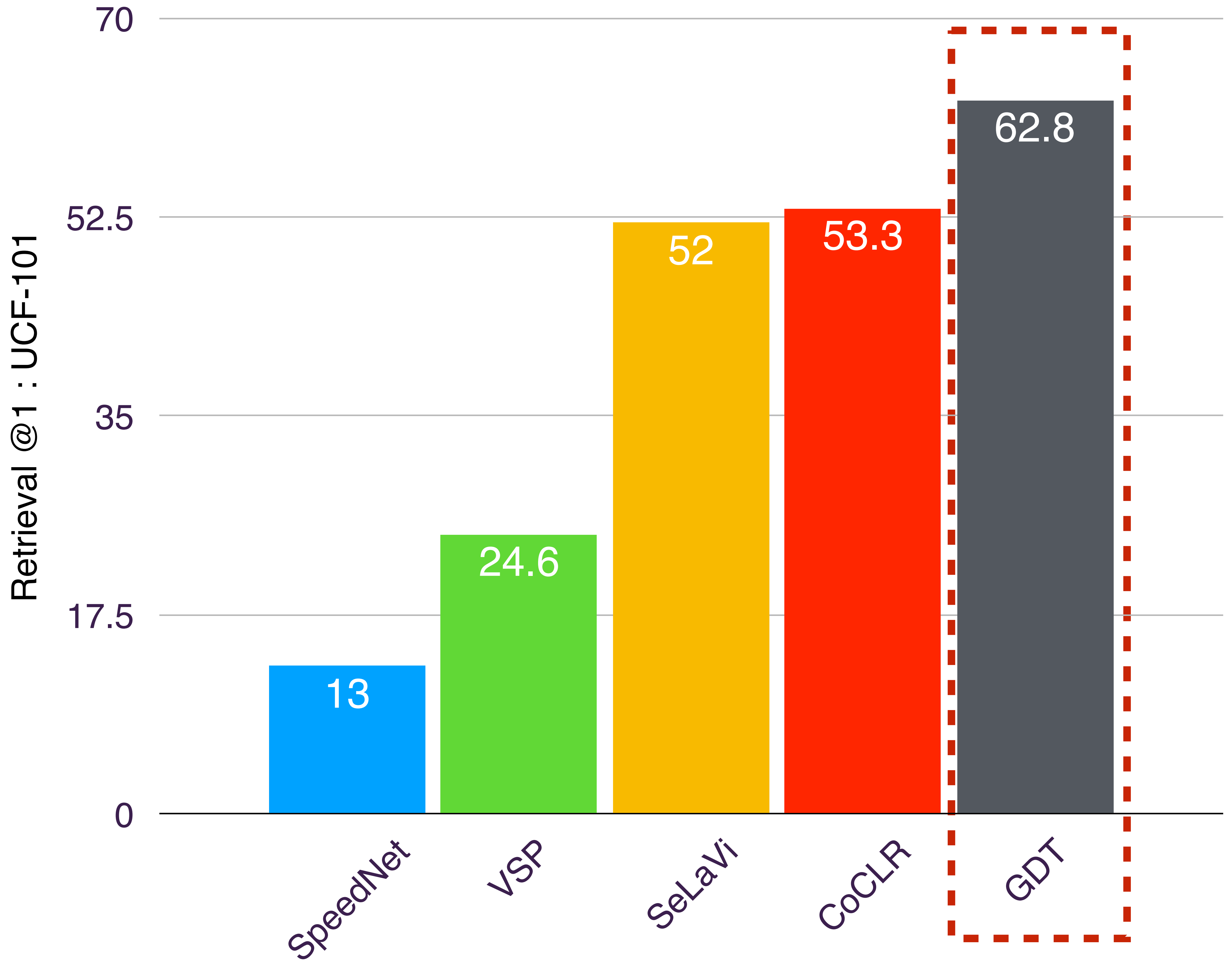


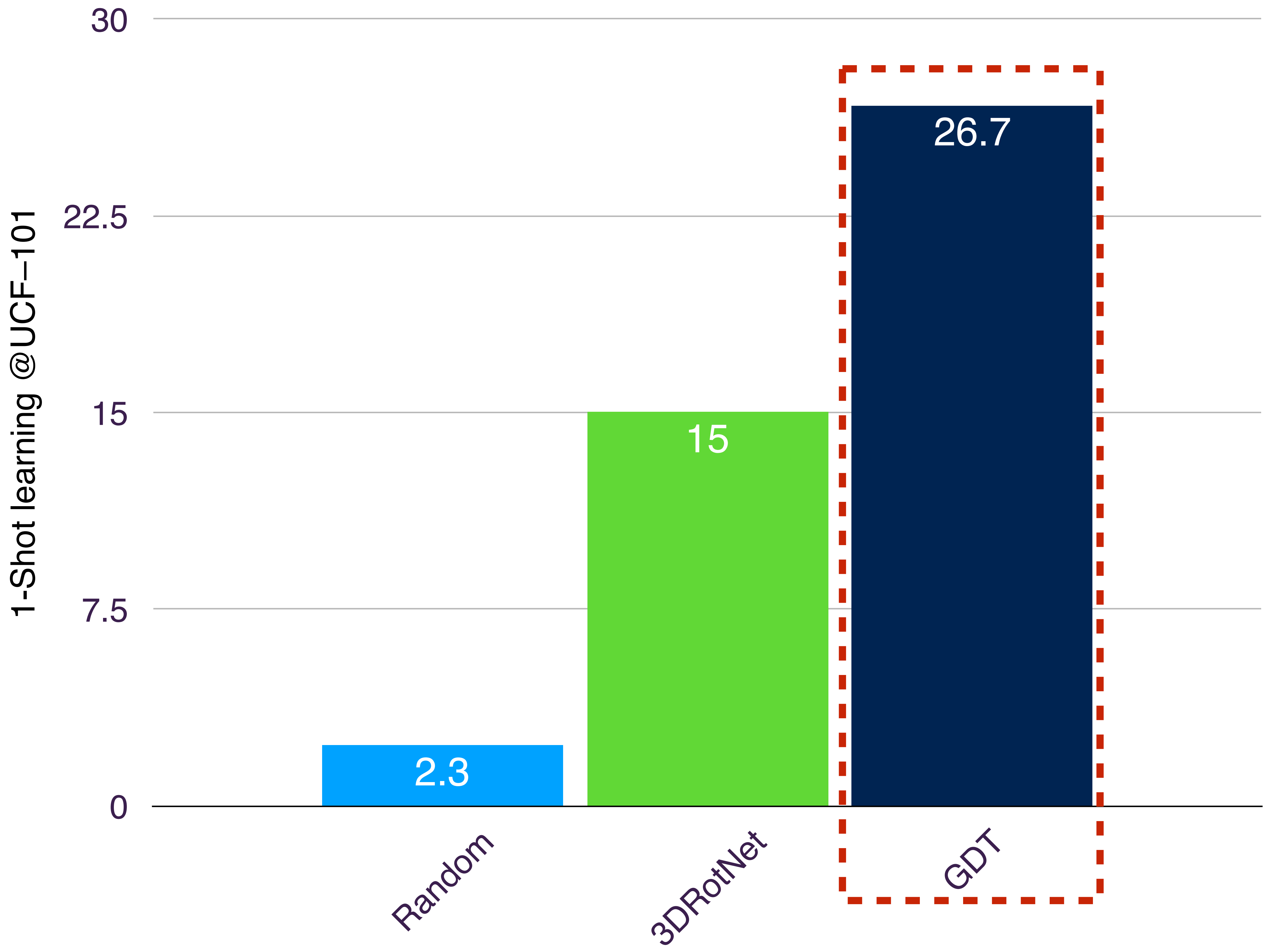
- The scene-level semantics match, but the timing doesn't
- This requires fine-grained recognition
- However, there's work* that shows time can also be used as an augmentation to some degree

*A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. Feichtenhofer et al. CVPR 2021

Gains from hypotheses

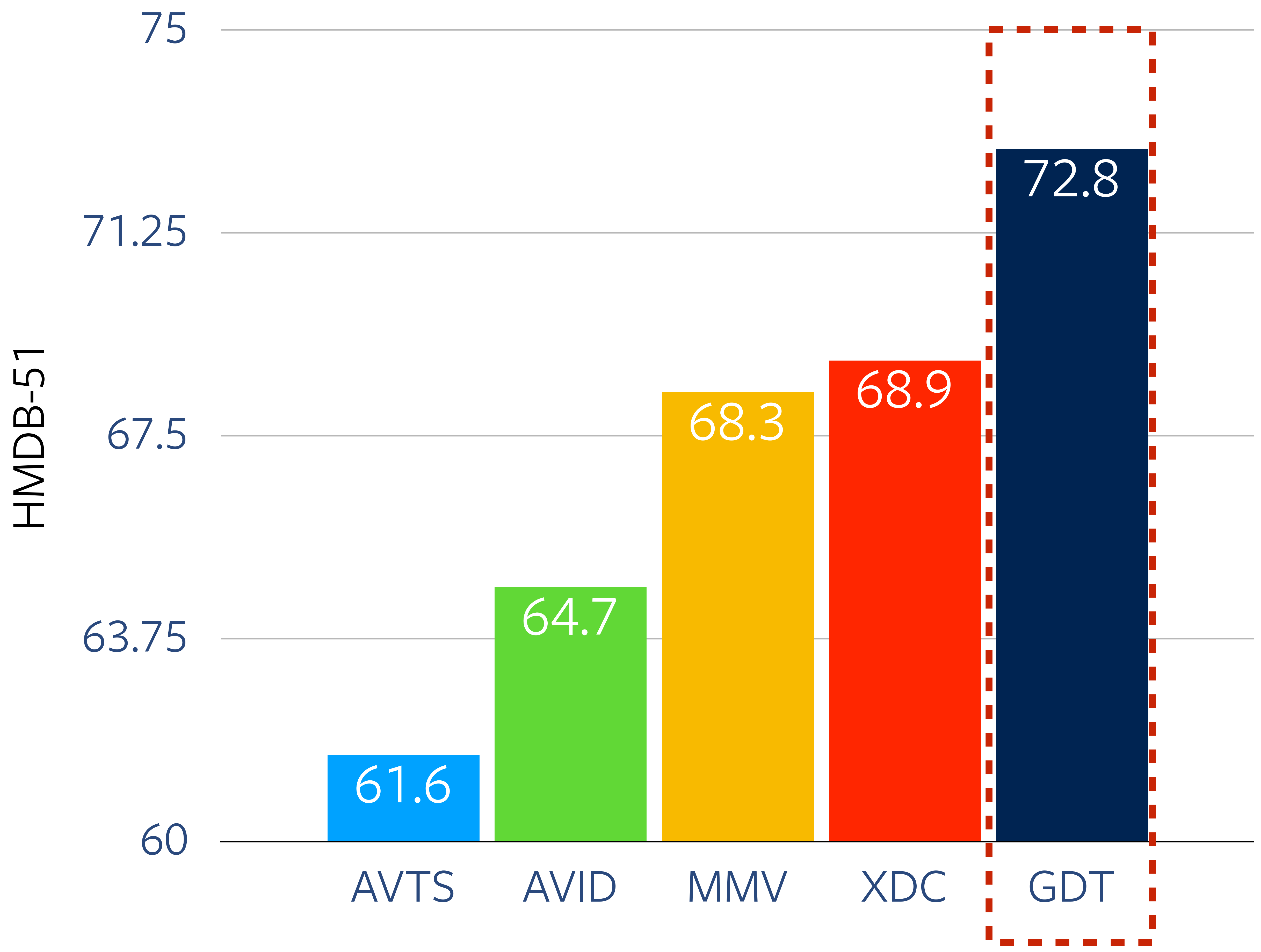
SOTA video action retrieval and few-shot learning results

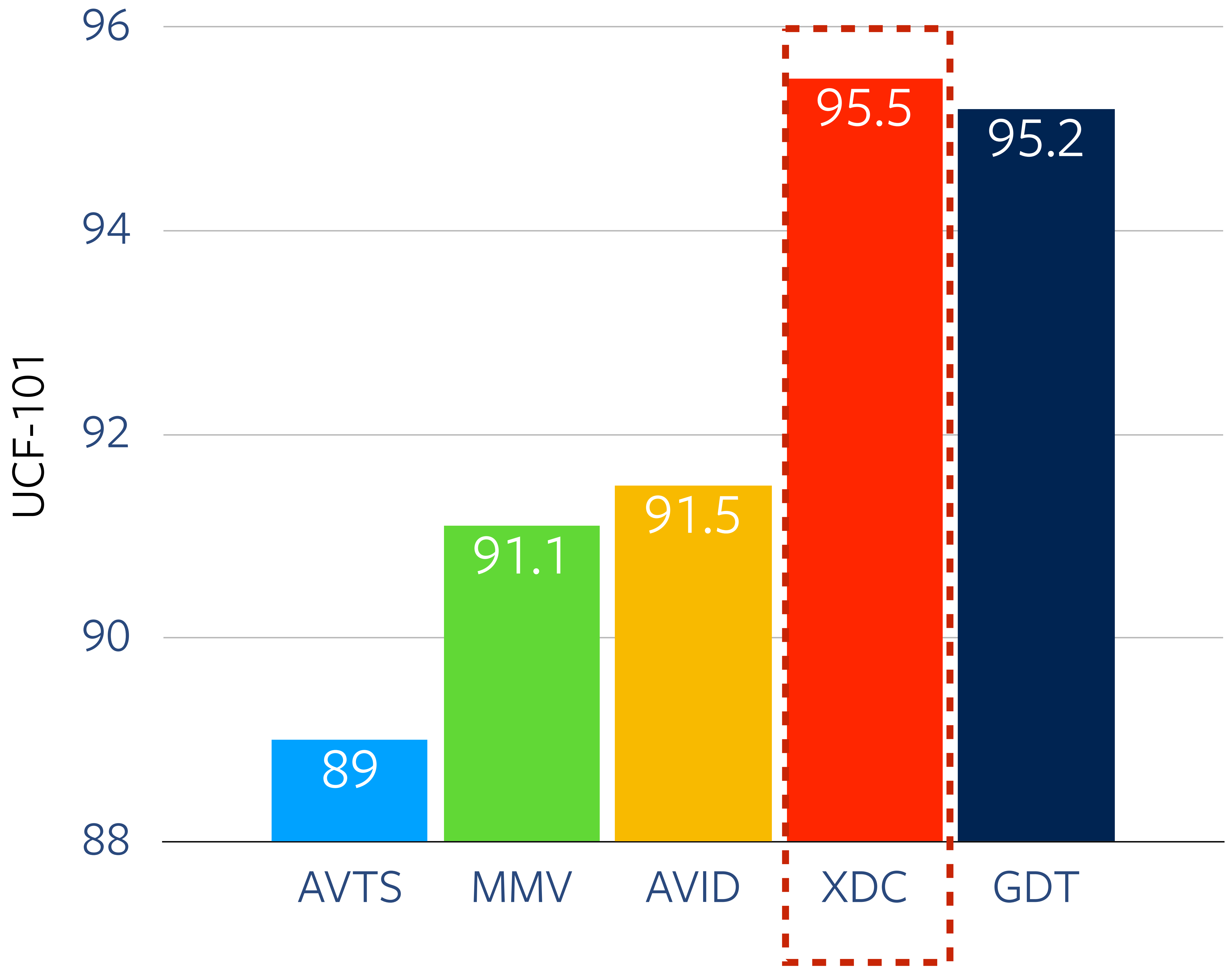




[Benaim et al, CVPR 2020; Cho et al., ArXiv; Asano et al., NeurIPS 2020; Han et al., NeurIPS 2020]
[Jing and Tian, arXiv]

