Action recognition in videos

Cordelia Schmid

Inria

Automatic video understanding

- Huge amount of video is available and growing daily

TV-channels recorded since 60's

30k hours of videos uploaded every hour

770M surveillance cameras world-wide

Automatic video understanding

- Classification of short clips, i.e. answer phone, shake hands

answer phone

hand shake

Hollywood dataset

Automatic video understanding

- Classification of activities, i.e. birthday party, groom an animal

Birthday party

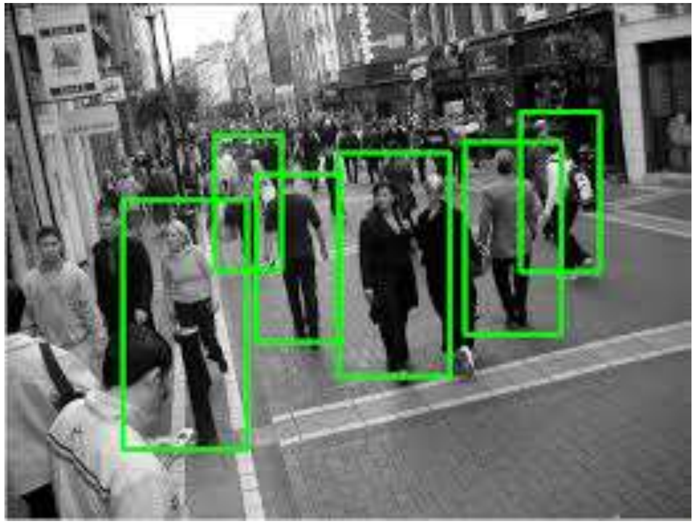
Grooming an animal

TrecVid Multi-media event detection task (MED)

Automatic video understanding

- Car safety & self-driving and video surveillance
  - Detection of humans (pedestrians) and their motion, detection of unusual behavior

Courtesy Volvo



Courtesy Embedded Vision Alliance

Automatic video understanding

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Automatic video understanding

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Automatic video understanding

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Automatic video understanding

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Action recognition - difficulties

- Large variations in appearance
  - Viewpoint changes
  - Intra-class variation
  - Camera motion

Variation in appearance: viewpoint change

Variation in appearance: intra-class variation

Variation in appearance: camera motion

Action recognition - difficulties

- Large variations in appearance
  - Viewpoint changes
  - Intra-class variation
  - Camera motion

- Manual collection of training data is difficult
  - Many action classes, rare occurrence
  - Pose, object and interaction annotation often a plus

- Action vocabulary is not well defined
  - What is the action granularity?
  - How to represent composite actions?

Action recognition – approaches

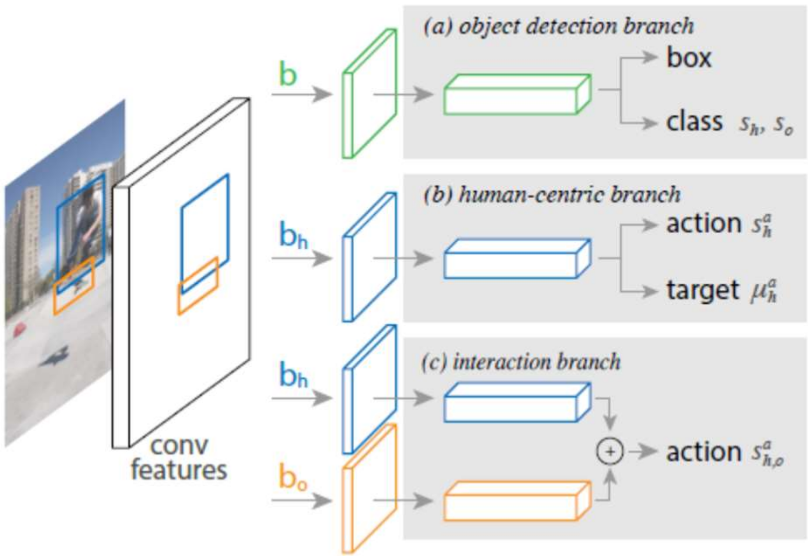
- Action recognition from still images
  - Detect human pose + interaction with objects

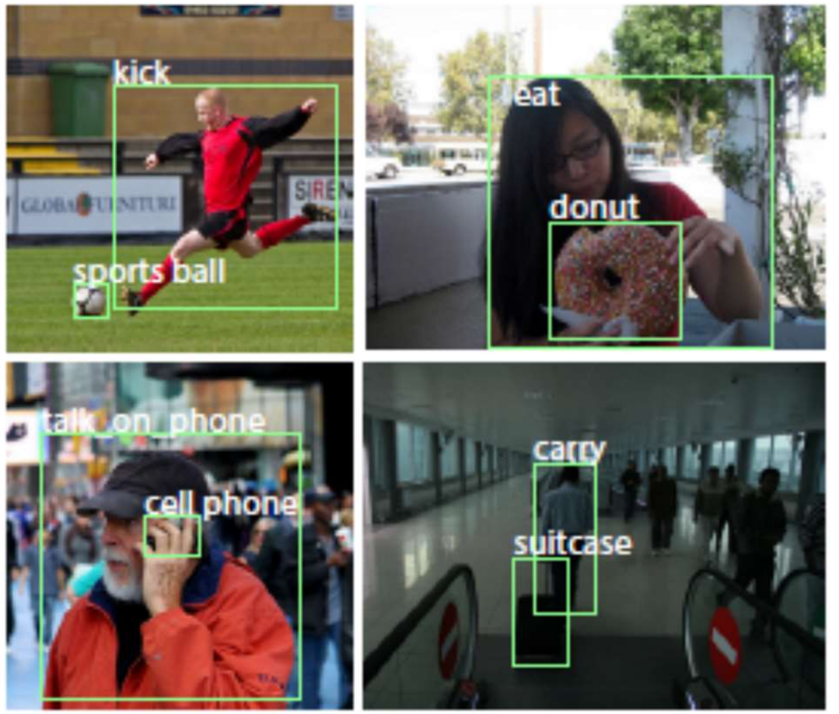
PASCAL VOC Human action classification dataset

[Weakly Supervised Learning of Interactions between Humans and Objects, Prest et al., PAMI 2012]

Action recognition – approaches

- Action recognition from still images
  - Human pose + interaction with objects





[Detecting and Recognizing Human-Object Interactions. G. Gkioxari, R. Girshick, P. Dollar and K. He. CVPR 2018]

Action recognition – approaches

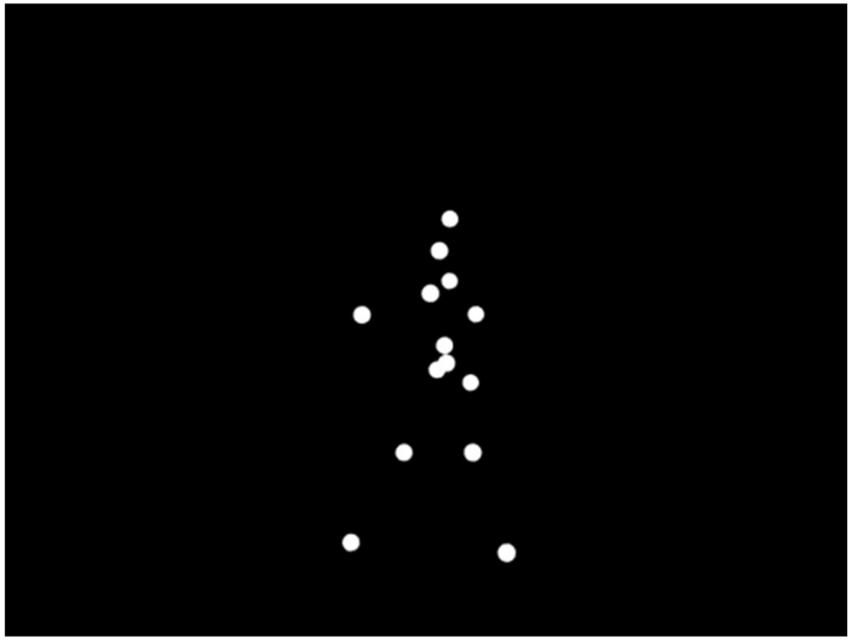
- Motion information necessary to disambiguate actions

Open or close door?

- Motion often sufficient by itself

Motion perception

- Johansson [1973] pioneered studies on sequence based human motion analysis
- Moving light displays enable identification of motion, familiar people and gender



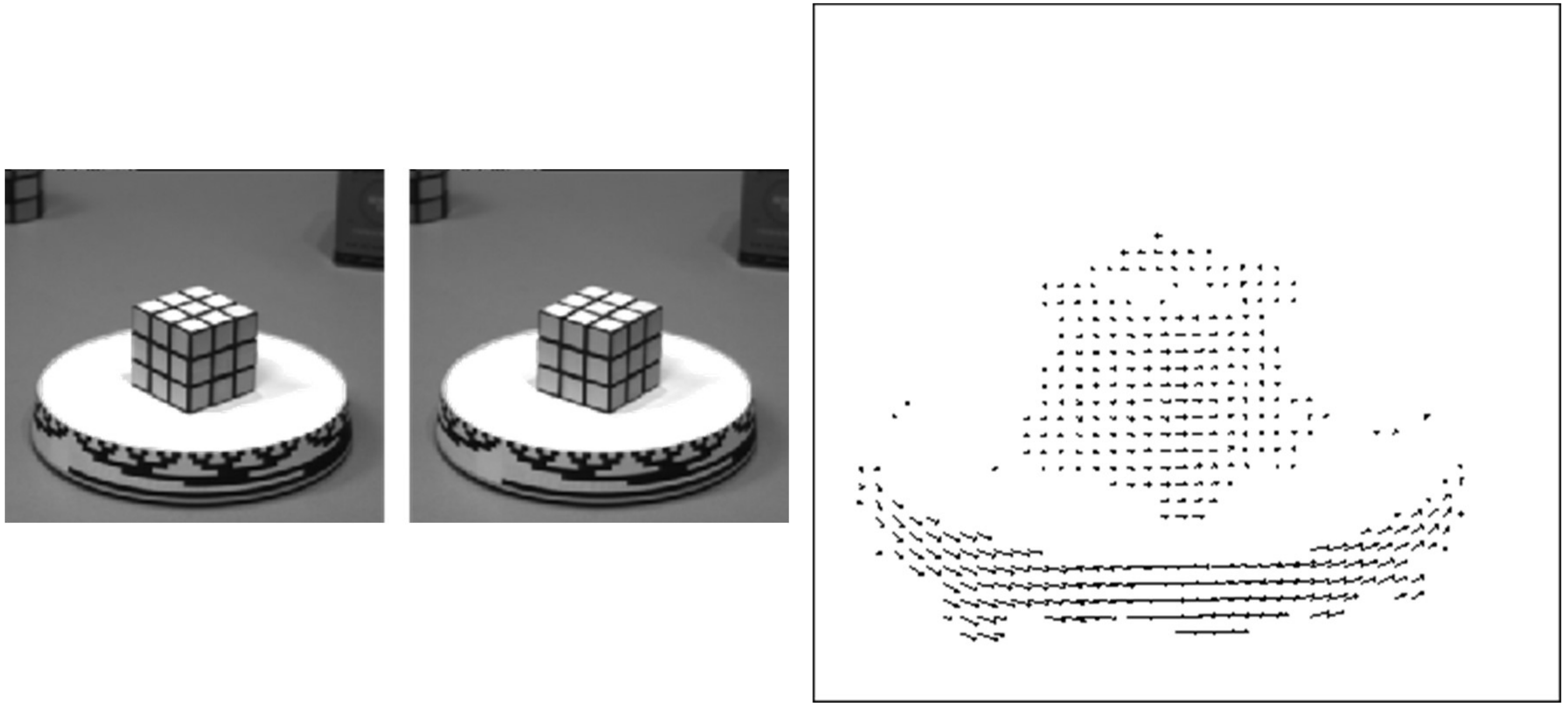
male walker

Overview

- *Optical flow*
- Video classification
- Multi-modal / LLM-based video understanding

Motion field

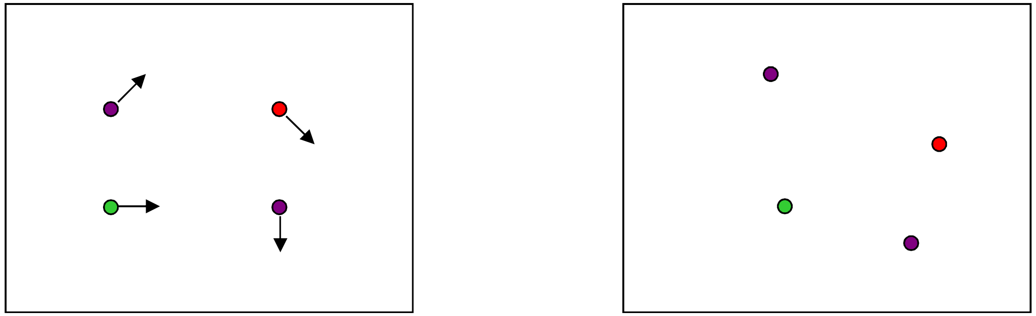
- The motion field is the projection of the 3D scene motion into the image



Optical flow

- Definition:
  - optical flow is the *apparent* motion of brightness patterns in the image
- Ideally, optical flow would be the same as the motion field
  - However, apparent motion can be caused by lighting changes without any actual motion
  - For example: a uniform rotating sphere under fixed lighting
    vs. a stationary sphere under moving illumination

Estimating optical flow



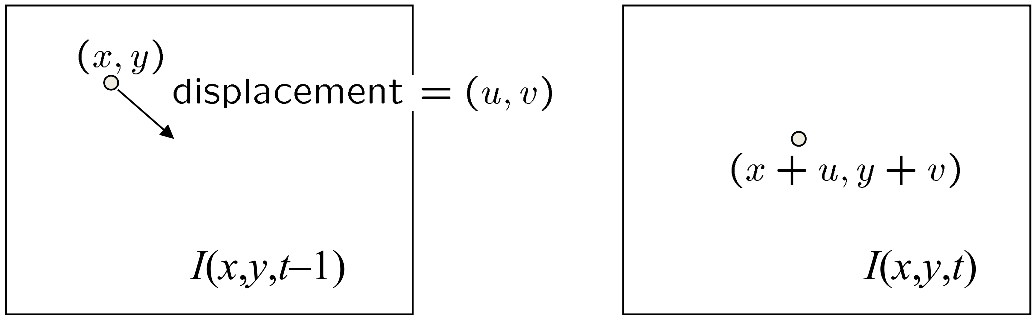
$I(x,y,t-1)$ $I(x,y,t)$

Given two subsequent frames, estimate the apparent motion field $u(x,y)$ and $v(x,y)$ between them

Key assumptions for the flow estimation in “classical” approaches

- Brightness constancy: projection of the same point looks the same in every frame
- Small motion: points do not move very far
- Spatial coherence: points move like their neighbors

The brightness constancy constraint



Brightness Constancy Equation:

$$I(x, y, t-1) = I(x + u(x, y), y + v(x, y), t)$$

Linearizing the right side using Taylor expansion (small motion):

$$I(x, y, t-1) \approx I(x, y, t) + I_x\, u(x, y) + I_y\, v(x, y)$$

Hence, $I_x u + I_y v + I_t \approx 0$

The brightness constancy constraint

$$I_x u + I_y v + I_t = 0$$

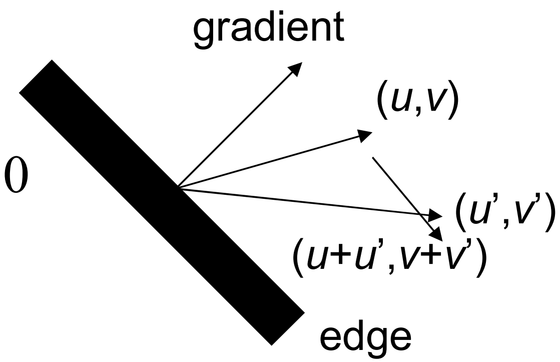
- How many equations and unknowns per pixel?
  - One equation, two unknowns

- What does this constraint mean?

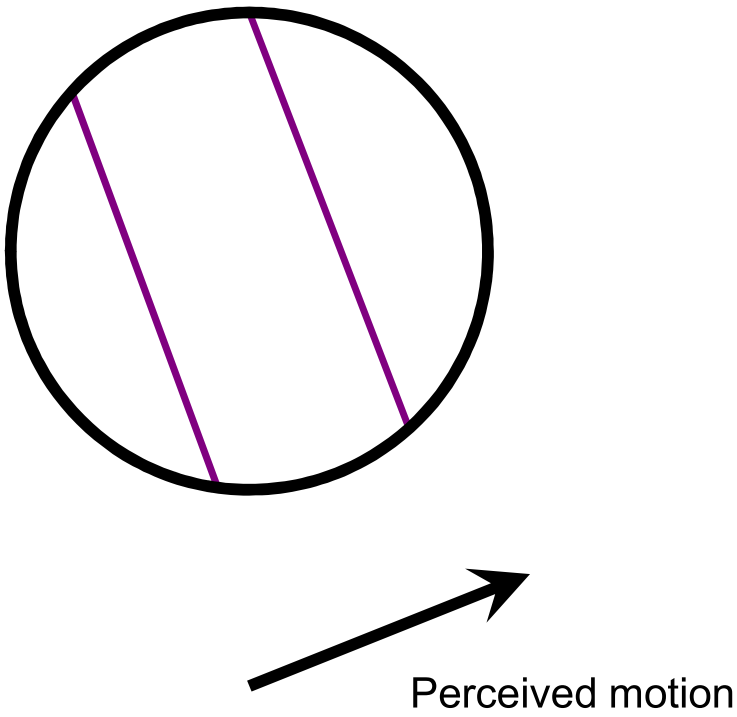
$$\nabla I \cdot (u, v) + I_t = 0$$

- The component of the flow perpendicular to the gradient (i.e., parallel to the edge) is unknown

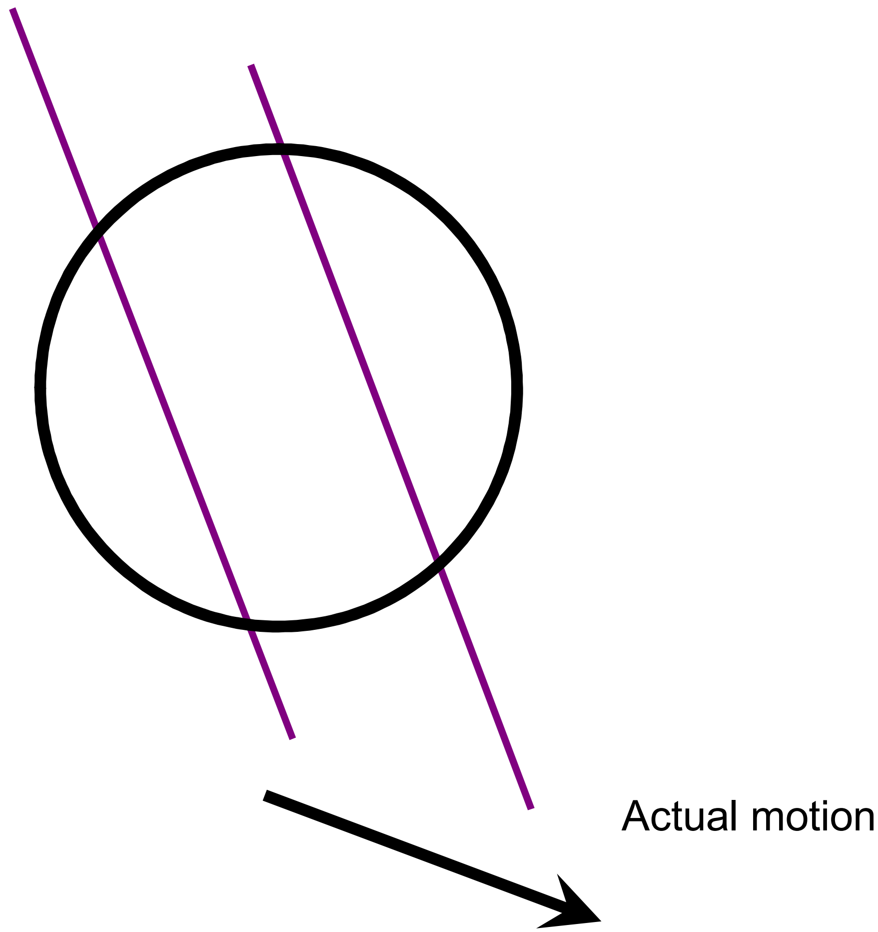
If (*u*, *v*) satisfies the equation,
so does (*u*+*u*', *v*+*v*') if $\nabla I \cdot (u', v') = 0$



The aperture problem



The aperture problem



Solving the aperture problem

- How to get more equations for a pixel?
- **Spatial coherence constraint:** pretend the pixel's neighbors have the same (u,v)
  - E.g., if we use a 5x5 window, that gives us 25 equations per pixel

$$\begin{bmatrix} I_x(\mathbf{x}_1) & I_y(\mathbf{x}_1) \\ I_x(\mathbf{x}_2) & I_y(\mathbf{x}_2) \\ \vdots & \vdots \\ I_x(\mathbf{x}_n) & I_y(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -\begin{bmatrix} I_t(\mathbf{x}_1) \\ I_t(\mathbf{x}_2) \\ \vdots \\ I_t(\mathbf{x}_n) \end{bmatrix}$$

B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*,1981.

Lucas-Kanade flow

- Linear least squares problem

$$\begin{bmatrix} I_x(\mathbf{x}_1) & I_y(\mathbf{x}_1) \\ I_x(\mathbf{x}_2) & I_y(\mathbf{x}_2) \\ \vdots & \vdots \\ I_x(\mathbf{x}_n) & I_y(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(\mathbf{x}_1) \\ I_t(\mathbf{x}_2) \\ \vdots \\ I_t(\mathbf{x}_n) \end{bmatrix}$$

$$\underset{n\times 2}{\mathbf{A}} \; \underset{2\times 1}{\mathbf{d}} = \underset{n\times 1}{\mathbf{b}}$$

Solution given by $(\mathbf{A}^T\mathbf{A})\mathbf{d} = \mathbf{A}^T\mathbf{b}$

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

The summations are over all pixels in the window

Lucas-Kanade flow

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -\begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

- Recall the Harris corner detector: M = $A^T A$ is the *second moment matrix*
- When is the system solvable?
    - By looking at the eigenvalues of the second moment matrix
    - The eigenvectors and eigenvalues of M relate to edge direction and magnitude
    - The eigenvector associated with the larger eigenvalue points in the direction of fastest intensity change, and the other eigenvector is orthogonal to it

Uniform region

– gradients have small magnitude
– small λ_1, small λ_2
– system is ill-conditioned

Edge

– gradients have one dominant direction

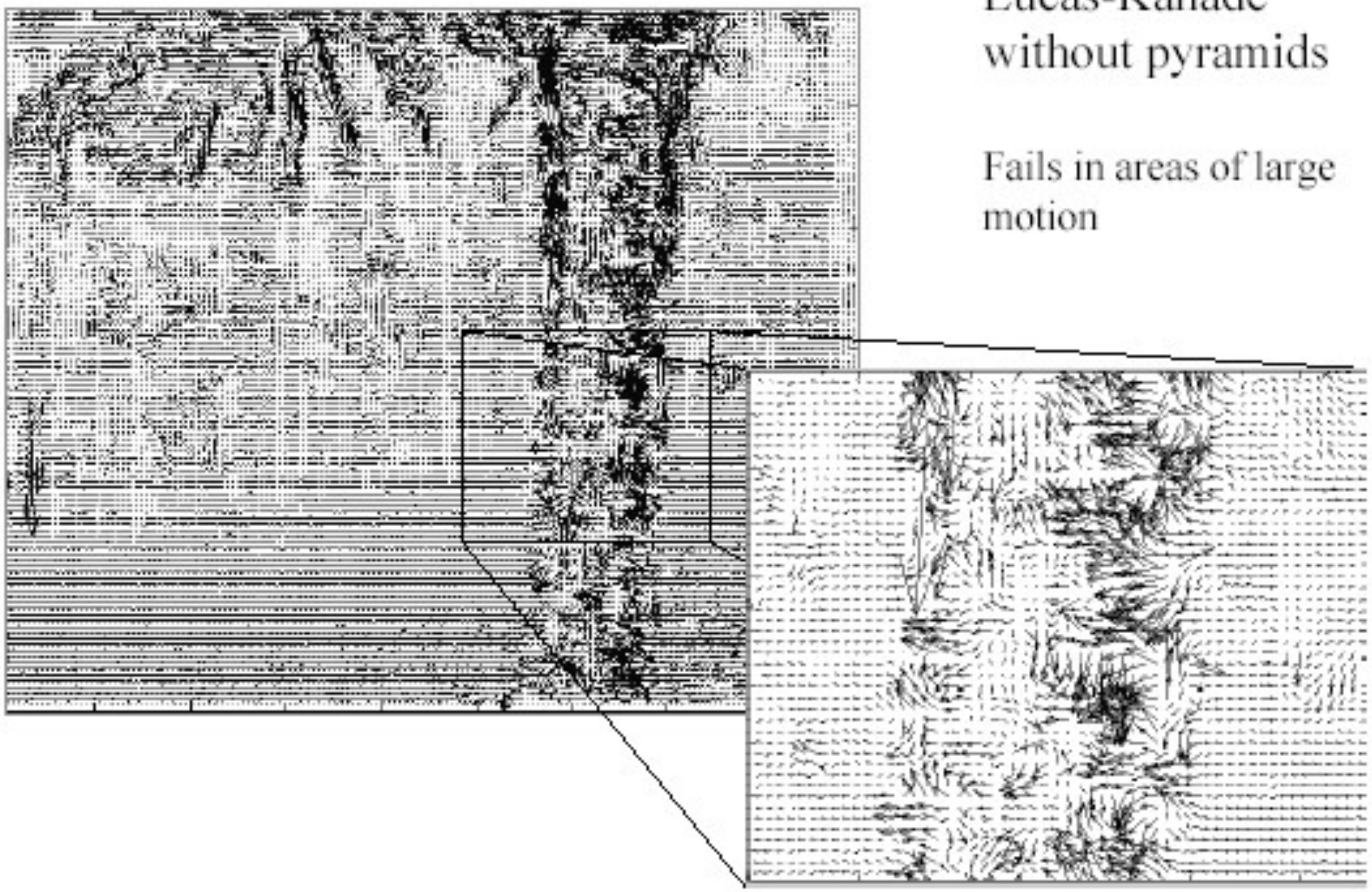
– large λ_1, small λ_2

– system is ill-conditioned

High-texture or corner region

– gradients have different directions, large magnitudes
– large λ_1, large λ_2
– system is well-conditioned