SOTA finetuning video-action recognition results



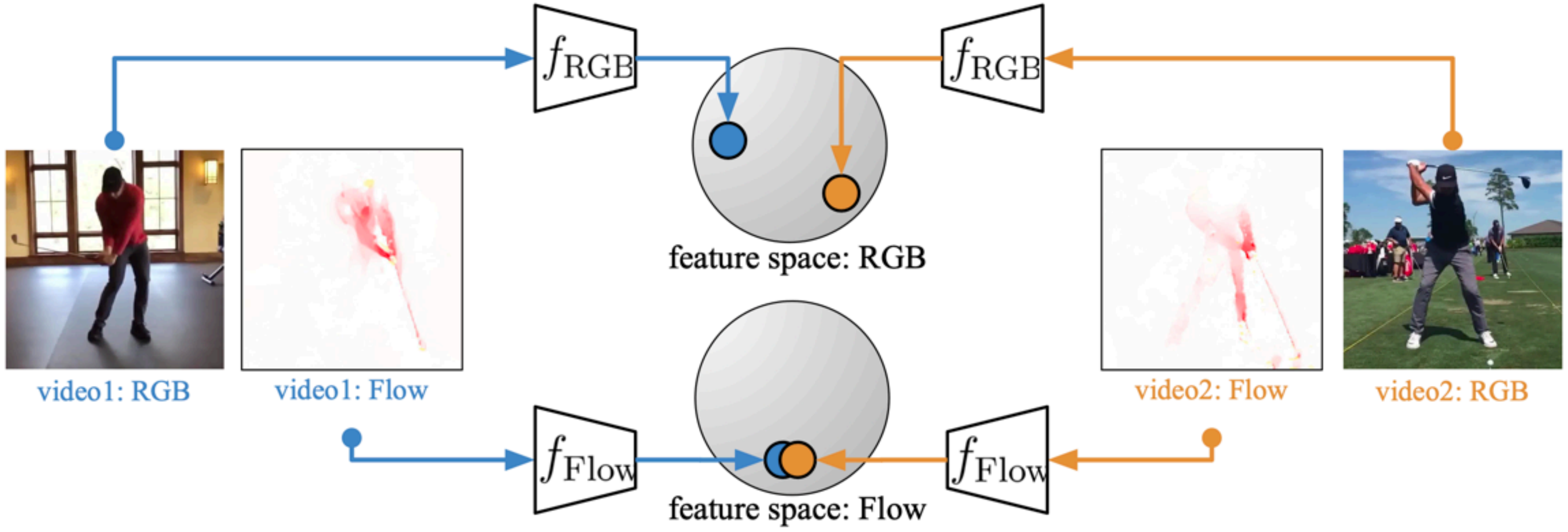


[Bruno et al., NeurIPS 2018; Morgado et al., CVPR 2021; Alayrac el al., NeurIPS 2020; Alwassel et al., NeurIPS 2020]

Another work that generalises SimCLR to more modalities: CoCLR

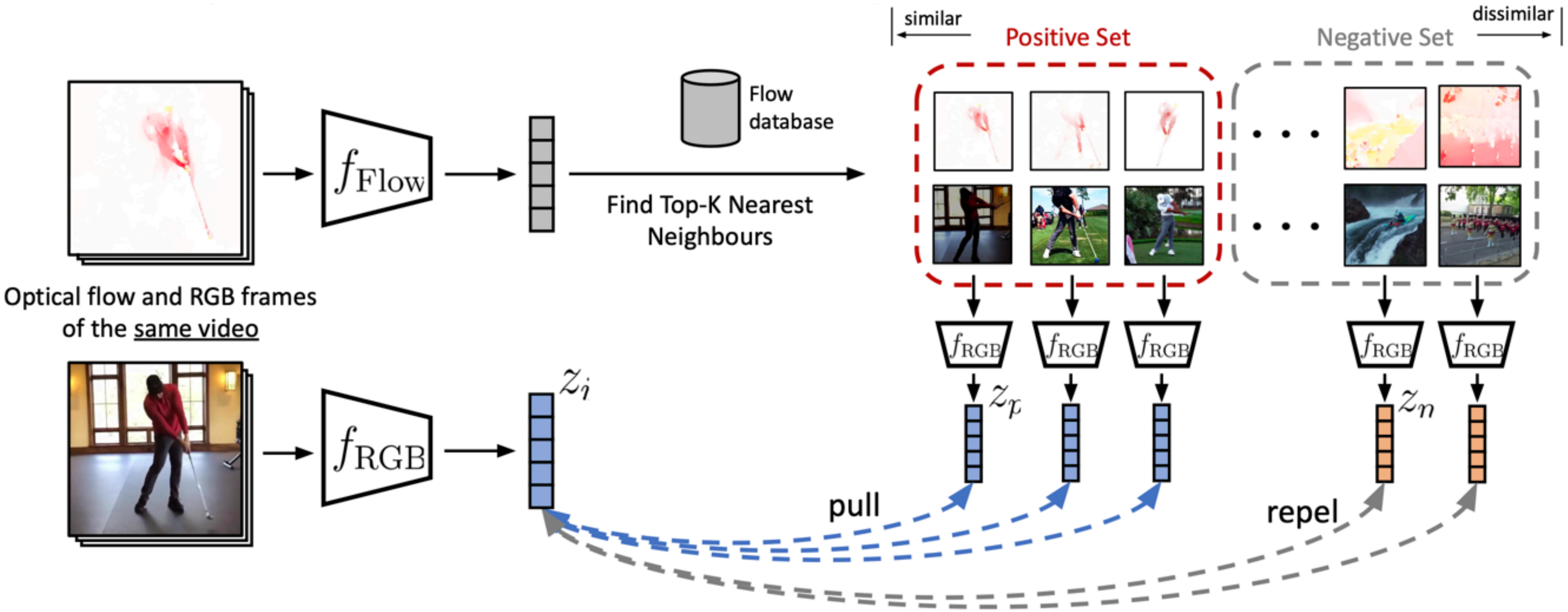
- Improving the sampling mechanism:

- 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: Self-supervised Co-Training for Video Representation Learning. *Han et al.* NeurIPS 2021

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

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

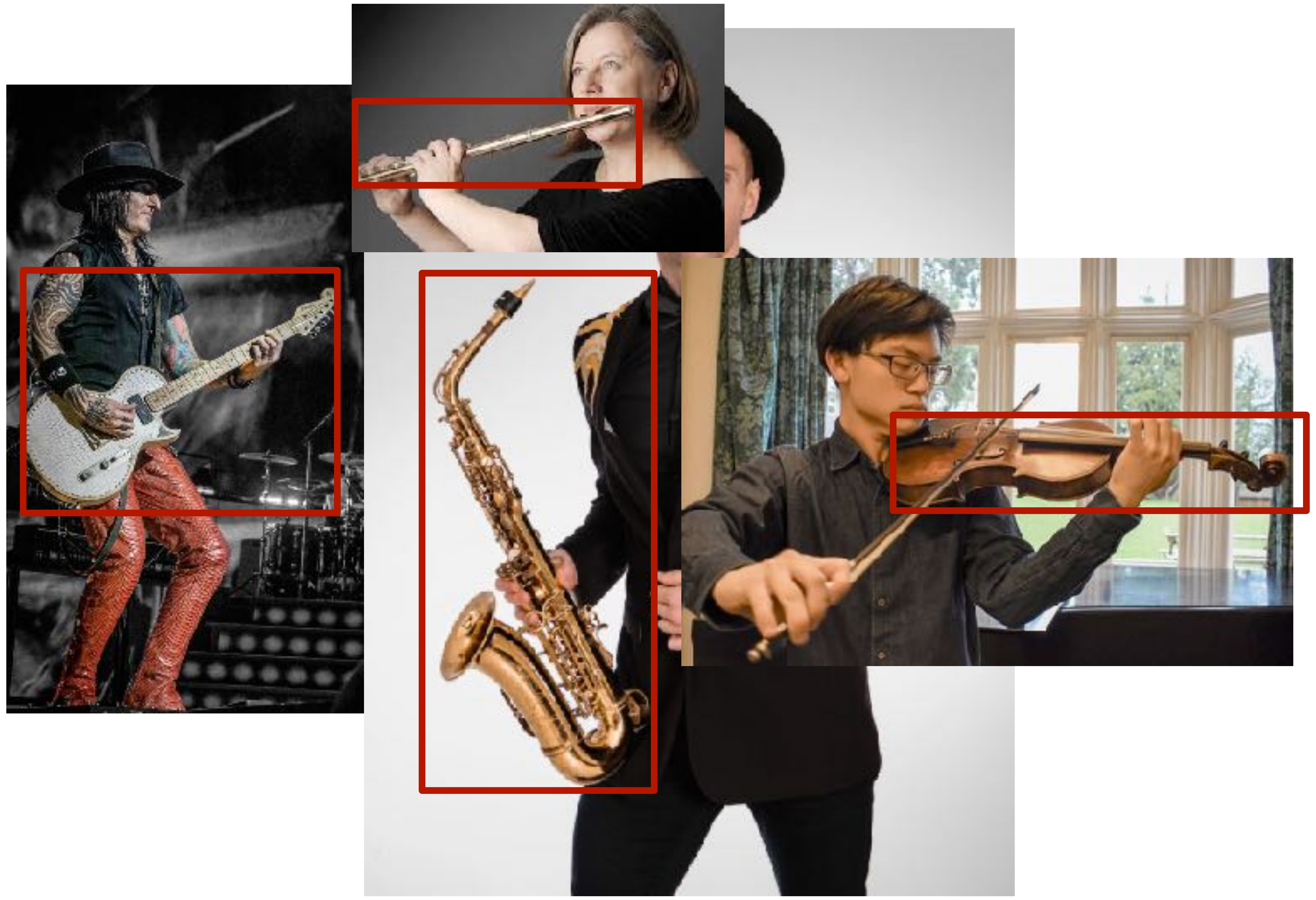
Self-supervised object detection from audio visual correspondence CVPR'22

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

Object detection - supervised training

Detector

Supervised
training

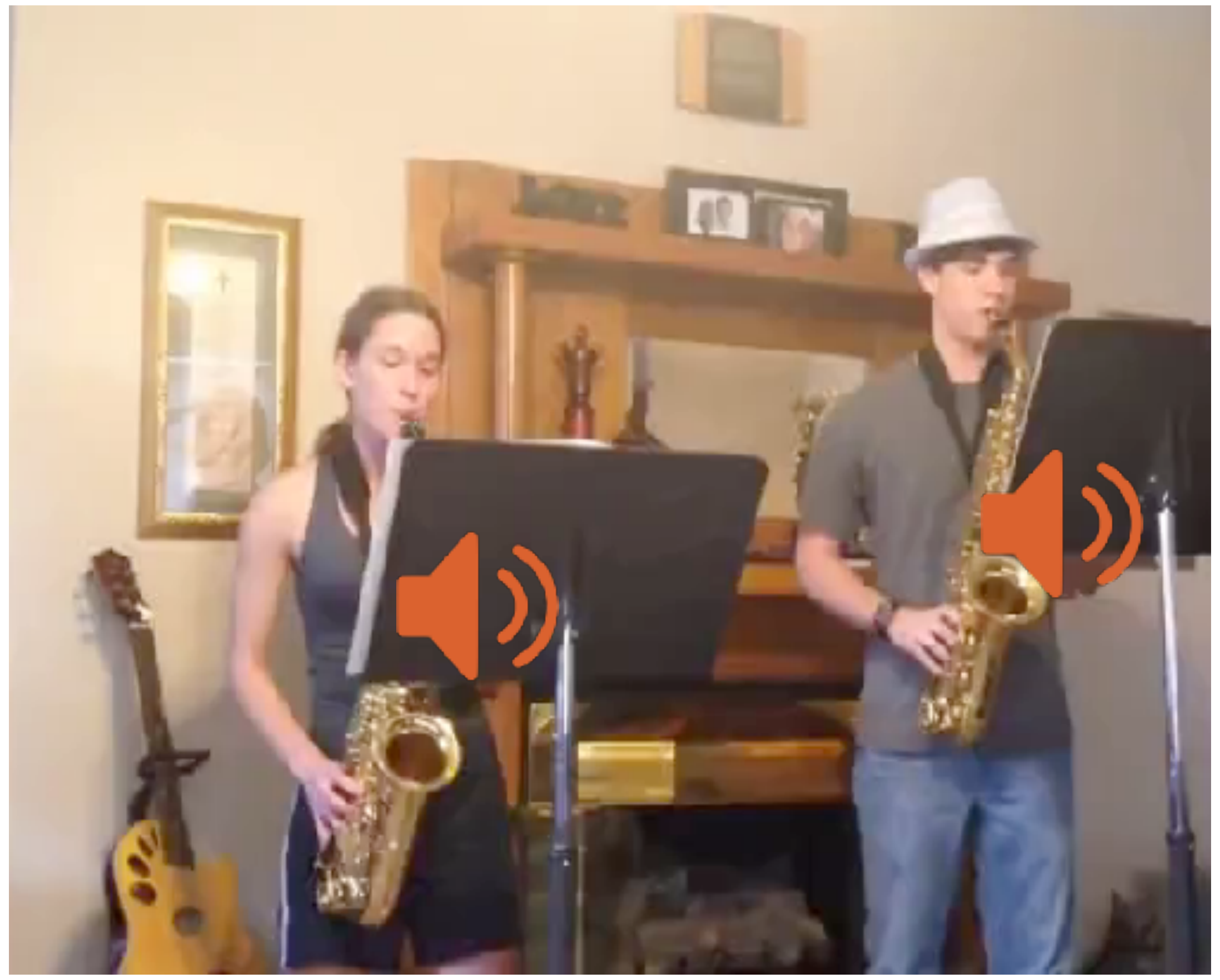


Lots of
annotated samples

Data annotation expensive

Process hard to generalise

What we propose instead:

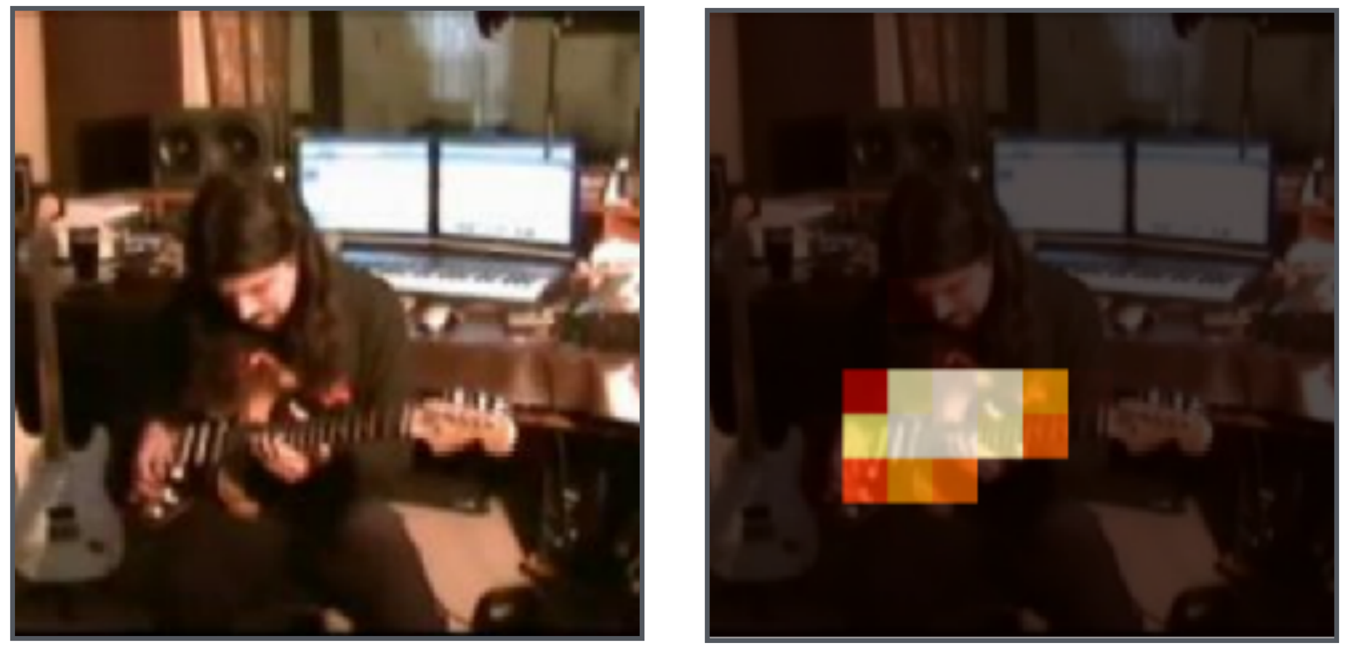


Use sound as supervision for detection

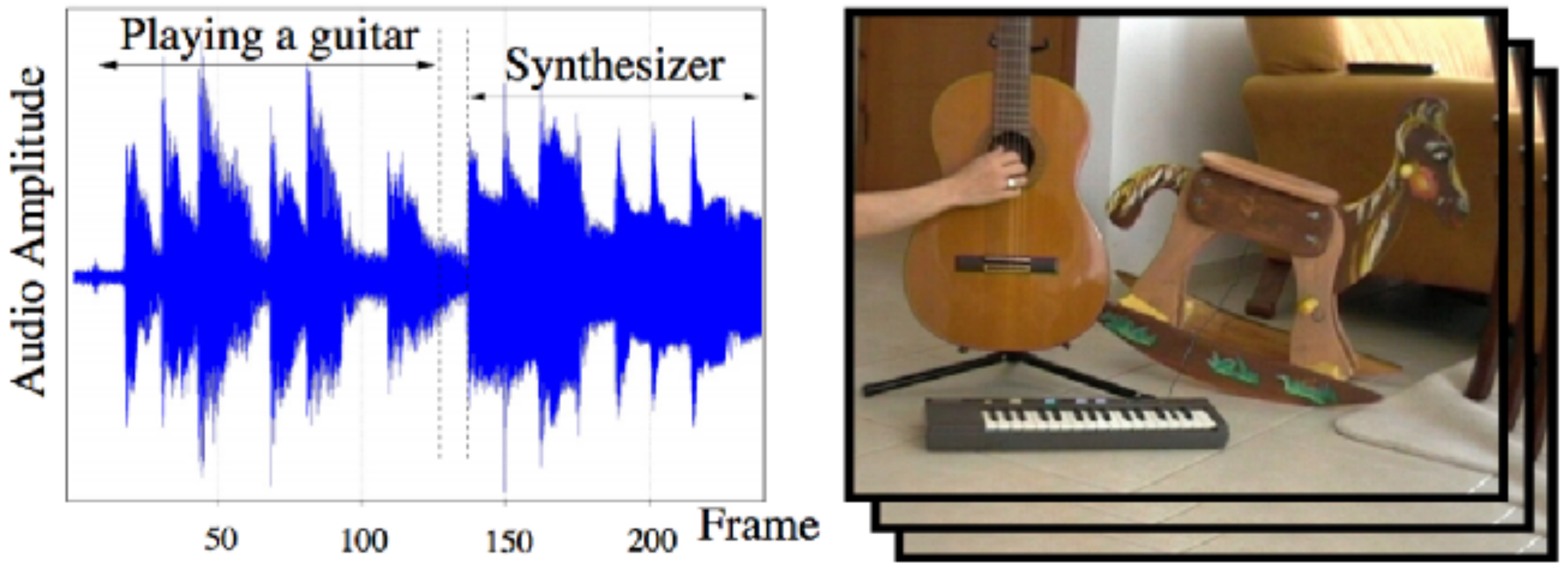
e.g. 1 second window
around frame

Related work: sound source localisation

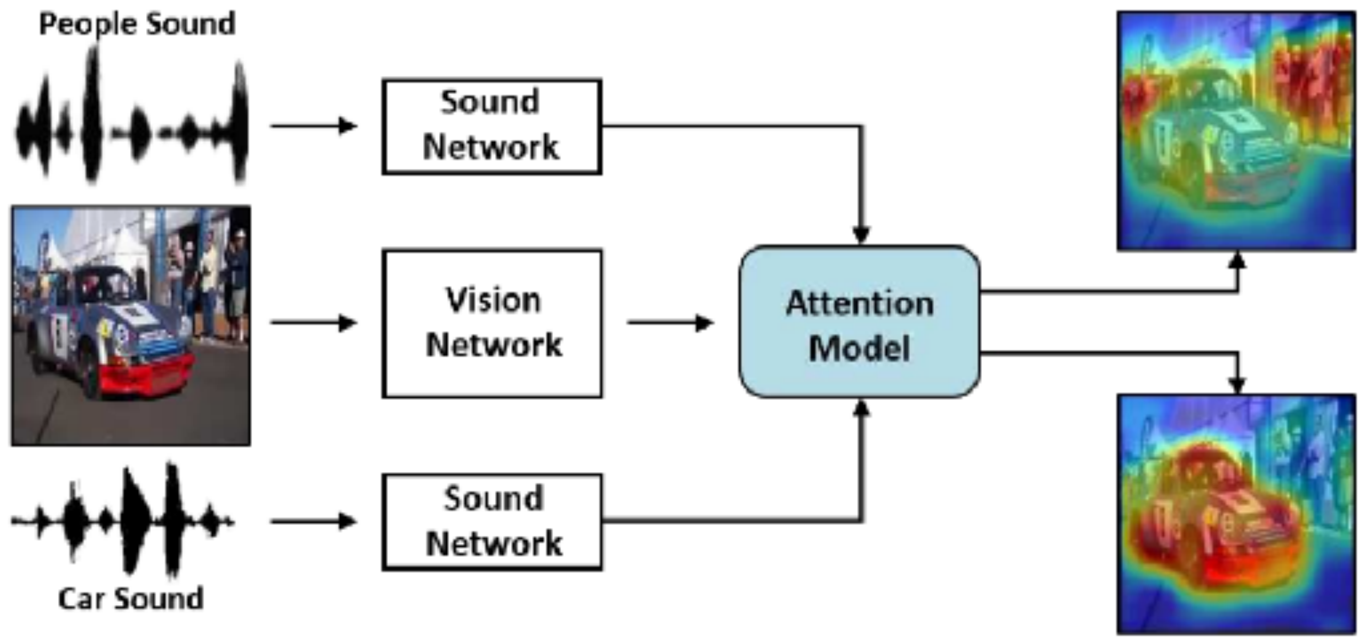
Audio vision: Using audio-visual synchrony to locate sounds. Hershey and Movellan, NeurIPS 2000.



Objects that Sound. Arandjelović and Zisserman, ECCV 2018.

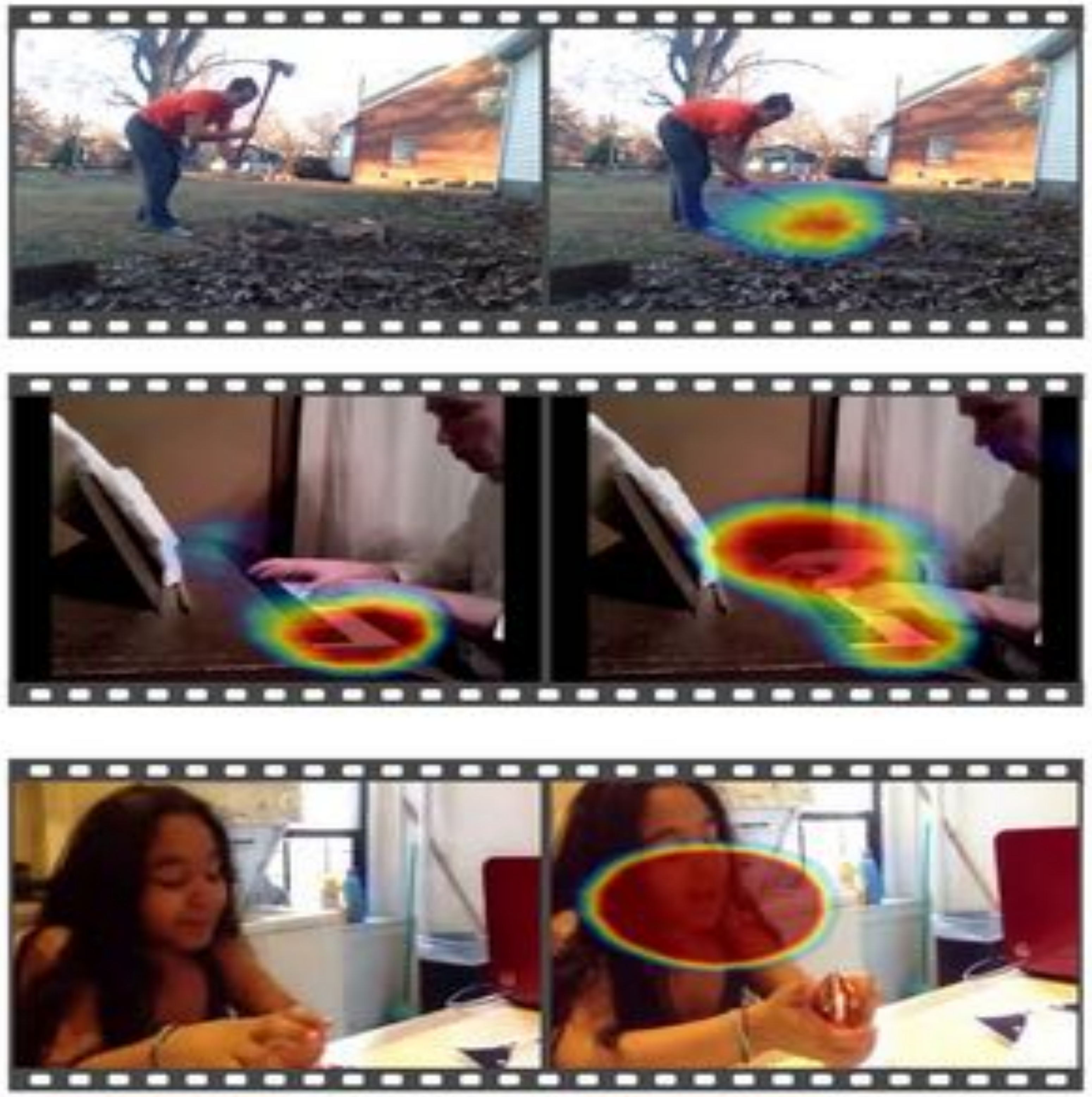


Pixels that Sound. Kidron et al., CVPR 2005.



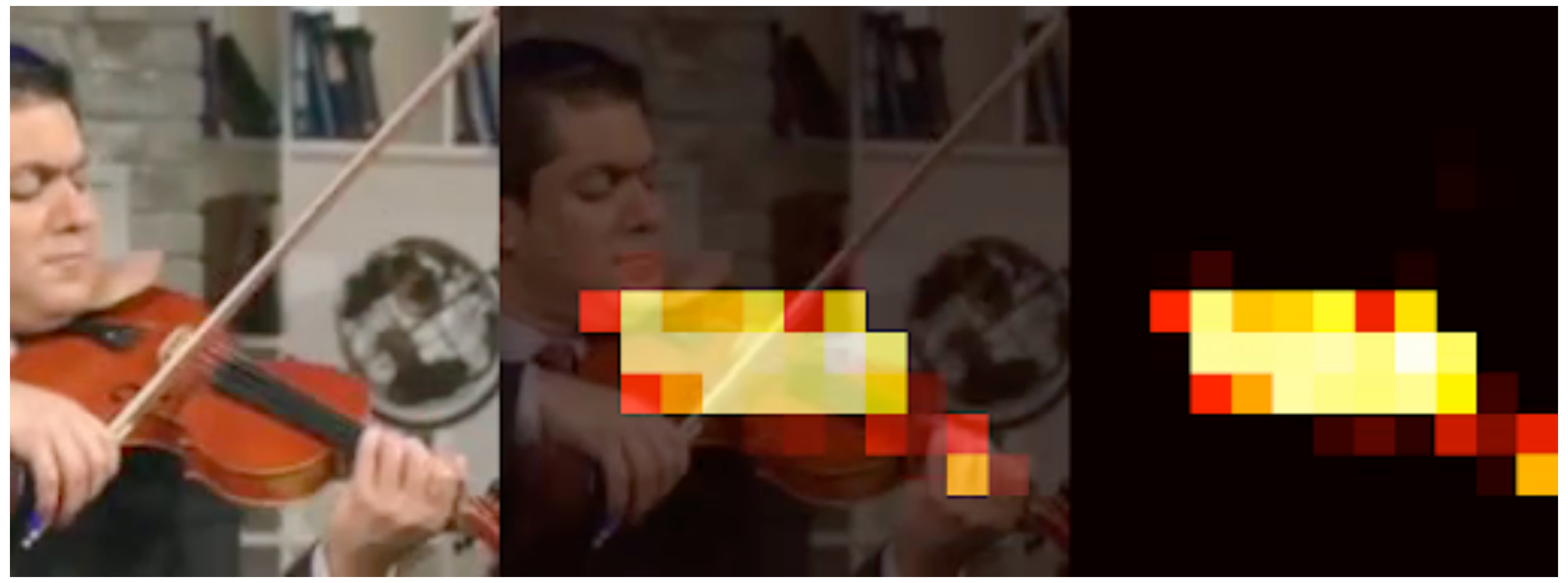
Learning to Localize Sound Source in Visual Scenes. Senocak et al., CVPR 2018.

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.

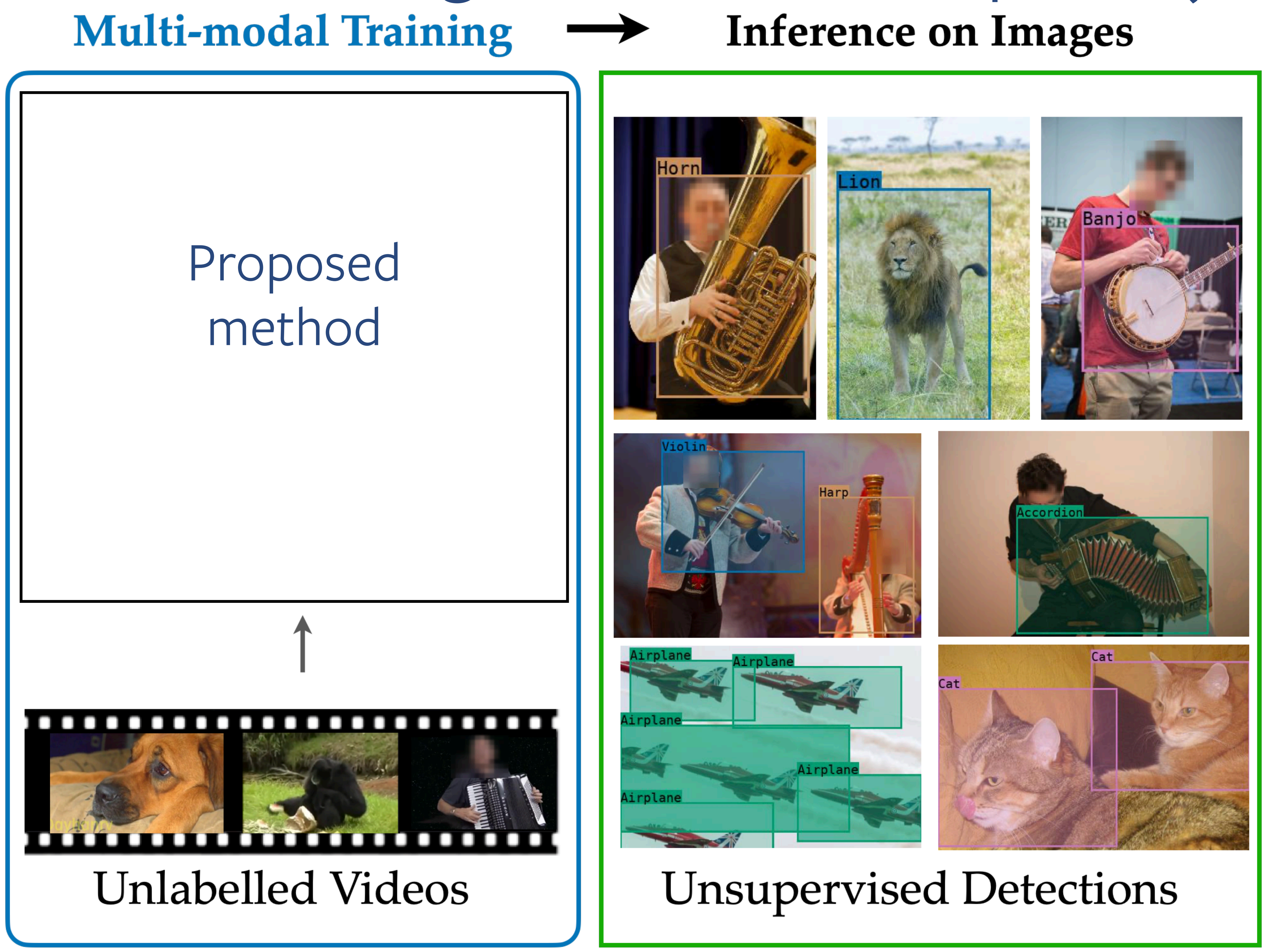
No class labels

Inference requires audio

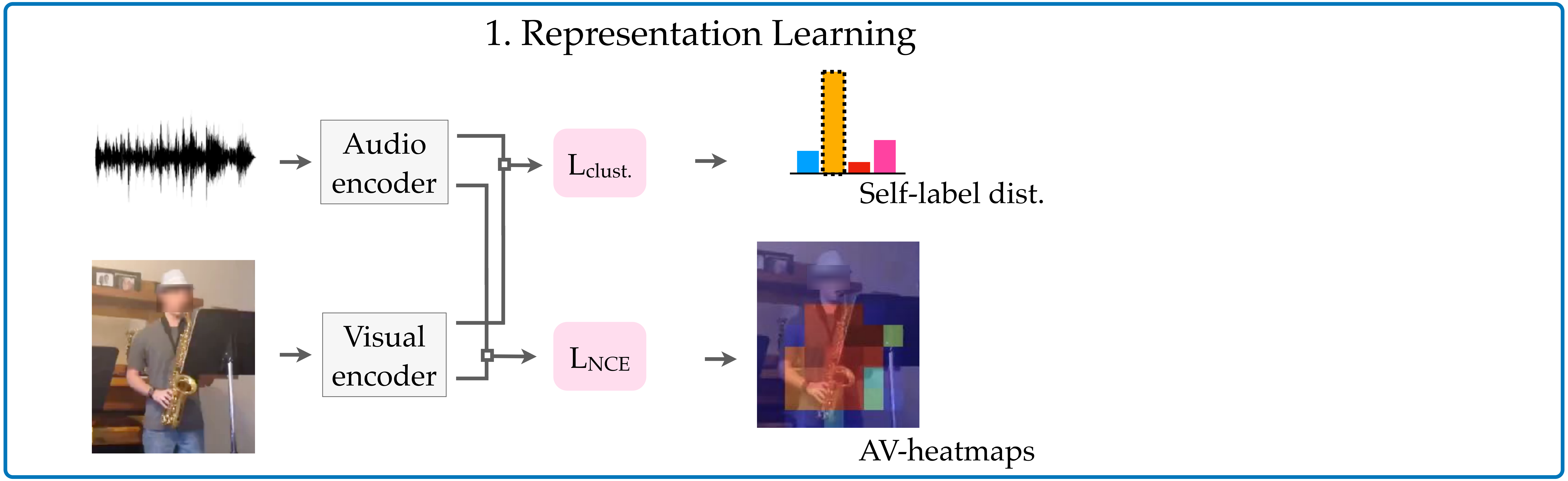
Goal of this paper:
Combining self-labelling and multi-modal learning to detect multiple objects

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

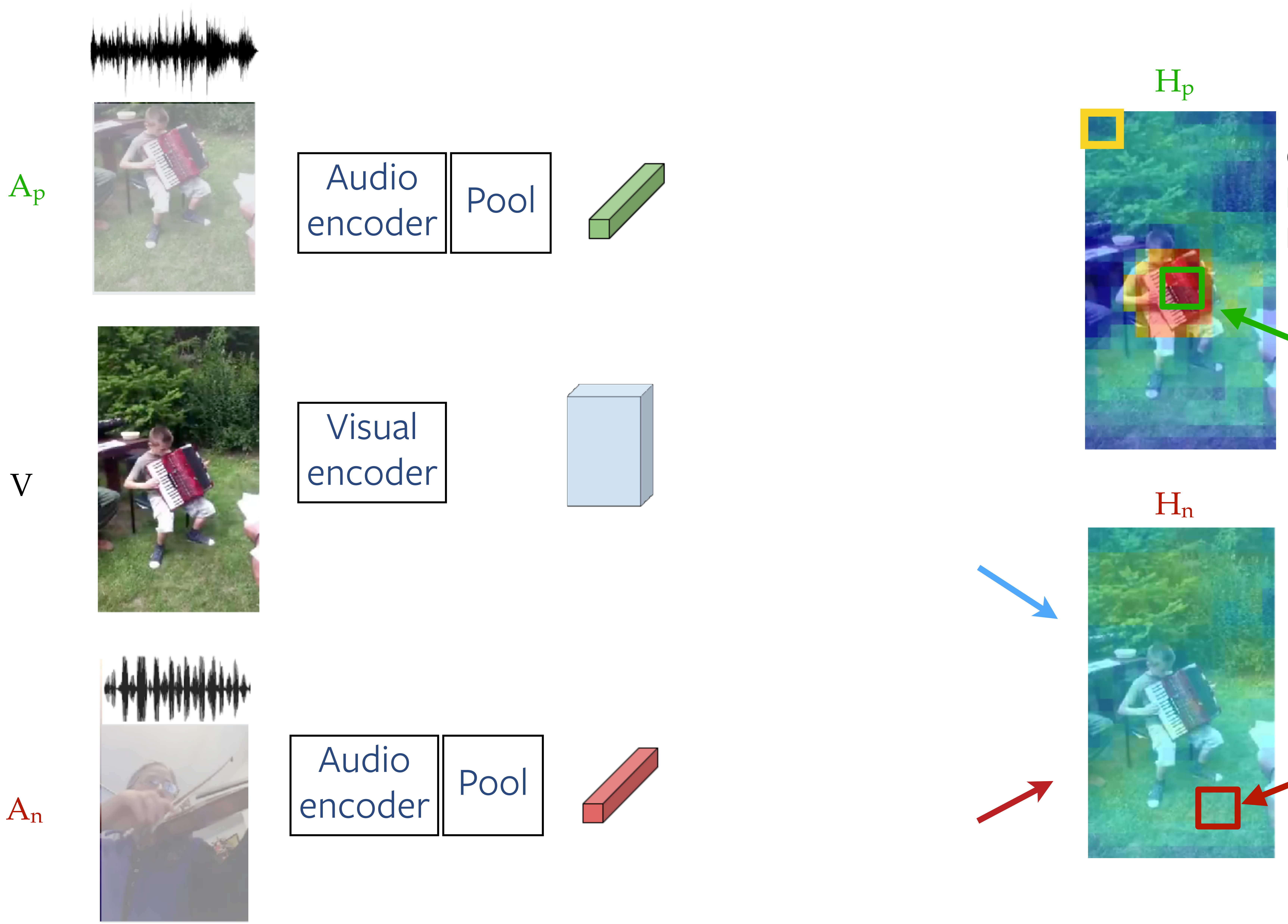
"Useful"
Self-supervised Learning



Framework overview



Ingredient 1: training heat maps



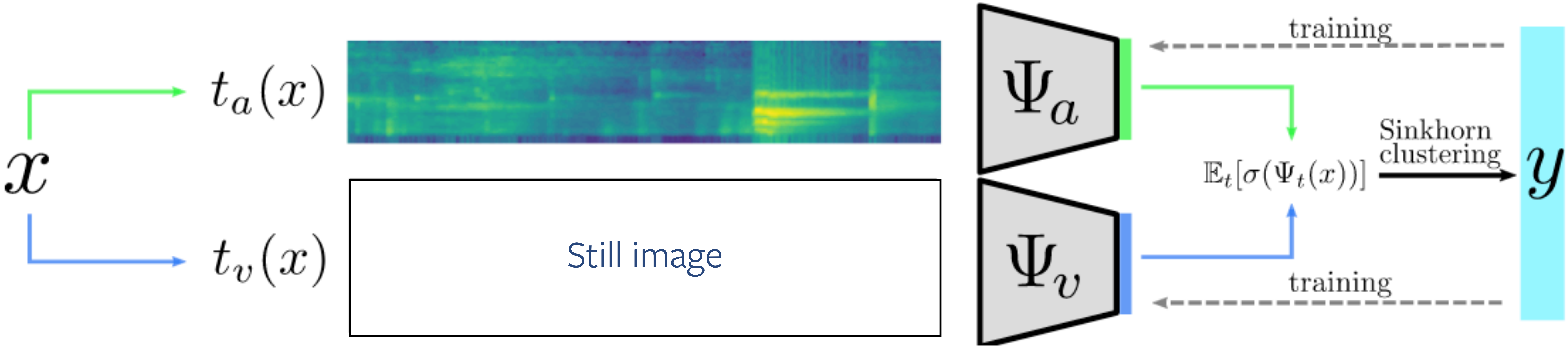
Contrastive loss:

Positive pairs: V & A_p from same clip

Negative pairs: V & A_n from different clip

$$\mathcal{L}_{NCE} = \log\left(\frac{\max\limits_{h_i \in \mathcal{H}_p} e^{h_i}}{\max\limits_{h_i \in \mathcal{H}_p} e^{h_i} + \max\limits_{h'_i \in \mathcal{H}_n} e^{h'_i}}\right)$$

Ingredient 2: Self-labels training ($L_{clust.}$)

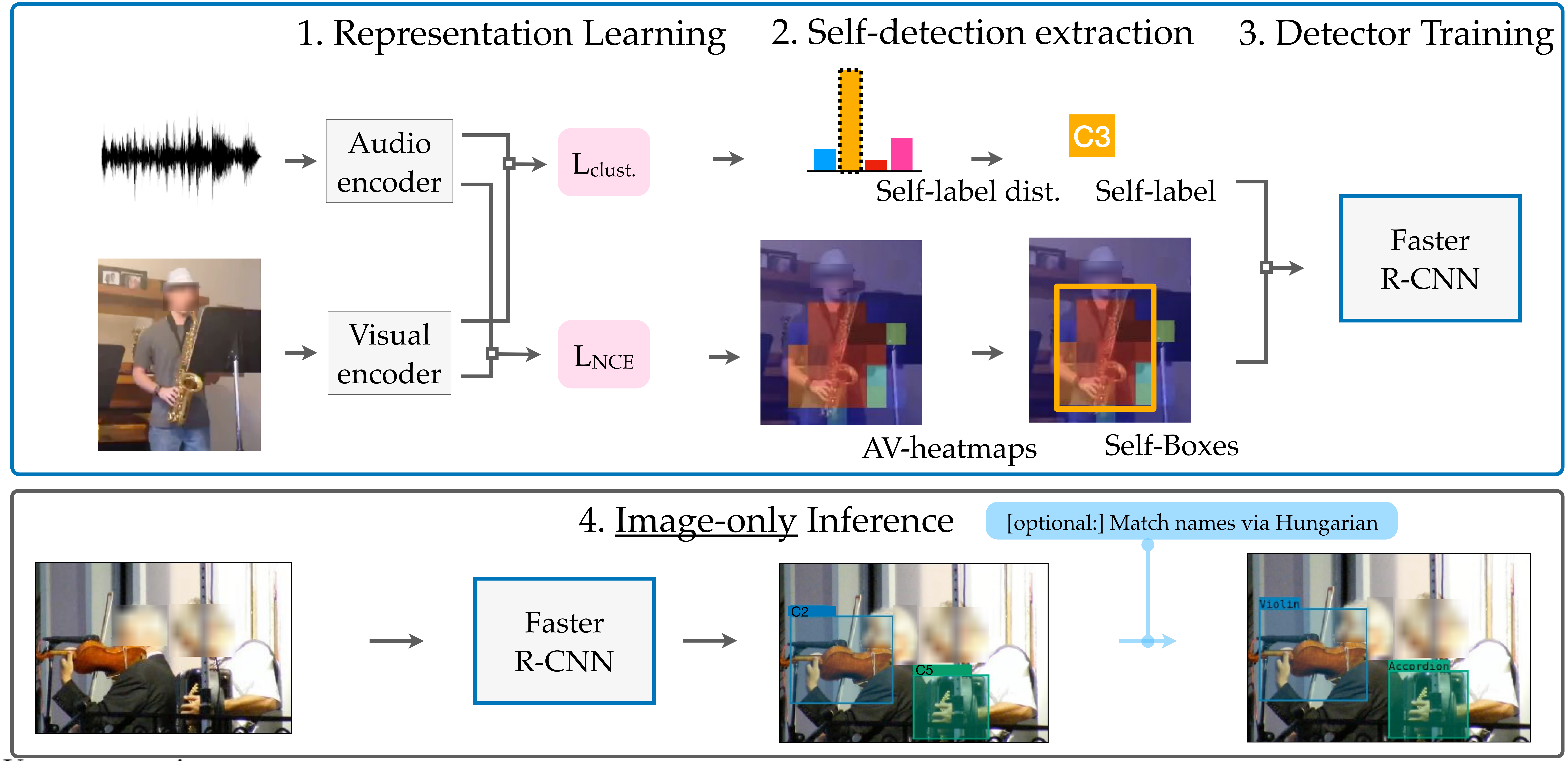


$$\mathcal{L}_v(\mathcal{B}|y) = -\frac{1}{|\mathcal{B}|} \sum_{(v,a)\in\mathcal{B}} \log \operatorname{softmax}(y(v,a) \,|\Psi_v(v))$$

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

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

Framework overview



Qualitative results compared to weakly-supervised

Detection examples and failure cases

What about more general objects, beyond instruments?

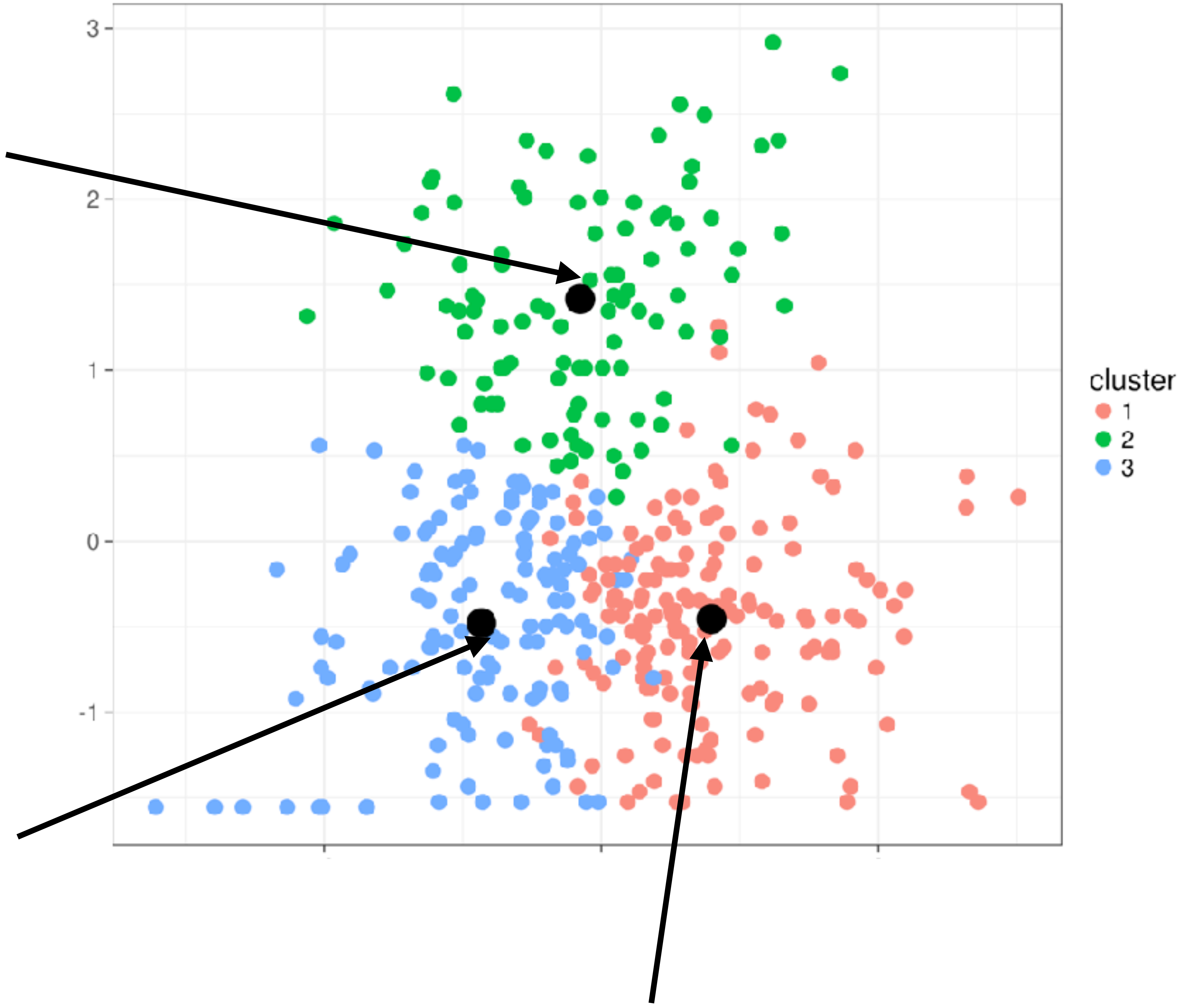
We train on all ~300 VGGSound classes, learning 300 clusters.