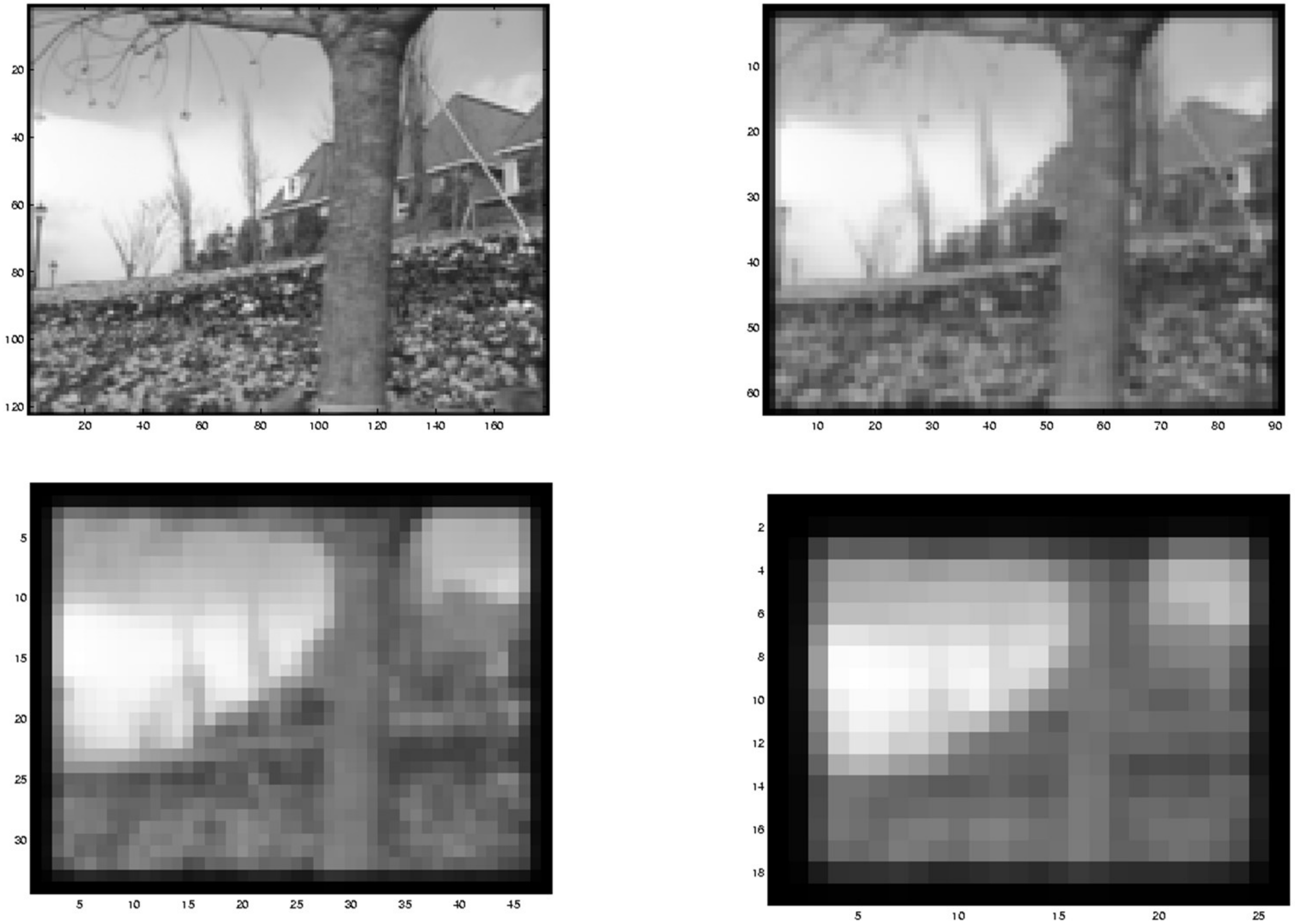
Optical Flow Results



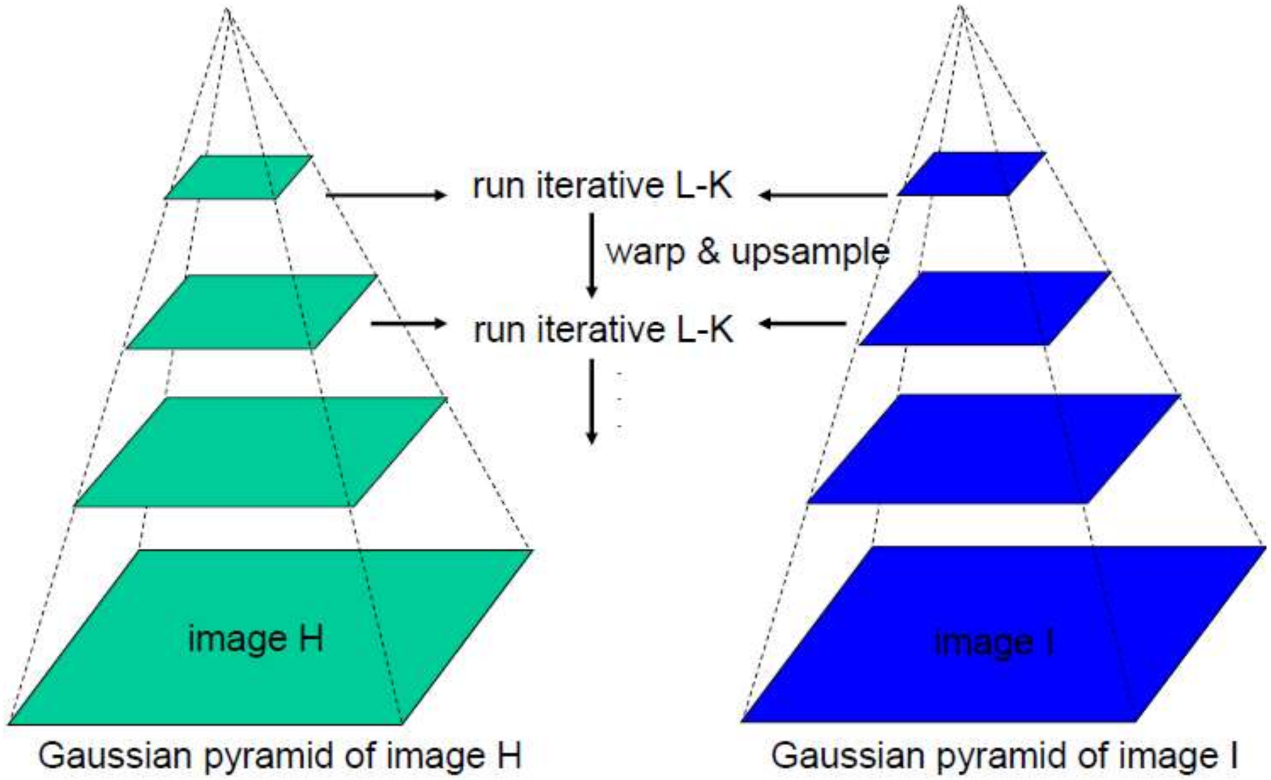
Lucas-Kanade without pyramids

Fails in areas of large motion

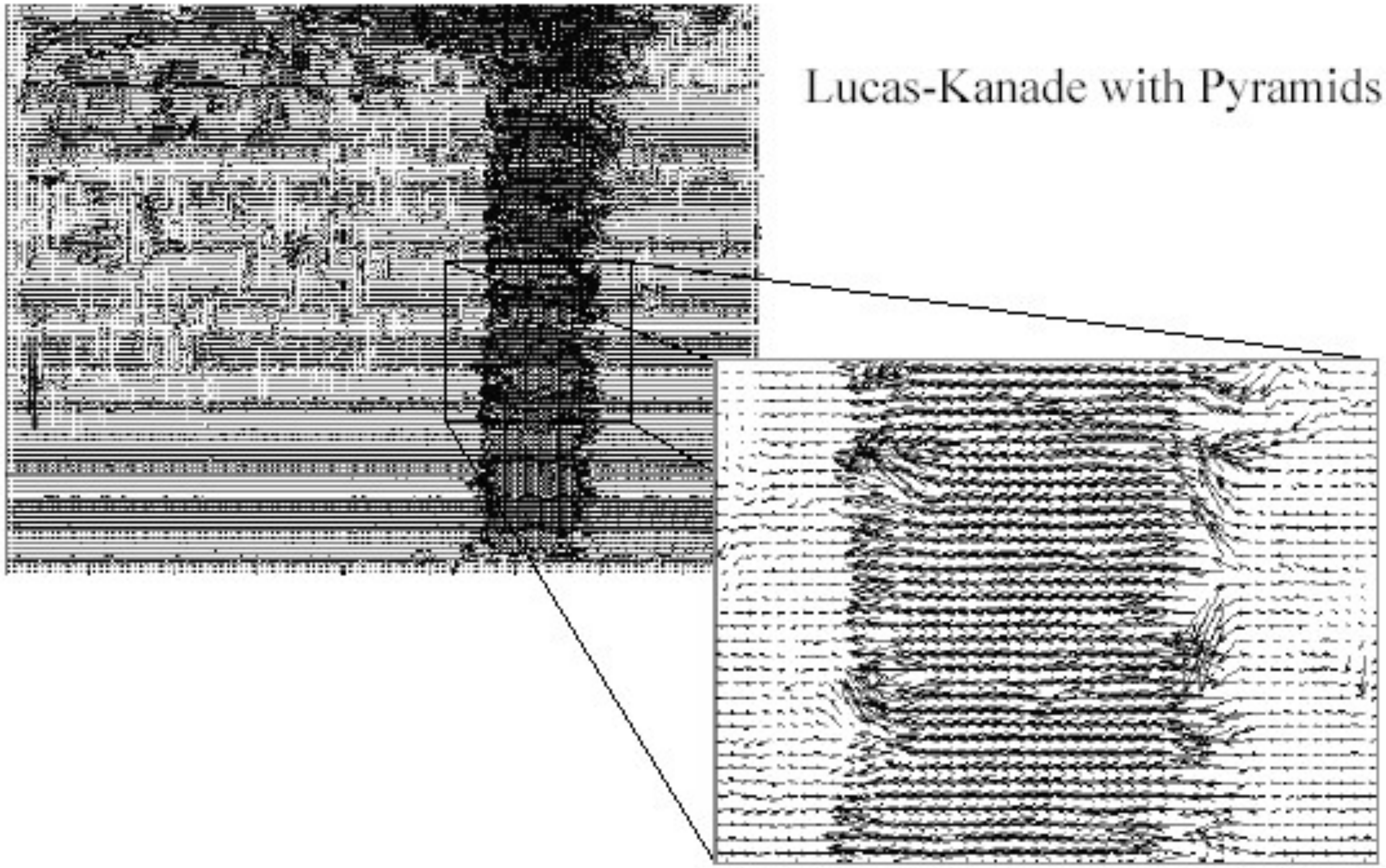
Multi-resolution registration



Coarse to fine optical flow estimation



Optical Flow Results



Lucas-Kanade with Pyramids

Horn & Schunck algorithm

Additional smoothness constraint :

- nearby point have similar optical flow
- additional constraint $||\nabla u||^2, \ ||\nabla v||^2$ small

$$e_s = \iint ((u_x^2 + u_y^2) + (v_x^2 + v_y^2))dxdy,$$

In addition to OF constraint equation term

$$e_c = \iint (I_x u + I_y v + I_t)^2 dxdy,$$

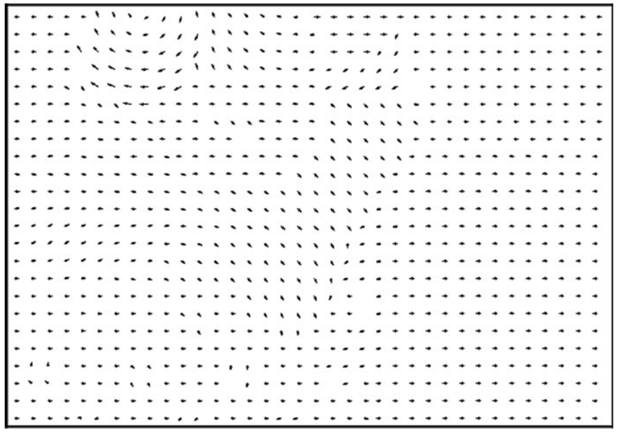
$\boxed{\text{minimize } e_s+\lambda e_c}$ λ regularization parameter

Coupled PDEs solved with iterative methods + finite differences

B.K.P. Horn and B.G. Schunck, "Determining optical flow." *Artificial Intelligence*,1981

Horn & Schunck

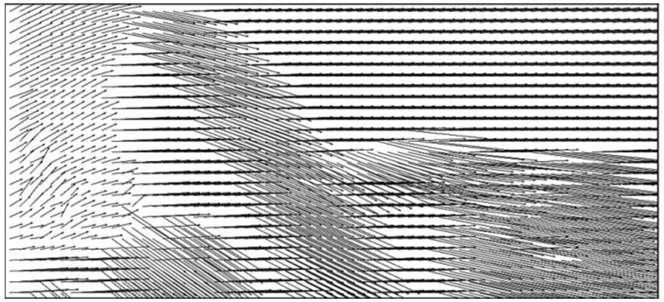
- Works well for small displacements
  - For example Middlebury sequence



Large displacement estimation in optical flow

Large displacement is difficult for optical flow estimation due to:

- locality and smoothness constraints



MPI Sintel dataset

Large displacement optical flow

- Classical optical flow [Horn and Schunck 1981]
  - energy: $$E(\mathbf{w}) = \iint \underbrace{E_{data}}_{\text{color/gradient constancy}} + \alpha \underbrace{E_{smooth}}_{\text{smoothness constraint}} \mathbf{dx}$$
  - minimization using a coarse-to-fine scheme

- Large displacement approaches:
  - LDOF [Brox and Malik 2011]
    a matching term, penalizing the difference between flow and HOG matches
    $$E(\mathbf{w}) = \iint E_{data} + \alpha E_{smooth} + \beta E_{match} \mathbf{dx}$$
  - MDP-Flow2 [Xu *et al.* 2012]
    expensive fusion of matches (SIFT + PatchMatch) and estimated flow at each level
  - DeepFlow [Weinzaepfel *et al.* 2013]
    deep matching + flow refinement with variational approach

Experimental results: datasets

- MPI-Sintel [Butler *et al.* 2012]
  - ► sequences from a realistic animated movie
  - ► large displacements (>20px for 17.5% of pixels)
  - ► atmospheric effects and motion blur

Experimental results: datasets

- KITTI [Geiger *et al.* 2013]
  - ► sequences captured from a driving platform
  - ► large displacements (>20px for 16% of pixels)
  - ► real-world: lightings, surfaces, materials




Experimental results: sample results

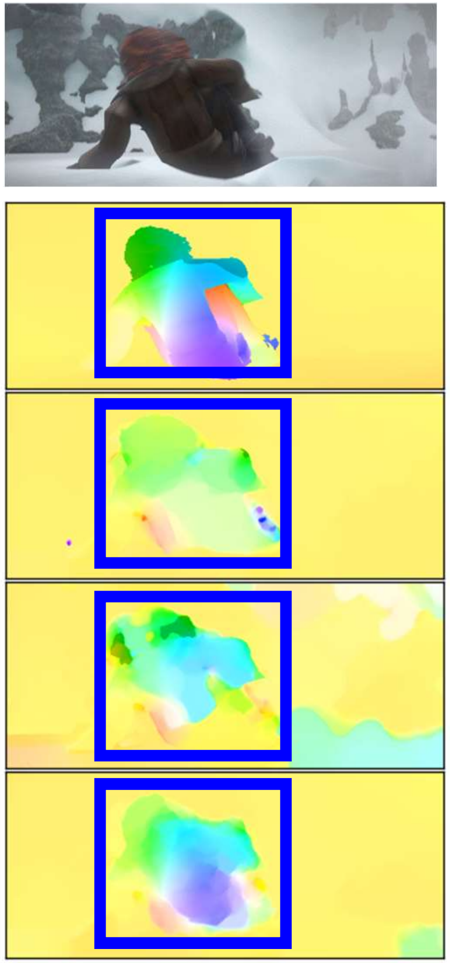
Ground-truth

LDOF [Brox & Malik 2011]

MDP-Flow2 [Xu *et al.* 2012]

DeepFlow [Weinzaepfel *et al.* 2013]

Experimental results: sample results



Ground-truth

LDOF [Brox & Malik 2011]

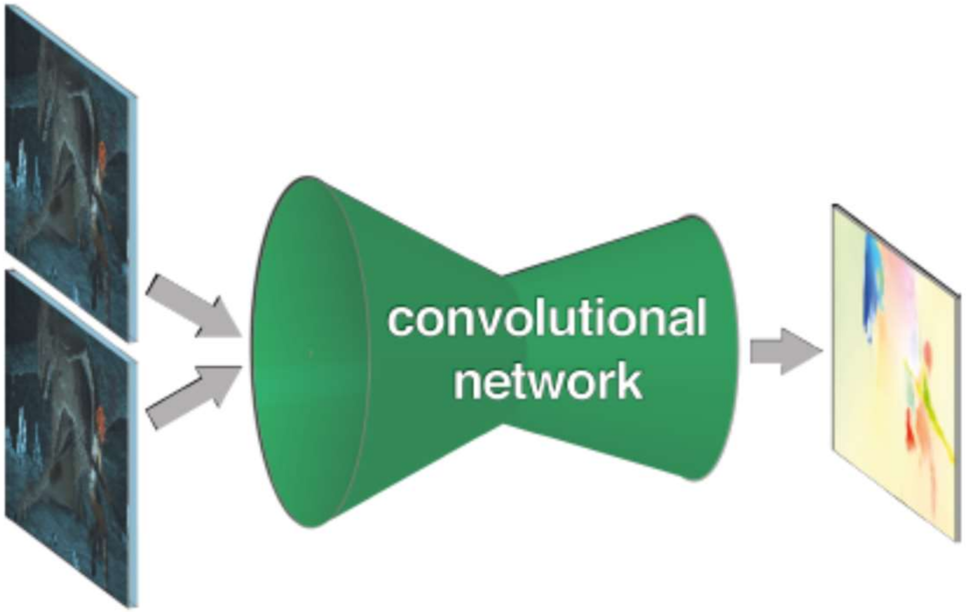
MDP-Flow2 [Xu *et al.* 2012]

DeepFlow [Weinzaepfel *et al.* 2013]

Methods – overview

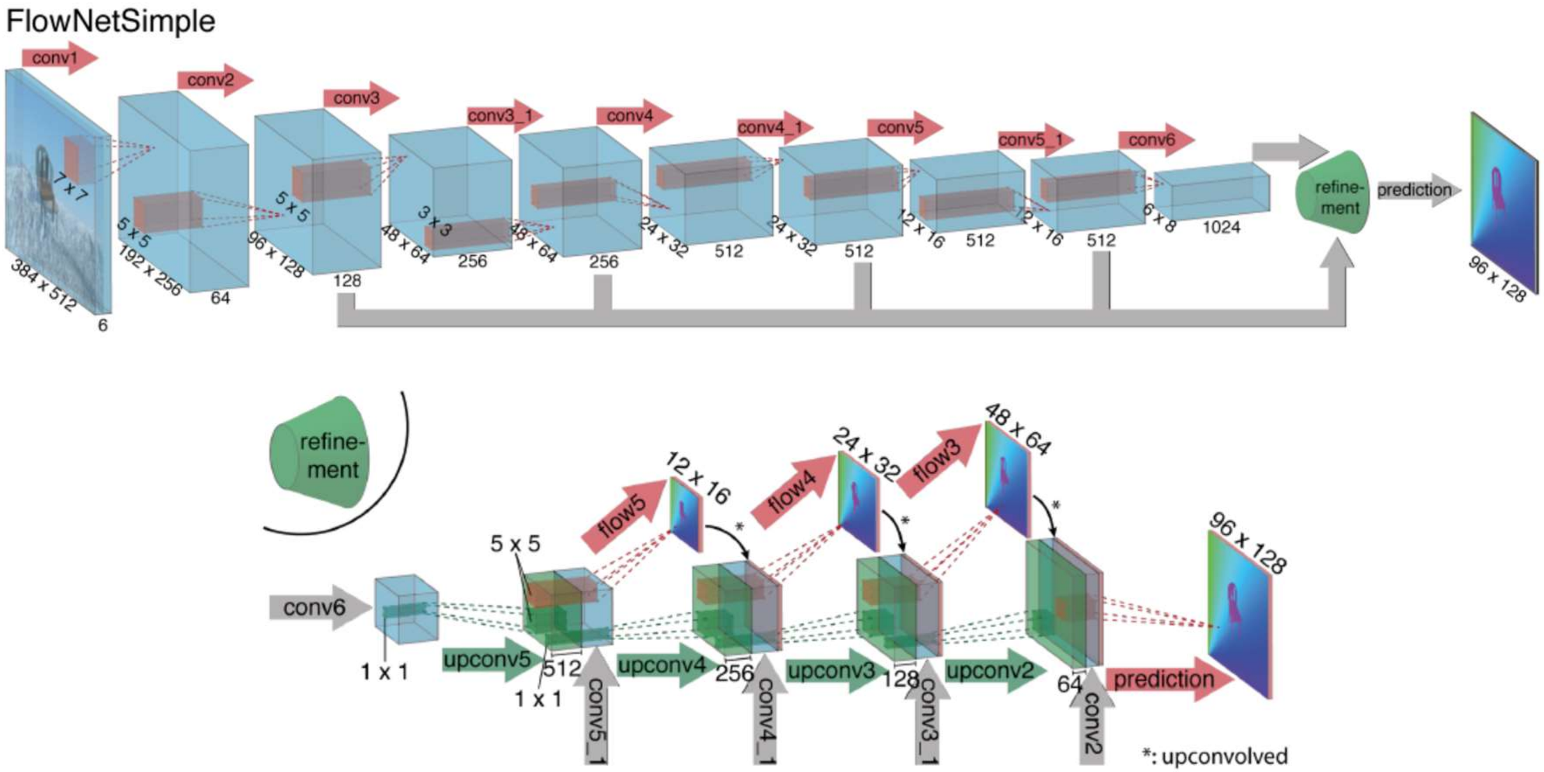
- Brightness constancy assumption
- + spatial coherence constraint: Lucas & Kanade, IJCAI'81
- + smoothness constraint: Horn & Schunk, AI'81
- + addition of matching term: Brox & Malik, PAMI'10
- recently: deep CNN based approaches

CNN to estimate optical flow: FlowNet

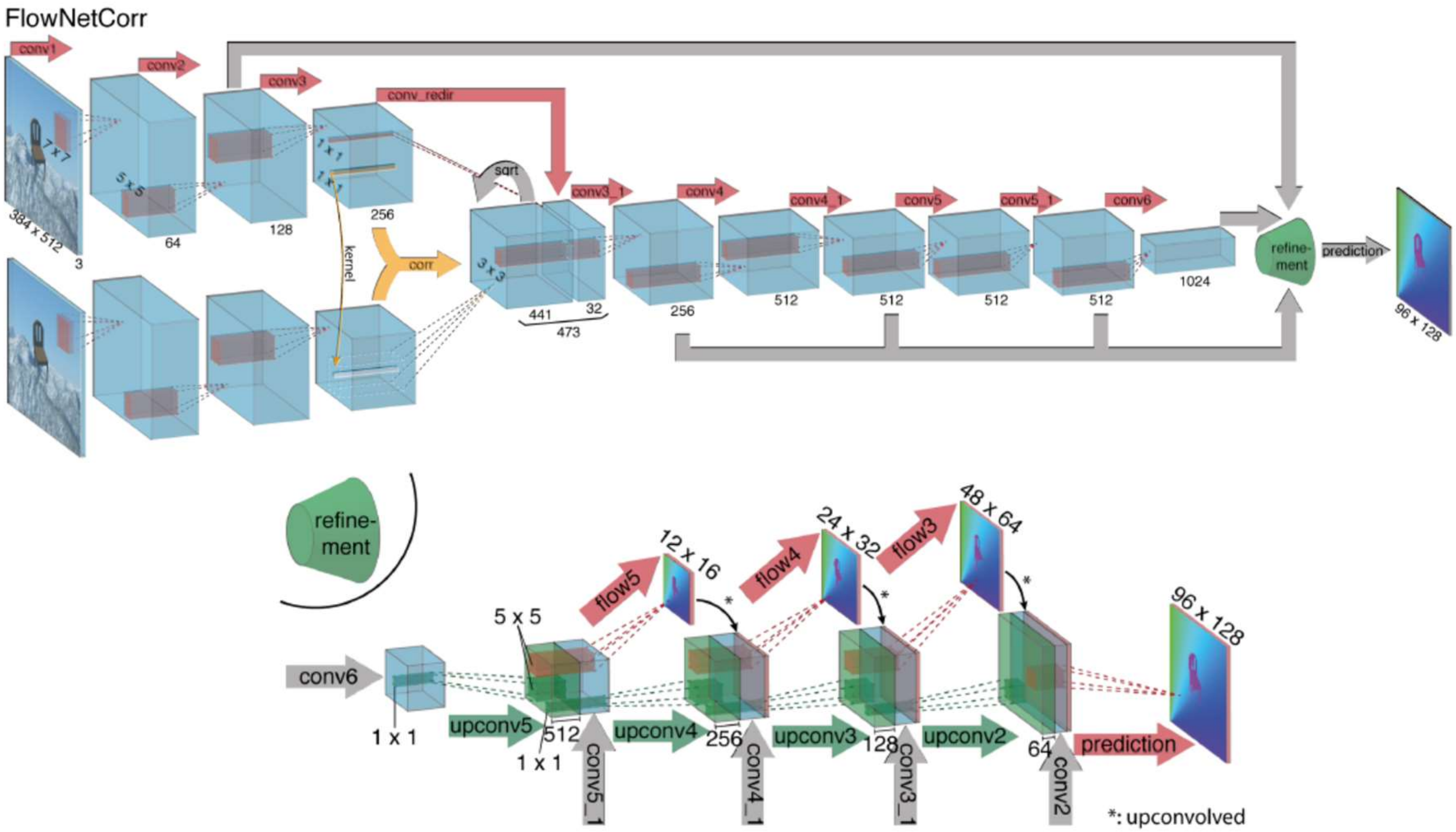


[A. Dosovitskiy et al. ICCV'15]

Architecture FlowNetSimple



Architecture FlowNetCorrelation



Synthetic dataset for training: Flying chairs

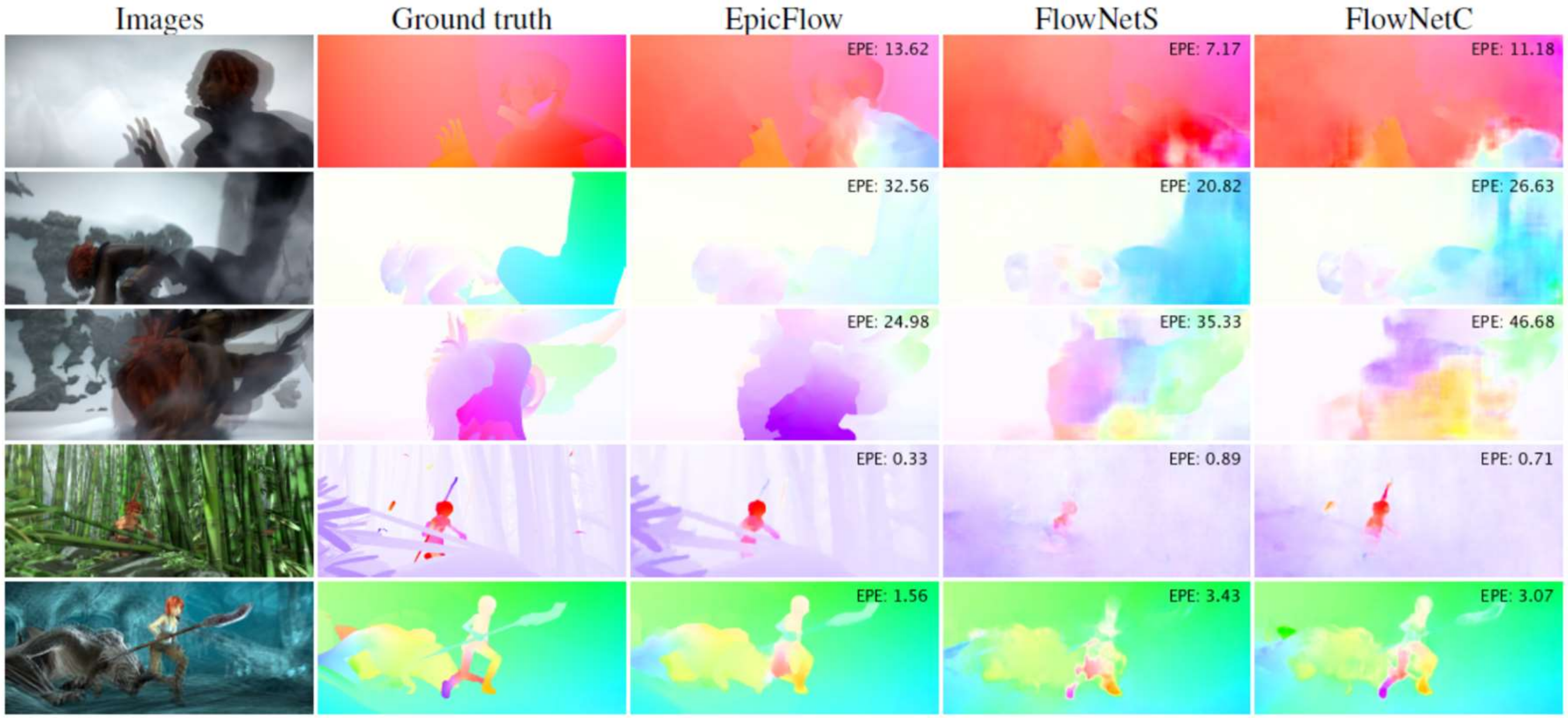
A dataset of approx. 23k image pairs

Experimental results

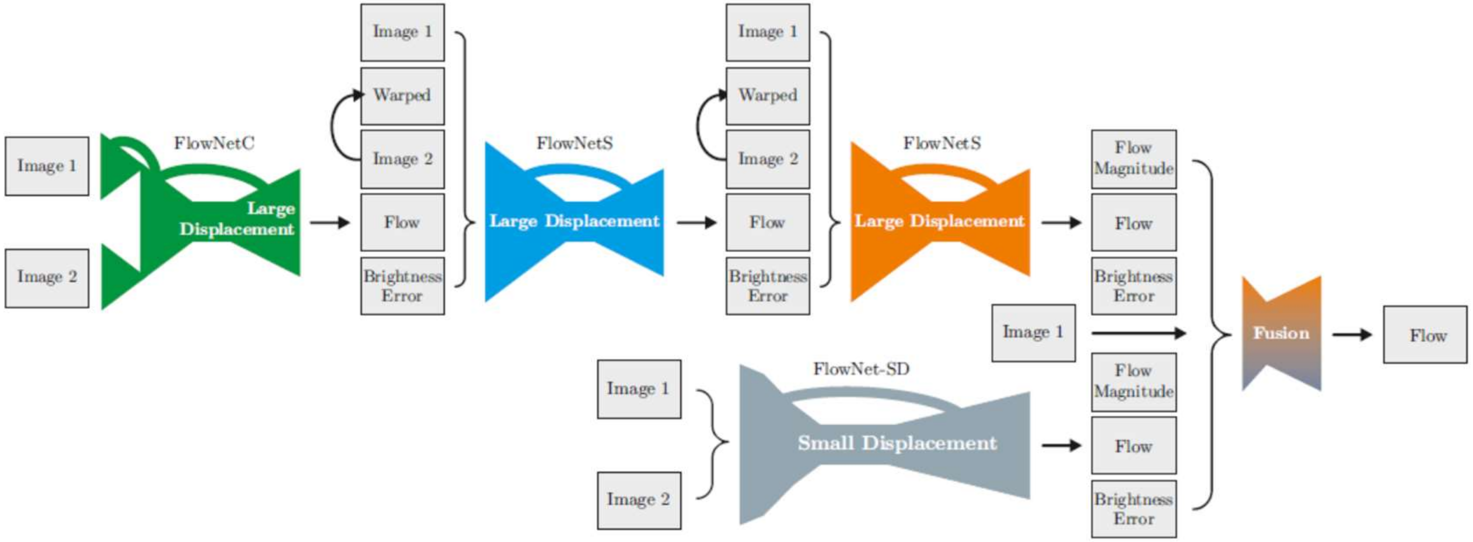
Method	Sintel Clean		Sintel Final	
	train	test	train	test
EpicFlow [30]	2.27	4.12	3.57	6.29
DeepFlow [35]	3.19	5.38	4.40	7.21
EPPM [3]	-	6.49	-	8.38
LDOF [6]	4.19	7.56	6.28	9.12
FlowNetS	4.50	7.42	5.45	8.43
FlowNetS+v	3.66	6.45	4.76	7.67
FlowNetS+ft	(3.66)	6.96	(4.44)	7.76
FlowNetS+ft+v	(2.97)	6.16	(4.07)	7.22
FlowNetC	4.31	7.28	5.87	8.81
FlowNetC+v	3.57	6.27	5.25	8.01
FlowNetC+ft	(3.78)	6.85	(5.28)	8.51
FlowNetC+ft+v	(3.20)	6.08	(4.83)	7.88