- 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”)

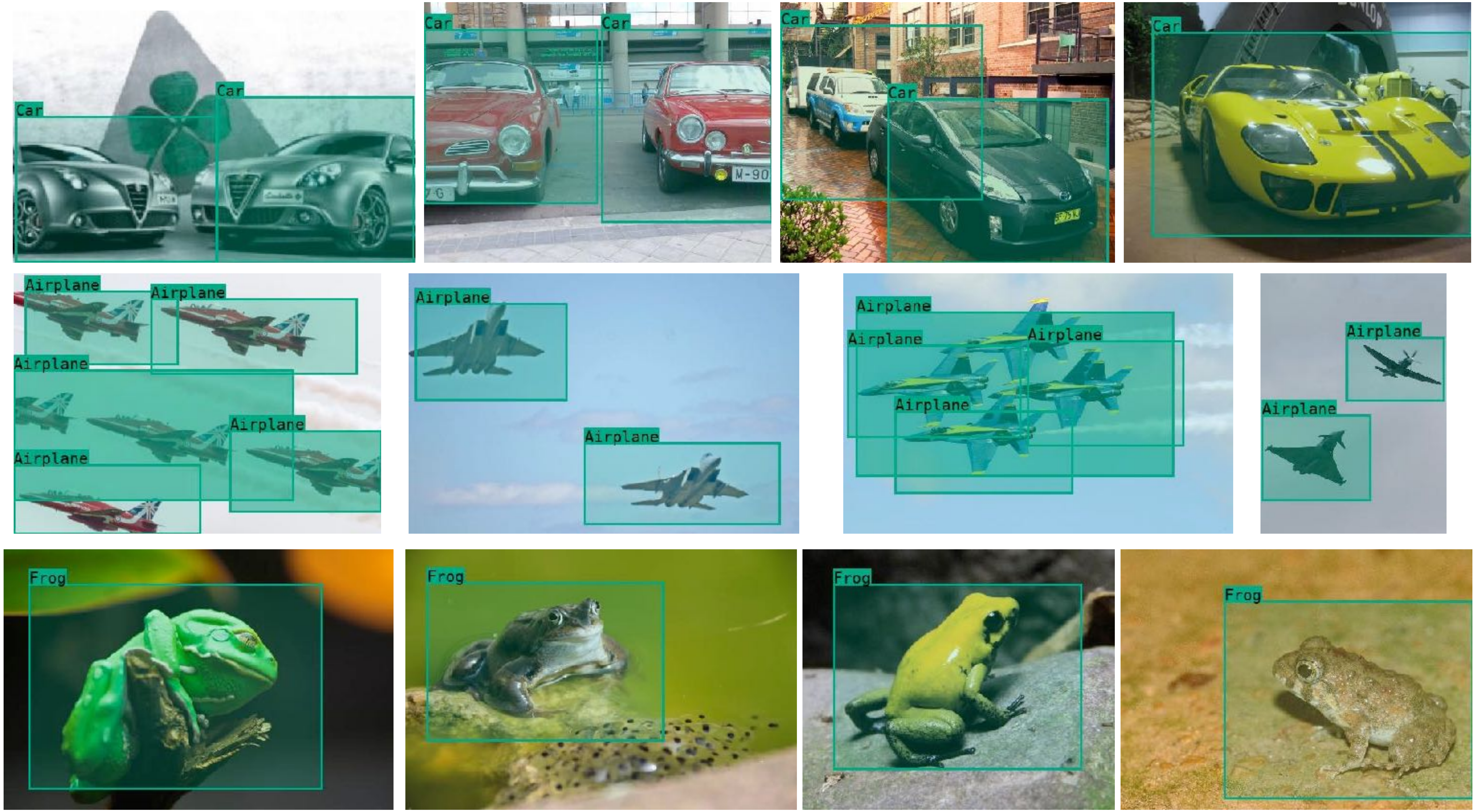
Prototype efficient matching: obtain labels of samples closest to centroids

	mAP$_{50}$	
Matching	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

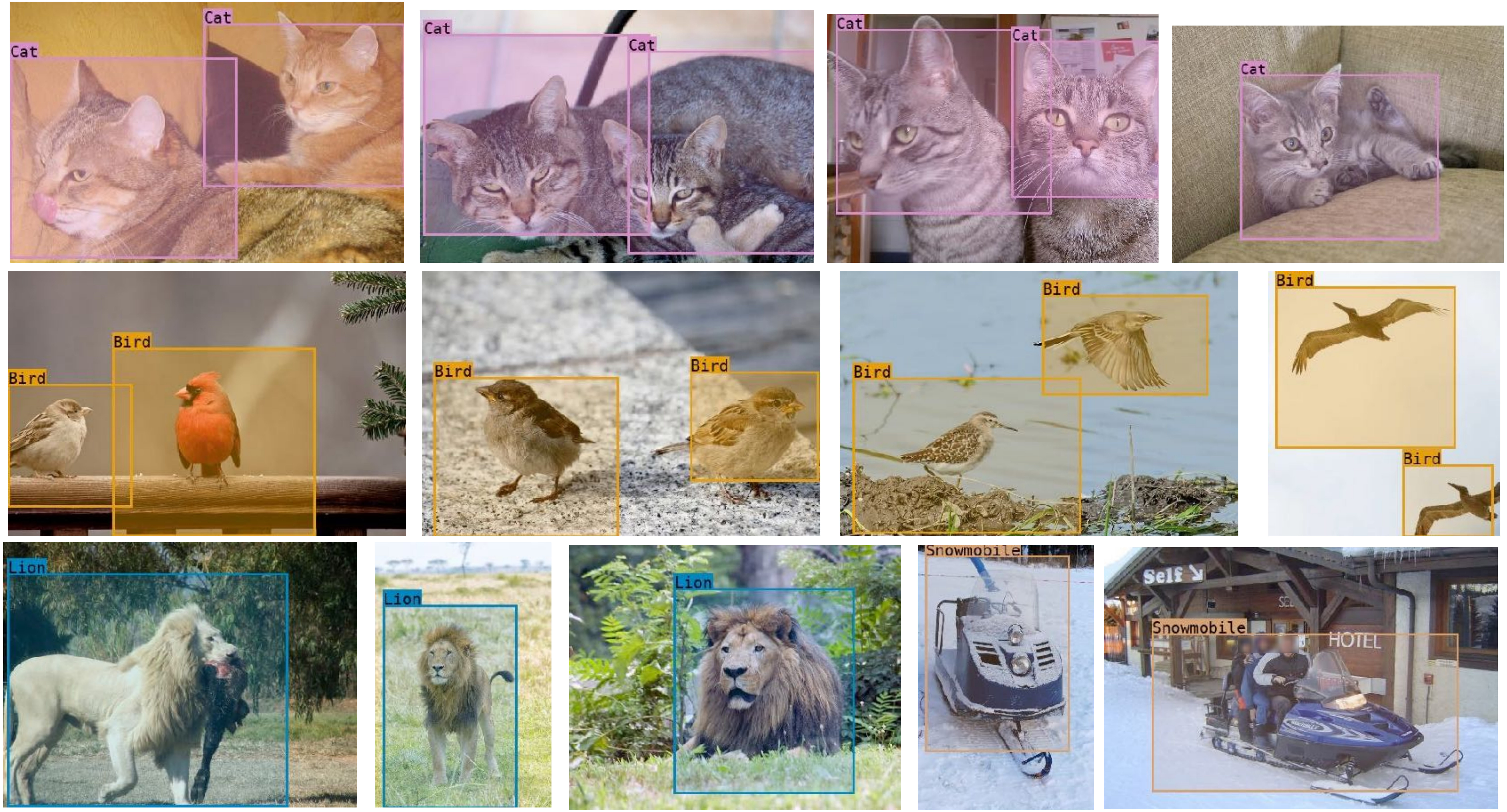
Table 6: **Matching strategies.** Even with as little as 39 labels, our method can detect and classify objects accurately.



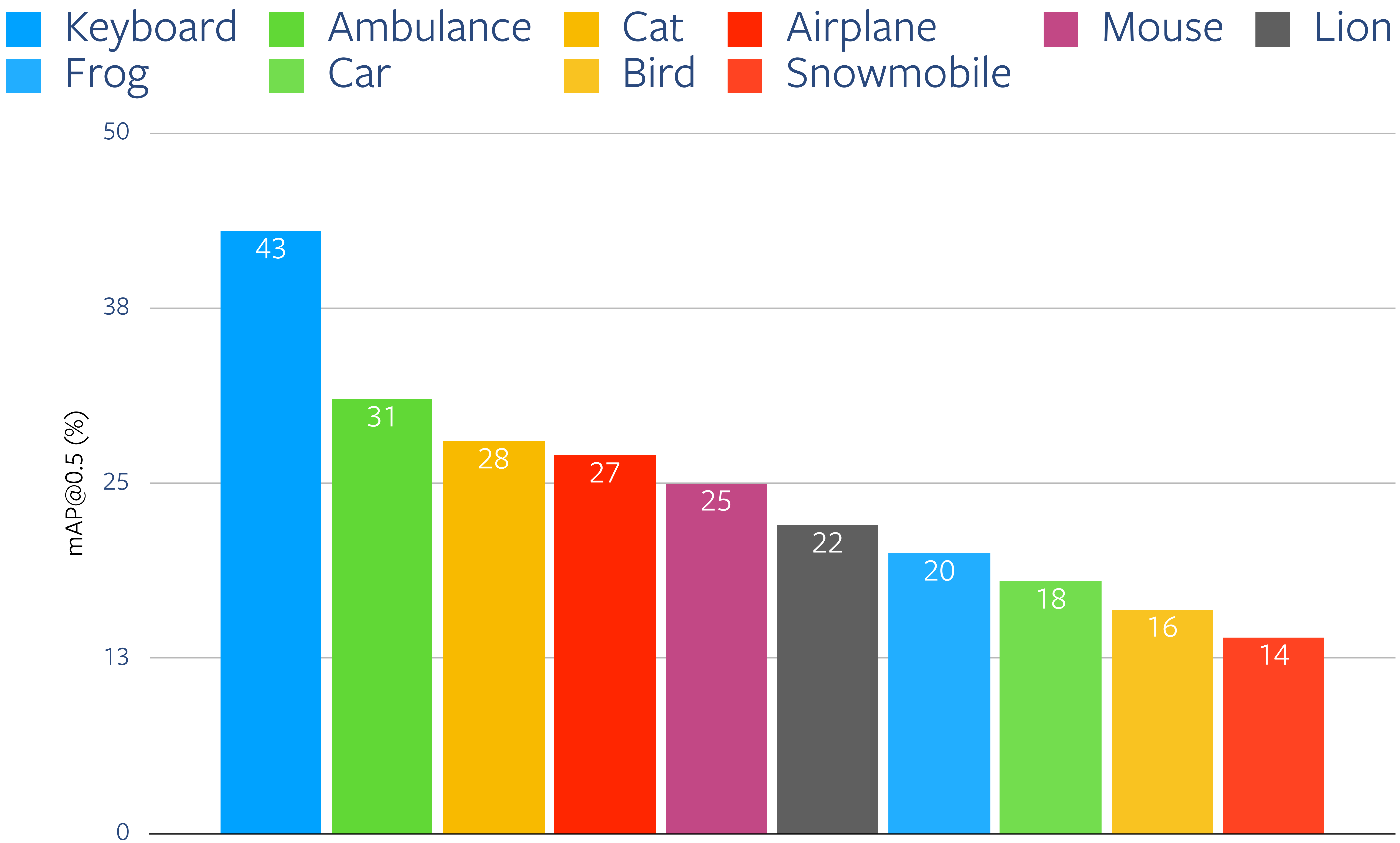
VGGSound object detections



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

You

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

https://twitter.com/osanseviero/status/1546224159553986561?s=20&t=rYbQ2MWKaAPiZ13MUTxPBQ

MMV

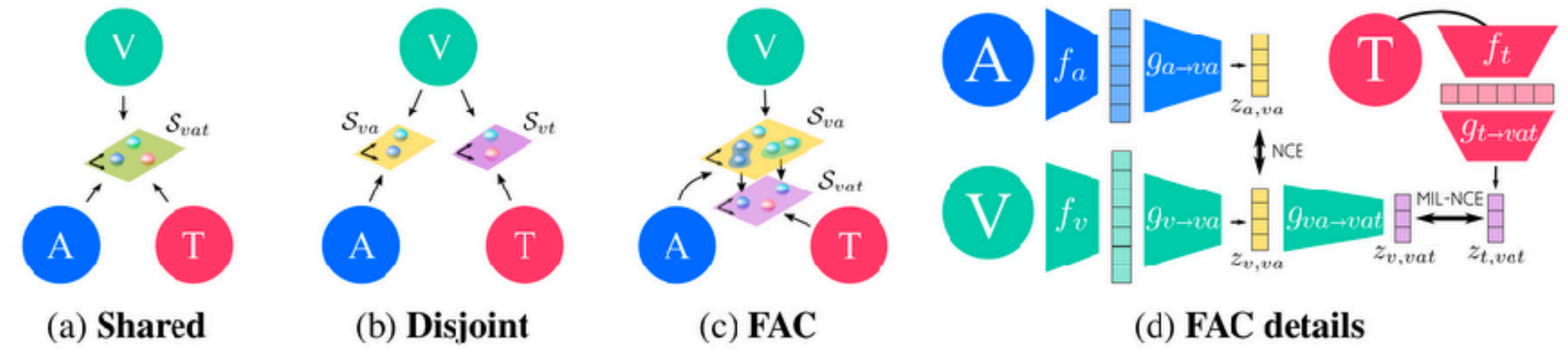


Figure 1: (a)-(c) Modality Embedding Graphs, (d) Projection heads and losses for the FAC graph. V=Vision, A=Audio, T=Text.

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

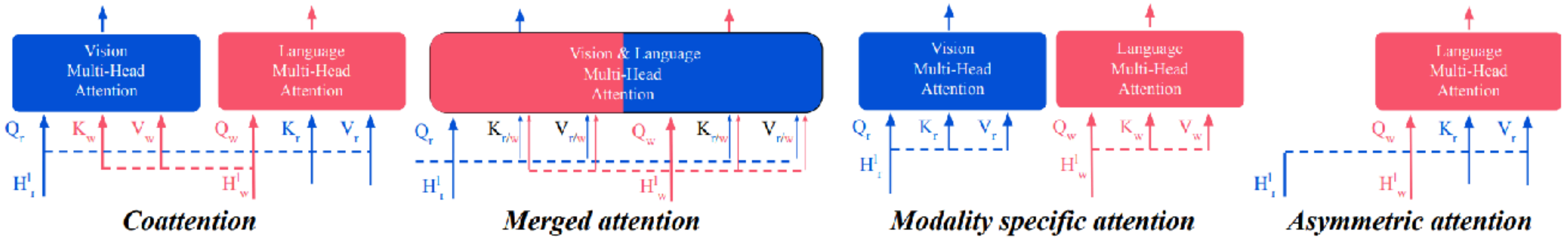
Table 2: **Comparison of learnt representations versus the state-of-the-art.** Results are averaged over all splits. The "Mod." column shows which combinations of modalities are used by the methods, possibilities: Vision, Audio, Text, Flow. Dataset abbreviations: AudioSet, HowTo100M, Instagram65M [23], 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.

					UCF101		HMDB51		ESC-50	AS	K600
Method	f_v (#params)	Train data	years	Mod.	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.

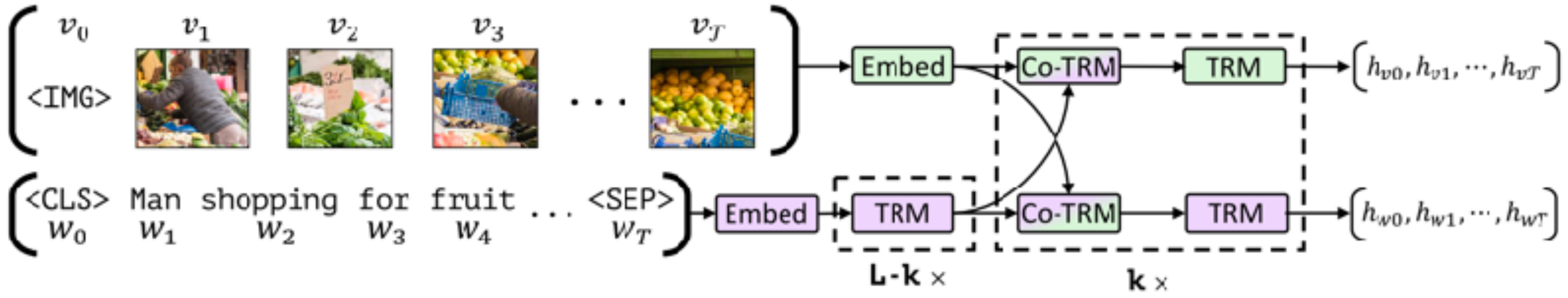


[1] Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers, Hendricks et al. (2018)

Multimodal Transformers

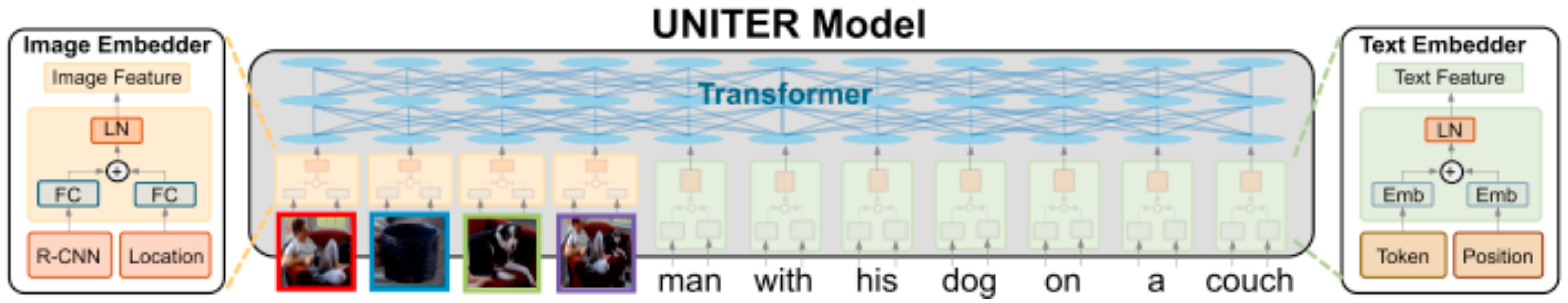
Two-stream Multimodal Transformers: two Transformers are applied to images and text independently, which is fused by a third Transformer in a later stage.

Vision & **L**anguage BERT (ViLBERT)[1]



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

UNiversal **I**mage-**TE**xt **R**epresentation (UNITER)[2]



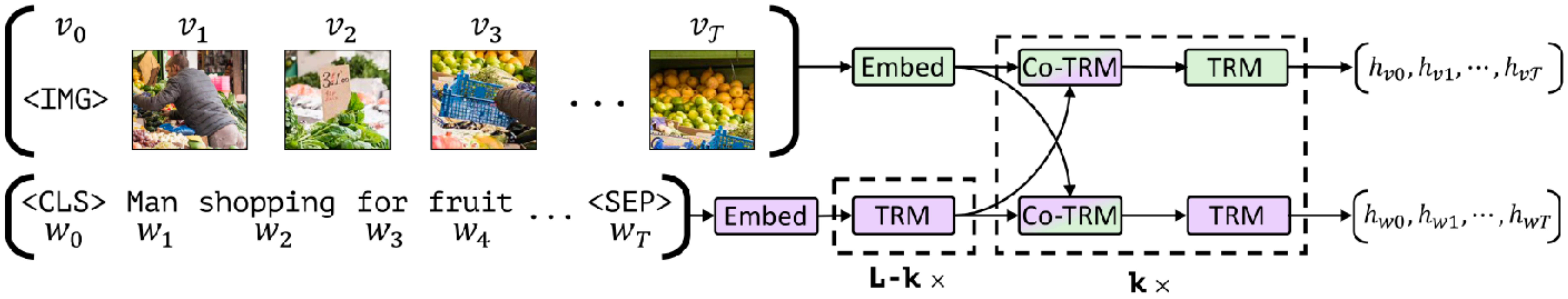
[1] ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, Lu et al. (2019)
[2] UNITER: UNiversal Image-TExt Representation Learning, Chen et al. (2020)

Multimodal Transformers: ViLBERT

ViLBERT[1]: 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.

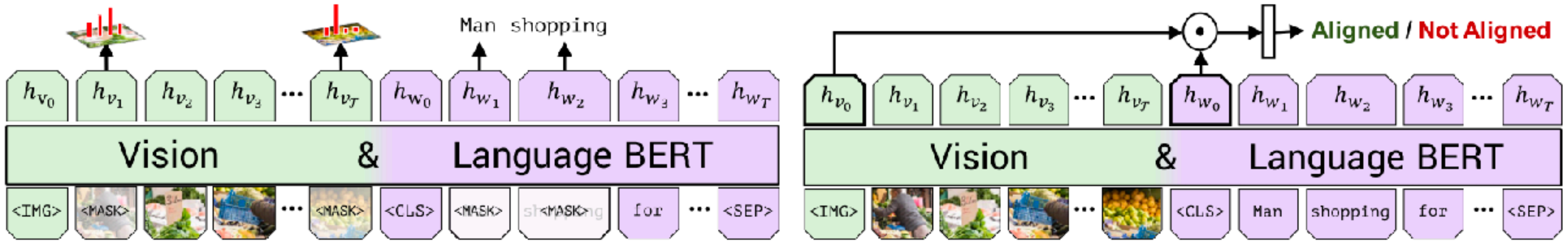


[1] ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, Lu et al. (2019)

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.

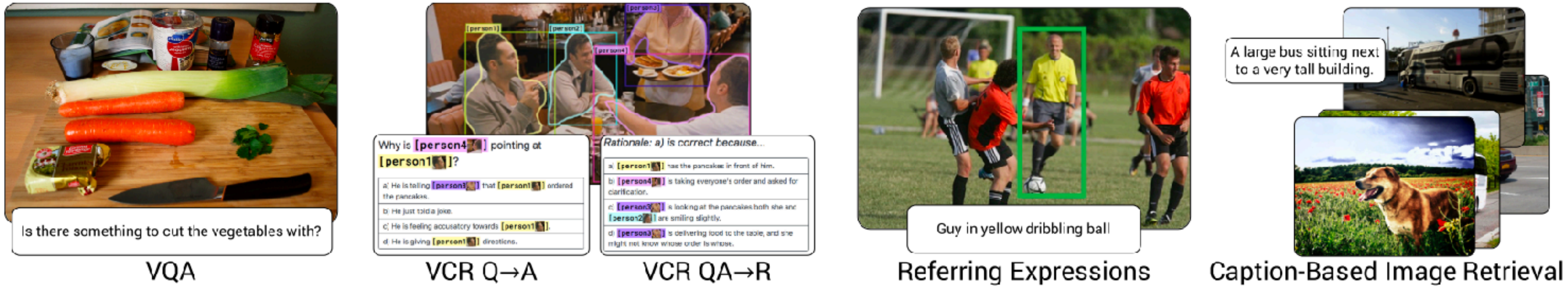


(a) Masked multi-modal learning (b) Multi-modal alignment prediction

Multimodal Transformers: ViLBERT

For the downstream tasks: learning a classification layer.

Wide range of tasks



answering natural language questions about images

posed as multiple- choice problems

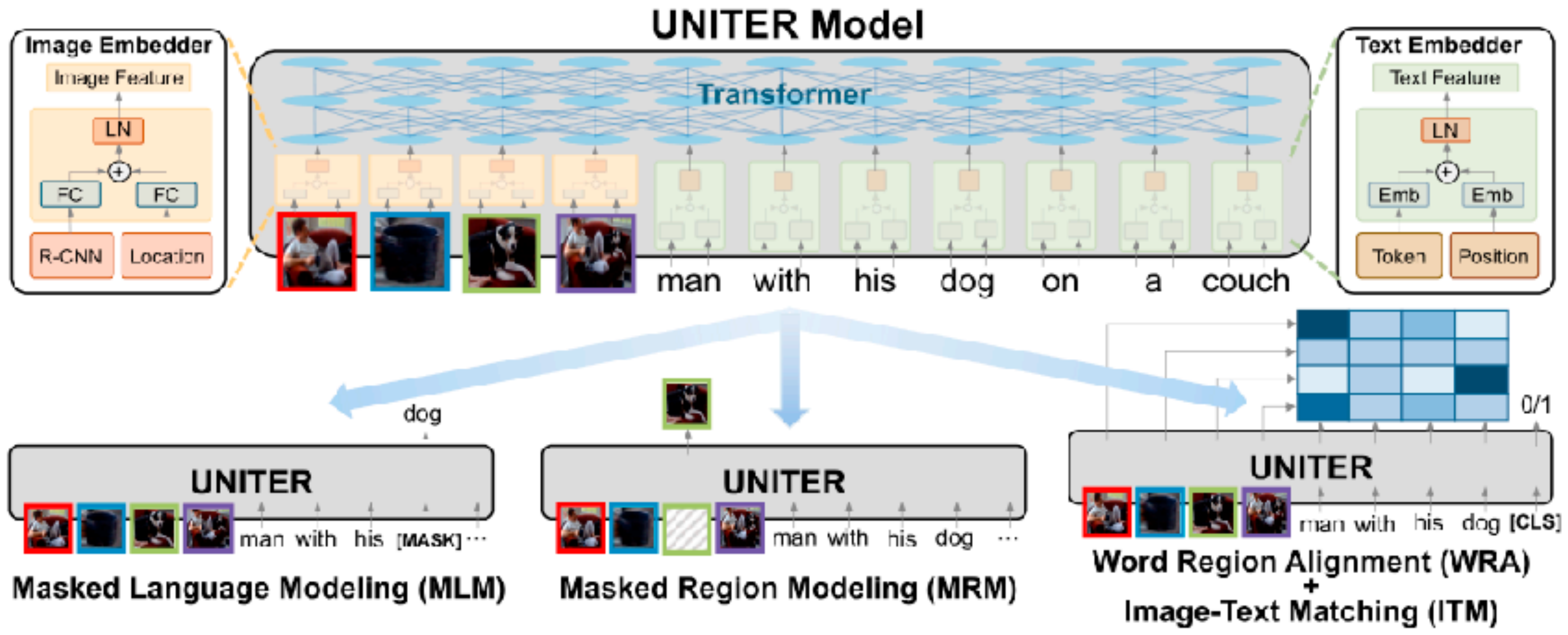
localize an image region given a natural language reference

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

Multimodal Transformers: UNITER

UNITER[1] 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

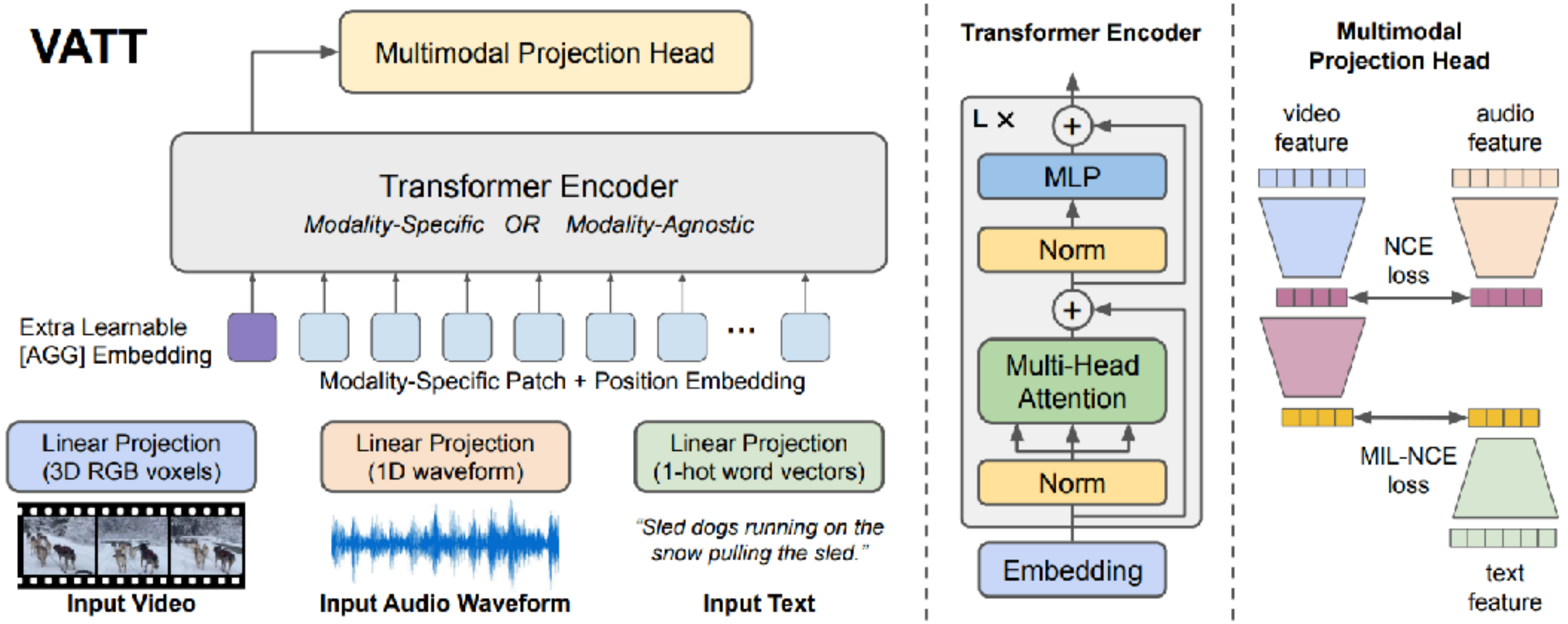
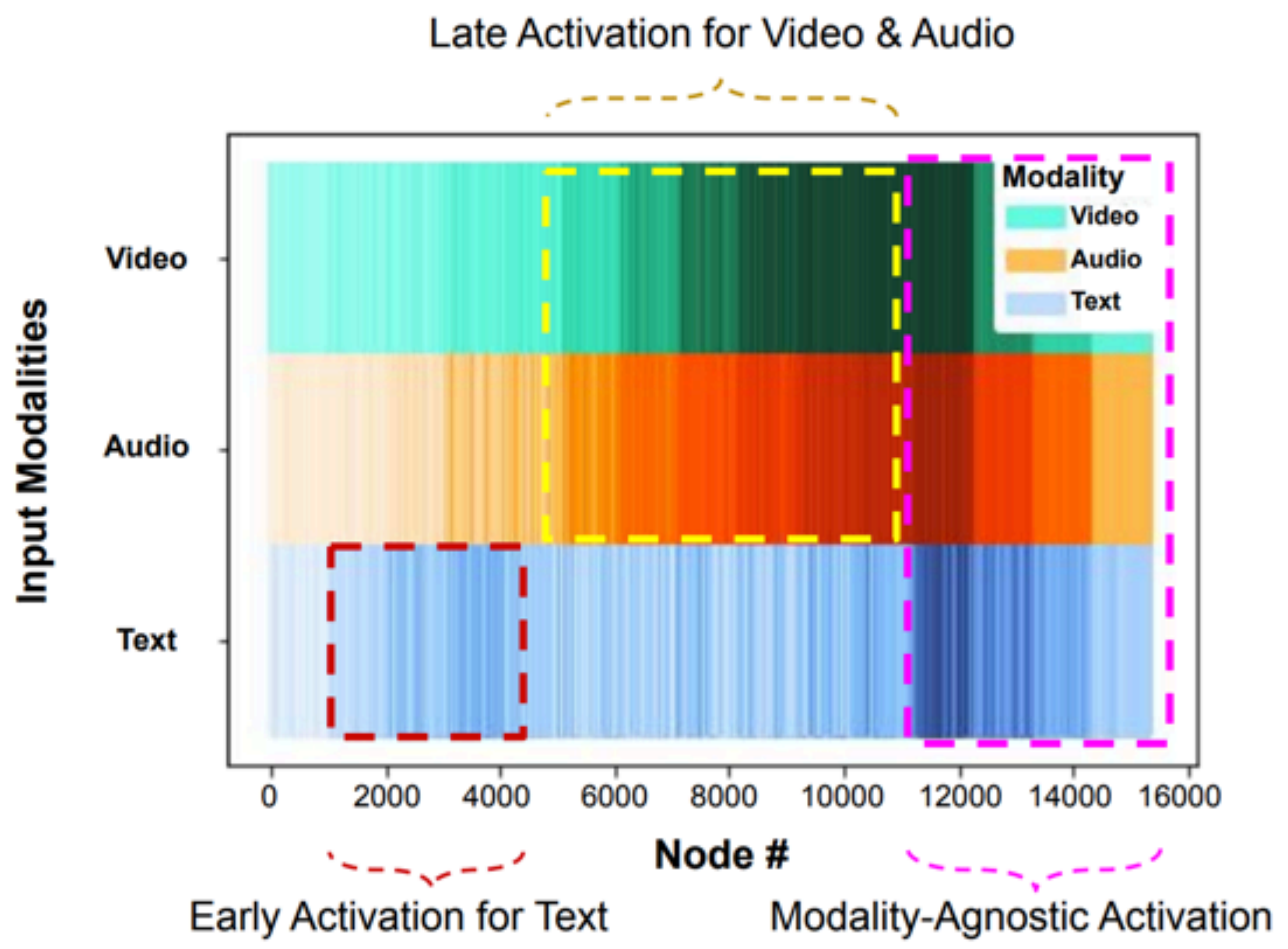