S: simple, C: correlation, v: variational refinement, ft:fine-tuning

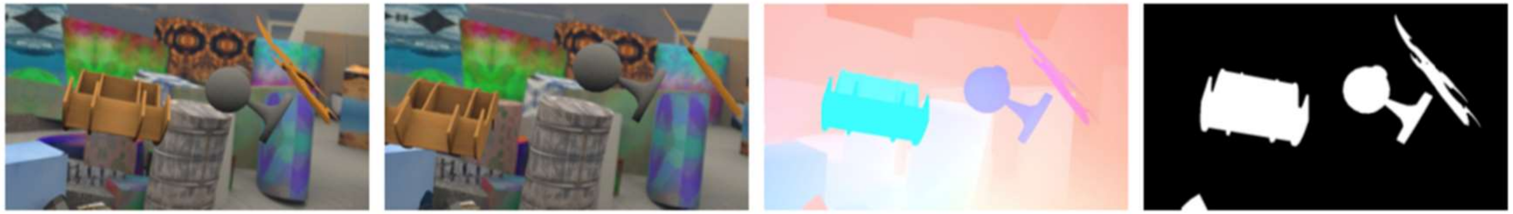
Experimental results



FlowNet2.0 [Ilg et al. CVPR'17]



FlyingThings3D [Mayer et al., CVPR'16]

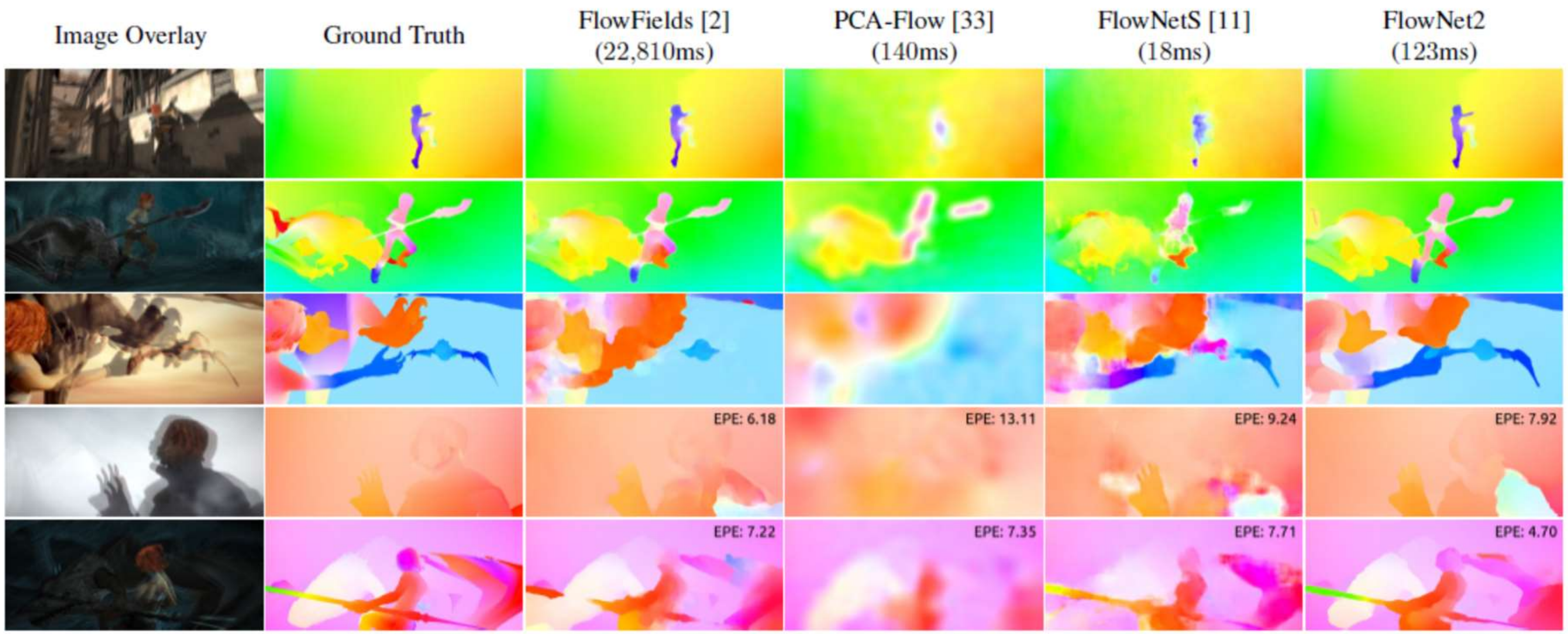


Stacking of networks

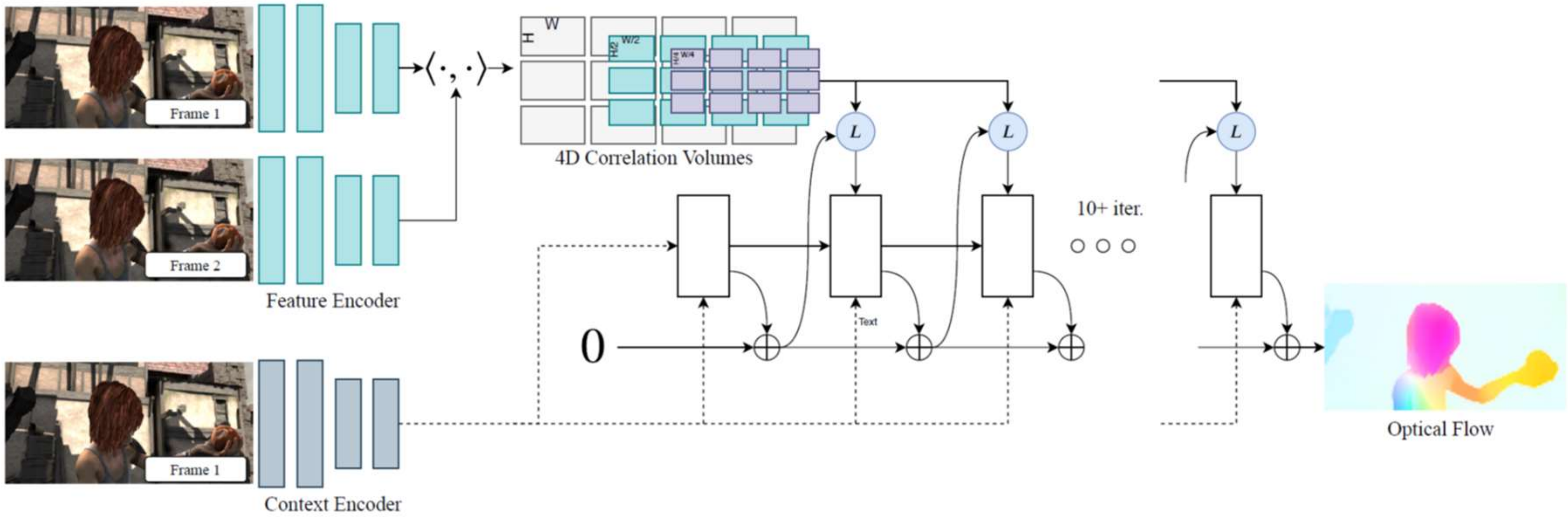
Stack architecture	Training enabled		Warping included	Warping gradient enabled	Loss after		EPE on Chairs test	EPE on Sintel train *clean*
	Net1	Net2			Net1	Net2		
Net1	✓	–	–	–	✓	–	3.01	3.79
Net1 + Net2	✗	✓	✗	–	–	✓	2.60	4.29
Net1 + Net2	✓	✓	✗	–	✗	✓	2.55	4.29
Net1 + Net2	✓	✓	✗	–	✓	✓	2.38	3.94
Net1 + W + Net2	✗	✓	✓	–	–	✓	1.94	**2.93**
Net1 + W + Net2	✓	✓	✓	✓	✗	✓	1.96	3.49
Net1 + W + Net2	✓	✓	✓	✓	✓	✓	**1.78**	3.33

Importance of warping

Optical flow results on Sintel



RAFT optical flow



- Feature extraction with CNNs
- Comparison between all features in the 2 images → 4D correlation volume
- Multi-scale representation of the 4D correlation volume
- Matching to the features of image 1
- Iterative updates which refine the current flow

[RAFT, Z. Teed and J. Deng, ECCV 2020]

RAFT optical flow – results

Fig. 3: Flow predictions on the Sintel test set.

Video object segmentation

- Segment the moving object in all the frames of a video

DAVIS (ground-truth)

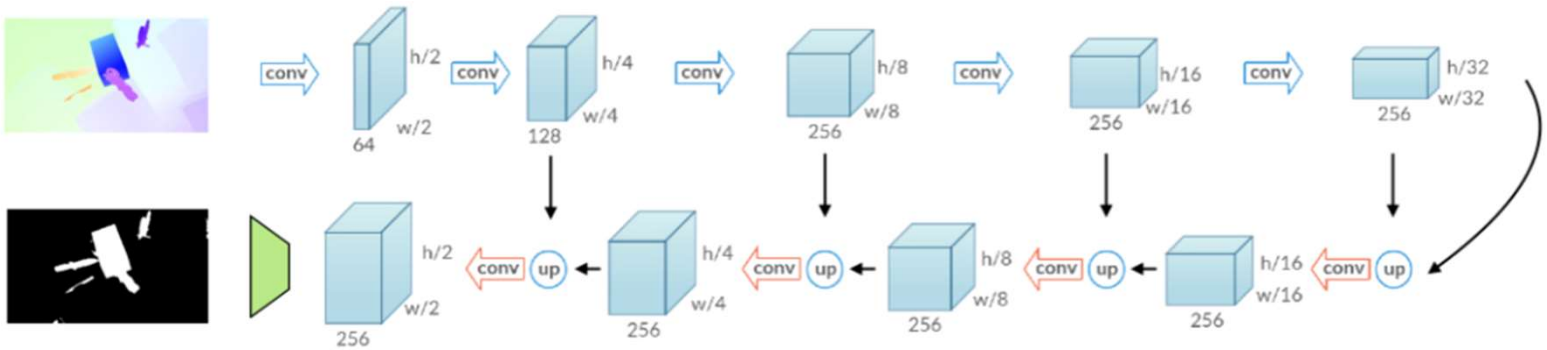
Challenges

- Strong camera or background motion

LDOF flow

DAVIS

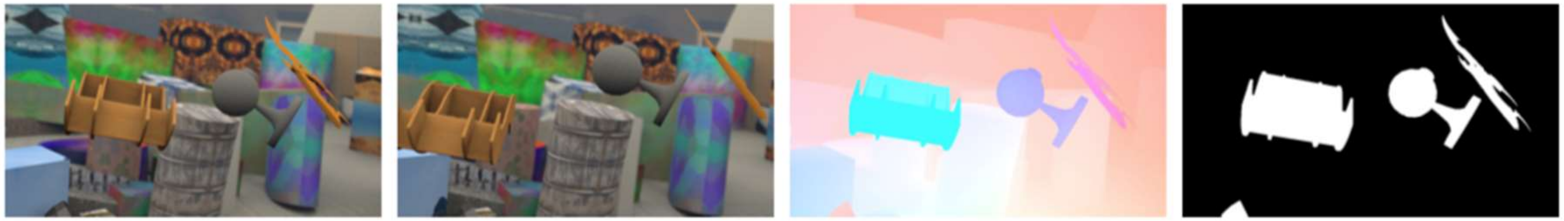
Network architecture – MP-Net



Convolutional/deconvolutional network, similar to U-Net

Training data

- FlyingThings3D dataset [Mayer et al., CVPR'16]
- 2700 synthetic, 10-frame stereo videos of random object flying in random trajectories (2250/450 training/test split)
- Ground-truth optical flow and camera data available
- Labels for moving object can be obtained from the data



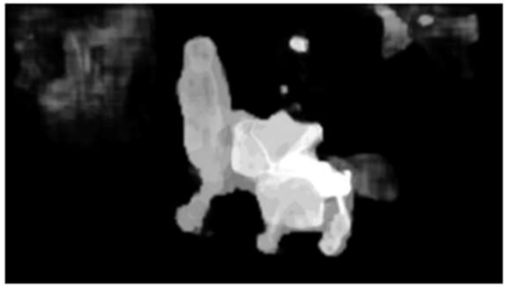
Results on FlyingThings3D test set

Motion estimation in real videos

- Flow estimation inaccuracies


DAVIS

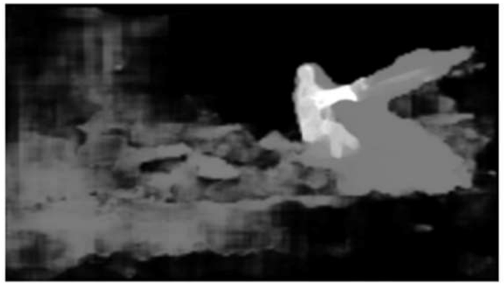

LDOF


MP-Net

- Background motion

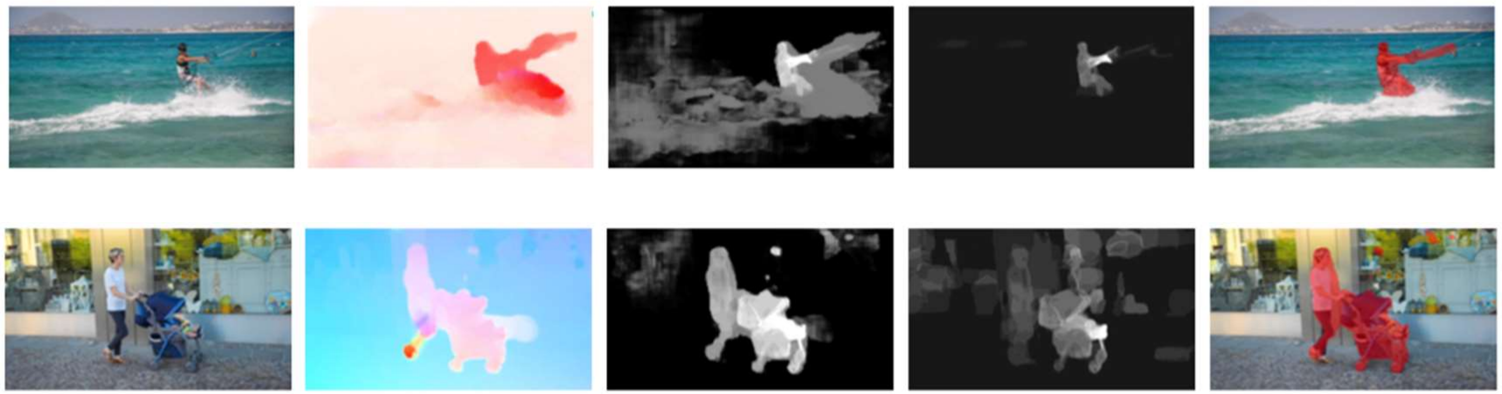

DAVIS


LDOF


MP-Net

Addition of an objectness measure

- Extract 100 object proposals per frame with SharpMask [Pinheiro et al., ECCV'16]
- Aggregate to obtain pixel-level objectness scores o_i
- Combine with the motion predictions m_i



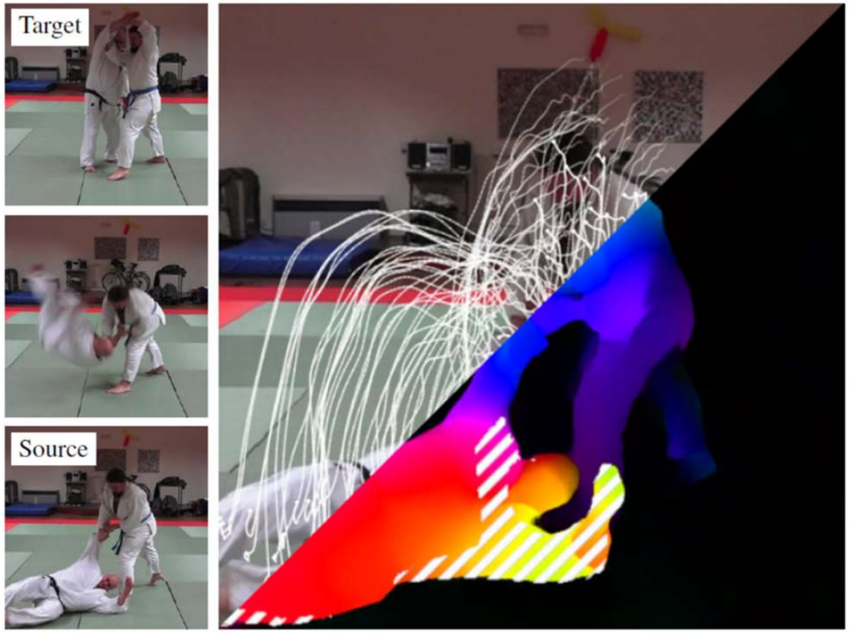
DAVIS LDOF MP-Net Objectness Result

FlowNet 2.0 Evaluation

Setting	LDOF flow	FLowNet 2.0 flow
MP-Net	52.4	62.6
MP-Net + Obj	63.3	69.0
MP-Net + Obj + CRF	69.7	72.5

Mean IoU on DAVIS trainval set

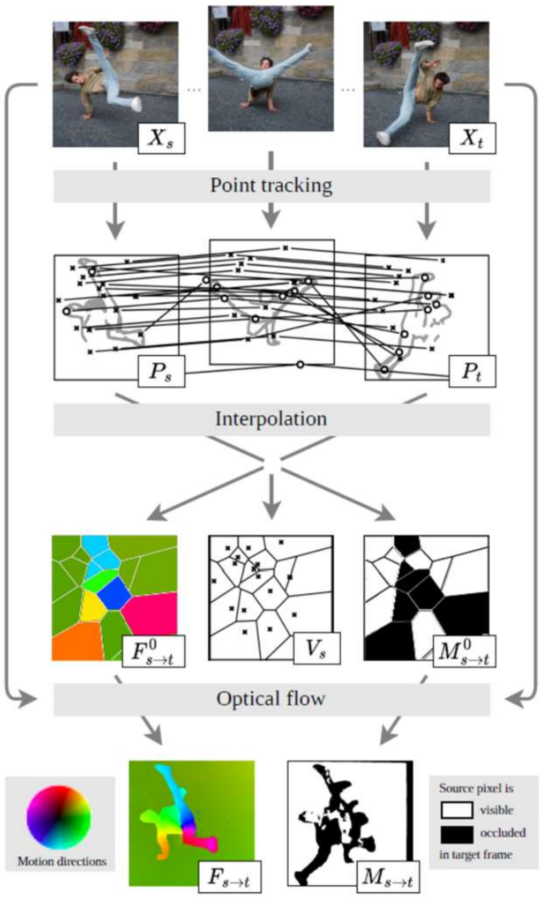
Dense point tracking



- Dense motion from source to target frames
- From a few point tracks (white)
  → dense flow (colors for directions, occlusion with stripes)

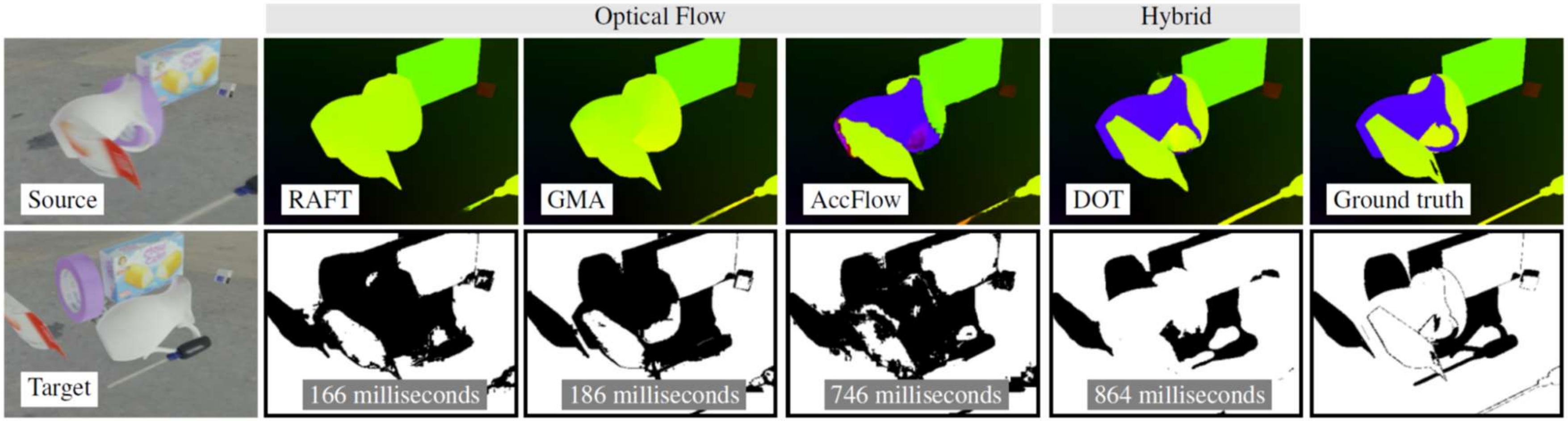
[Le Moing et al., Dense Optical Tracking: Connecting the Dots, arXiv'23]

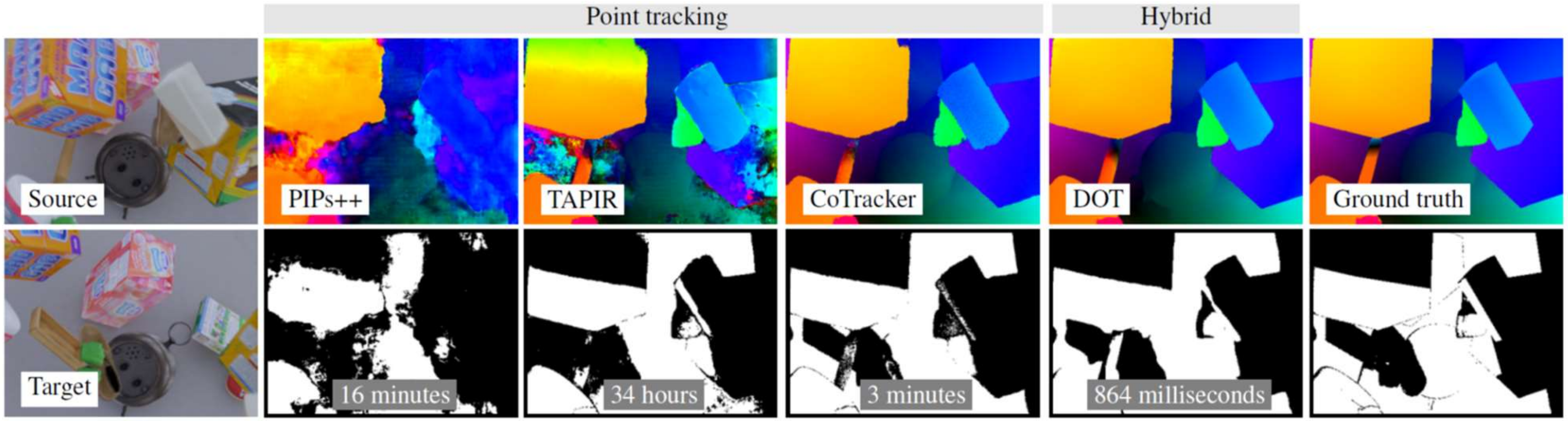
Dense point tracking



- Sparse point tracks (TAPIR, Co-Tracker)
- Near neighbor point interpolation
- Optical flow estimation to refine local neighborhood (RAFT)

Dense point tracking – results





Dense point tracking – results

	Method	N	CVO (*Clean*) EPE ↓ (all / vis / occ)	CVO (*Clean*) IoU ↑	CVO (*Final*) EPE ↓ (all / vis / occ)	CVO (*Final*) IoU ↑	CVO (*Final*) Time* ↓	CVO (*Extended*) EPE ↓ (all / vis / occ)	CVO (*Extended*) IoU ↑	CVO (*Extended*) Time ↓
Optical flow	RAFT [57]	-	2.82 / 1.70 / 8.01	58.1	2.88 / 1.79 / 7.89	57.2	**0.166**	28.6 / 21.6 / 41.0	61.7	**0.166**
	GMA [28]	-	2.90 / 1.91 / 7.63	60.9	2.92 / 1.89 / 7.48	60.1	0.186	30.0 / 22.8 / 42.6	61.5	0.186
	RAFT (🔥) [57]	-	2.48 / 1.40 / 7.42	57.6	2.63 / 1.57 / 7.50	56.7	0.634	21.8 / 15.4 / 33.4	65.0	4.142
	GMA (🔥) [28]	-	2.42 / 1.38 / 7.14	60.5	2.57 / 1.52 / 7.22	59.7	0.708	21.8 / 15.7 / 32.8	65.6	4.796
	MFT [47]	-	2.91 / 1.39 / 9.93	19.4	3.16 / 1.56 / 10.3	19.5	1.350	21.4 / 9.20 / 41.8	37.6	18.69
	AccFlow [61]	-	1.69 / 1.08 / 4.70	48.1	1.73 / 1.15 / 4.63	47.5	0.746	36.7 / 28.1 / 52.9	36.5	5.598
Point tracking	PIPs++ [68]	262144	9.05 / 6.62 / 21.5	33.3	9.49 / 7.06 / 22.0	32.7	974.3	18.4 / 10.0 / 32.1	58.7	1922.
	TAPIR† [17]	262144	3.55 / 1.34 / 15.2	74.0	4.36 / 2.04 / 16.1	72.5	$\sim 10^5$	- / - / -	-	$\sim 10^6$
	CoTracker [30]	262144	1.51 / 0.88 / 4.57	75.5	1.52 / 0.93 / 4.38	75.3	191.5	5.20 / 3.84 / 7.70	70.4	1737.
Hybrid	*Dense optical tracking* (DOT)	1024	1.36 / 0.76 / 4.26	80.0	1.43 / 0.85 / 4.29	79.7	0.864	5.28 / 3.78 / 7.71	70.8	5.234
		2048	1.32 / 0.74 / 4.12	80.4	1.38 / 0.82 / 4.10	80.2	1.652	5.07 / 3.67 / 7.34	71.0	9.860
		4096	**1.29 / 0.72 / 4.03**	**80.4**	**1.34 / 0.80 / 3.99**	**80.4**	3.152	**4.98 / 3.59 / 7.17**	**71.1**	19.73

"†": evaluation is only performed on a random subset of 2% of the test videos due to extremely slow inference speed. "∗": the time is the same for *Clean* and *Final* sets.

Overview

- Optical flow
- *Video classification*
- Multi-modal / LLM-based video understanding

Action recognition - tasks

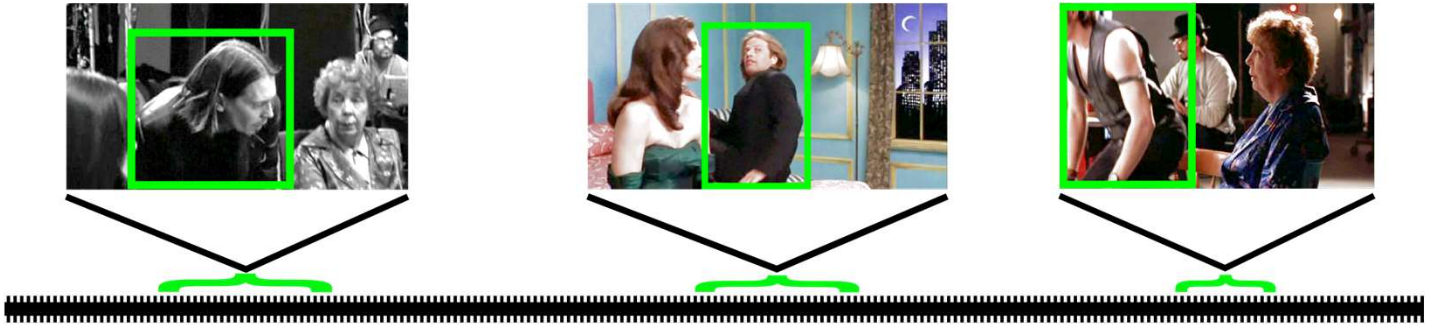
- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present
...