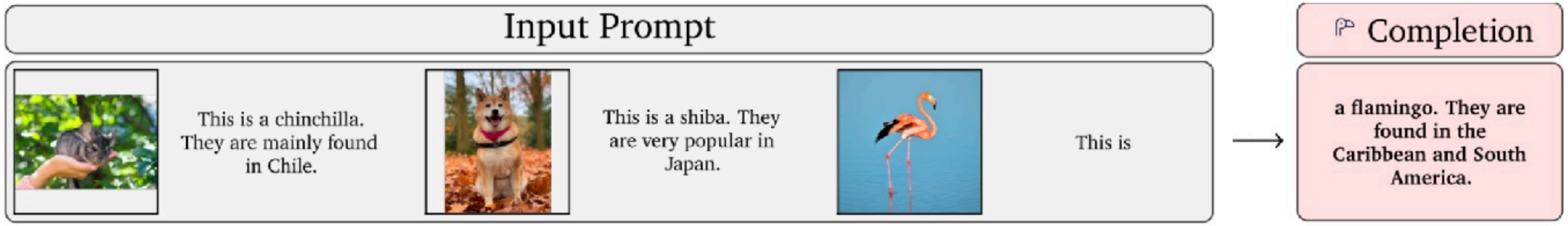
Figure 1: **Overview of the VATT architecture and the self-supervised, multimodal learning strategy**. VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. We define a semantically hierarchical common space to account for the granularity of different modalities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



VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. *Akbari et al.* NeurIPS 2021

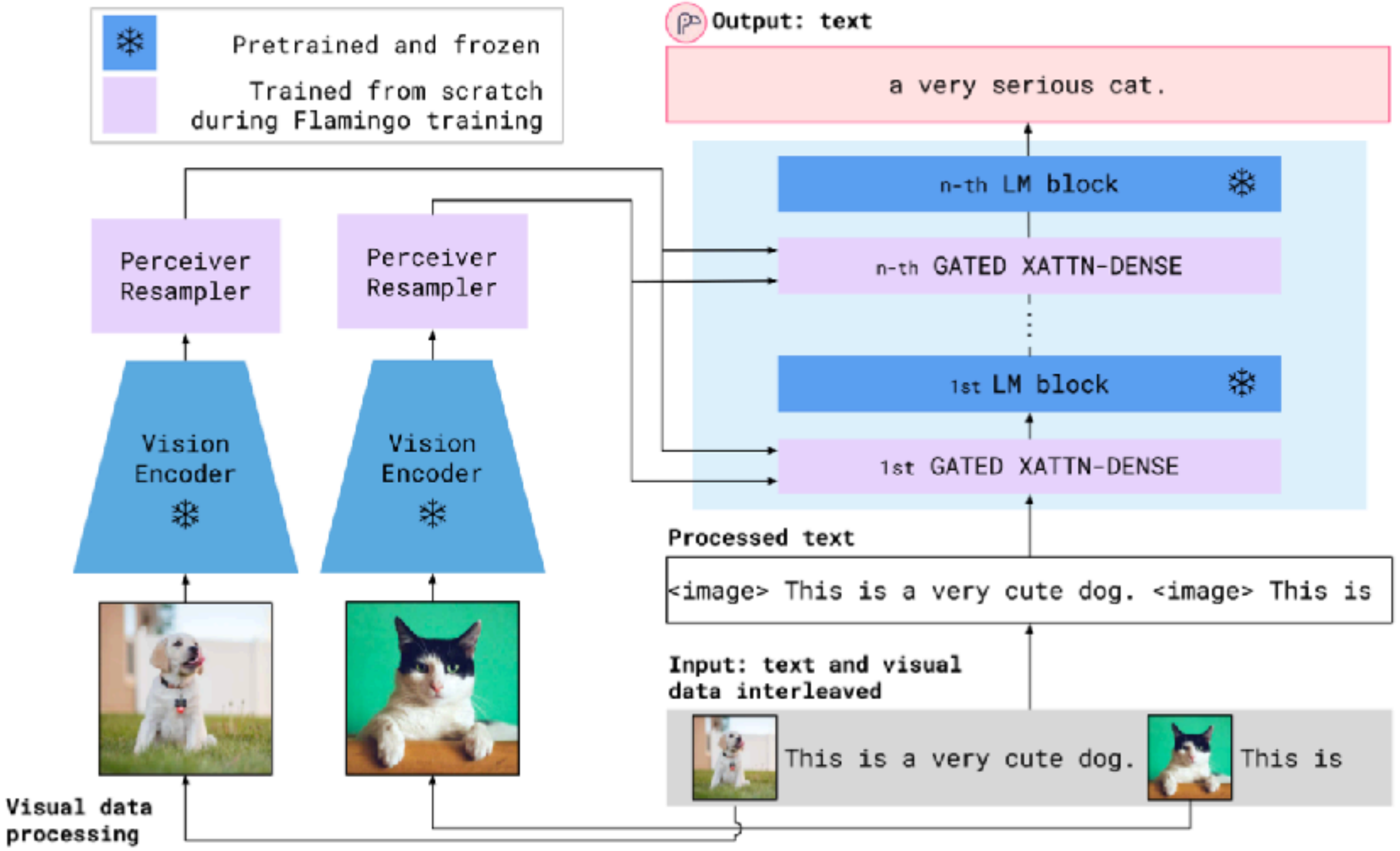
Flamingo

- 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.

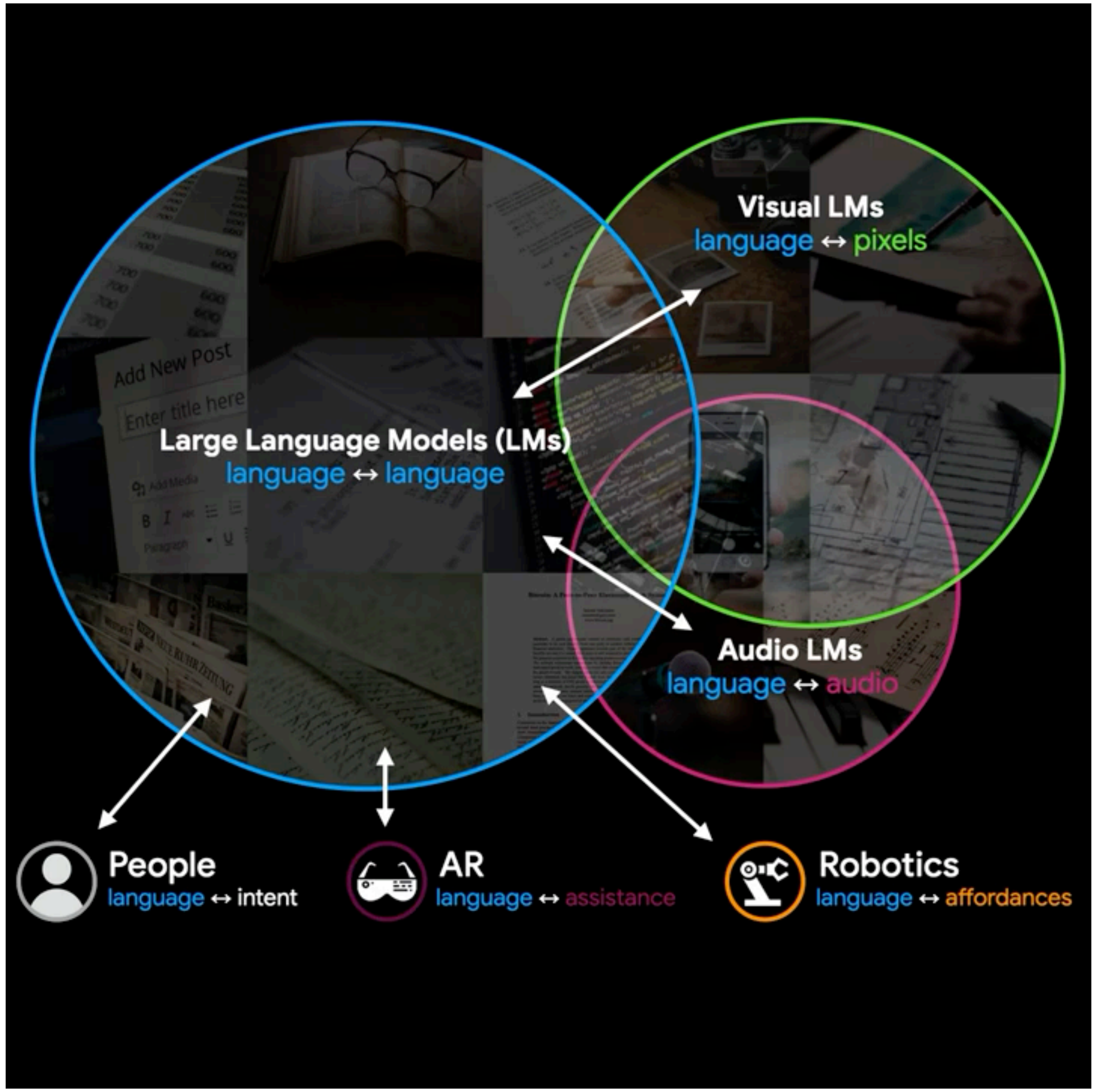


Socratic Models

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

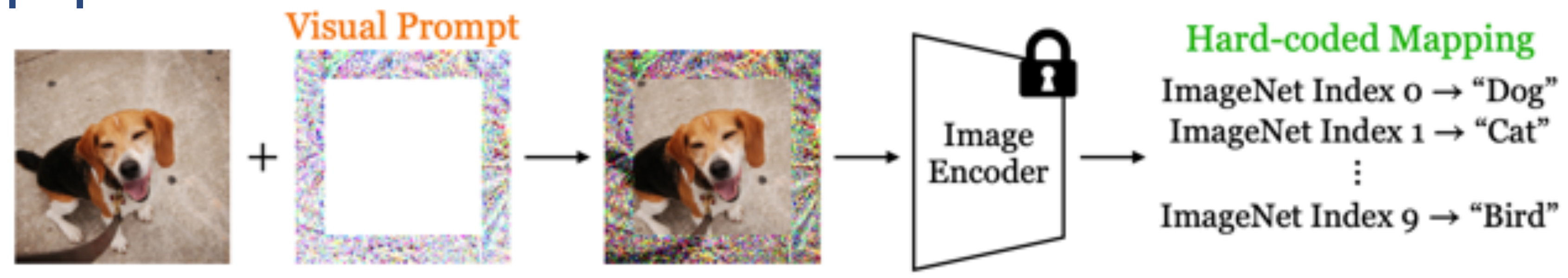
(i) the VLM detects visual entities,
(ii) the LM suggests sounds that may be heard,
(iii) the ALM chooses the most likely sound,
(iv) the LM suggests possible activities,
(v) the VLM ranks the most likely activity,
(vi) the LM generates a summary of the Socratic interaction.