Action recognition - tasks

- Action classification: assigning an action label to a video clip

- Action localization: search locations of an action in a video



Action classification in videos

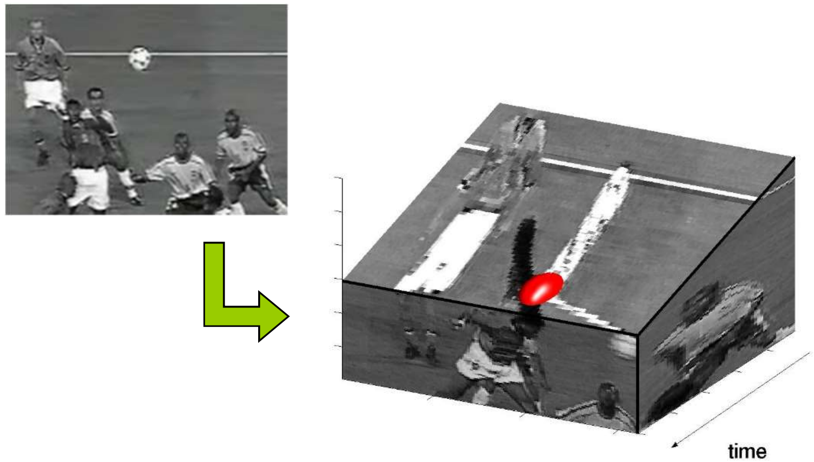
- Space-time interest points
- Dense trajectories
- Video-level CNN features
- Transformer-based approaches

Space-time interest points (STIP) [Laptev'05]

- Space-time corner detector
  [Laptev, IJCV 2005]

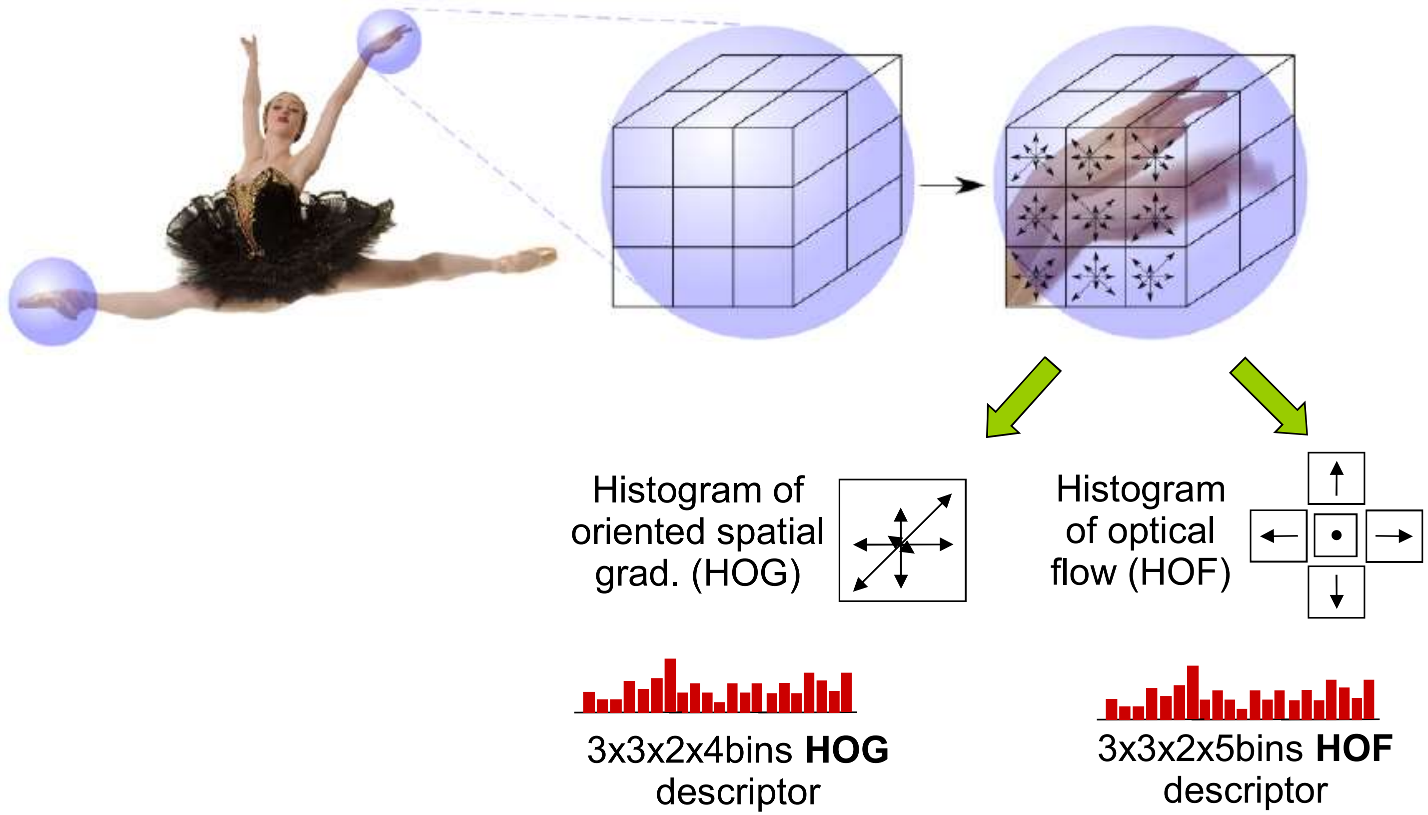
$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

$$\mu = \begin{pmatrix} I_xI_x & I_xI_y & I_xI_t \\ I_xI_y & I_yI_y & I_yI_t \\ I_xI_t & I_yI_t & I_tI_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$



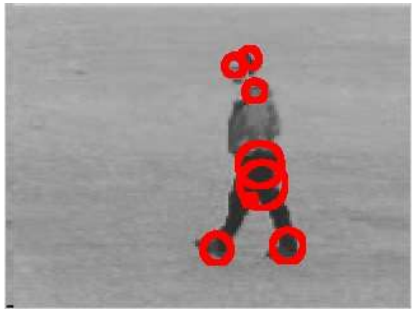
STIP descriptors

Space-time interest points

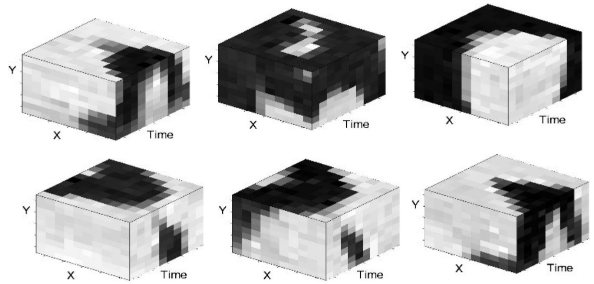


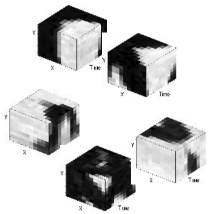
Action classification

- Bag of space-time features + support vector machine (SVM)
  [Schuldt'04, Niebles'06, Zhang'07]



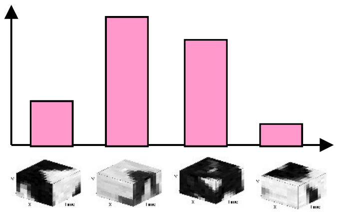
Collection of space-time patches





HOG & HOF patch descriptors

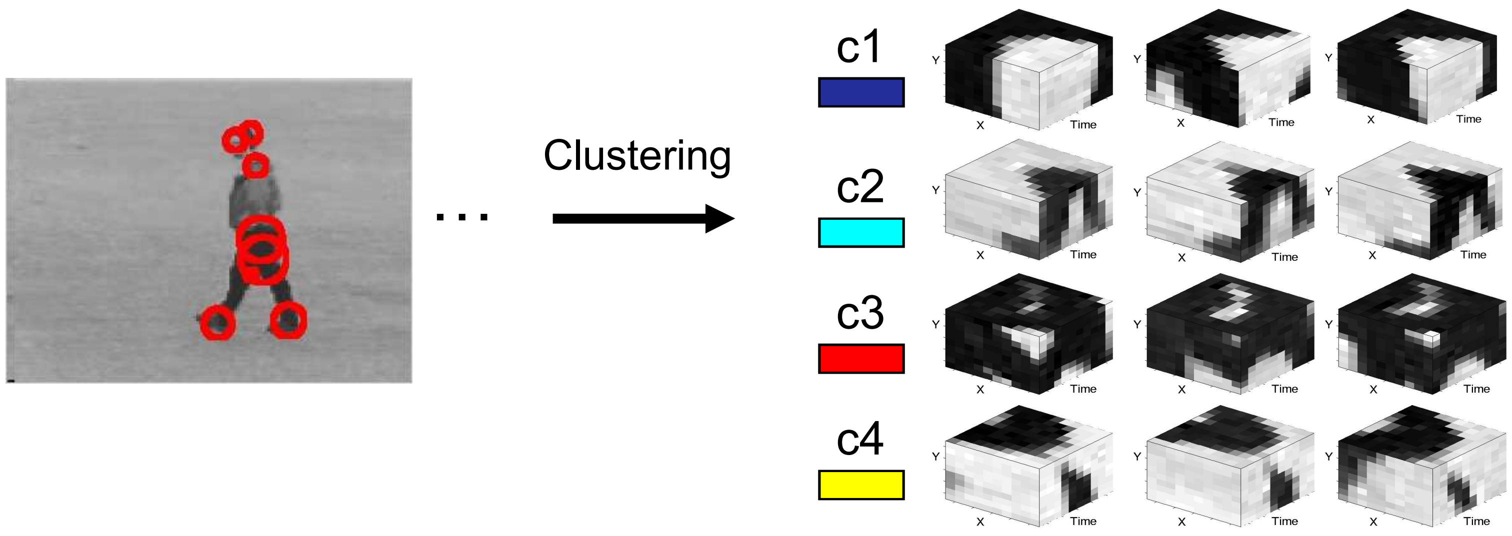
Histogram of visual words



SVM Classifier

Visual words: k-means clustering

- Group similar STIP descriptors together with k-means

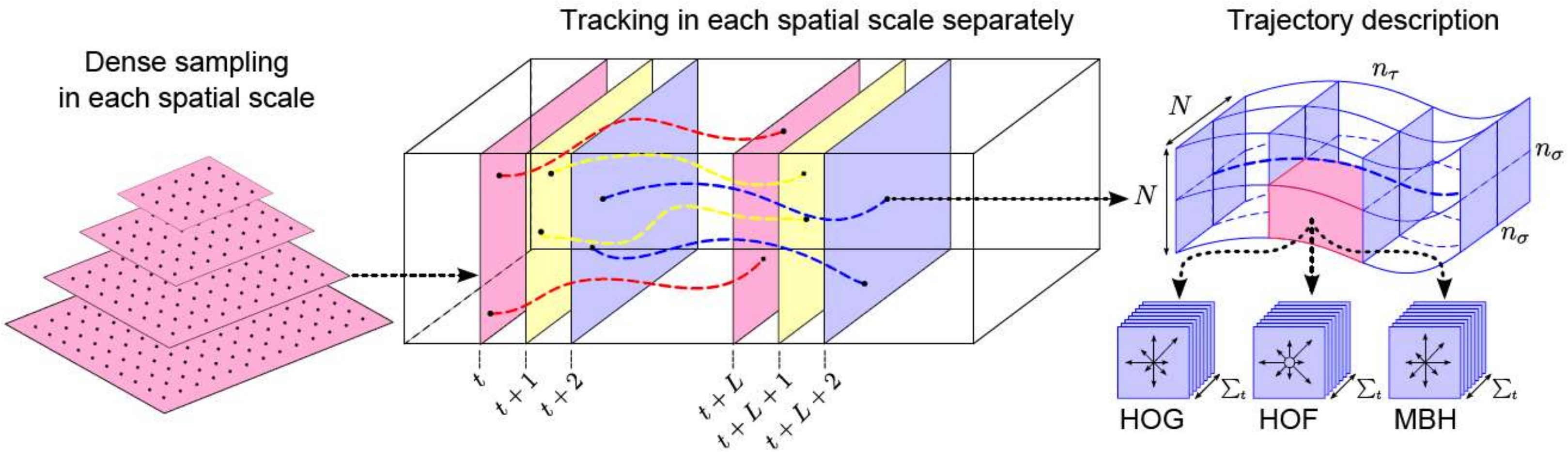


Action classification

Test episodes from movies “The Graduate”, “It’s a Wonderful Life”,
“Indiana Jones and the Last Crusade”

Dense trajectories [Wang et al., IJCV'13]

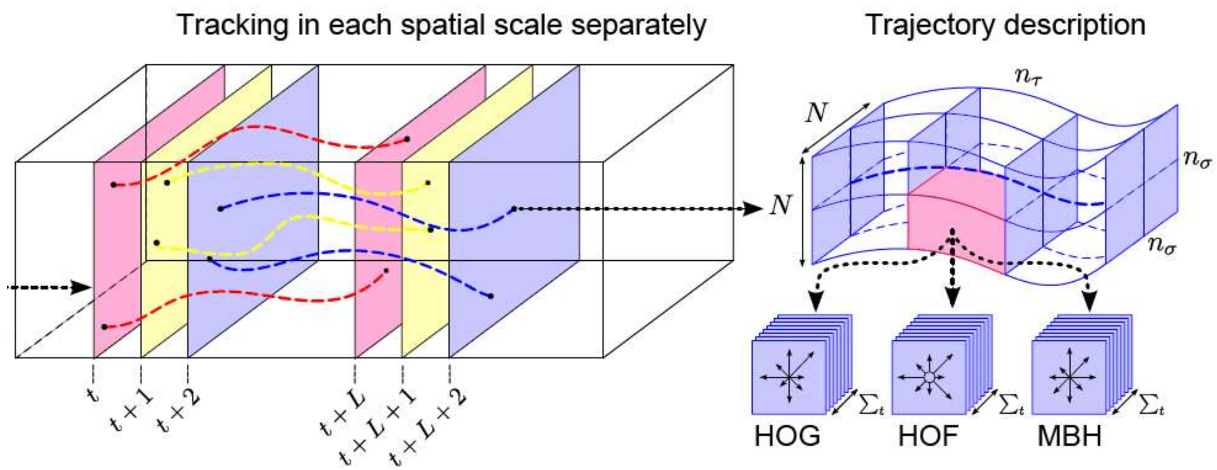
- Dense trajectories [Wang et al., IJCV'13] and Fisher vector encoding [Perronnin et al. ECCV'10]
  - Dense sampling at several scales
  - Feature tracking based on optical flow for several scales
  - Length 15 frames, to avoid drift



Example for dense trajectories

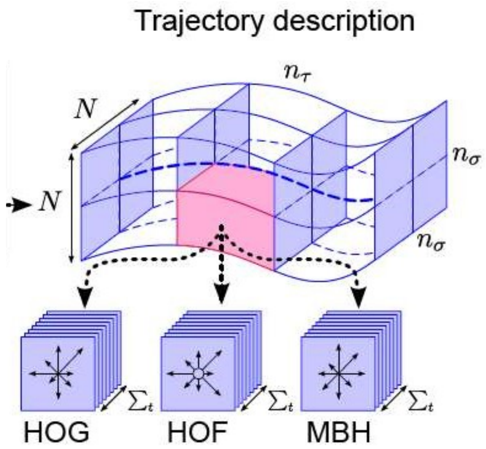
Descriptors for dense trajectory

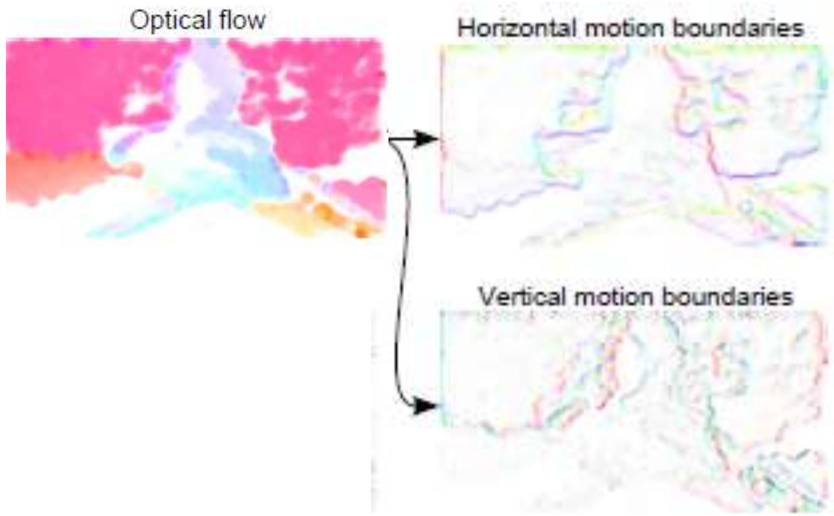
- Histogram of gradients (HOG: 2x2x3x8)
- Histogram of optical flow (HOF: 2x2x3x9)
- Motion-boundary histogram (MBHx + MBHy: 2x2x3x8)



Descriptors for dense trajectory

- Motion-boundary histogram (MBHx + MBHy: 2x2x3x8)
  - spatial derivatives are calculated separately for optical flow in x and y, quantized into a histogram
  - captures relative dynamics of different regions
  - suppresses constant motions





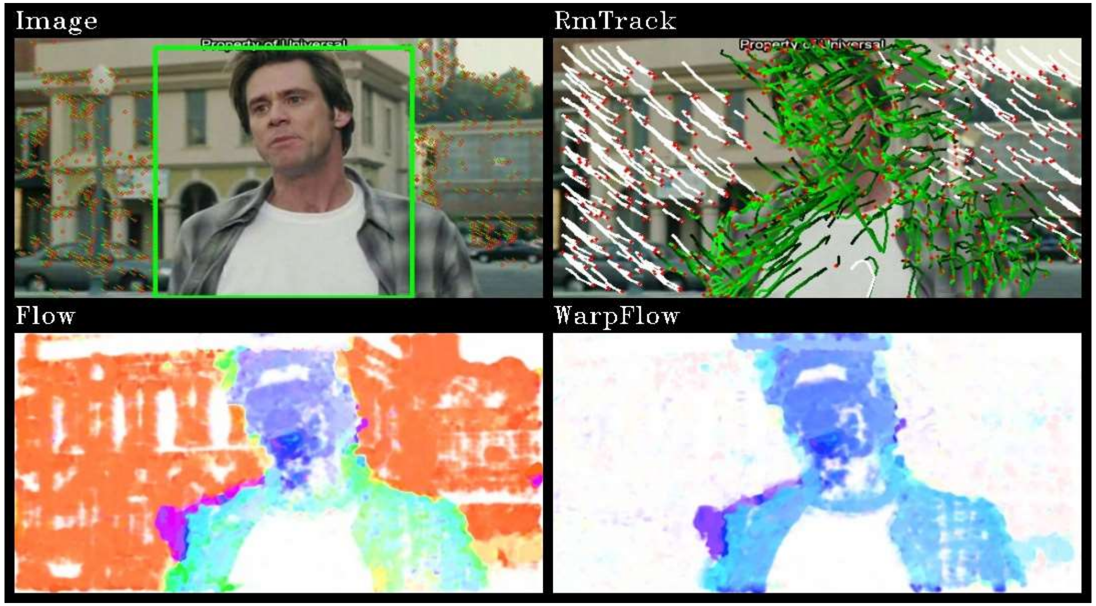
Dense trajectories

- Advantages:
  - Captures the intrinsic dynamic structures in videos
  - MBH is robust to certain camera motion

- Disadvantages:
  - Generates irrelevant trajectories in background due to camera motion
  - Motion descriptors are modified by camera motion, e.g., HOF, MBH

Improved dense trajectories

- Improve dense trajectories by explicit camera motion estimation
- Detect humans to remove outlier matches for homography estimation
- Stabilize optical flow to eliminate camera motion



[Wang and Schmid. Action recognition with improved trajectories. ICCV'13]

Camera motion estimation

- Find the correspondences between two consecutive frames:
  - Extract and match SURF features (robust to motion blur)
  - Use optical flow, remove uninformative points
- Combine SURF (green) and optical flow (red) results in a more balanced distribution
- Use RANSAC to estimate a homography from all feature matches

Inlier matches of the homography

Remove inconsistent matches due to humans

- Human motion is not constrained by camera motion, thus generates outlier matches
- Apply a human detector in each frame, and track the human bounding box forward and backward to join detections
- Remove feature matches inside the human bounding box during homography estimation

Inlier matches and warped flow, without or with HD

Remove background trajectories

- Remove trajectories by thresholding the maximal magnitude of stabilized motion vectors

- Our method works well under various camera motions, such as pan, zoom, tilt

Successful examples Failure cases

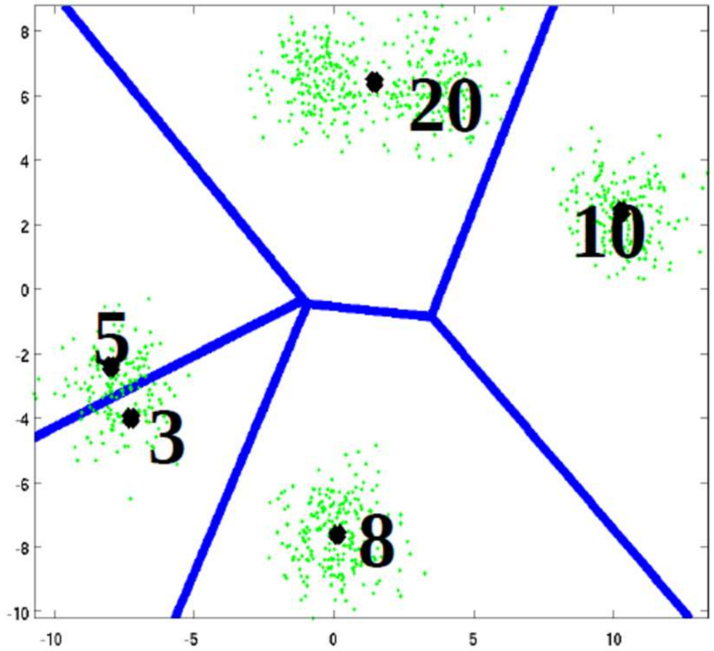
Removed trajectories (white) and foreground ones (green)

- Failure due to severe motion blur; the homography is not correctly estimated due to unreliable feature matches

Fisher Vector [Sanchez et al, 2013]

- Bag of features: stores the number of features assigned to each cluster center

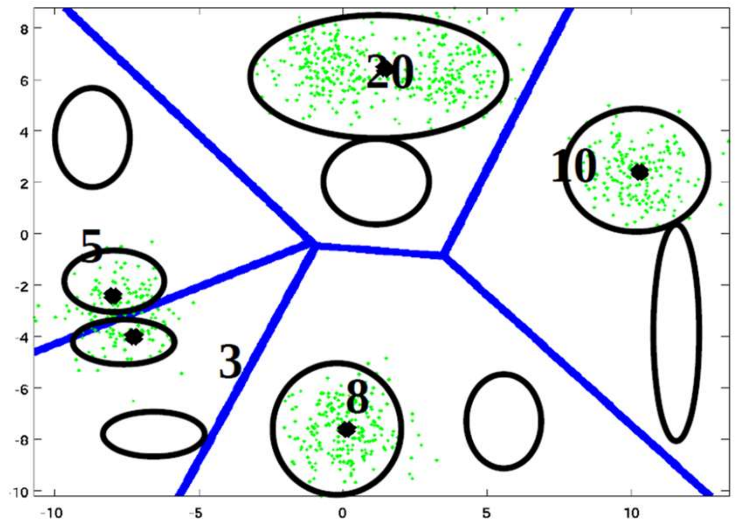
- Drawbacks:
  - Needs more words to refine the representation
  - This directly increases the computational cost
  - Also leads to many empty bins: redundancy



Fisher Vector [Sanchez et al, 2013]

- Fisher vector: also stores mean and variance of the features per cluster

- Even when the counts are the same,
the position can vary



- Advantages:
  - More information for the same visual word
  - Does not increase compute significantly
  - Leads for high dimensional features vectors

Evaluation datasets

Hollywood dataset [Marszalek et al.'09]

answer phone

get out of car

fight person

Hollywood2: 12 classes from 69 movies, report mAP

Evaluation datasets

HMDB 51 dataset [Kuehne et al.'11]

push-up

cartwheel

sword-exercice

HMDB51: 51 classes, report accuracy on three splits

Evaluation datasets

UCF 101 dataset [Soomro et al.'12]

haircut

archery

ice-dancing

UCF101: 101 classes, report accuracy on three splits

Evaluation of the intermediate steps

	HOG	HOF	MBH	HOF+MBH	Combined
DTF	38.4%	39.5%	49.1%	49.8%	52.2%
ITF	40.2%	48.9%	52.1%	54.7%	57.2%

Results on HMDB51 using Fisher vector

- Baseline: DTF = "dense trajectory feature"
- ITF = "improved trajectory feature”

- HOF improves significantly and MBH somewhat
- Almost no impact on HOG

- HOF and MBH are complementary, as they represent zero and first order motion information

Impact of feature encoding on improved trajectories

Datasets	Fisher vector		
	DTF	ITF wo human	ITF w human
Hollywood2	63.6%	66.1%	66.8%
HMDB51	55.9%	59.3%	60.1%
UCF101	83.5%	85.7%	86.0%

Compare DTF and ITF with and without human detection using HOG+HOF+MBH and Fisher encoding

- IDT significantly improvement over DT
- Human detection always helps. For Hollywood2 and HMDB51, the difference is more significant, as there are more humans present.

TrecVid MED 2011

- 15 categories

Attempt a board trick

Feed an animal

Landing a fish

…

Wedding ceremony

Working on a
wood project

Birthday party

TrecVid MED 2011

- 15 categories
- ~100 positive video clips per event category, 9600 negative video clips
- Testing on 32000 videos clips, i.e., 1000 hours
- Videos come from publicly available, user-generated content on various Internet sites

- Descriptors: MBH, SIFT, audio, text & speech recognition

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP
Motion	44.65
Static	33.97
Audio	18.15
OCR	10.85
ASR	8.21
Visual=Motion+Static	47.22
Visual+Audio	50.41
Visual+OCR	48.97
Visual+ASR	48.28
Visual+Audio+OCR+ASR	52.28

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party
Motion	44.65	30.7
Static	33.97	25.9
Audio	18.15	33.3
OCR	10.85	10.1
ASR	8.21	3.6
Visual=Motion+Static	47.22	34.8
Visual+Audio	50.41	47.7
Visual+OCR	48.97	35.8
Visual+ASR	48.28	35.0
Visual+Audio+OCR+ASR	52.28	48.4

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party	Repair appliance
Motion	44.65	30.7	42.6
Static	33.97	25.9	43.6
Audio	18.15	33.3	43.3
OCR	10.85	10.1	32.1
ASR	8.21	3.6	39.2
Visual=Motion+Static	47.22	34.8	47.5
Visual+Audio	50.41	47.7	54.5
Visual+OCR	48.97	35.8	50.8
Visual+ASR	48.28	35.0	54.5
Visual+Audio+OCR+ASR	52.28	48.4	57.2

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party	Repair appliance	Make sandwich
Motion	44.65	30.7	42.6	22.5
Static	33.97	25.9	43.6	21.5
Audio	18.15	33.3	43.3	11.2
OCR	10.85	10.1	32.1	19.4
ASR	8.21	3.6	39.2	6.7
Visual=Motion+Static	47.22	34.8	47.5	27.8
Visual+Audio	50.41	47.7	54.5	27.3
Visual+OCR	48.97	35.8	50.8	35.7
Visual+ASR	48.28	35.0	54.5	28.8
Visual+Audio+OCR+ASR	52.28	48.4	57.2	35.4

Experimental results

- Example results

rank 1

rank 2

rank 3

Highest ranked results for the event «horse riding competition»

Experimental results

- Example results

rank 1

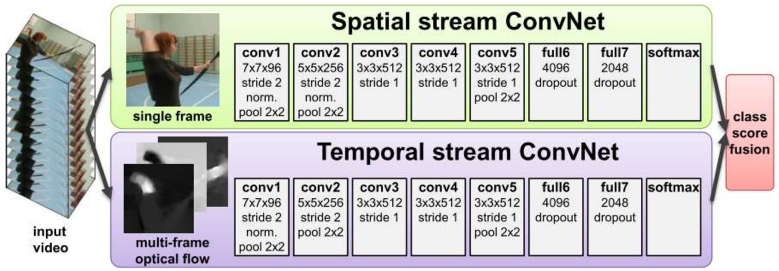
rank 2

rank 3

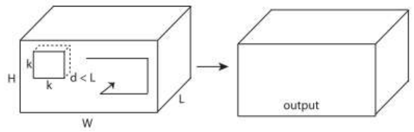
Highest ranked results for the event «tuning a musical instrument»

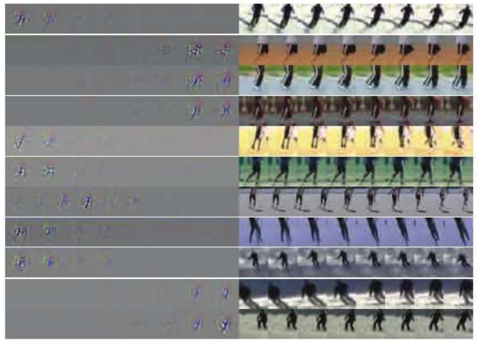
CNN based methods

Two-Stream Convolutional Networks
for Action Recognition in Videos
[Simonyan and Zisserman NIPS14]

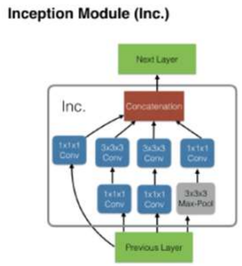


Learning Spatiotemporal Features with
3D Convolutional Networks
[Tran et al. ICCV15]



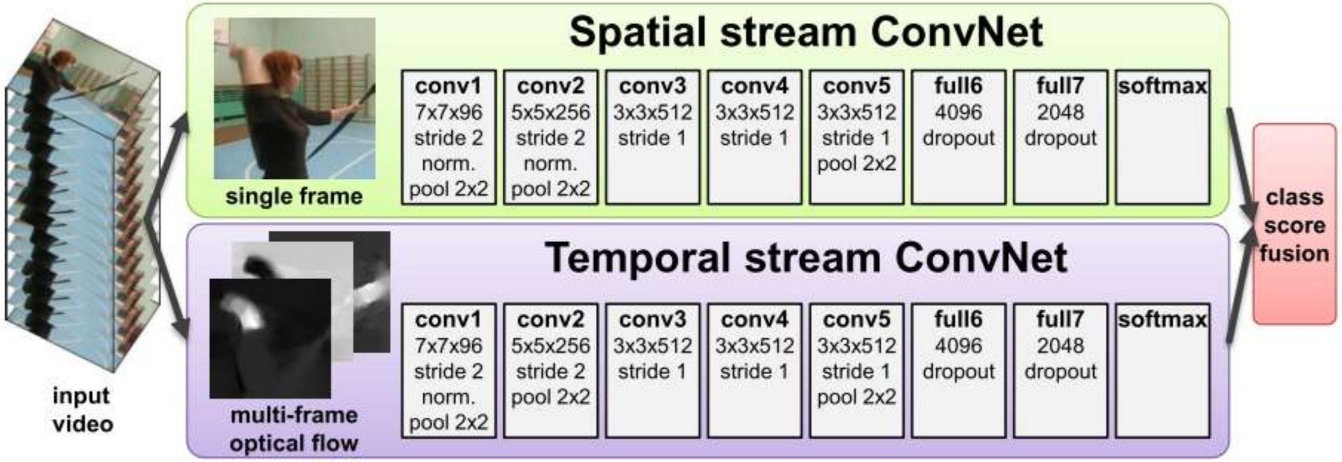


Quo vadis action recognition? A new
model and the Kinetics dataset
[Carreira et al. CVPR17]



Recent CNN methods

Two-Stream Convolutional Networks
for Action Recognition in Videos
[Simonyan and Zisserman NIPS14]



CNN based methods

Learning Spatiotemporal Features with 3D Convolutional Networks [Tran et al. ICCV15]

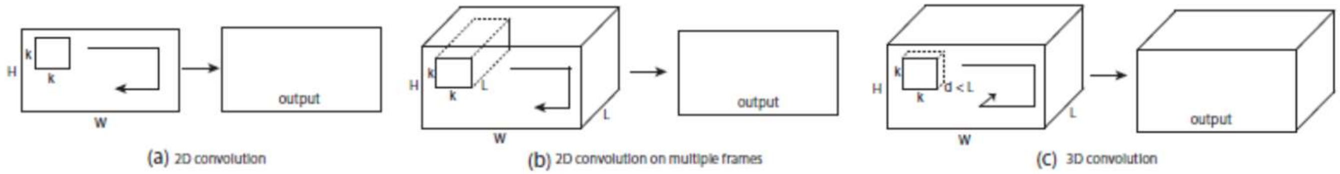
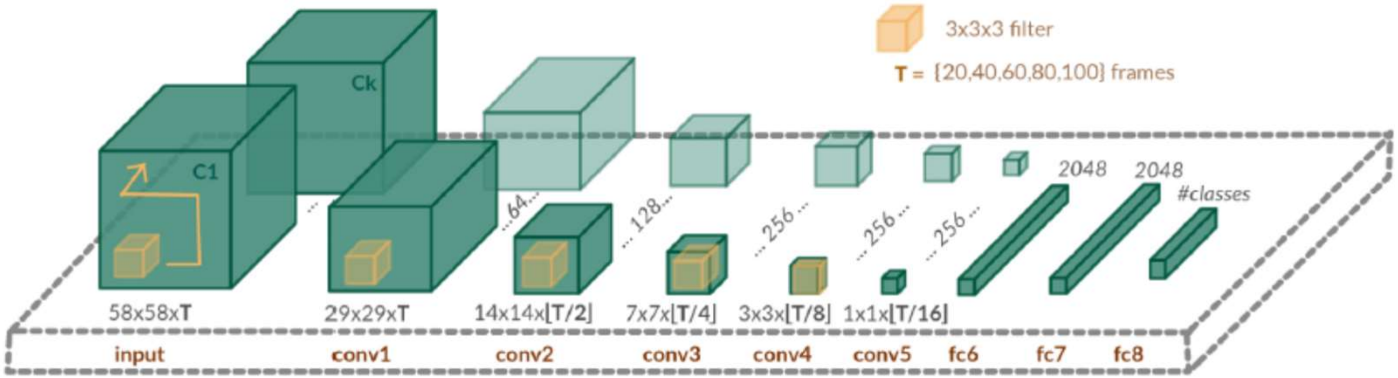
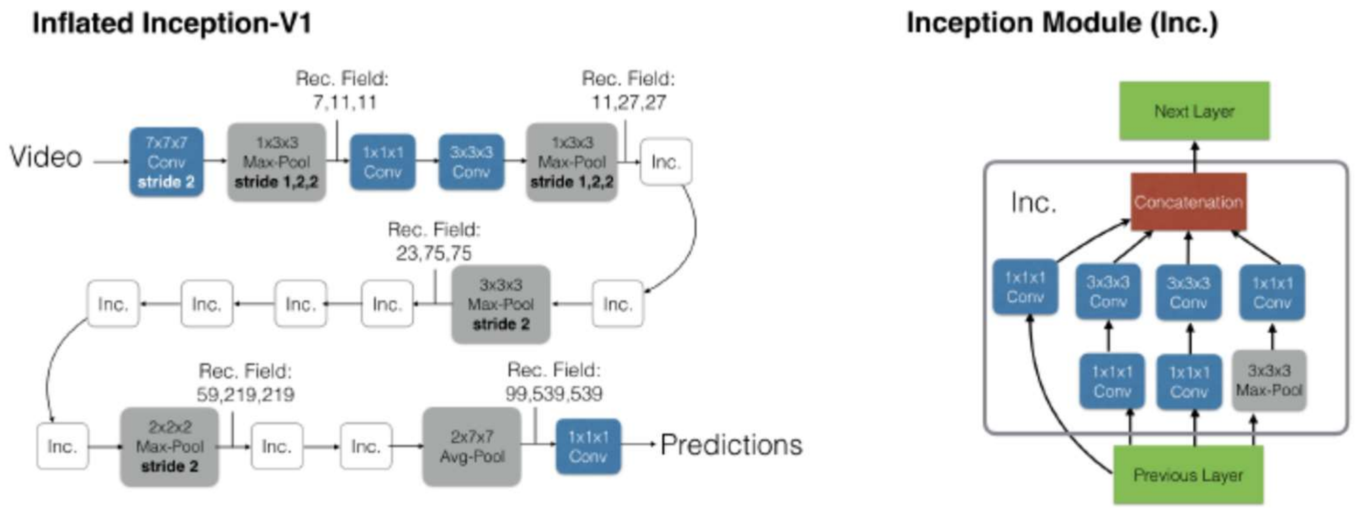


Figure 1. **2D and 3D convolution operations.** a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.



CNN based methods

Quo vadis, action recognition? A new model and the Kinetics dataset [Carreira et al. CVPR17]



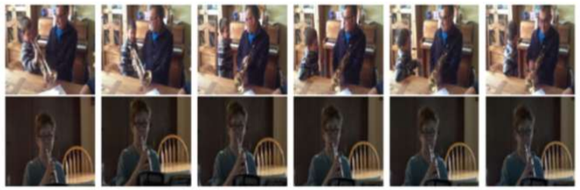
Pre-training on the large-scale Kinetics dataset 240k training videos
→ significant performance grain

Kinetics dataset

- Kinetics-700 dataset
  - 700 action classes
  - 650 00 clips
  - manual verification after automatic collection from YouTube

(c) shaking hands

(n) dunking basketball

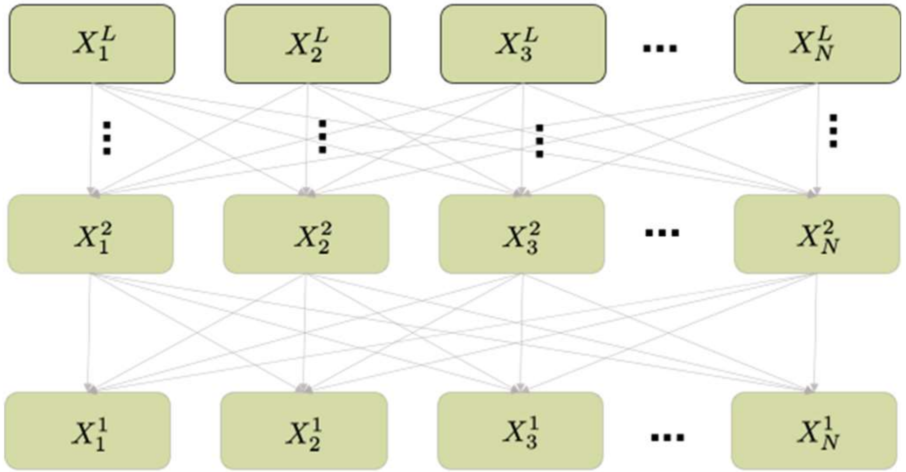


(j) playing trumpet

(l) brushing hair

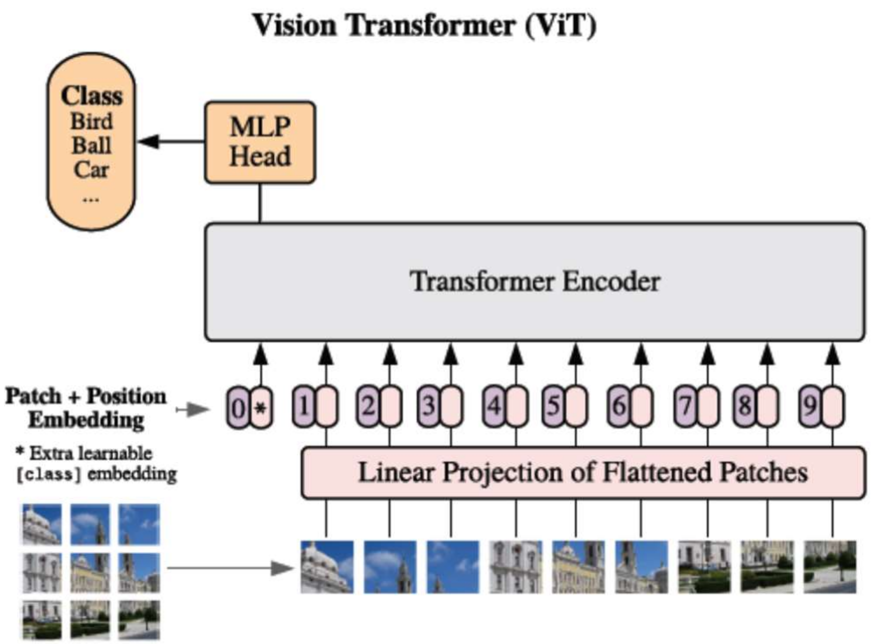
Transformer based models

- Transformer models are great for processing sequences
  - Text, images, videos can be expressed as sequences
  - Relies on self-attention between all tokens of a sequence [Vaswani et al., Neurips'17]



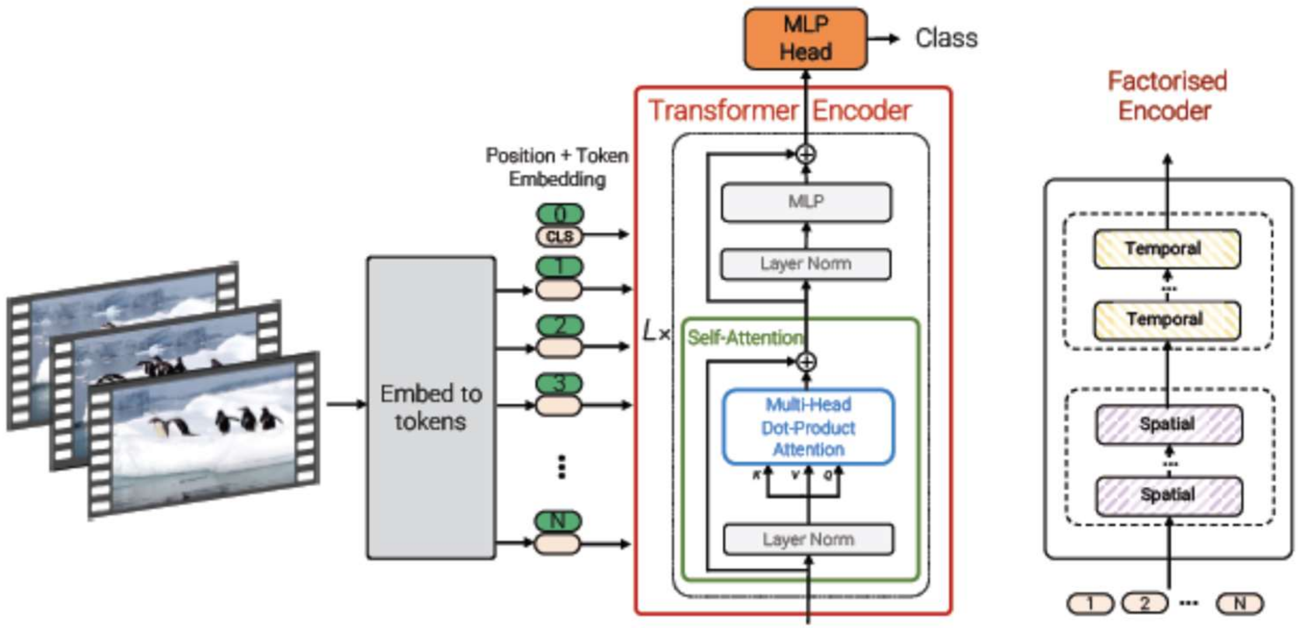
Vision Transformer (ViT)

- Fully transformer based architecture for image classification [A. Dosovitskiy et al., ICLR'21]
  - Image encoded as sequence of 16x16 patches
  - Tokenization by linear projection



ViViT: A Video Vision Transformer

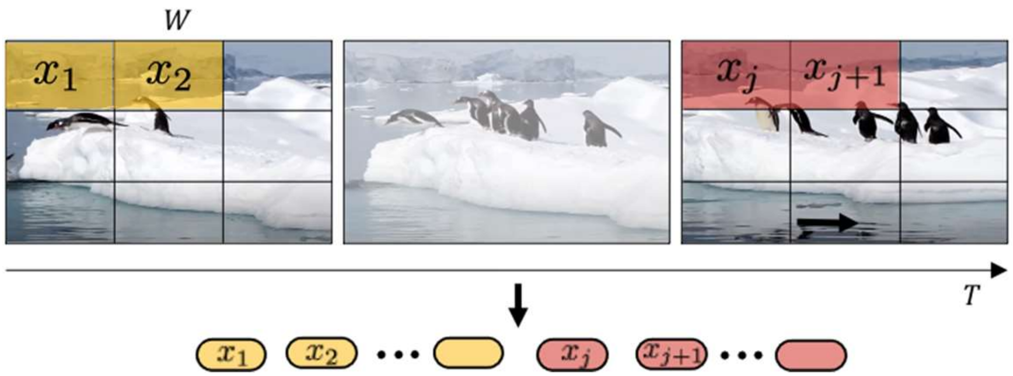
- Extend Vision Transformer ViT (for static images) to videos
- To handle large number of tokens, explore more efficient factorised attention variants



[ViViT, A. Arnab et al. ICCV'21]

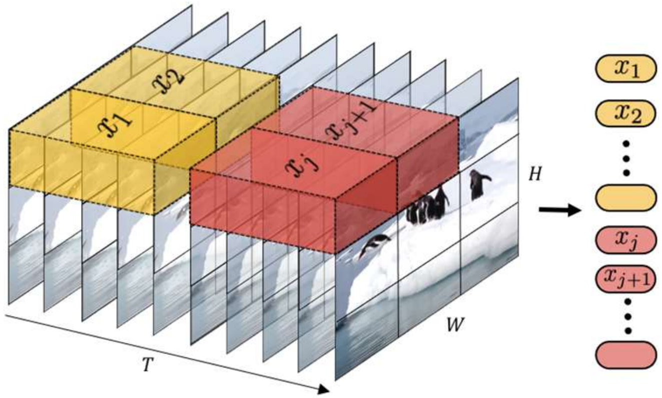
Input encoding – uniform frame sampling

- Sample frames, extract 2D patches and linearly project
- Effectively consider a video as a “big image”



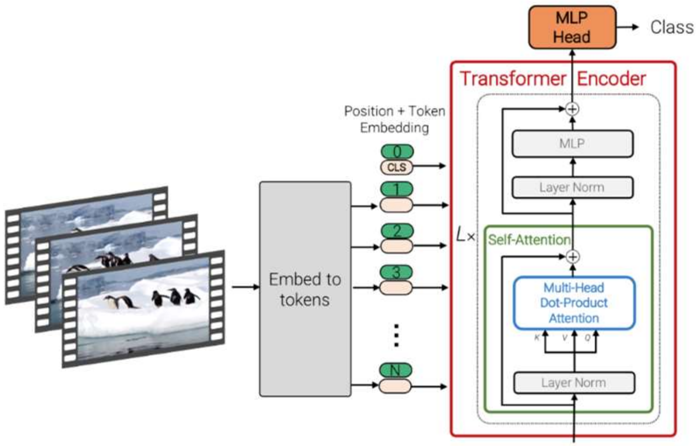
Input encoding – tubelet embedding

- Extract 3D spatio-temporal tubelets + linear project into tokens
- Captures temporal information in the tokenization stage
- Works better than uniform sampling



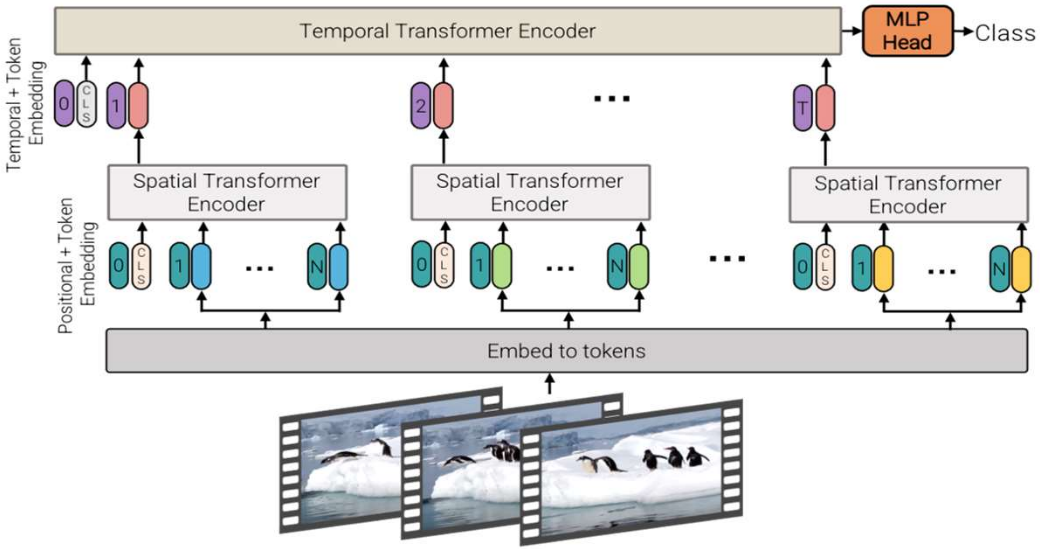
ViViT: A Video Vision Transformer

- An alternative to 3D convolutional neural networks
  - Extract 3D tubelets to encode spatio-temporal “tubes” into tokens
  - Encode tubes into embedding by linear project and add position
  - Train a transformer to predict classes
- Quadratic complexity in tokens



ViViT: Factorized Encoder

- Separate encoders for spatial and temporal information
  - Reduces complexity, compute, less overfitting
  - Spatial encoder is initialised from a pretrained-ViT model
  - "*Late fusion*" of spatial and temporal information



Comparison of model variants

	K400	EK	FLOPs ($\times 10^9$)	Params ($\times 10^6$)	Runtime (ms)
Model 1: Spatio-temporal	80.0	43.1	455.2	88.9	58.9
Model 2: Fact. encoder	78.8	43.7	284.4	100.7	17.4
Model 2: Ave. pool baseline	75.8	38.8	283.9	86.7	17.3

- Spatio-temporal model better for large datasets (K400)
- Factorized encoder faster than spatio-temporal model
- Factorized encoder better for small datasets (EK:EpicKitchen)
- Spatio-temporal model > average pooling

Impact of regularization

- Use pretrained ImageNet model for initialization
- Regularization with data augmentation and stochastic depth

	Top-1 accuracy
Random crop, flip, colour jitter	38.4
+ Kinetics 400 initialisation	39.6
+ Stochastic depth [28]	40.2
+ Random augment [10]	41.1
+ Label smoothing [58]	43.1
+ Mixup [79]	43.7

5.3% gain on Epic Kitchens

Comparison to state of the art

(a) Kinetics 400

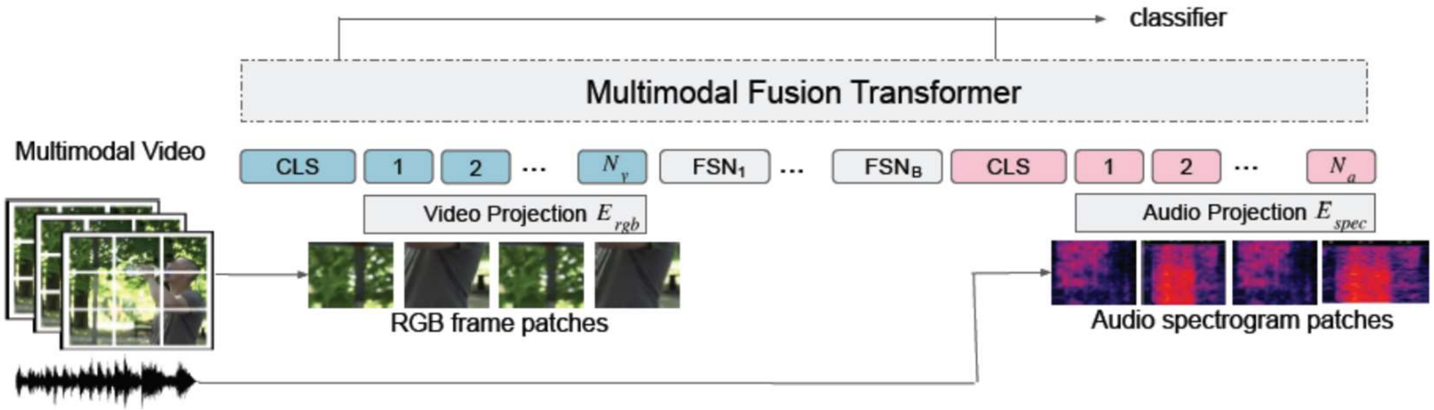
Method	Top 1	Top 5	Views
bIVNet [16]	73.5	91.2	–
STM [30]	73.7	91.6	–
TEA [39]	76.1	92.5	10×3
TSM-ResNeXt-101 [40]	76.3	–	–
I3D NL [72]	77.7	93.3	10×3
CorrNet-101 [67]	79.2	–	10×3
ip-CSN-152 [63]	79.2	93.8	10×3
LGD-3D R101 [48]	79.4	94.4	–
SlowFast R101-NL [18]	79.8	93.9	10×3
X3D-XXL [17]	80.4	94.6	10×3
TimeSformer-L [2]	80.7	94.7	1×3
ViViT-L/16x2	80.6	94.7	4×3
ViViT-L/16x2 320	**81.3**	**94.7**	4×3
Methods with large-scale pretraining			
ip-CSN-152 [63] (IG [41])	82.5	95.3	10×3
ViViT-L/16x2 (JFT)	82.8	95.5	4×3
ViViT-L/16x2 320 (JFT)	83.5	95.5	4×3
ViViT-H/16x2 (JFT)	**84.8**	**95.8**	4×3

(b) Kinetics 600

Method	Top 1	Top 5	Views
AttentionNAS [73]	79.8	94.4	–
LGD-3D R101 [48]	81.5	95.6	–
SlowFast R101-NL [18]	81.8	95.1	10×3
X3D-XL [17]	81.9	95.5	10×3
TimeSformer-HR [2]	82.4	**96.0**	–
ViViT-L/16x2	82.5	95.6	4×3
ViViT-L/16x2 320	**83.0**	95.7	4×3
ViViT-L/16x2 (JFT)	84.3	96.2	4×3
ViViT-H/16x2 (JFT)	**85.8**	**96.5**	4×3

A multimodal (audio-visual) transformer

- Extend ViViT to multimodal information by adding audio
- Audio is represented by a spectrogram

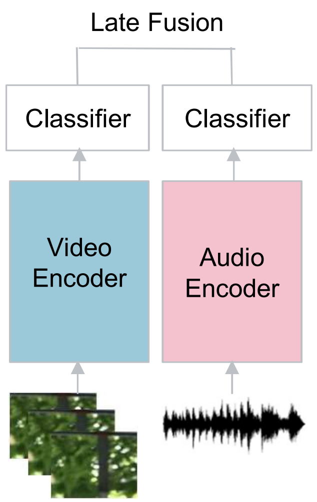


[Attention bottlenecks for multimodal fusion, A. Nagrani et al., Neurips'21l]

Late fusion

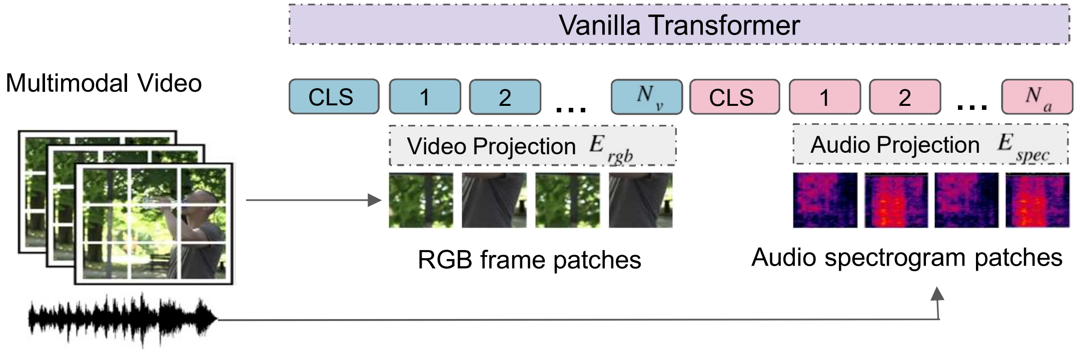
- Multimodal inputs
  - Heterogeneity of inputs (RGB frames, audio spectrograms)
  - Specialized architectures
  - Different datasets and evaluation benchmarks

- The “dominant” paradigm
  - Different encoders
  - Output scores a fused at the end



Vanilla Multimodal Transformer

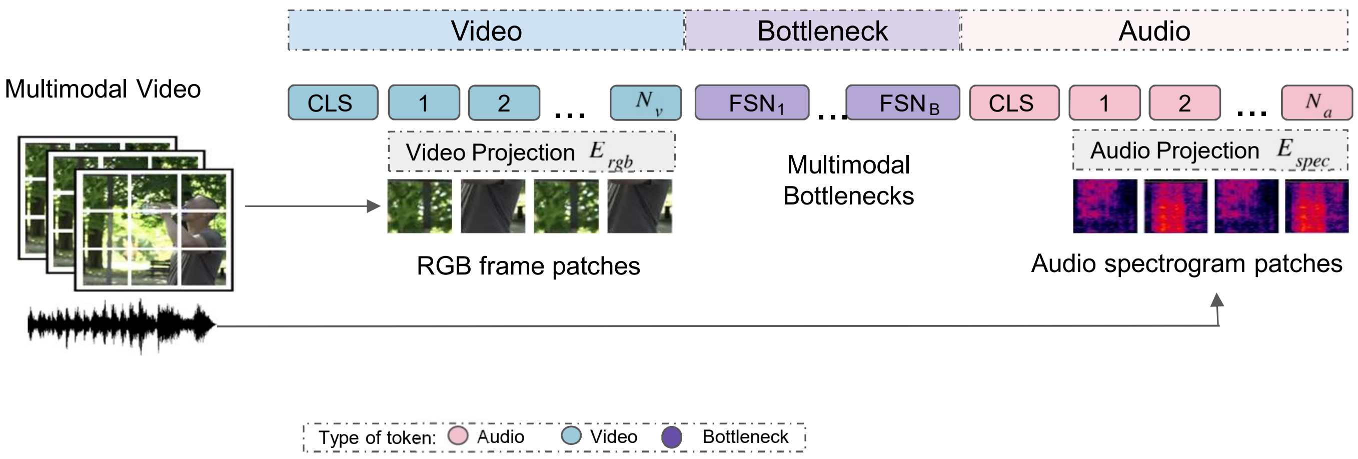
- Tokenize RGB frame and spectrogram patches
- Feed all tokens to a transformer
- Pairwise self-attention between all tokens (early fusion)



- Scales quadratically with sequence length
- Video has a lot of redundancy

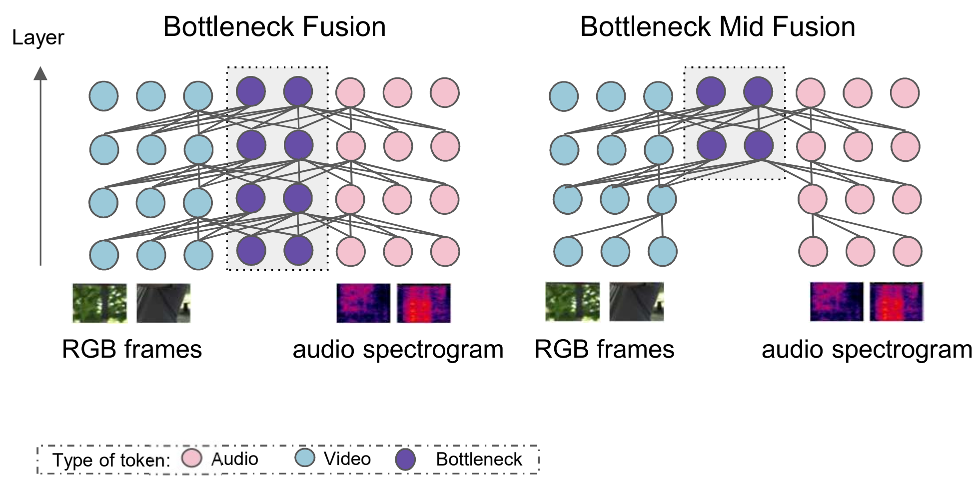
Multimodal Bottleneck Transformer

- Introduces a number of bottleneck tokens (B=4)
- Full pairwise self attention within a modality
- Attention between the vision/audio tokens and the bottleneck tokens



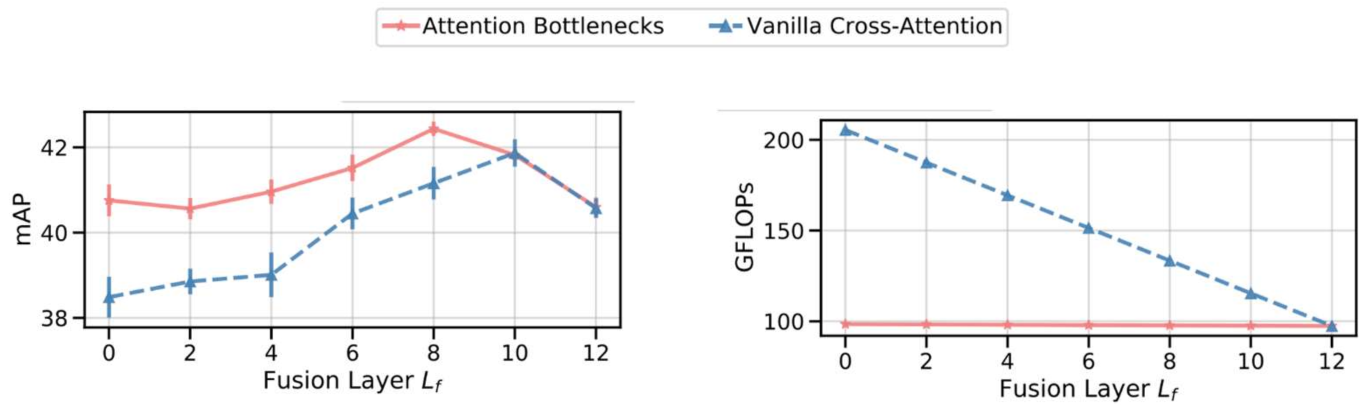
Do all layers need to be cross-modal?