Essentially language as the lingua franca

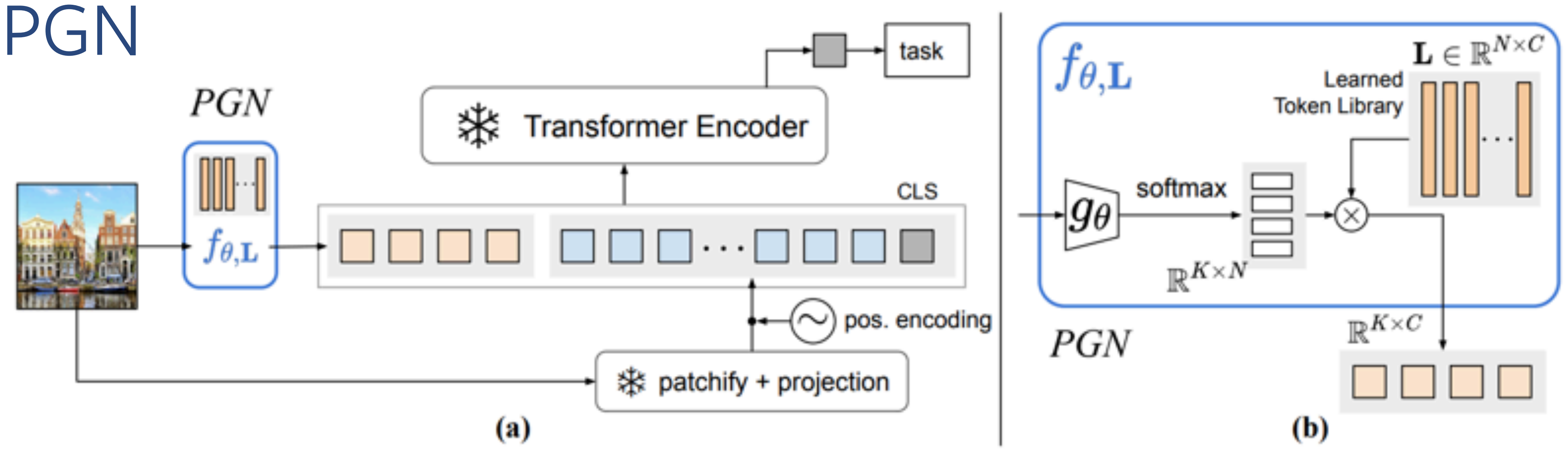


Simpler models that utilize pretrained models

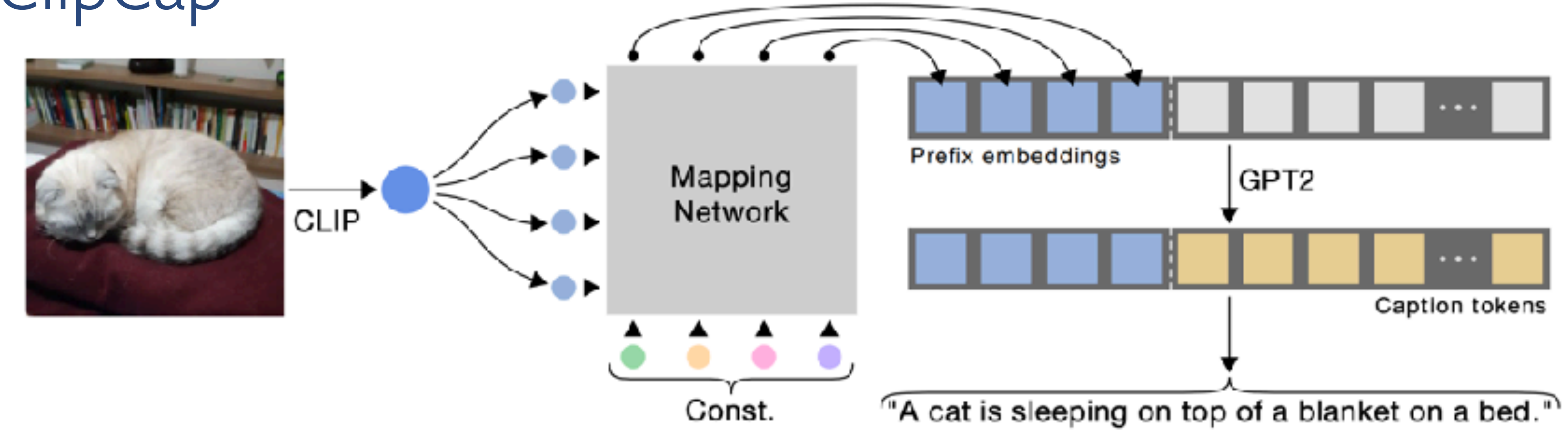
VPT



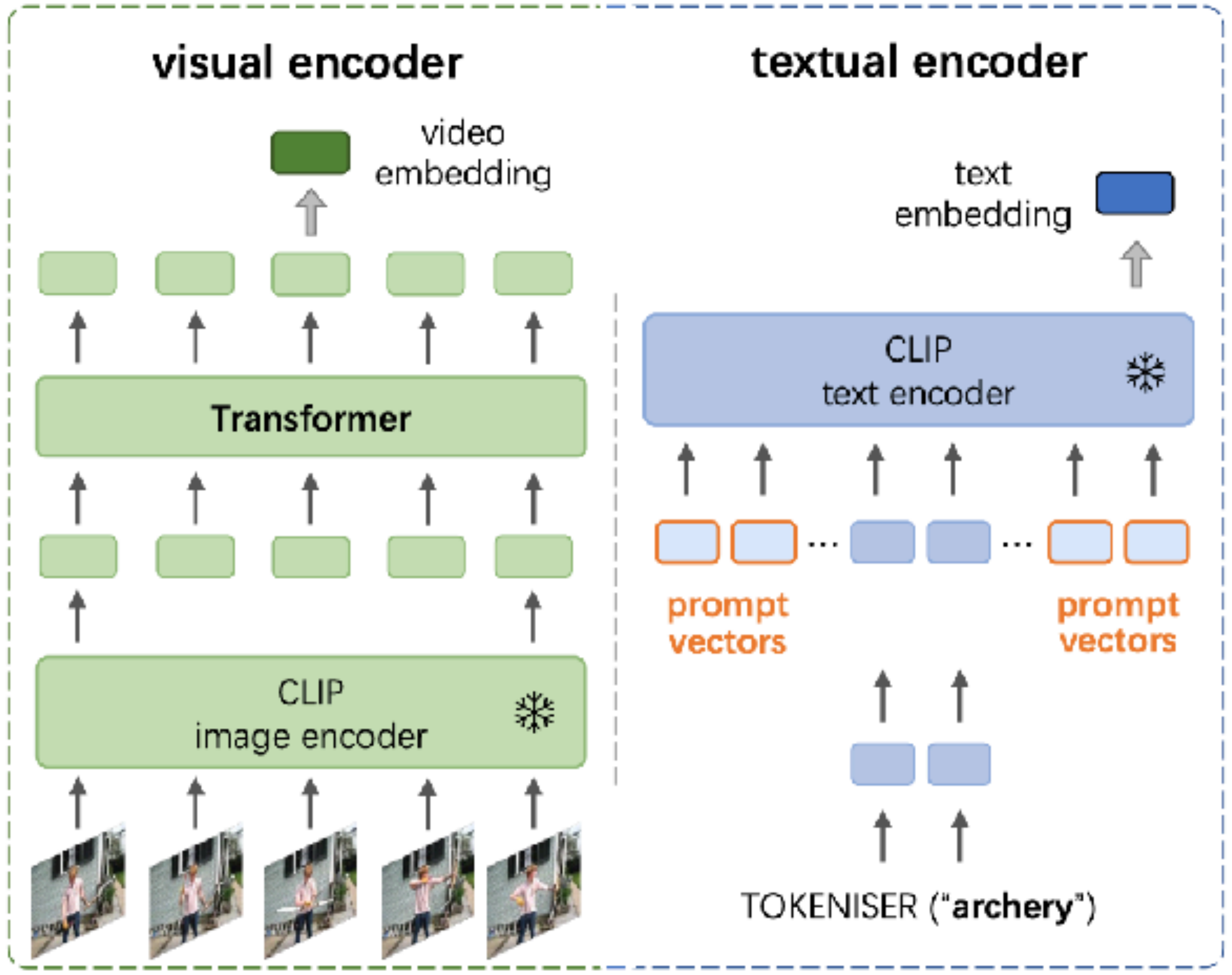
PGN



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)

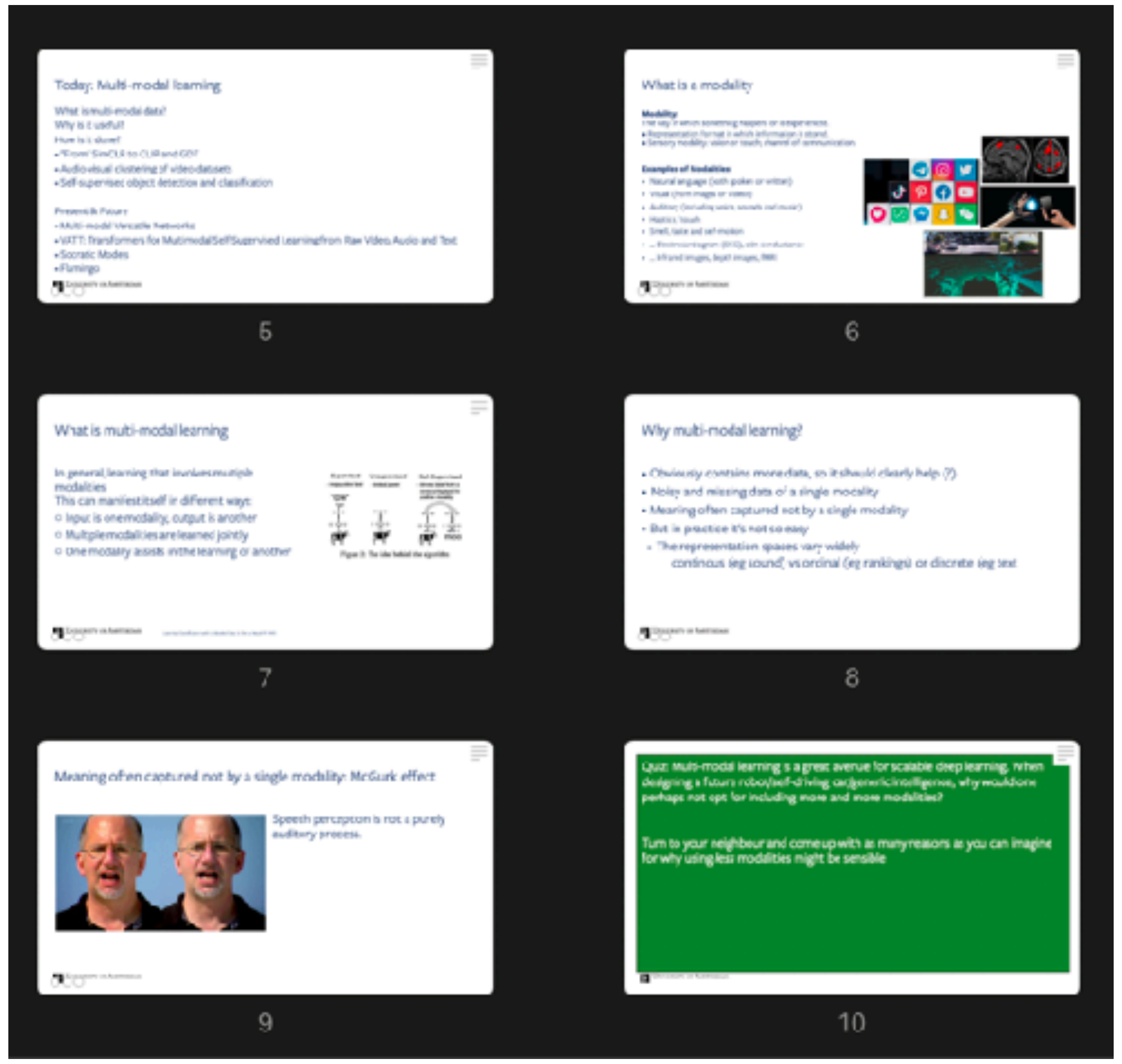
Combination of very large-scale "foundation" models

Move into more difficult domains like robotics

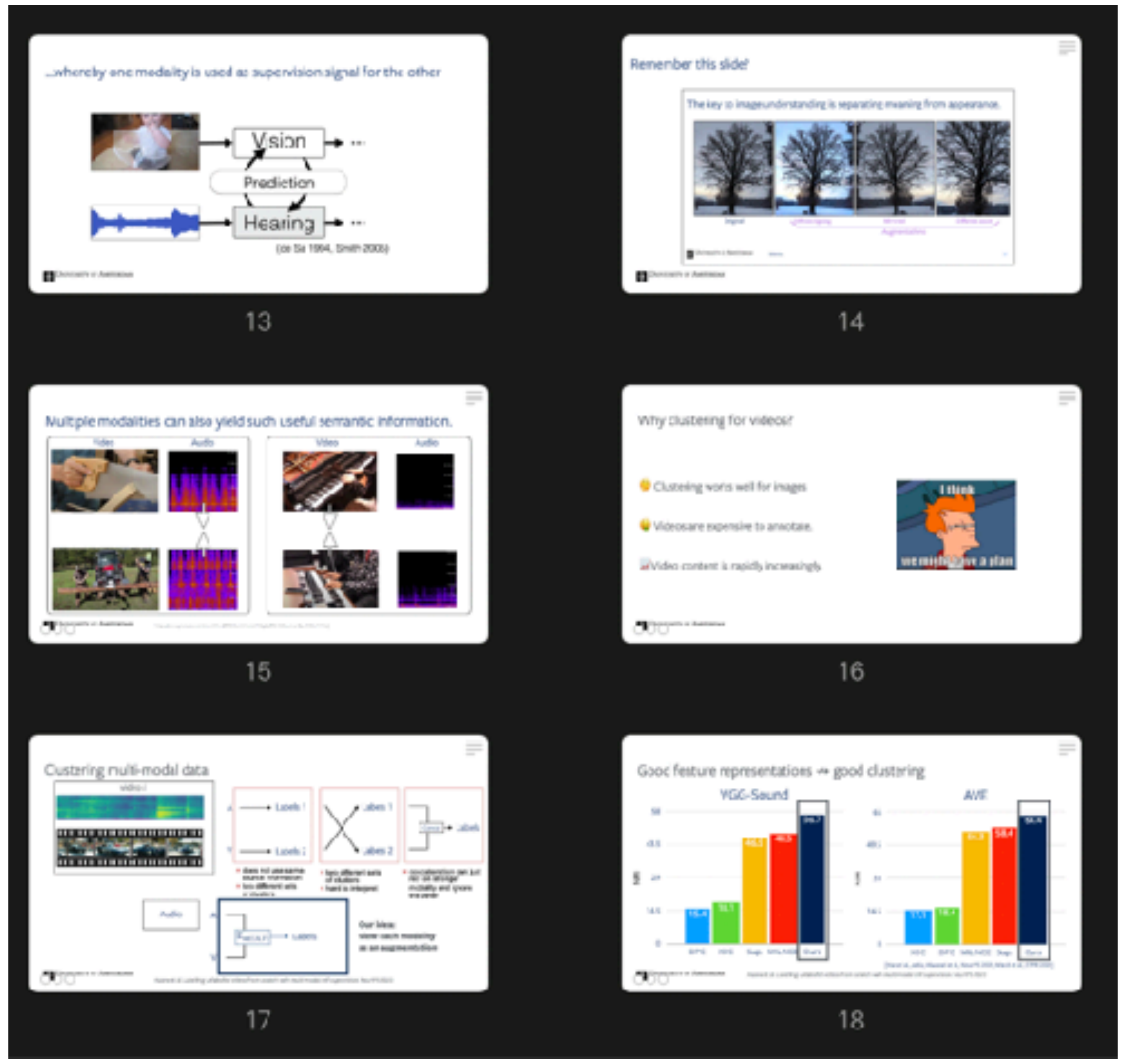
AI for science, self-coding language models

Renewed focus and work on privacy and bias (soon)

Summary



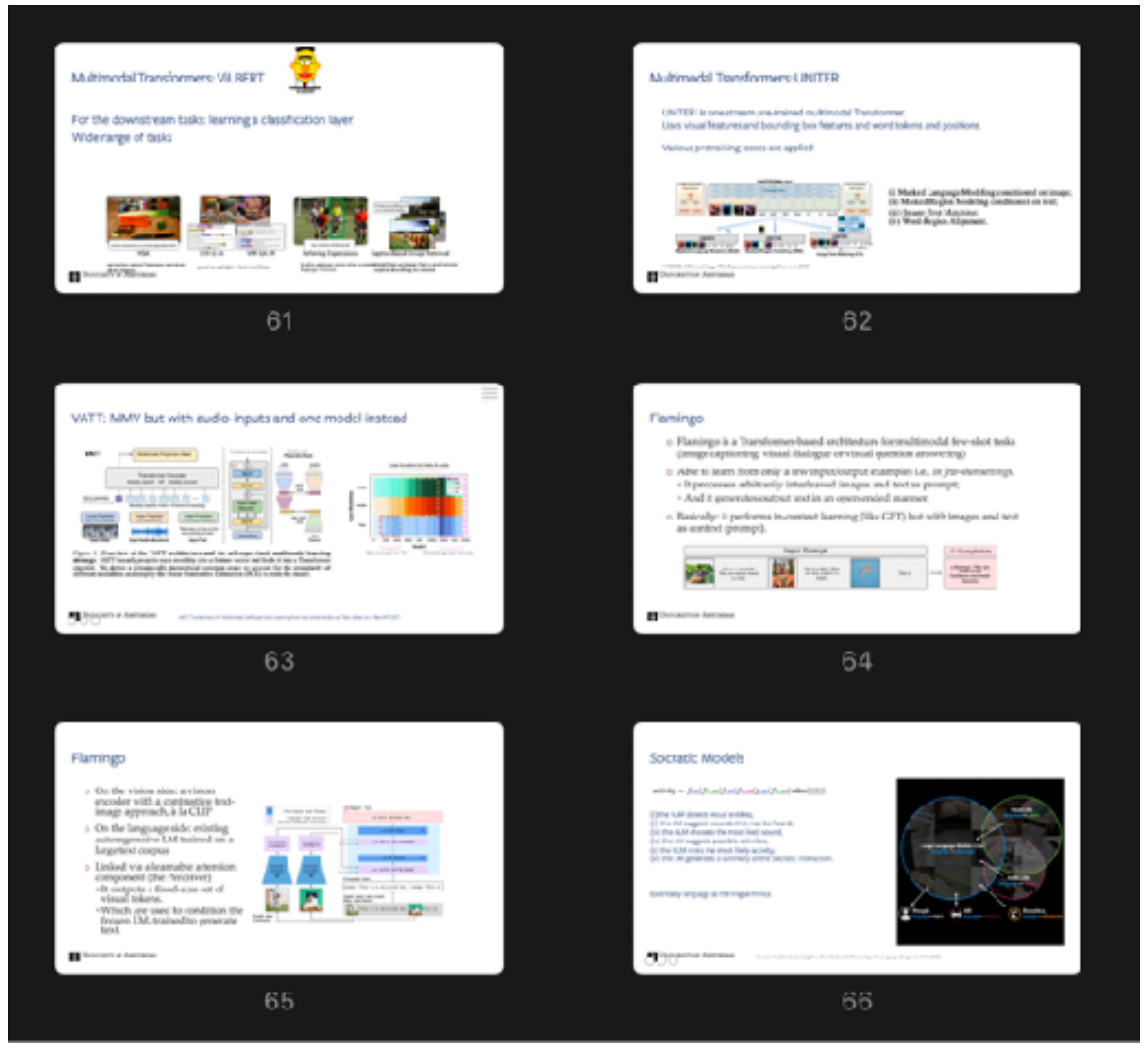
What and why multi-modal



Audio-visual clustering

Multi-modal variances learning

SSL object detection



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.