- Restrict cross-modal information to later layers (mid-fusion)
- The layer we introduce cross-modal interactions is called the “fusion layer”
- Allows early layers to “specialize” to unimodal patterns



Improved performance and efficiency

- Mid Fusion outperforms early and late fusion on most datasets



Results for Audio-Set and 4 bottleneck tokens

- Improved performance, lower compute

Experimental results

- Two different video classification tasks

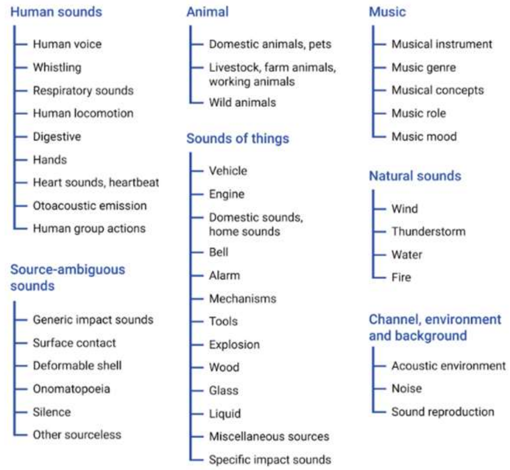
Action Recognition

Kinetics
Moments in Time

Epic Kitchens

Sound Event Classification

Audioset
VGGSound
Kinetics-Sou



Experimental results

Model	Training Set	A only	V only	AV Fusion
GBlend [58]	MiniAS	29.1	22.1	37.8
GBlend [58]	FullAS-2M	32.4	18.8	41.8
Attn Audio-Visual [19]	FullAS-2M	38.4	25.7	46.2
Perceiver [29]	FullAS-2M	38.4	25.8	44.2
MBT	MiniAS	31.3	27.7	43.9
MBT	AS-500K	**44.3**	**32.3**	**52.1**

Table 1: **Comparison to the state of the art on AudioSet [22].** We report mean average precision (mAP). For audio-visual fusion, our method outperforms others that use the entire AudioSet training set (almost 2M samples), while we train on only 500K.

Audioset

Late Fusion	49.2
MBT (ours)	**52.1**

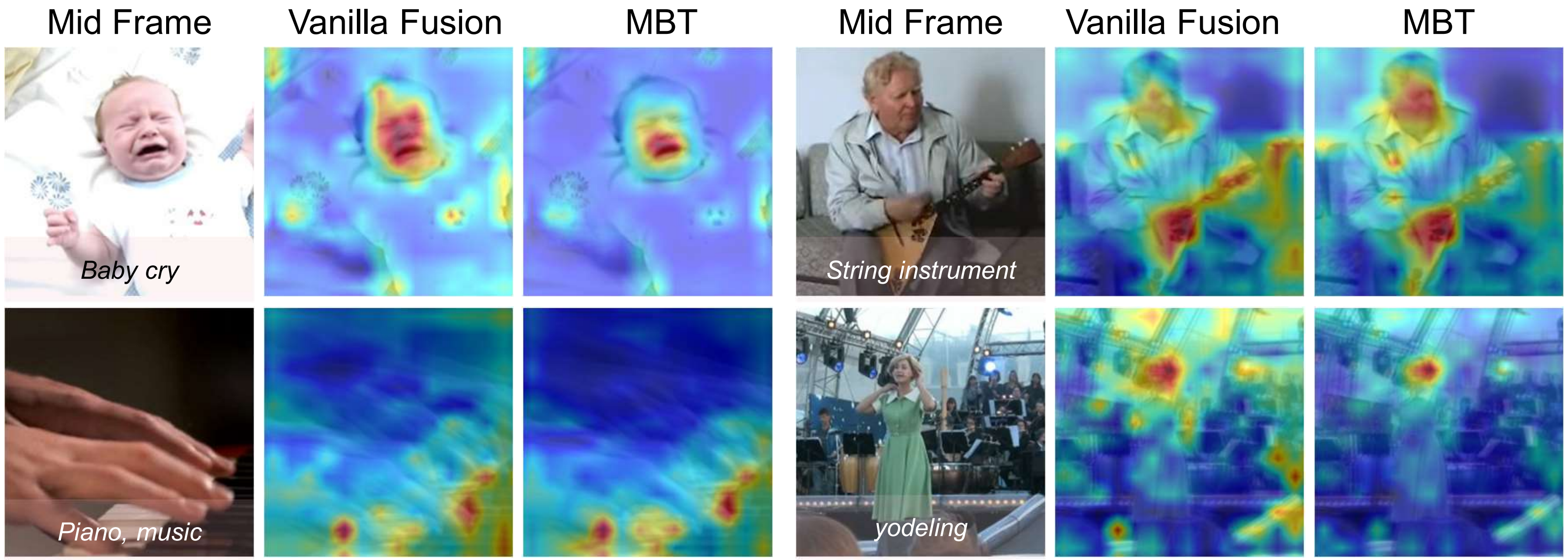
Model	Modalities	Verb	Noun	Action
Damen et al. [13]	A	42.1	21.5	14.8
AudioSlowFast [34]†	A	46.5	22.78	15.4
TSN [57]	V, F	60.2	46.0	33.2
TRN [63]	V, F	65.9	45.4	35.3
TBN [33]	A, V, F	66.0	47.2	36.7
TSM [42]	V, F	**67.9**	49.0	38.3
SlowFast [20]	V	65.6	50.0	38.5
MBT	A	44.3	22.4	13.0
MBT	V	62.0	56.4	40.7
MBT	A, V	64.8	**58.0**	**43.4**

Table 2: **Comparison to the state of the art on Epic Kitchens 100 [13].** Modalities (Mods) are **A:** Audio, **V:** Visual, **F:** Optical flow.

Epic-Kitchens

Late Fusion	37.9
MBT (ours)	**43.4**

Attention Heatmaps



MBT: focus on smaller regions, sound sources (mouth, fingertips)

Overview

- Optical flow
- Video classification
- *Multi-modal / LLM-based video understanding*

Why multimodal data?

- Precise understanding of the video content
  → Requires access to all modalities simultaneously

Is this Indian?

Why multimodal video representation?

- Large-scale cross-modal supervision
  → No manual annotation required

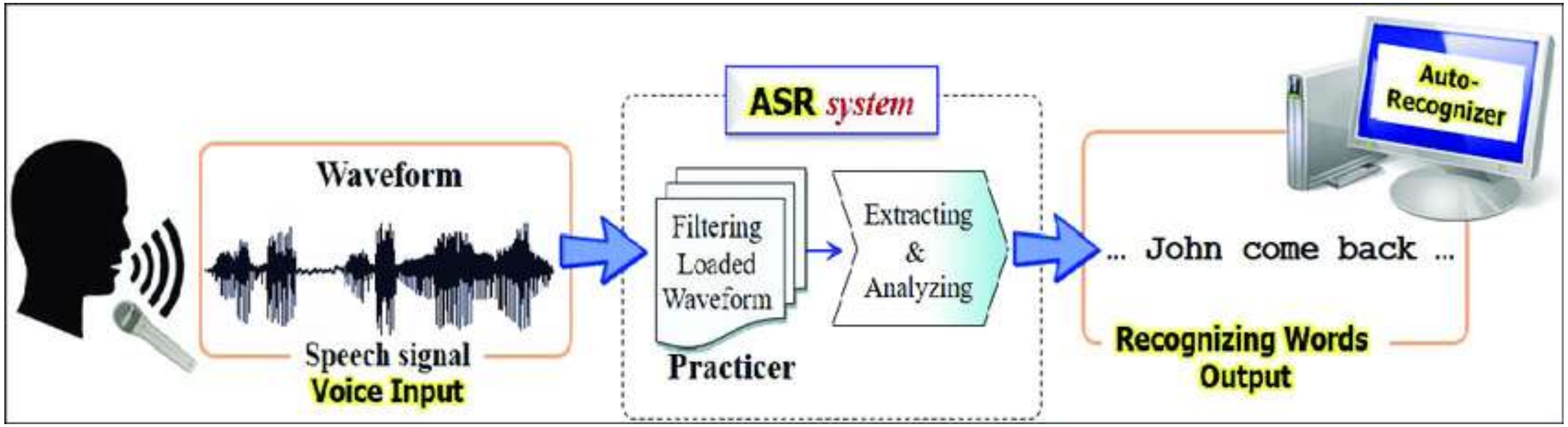
Training on the **HowTo100M** [1] dataset

→ + 120M pairs clip-narration → + 1,2M videos → Uncurated

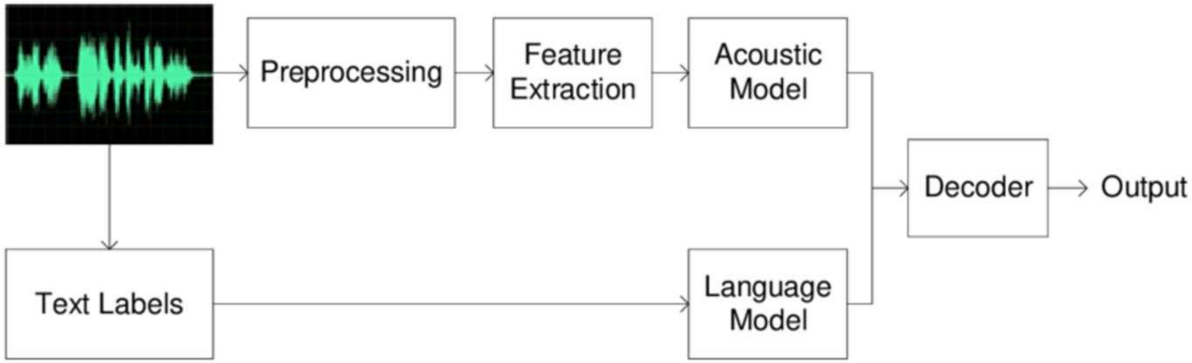
[HowTo100M. A. Miech, D. Zhukov, JB Alayrac, M. Tapaswi, I. Laptev and J. Sivic, ICCV 2019]

Automatic Speech Recognition (ASR)

- Takes as input human speech and turns it into text



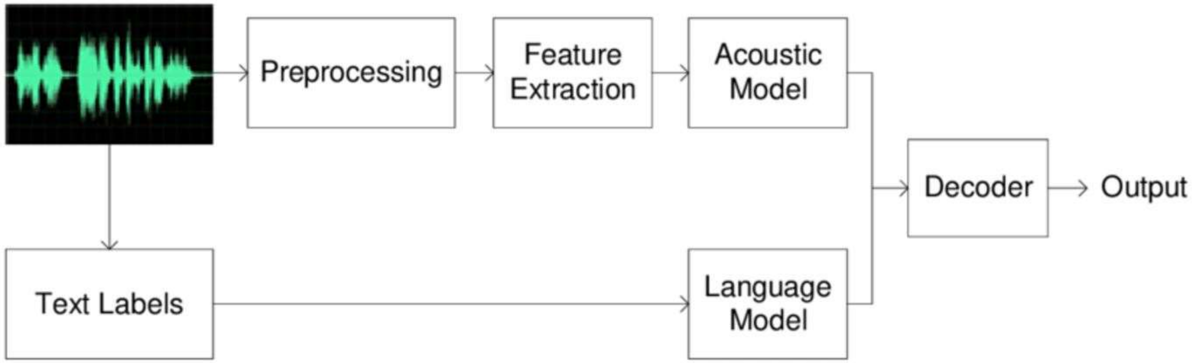
Automatic Speech Recognition (ASR)



Traditional systems use a sequence of steps

1. Preprocessing for noise reduction
2. Feature extraction from the raw audio signal to capture important characteristics of the sound, such as frequency, amplitude, and duration, for example Mel-frequency cepstral coefficients (MFCCs)
3. Acoustic modeling for training a statistical model that maps the extracted features to *phonemes*, the smallest units of sound in a language

Automatic Speech Recognition (ASR)

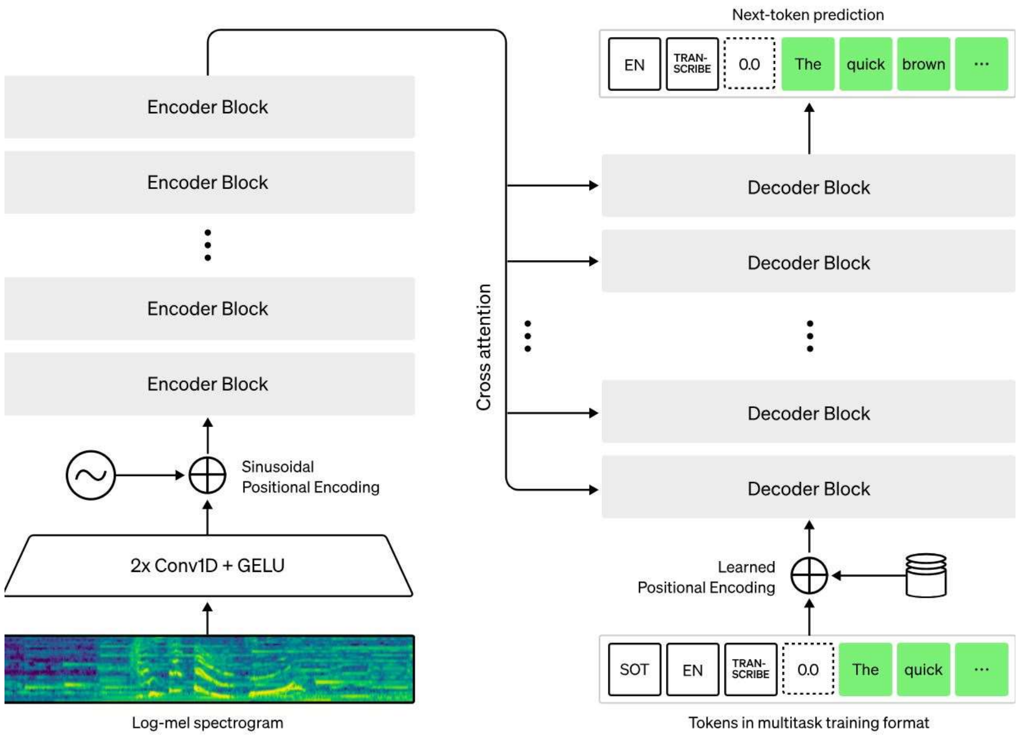


Traditional systems use a sequence of steps

4. Language modeling for creating a probabilistic model of how words and phrases are likely to appear in a particular language
5. Decoding uses the acoustic and language models to transcribe the audio into a sequence of words or tokens
6. Post-processing to improve accuracy and coherence, by including language constraints, grammar rules, and contextual analysis

Automatic Speech Recognition (ASR)

- End-to-end trained system: Whisper



- Trained 680,000 hours of multilingual and multitask supervised data collected from the web
- End-to-end training
- Features are represented with log-mel spectrum, input 30 second chunks
- Excellent results on main languages, worse on others
- Text includes more high-level information/semantics than audio and benefits from the large training corpus

VideoBERT: learning multimodal video representation

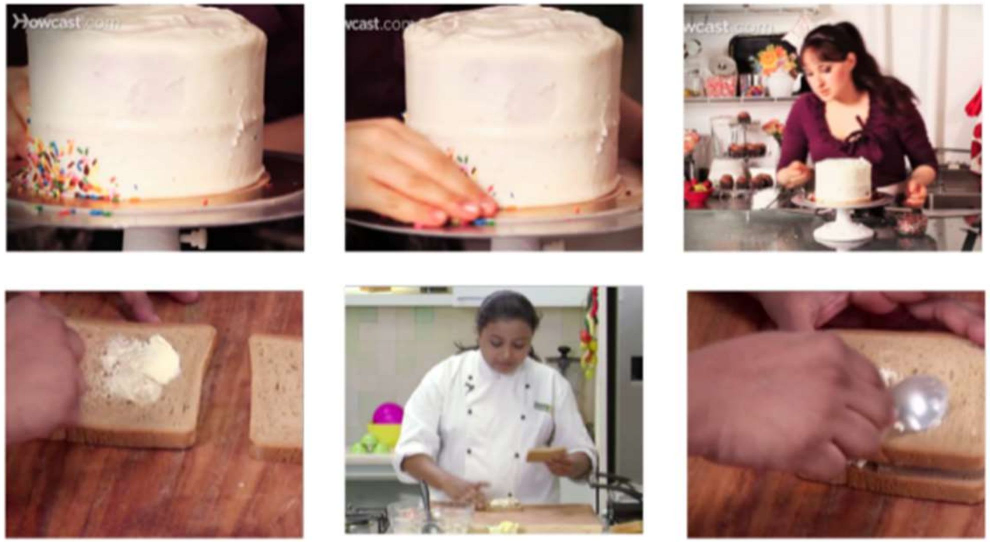
- Learning from visual video and speech transcribed with ASR

But in the meantime, you're just kind of moving around your cake board.

- BERT model learns correspondence between video and speech
- Learning from large-scale data without manual annotations

[VideoBERT, C. Sun et al., ICCV'19]

Large-scale training data without manual annotations

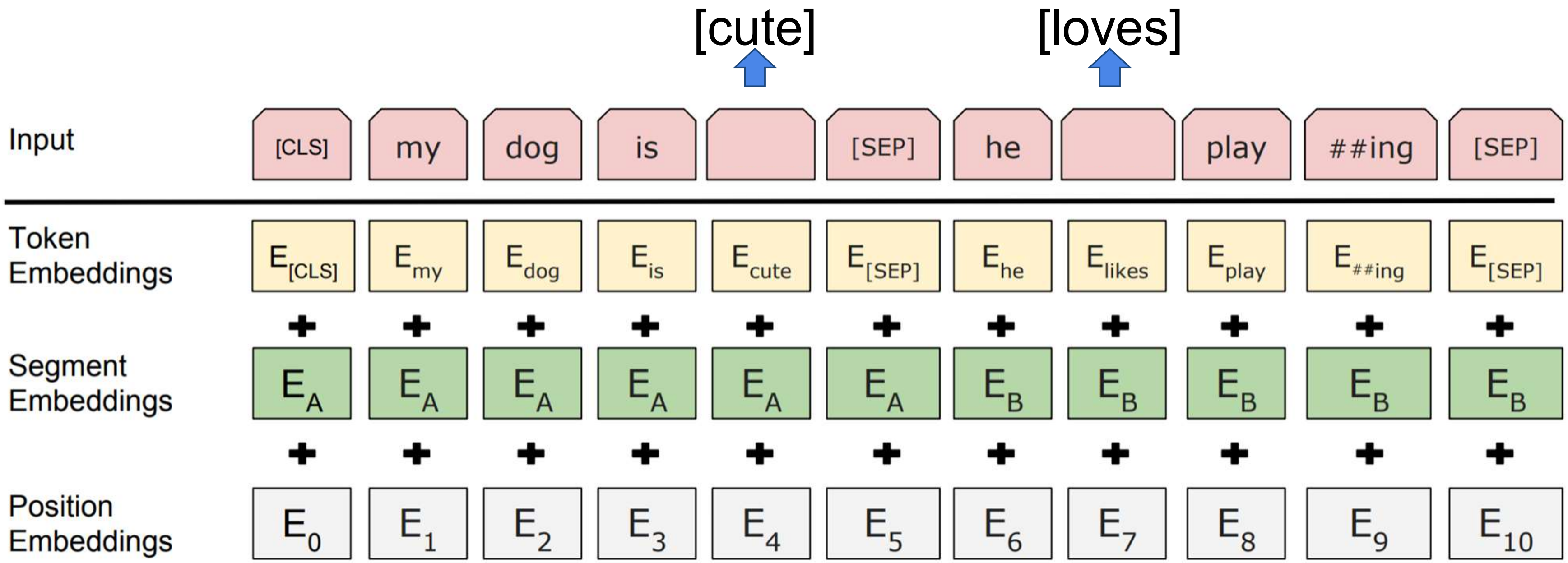


*“but in the meantime, you're just kind of **moving around** your **cake** board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party.”*

*“**apply** a little bit of **butter** on one side and place a portion of the stuffing and **spread** evenly **cover** with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches.”*

- ~320K *cooking/recipe* videos on YouTube
- ~1000 days in total, average length is ~4 mins
- ~120K videos with English ASR outputs

State-of-the-art for NLP: BERT



Two pre-training tasks:

- Masked language modeling
- Next sentence prediction

Network:

- Stacked Transformers
- Large amount of data

[1] Figure credit: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805

Self-supervised pre-training for NLP

Input corpus:

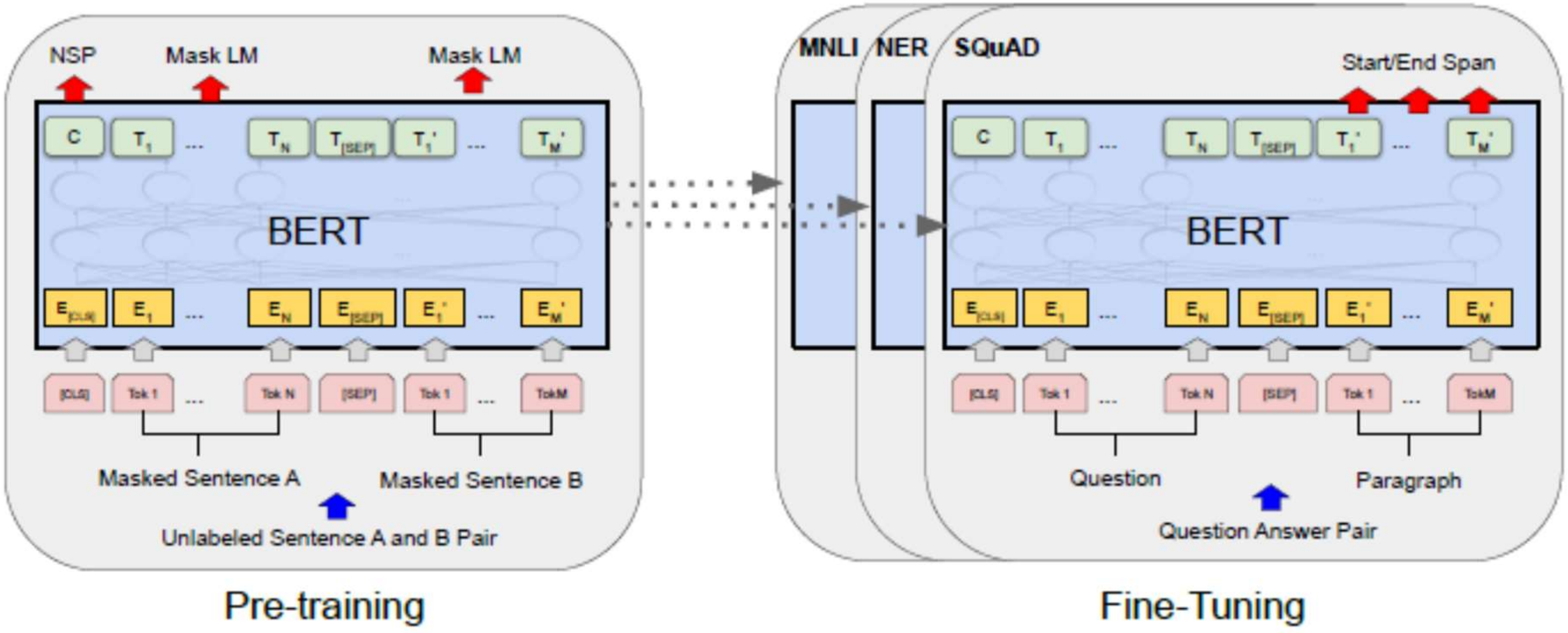
> Apply a little bit of butter on one side and place a portion of the stuffing. Spread evenly cover with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches.

Masked language modeling (MLM):

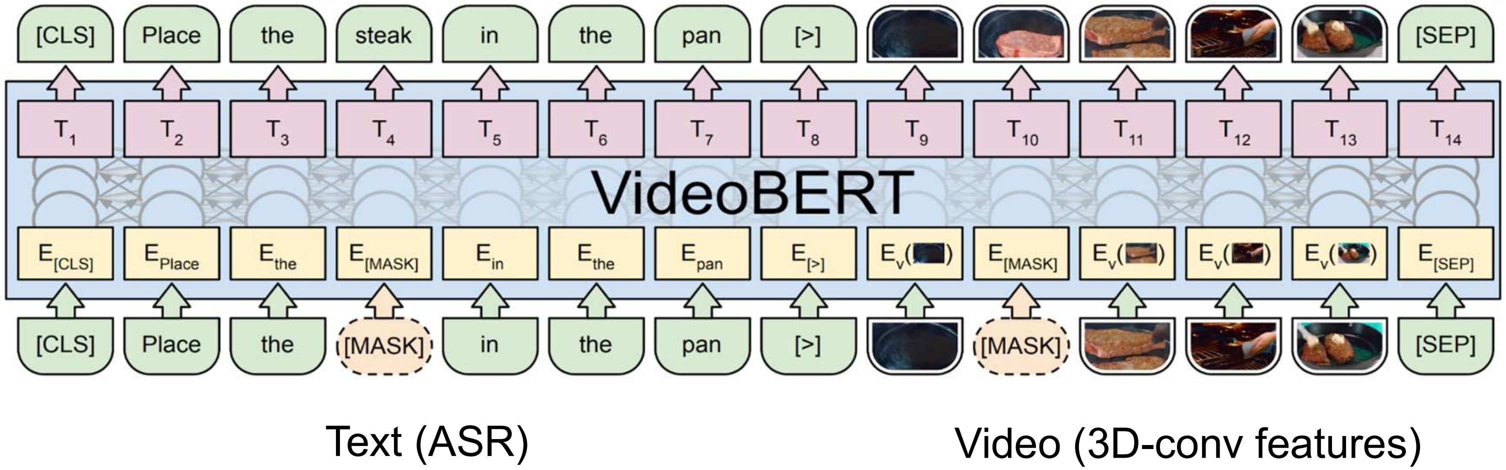
> Apply a little bit of [mask] on [mask] side and place a portion of the stuffing. Spread [mask] cover with another slice of the [mask] and apply some more butter on top since we're gonna grill the [mask].

BERT model

- BERT: Bidirectional Encoder Representations from Transformers [Devlin et al., NAACL'19]



VideoBERT



- Multimodal transformer: excellent way of combining multiple modalities
- Masked ‘language’ modeling as in BERT, video-speech alignment
- Video representation with 3D-convolutions + clustering

Video representation

- 3D convolutions for 1.5 second video clips (S3D), 1024-dim features vector
- Video tokenization by clustering
- Hierarchical k-means: depth of 4, branch size of 12 (20736 clusters)
- High-level semantics preserved after tokenization

Original:

Centroids:

VideoBERT

Training on 300k cooking videos

"Keep rolling tight and squeeze the air out to its side"

Zero-shot prediction

Verb: make, **Noun:** pizza

Zero-shot prediction

Method	Verb (top-5 %)	Object (top-5 %)
S3D (supervised)	**46.9**	30.9
VideoBERT	43.3	**33.7**

Results on YouCook II dataset

Pre-training size	Verb (top-5 %)	Object (top-5 %)
10K	15.5	17.8
50K	15.7	27.3
100K	24.5	30.6
300K	**43.3**	**33.7**

- VideoBERT learns video-language correspondence
- Close to fully-supervised accuracy
- More data improves the performance (not saturated yet)

Fine-tuning on downstream tasks

- For captioning cooking video on YouCook2

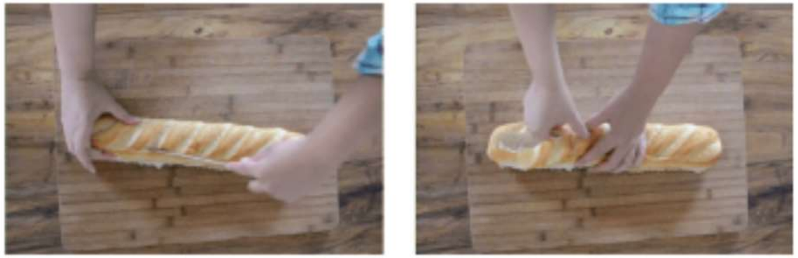
Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Zhou et al. (CVPR'18)	-	1.42	11.20	-	-
S3D	6.12	3.24	10.00	26.05	0.35
VideoBERT	6.80	4.07	10.99	27.51	0.50

- Effective and outperforms S3D features
- Pre-training helps!

Video captioning - examples

GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl



GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

GT: cut yu choy into diagonally medium pieces

VideoBERT: chop the cabbage

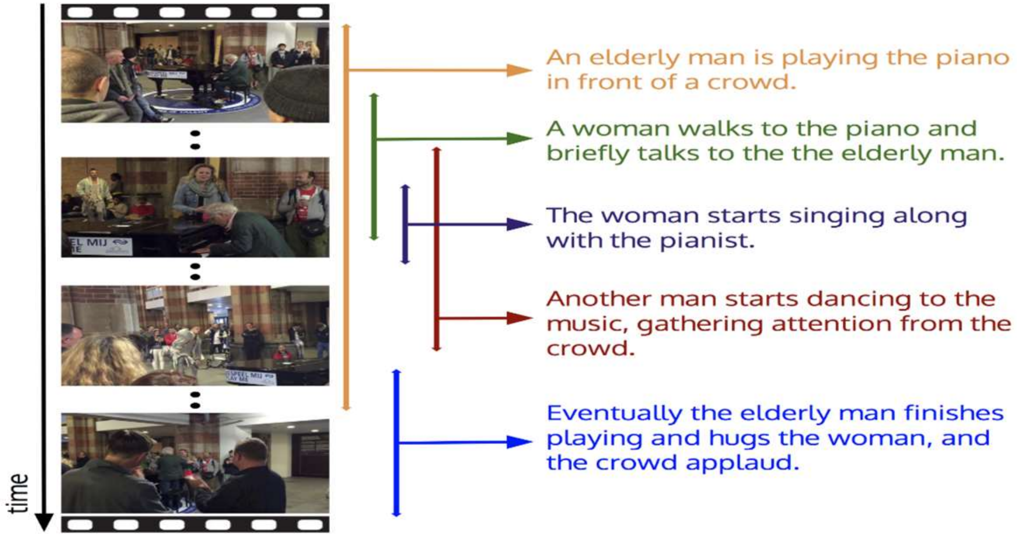
GT: remove the calamari and set it on paper towel

VideoBERT: fry the squid in the pan

Dense video captioning - task

Video captioning models for long videos with multiple events

- Captions are grounded in the video
- Combines localization and text generation



Example of dense, overlapping captions from the ActivityNet dataset

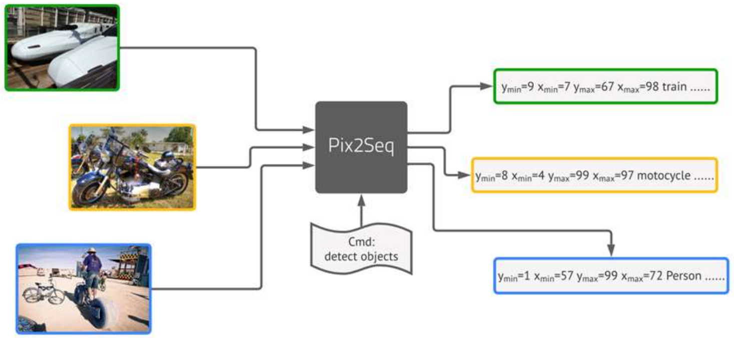
Dense video captioning – SOTA

Current approaches for dense video captioning

- Train separate networks for localization and captioning
- Require task-specific components like event counters
- Train on manually annotated datasets (small)
- Cannot reason over *long* videos

Localization as language modeling

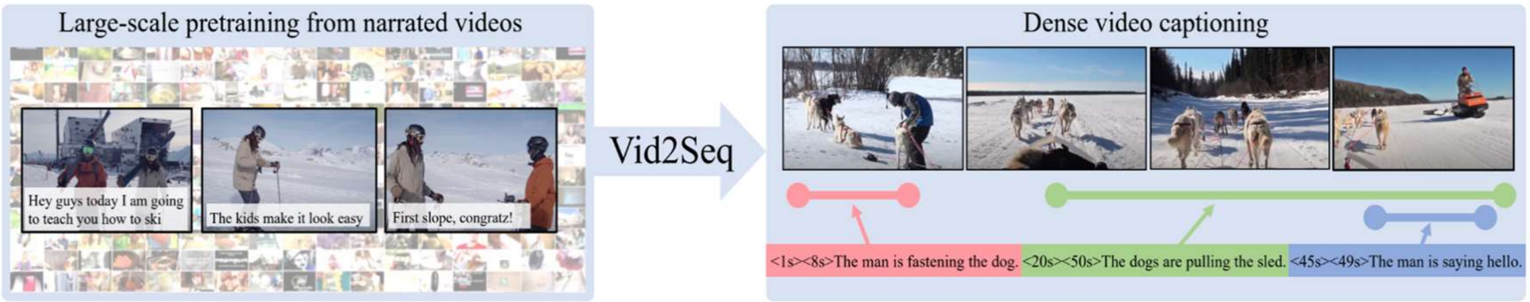
- Pix2seq casts object detection as sequence generation
- Spatial coordinates are quantized and tokenized



Vid2Seq approach

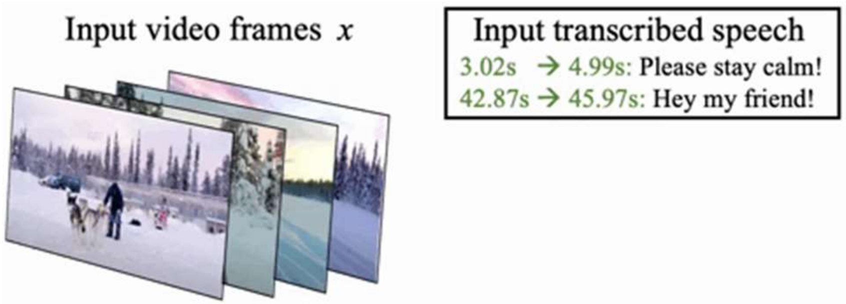
- Single target sequence consists of **Text + Time tokens combining localization + captioning**

- Large-scale pretraining from narrated untrimmed videos

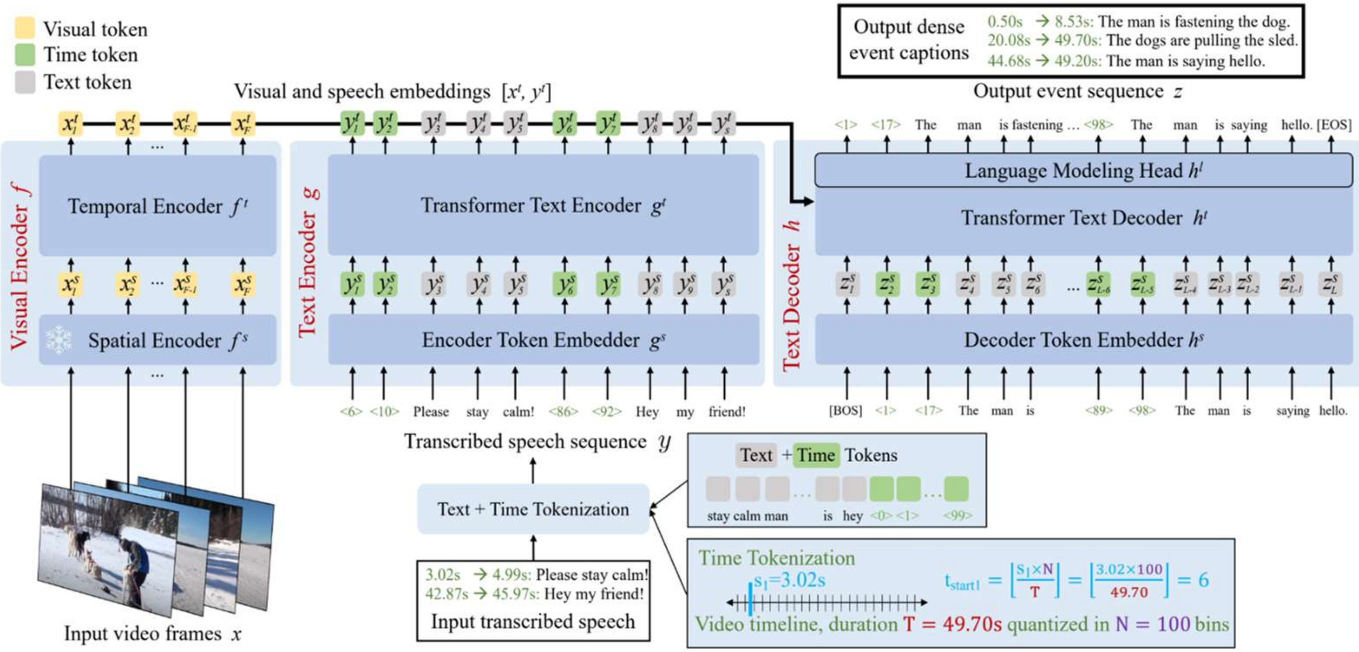


[Vid2Seq, A. Yang et al., CVPR 2023]

Vid2Seq – model



Vid2Seq – model



- Frozen Visual backbone (CLIP)
- Temporal Encoder for video
- Speech is cast as a single sequence of text and time tokens
- T5 Encoder & Decoder

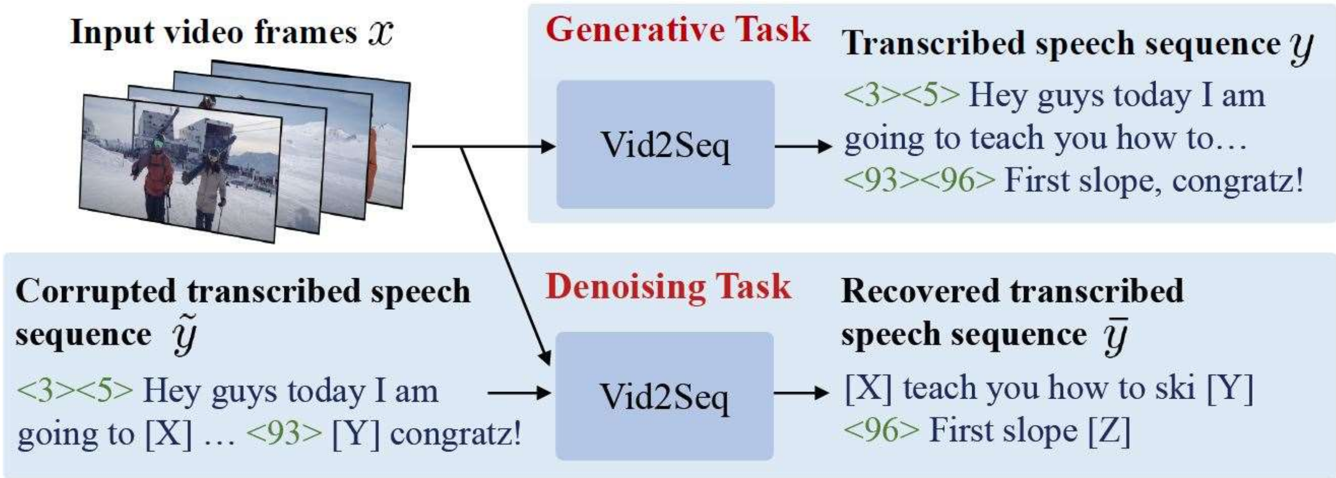
Vid2Seq – large-scale pretraining

- Pretraining data: 15 million videos from YT-Temporal-1B

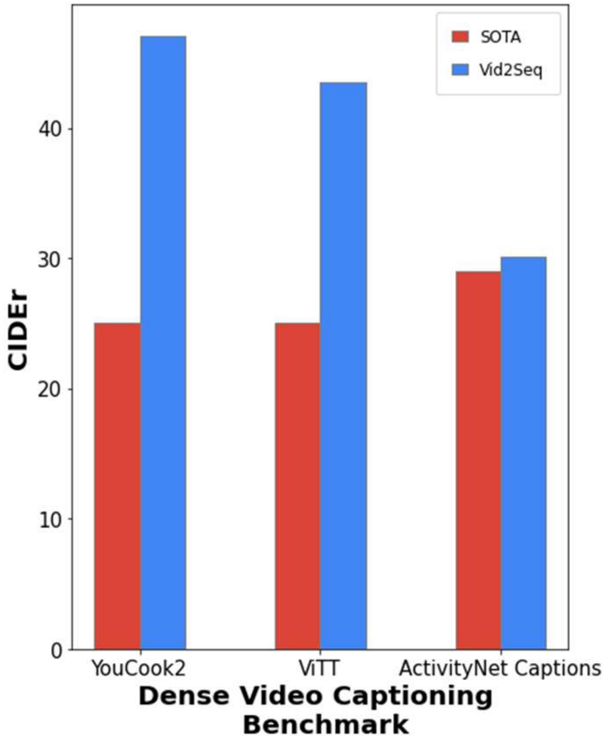
- ASR sentence boundaries used as event boundaries

Vid2Seq – large-scale pretraining

- Generative loss: given visual input predict speech
- Denoising loss: given visual input and corrupted ASR, predict the missing parts; training on visual + ASR input



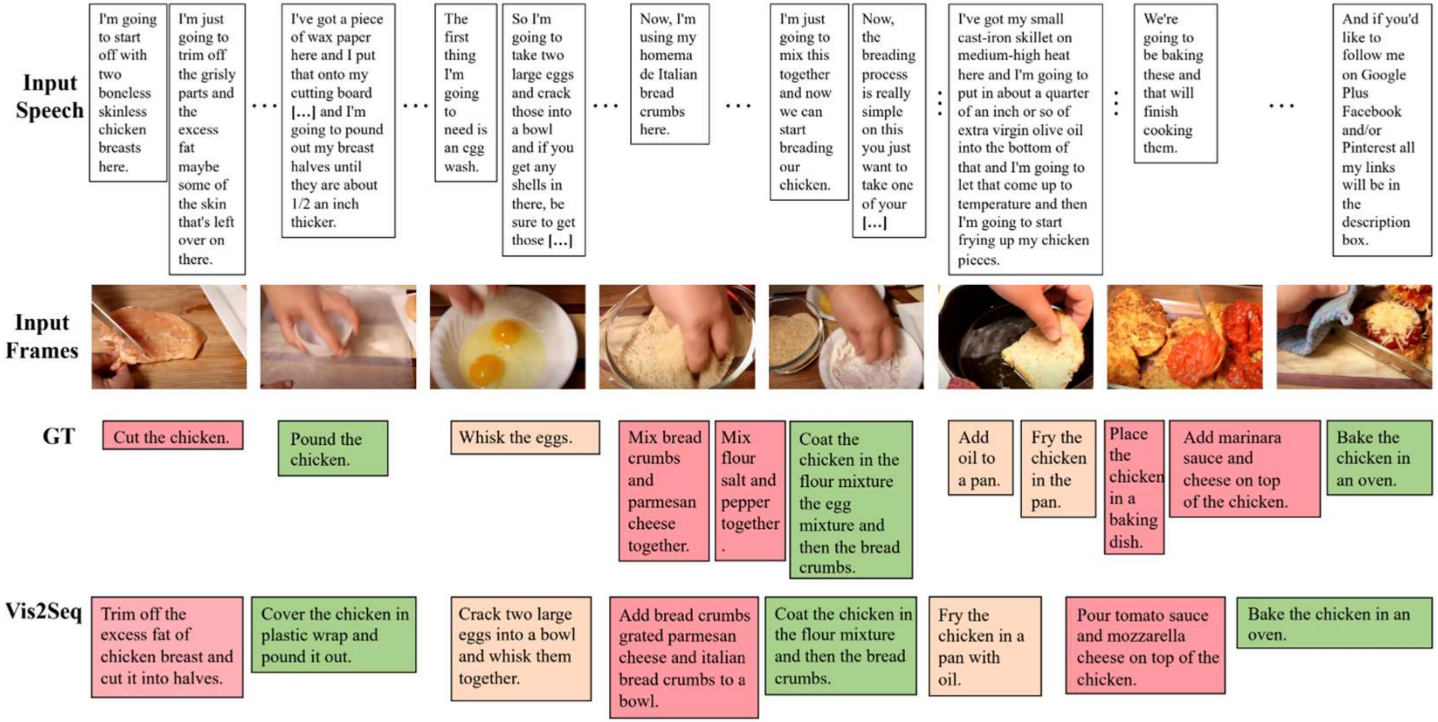
Vid2Seq – SOTA results



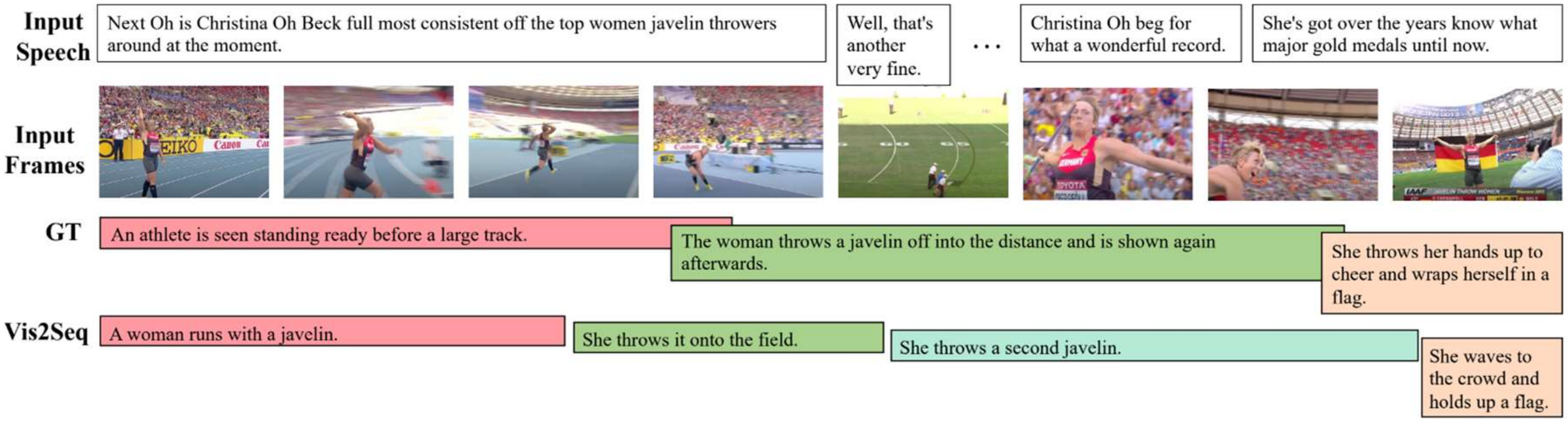
Ablation studies

- Pretraining is important, datasize and quality matter
- Time tokens help when pretraining on untrimmed videos
- Visual and speech information is complementary
- Importance of losses: denoising loss is important if we use speech during pretraining

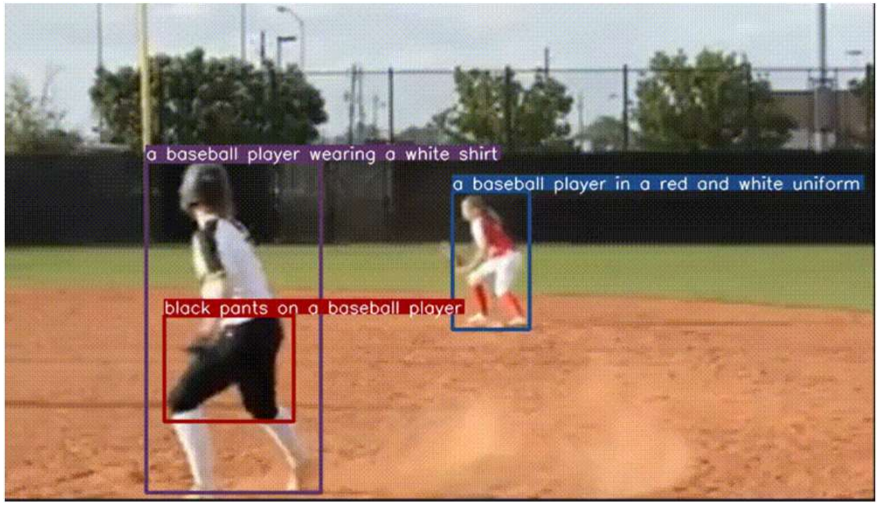
Qualitative results



Qualitative results



Dense Video Object Captioning



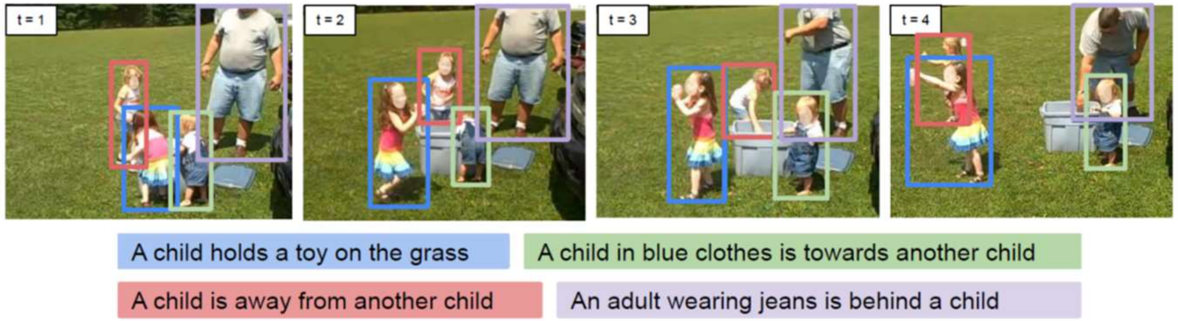
Detect, track and describe all objects in a video

→Object-centric video description / captioning

→Video object grounding

Dense video object captioning - task definition

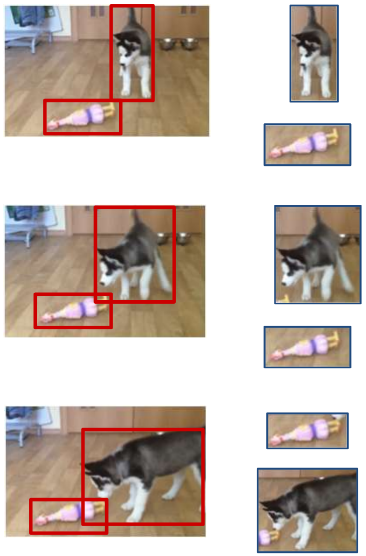
- Detect, track and caption objects



- Extension of the state-of-the-art multi-object tracking metric HOTA to include a captioning accuracy

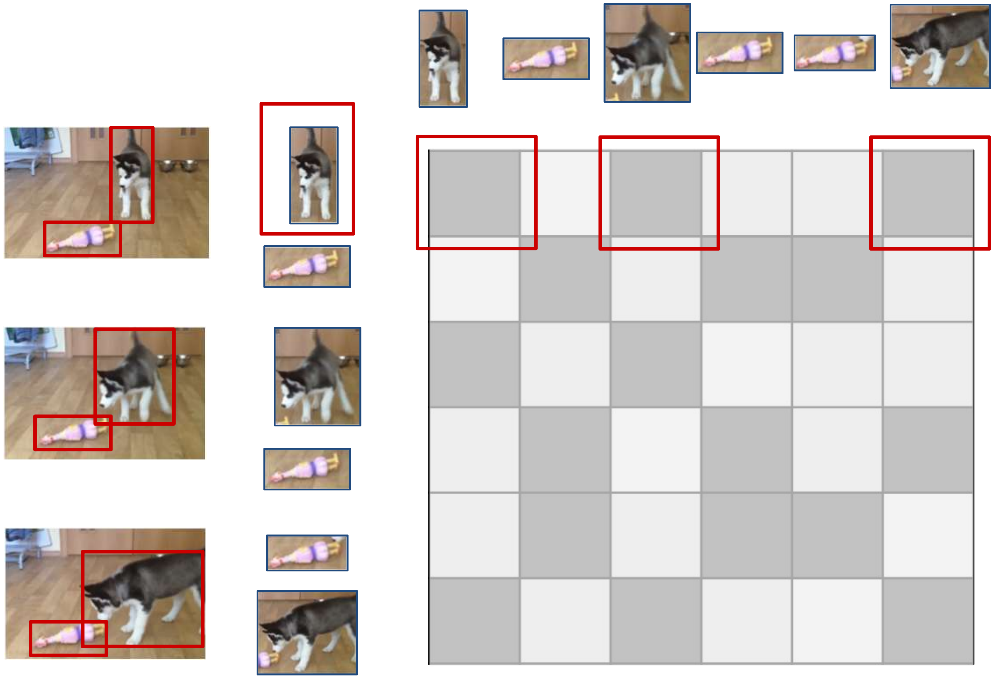
[Dense Video Object Captioning from Disjoint Supervision, X. Zhou et al., arXiv'24]

End-to-end video object tracking & captioning



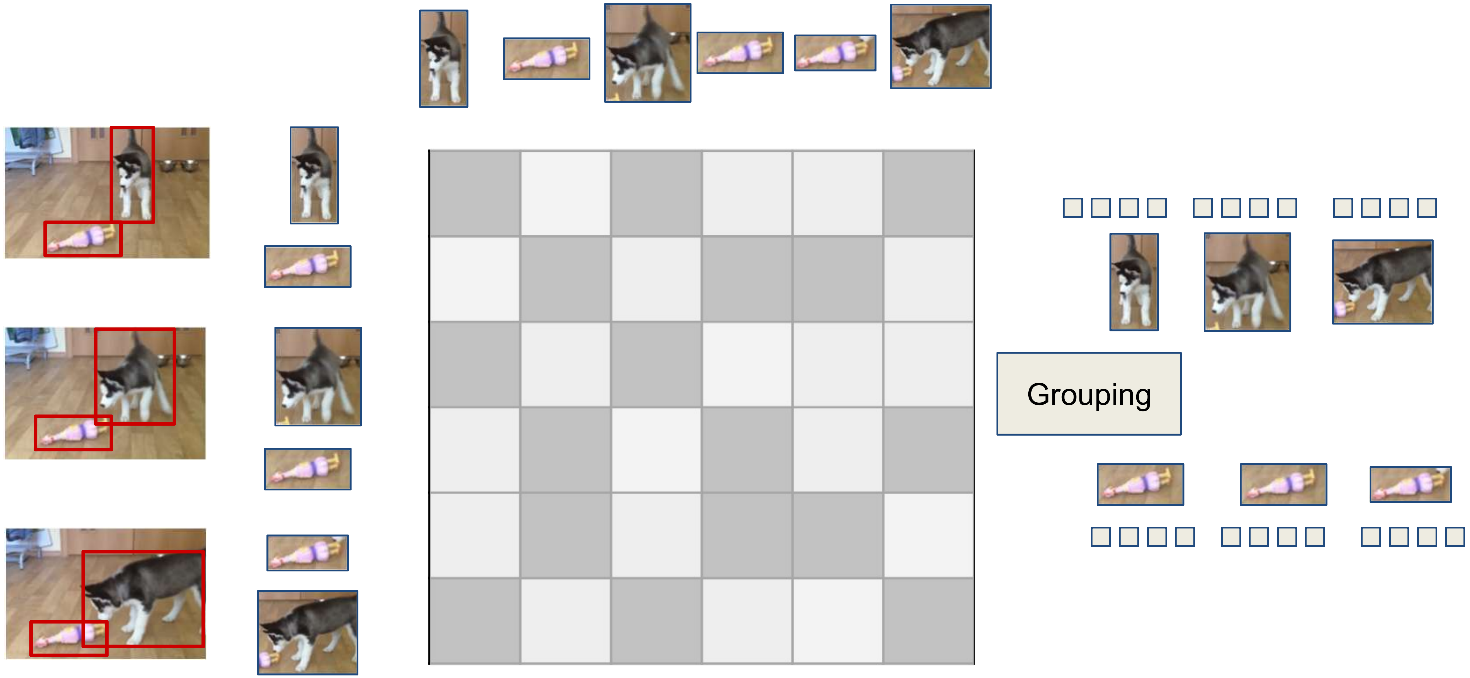
CenterNet to detect object proposals

End-to-end video object tracking & captioning

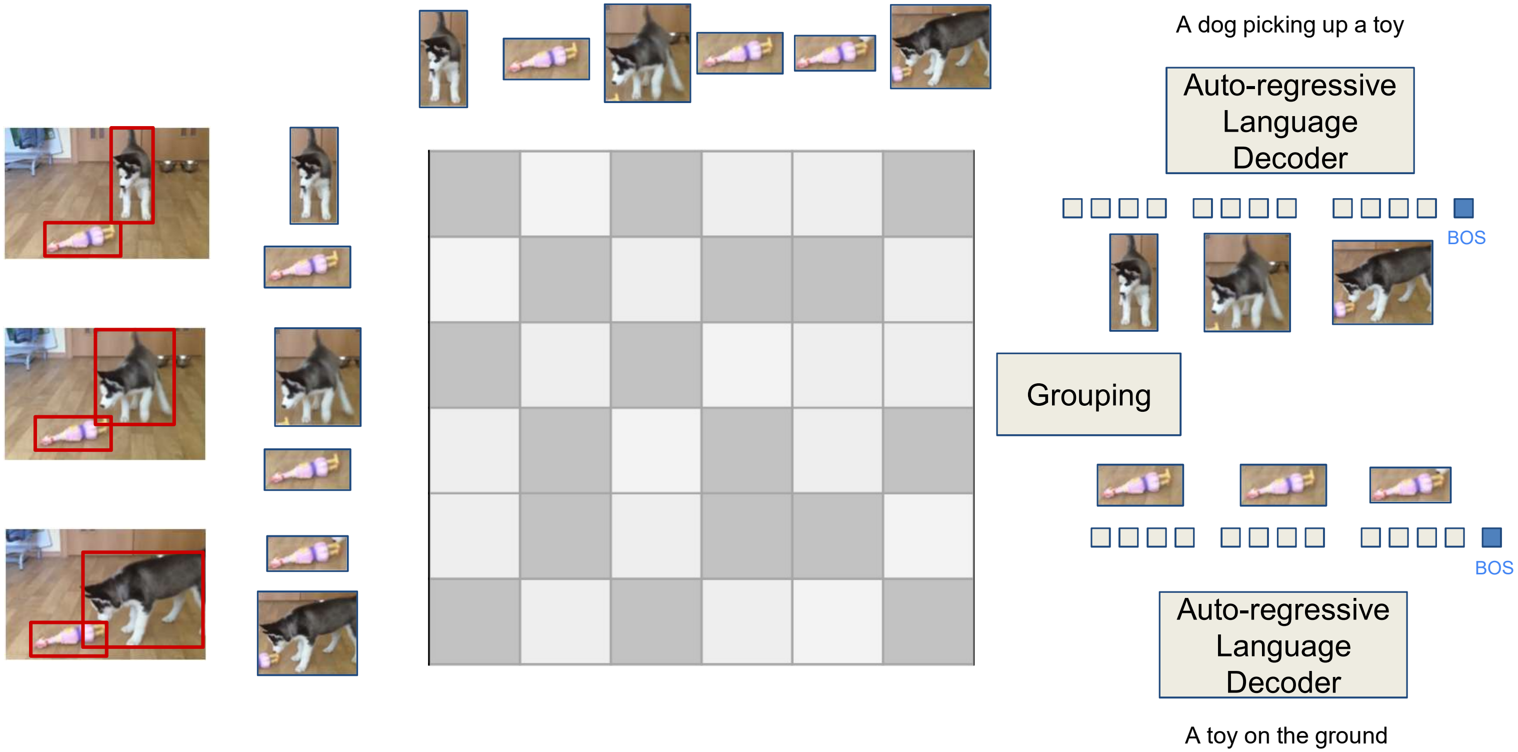


Feature association for tracking objects

End-to-end video object tracking & captioning

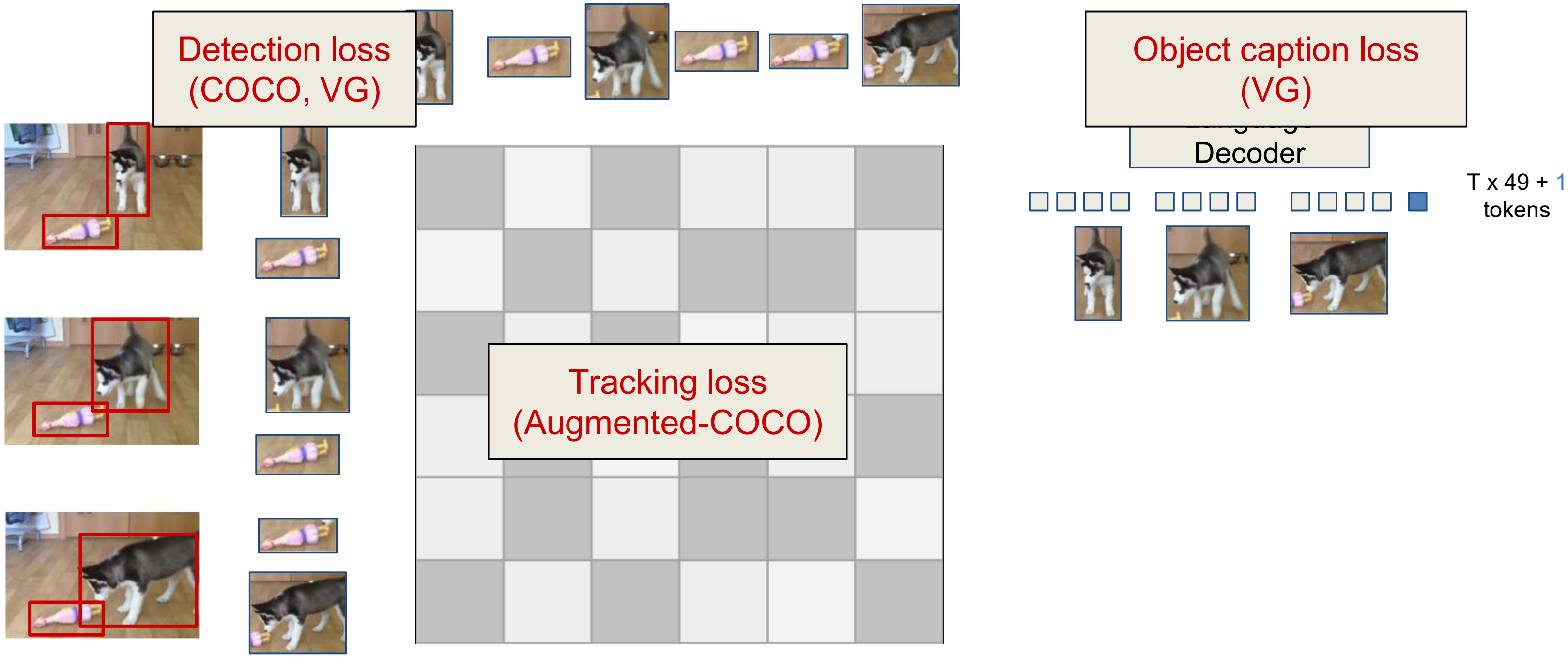


End-to-end video object tracking & captioning

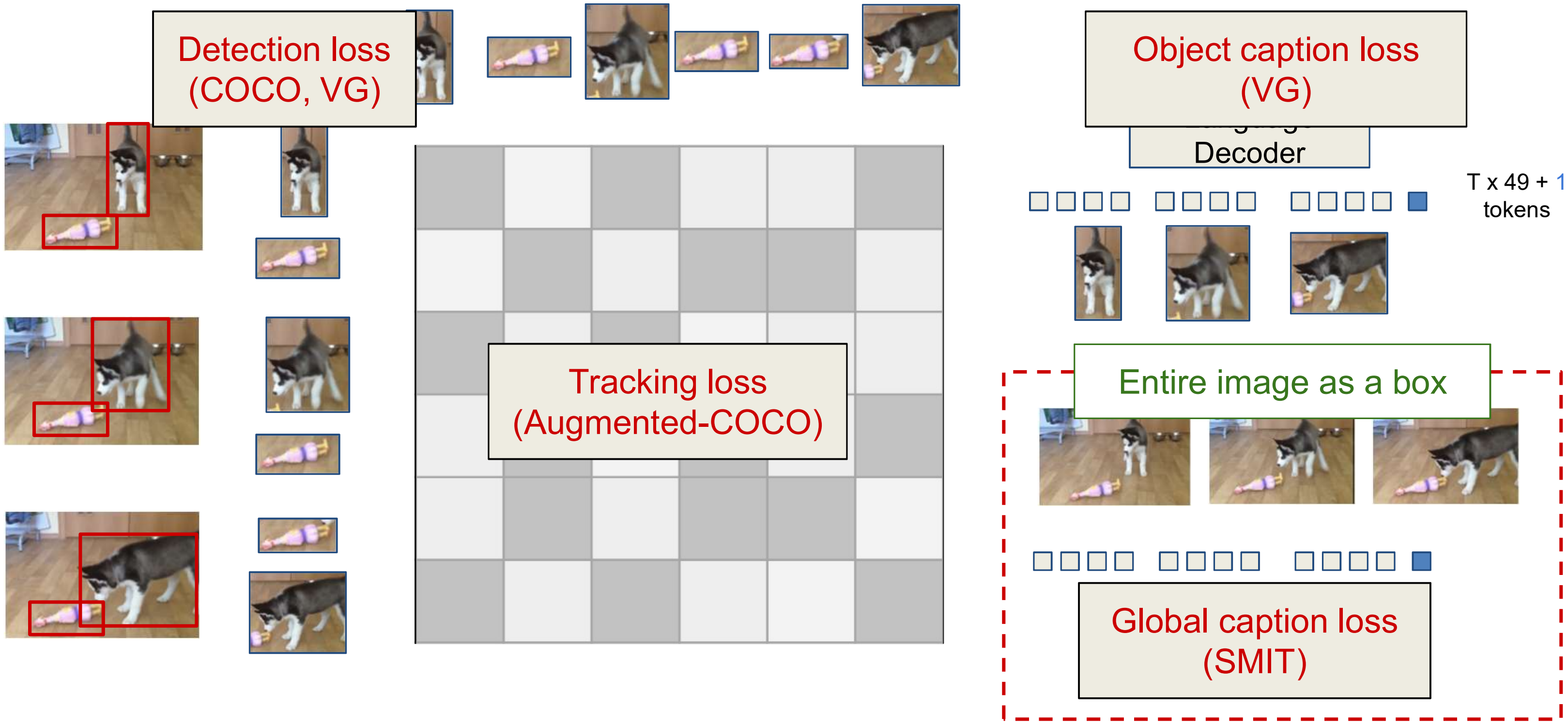


[Wu et al, GRiT: A Generative Region-to-text Transformer for Object Understanding, arXiv 2022]

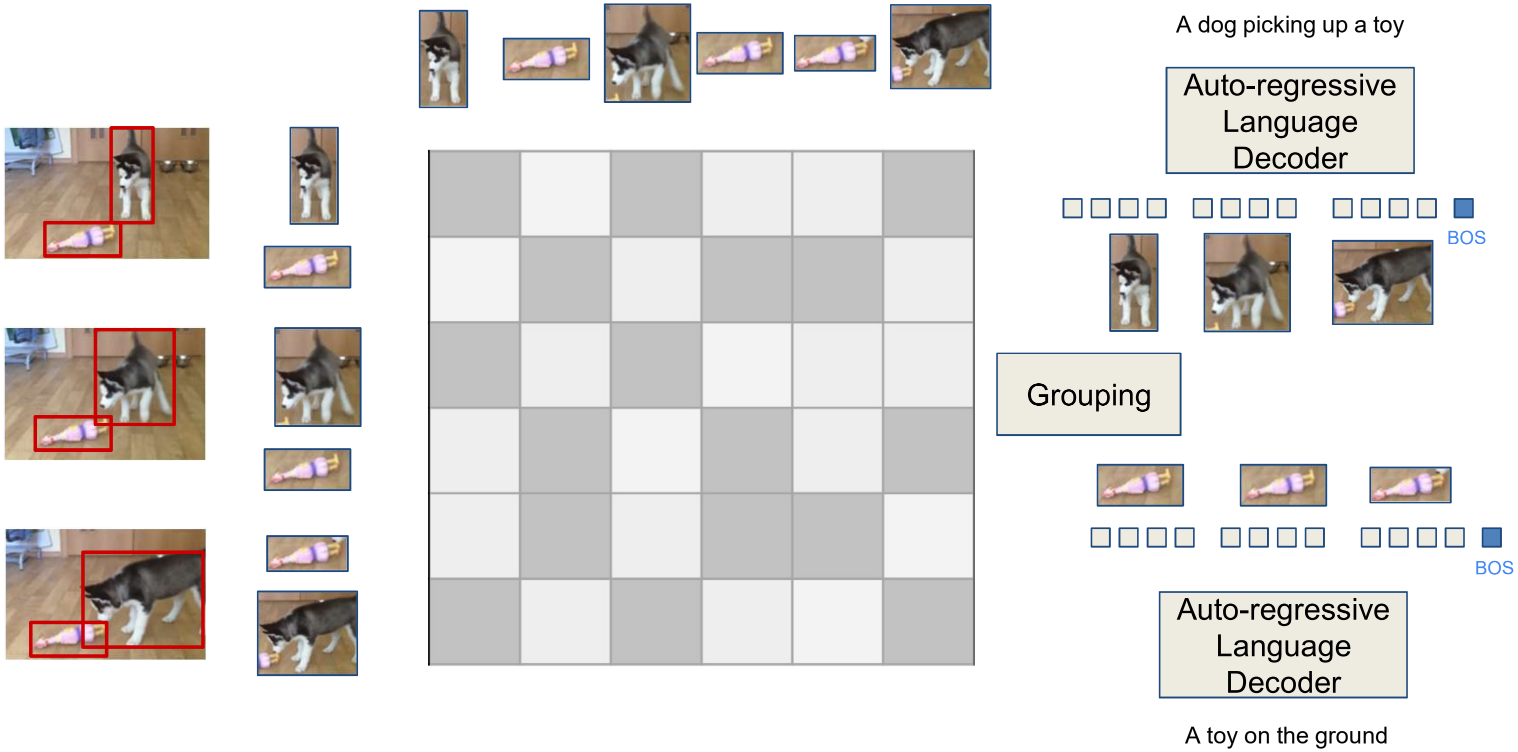
End-to-end video object tracking & captioning



End-to-end video object tracking & captioning



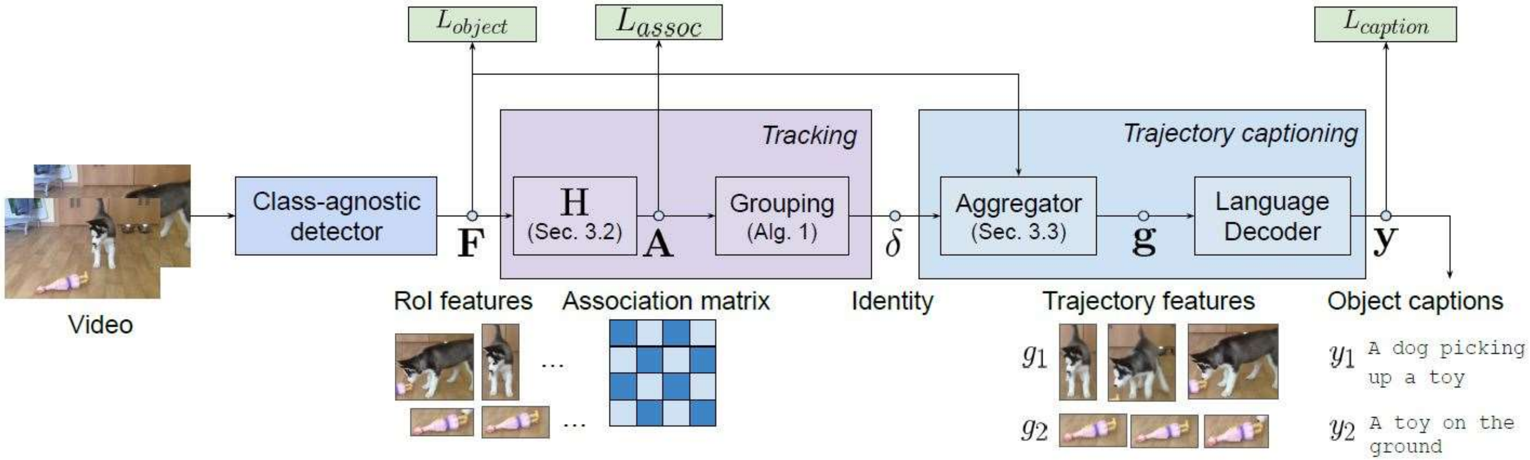
End-to-end video object tracking & captioning



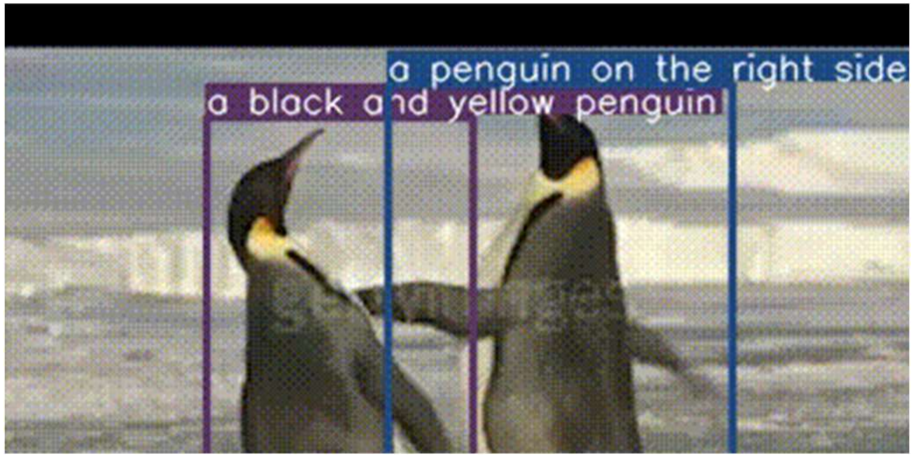
[Wu et al, GRiT: A Generative Region-to-text Transformer for Object Understanding, arXiv 2022]

Dense Video Object Captioning

Training losses



Qualitative results



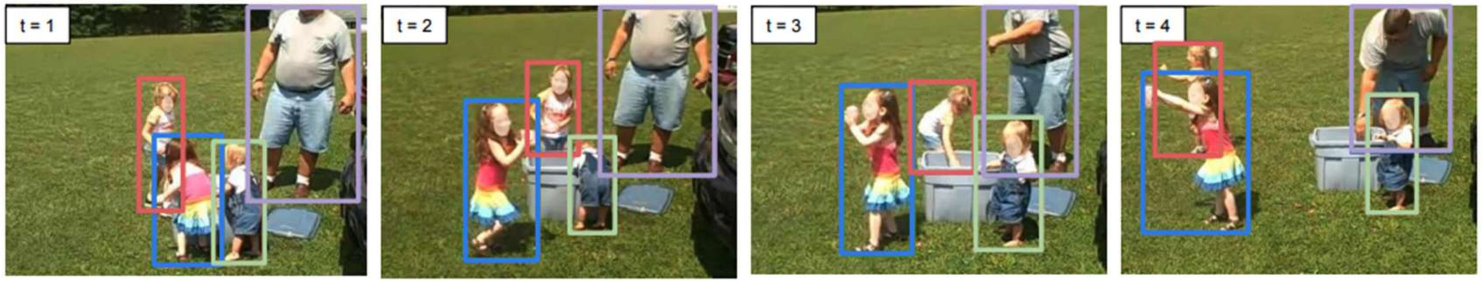
Quantitative results

#	COCO	VG	SMiT	Aug COCO	VidSTG (zero-shot) CHOTA	DetA	AssA	CapA	VLN (zero-shot) CHOTA	DetA	AssA	CapA	VidSTG (finetuned) CHOTA	DetA	AssA	CapA	VLN (finetuned) CHOTA	DetA	AssA	CapA
0					-	-	-	-	-	-	-	-	47.8	54.6	57.8	34.5	29.7	35.3	85.4	8.7
1	✓				-	48.9	-	-	-	27.8	-	-	52.3	64.9	63.0	34.9	31.8	43.9	88.7	8.2
2		✓			-	17.8	-	7.8	-	12.1	-	7.4	54.9	64.2	65.9	39.1	40.6	**45.1**	88.4	16.7
3			✓		-	-	-	-	-	-	-	-	45.4	51.9	56.9	31.6	37.4	41.2	87.7	14.5
4		✓	✓		-	19.1	-	8.5	-	14.3	-	8.5	55.2	64.0	67.1	39.2	41.0	44.2	88.4	**17.8**
5	✓	✓			-	49.9	-	8.1	-	28.0	-	7.8	55.6	65.7	68.9	38.4	40.9	44.1	88.8	17.4
6	✓		✓		-	50.4	-	4.9	-	28.7	-	7.5	54.4	64.9	63.9	38.8	35.6	43.7	88.5	11.6
7	✓	✓	✓		-	51.3	-	9.1	-	**29.9**	-	9.0	56.5	**65.8**	68.2	**40.1**	41.1	44.2	88.9	17.7
8	✓	✓	✓	✓	**31.1**	**51.4**	**59.6**	**9.8**	**29.2**	29.1	**88.0**	**9.7**	**56.9**	**65.8**	**70.4**	39.7	**41.3**	44.3	**89.5**	17.7

Measure for evaluation $\text{CHOTA} = \sqrt[3]{\text{DetA} \cdot \text{AssA} \cdot \text{CapA}}$

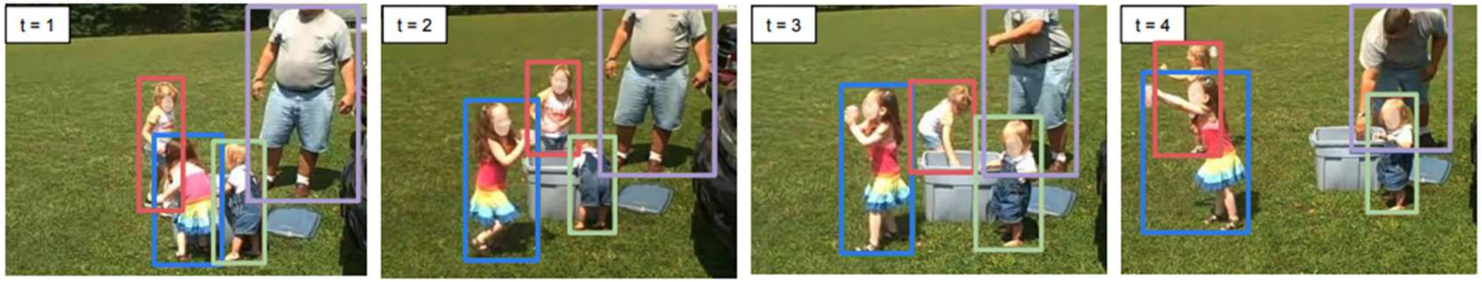
Application to video grounding

Query: q = “A child holds a toy on the grass”



Application to video grounding

Query: q = “A child holds a toy on the grass”

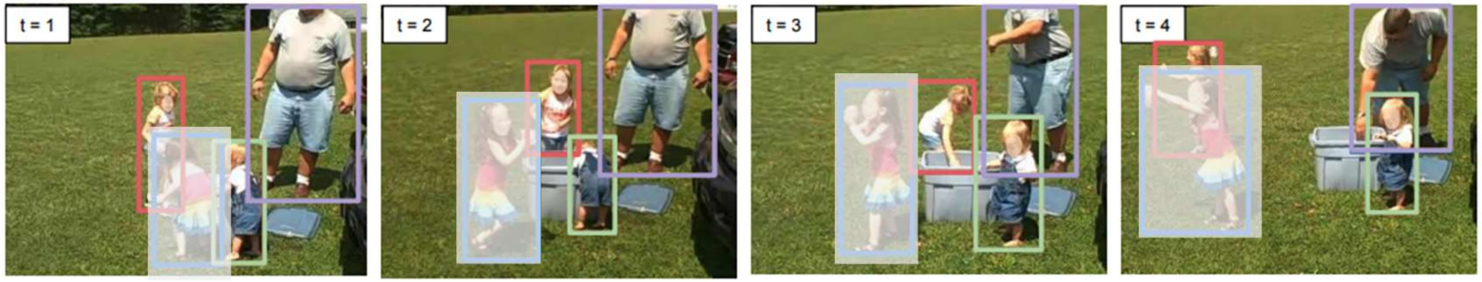


likelihood(, q) = 0.9 likelihood(, q) = 0.5

likelihood(, q) = 0.4 likelihood(, q) = 0.1

Application to video grounding

Query: q = “A child holds a toy on the grass”



likelihood( , q) = 0.9

likelihood( , q) = 0.5

likelihood(, q) = 0.4

likelihood( , q) = 0.1

Video grounding results

	Finetuned	Zero-shot
STVGBert [52]	47.3	-
TubeDETR [66]	59.0	-
STCAT [29]	61.7	-
Ours	**61.9**	54.1

VidSTG spatial-grounding

Average intersection over union with GT (IoU)

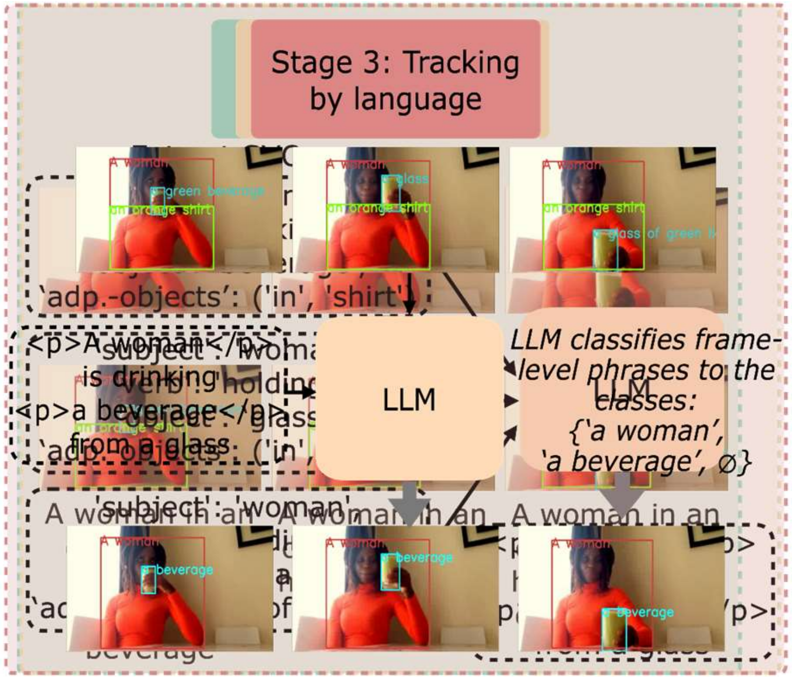
Grounded Video Caption

- Input: video
- Sub-task 1: captioning
- Sub-task 2: Identify noun phrases
- Sub-task 3: Grounding

[Grounded Video Object Captioning, E. Kazakos et al., arXiv'24]

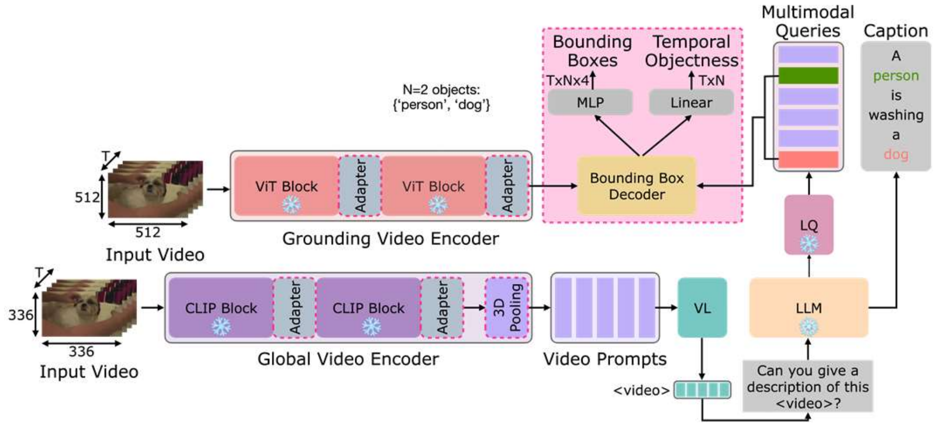
Automatic annotation method

- Run GLaMM for each frame
- Aggregate frame-level captions into video-level captions using extracted Subject-Verb-Object from the caption
- Tracking by language: Classify frame-level phrases into video-level phrases



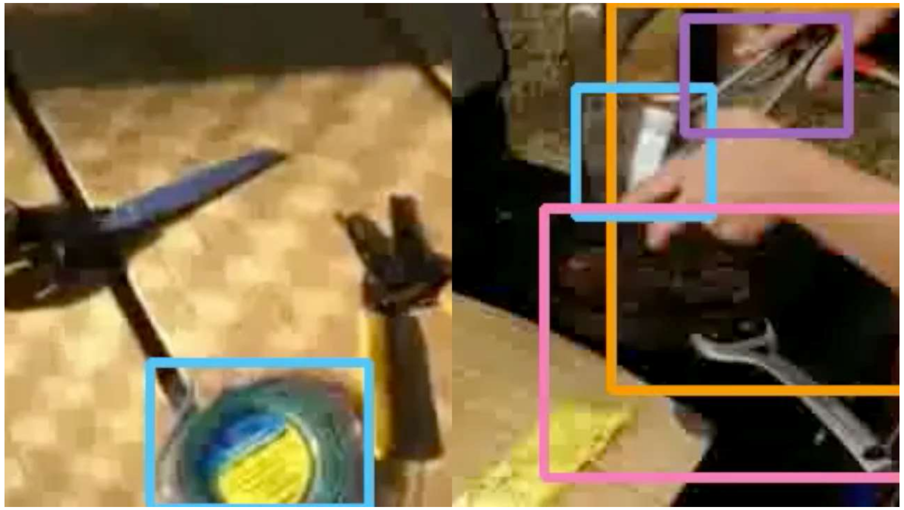
GROC model

- GROC = **GRO**unded Video **C**aptioning
- 1 encoder for captioning + 1 for grounding
- Adapters for spatio-temporal modelling
- LLM predicts caption and noun phrases locations
- Temporal objectness predicts the presence of an object in a frame



The GROC dataset

- 2100 examples
- Train/val/test: 1000/100/1000
- Multiple-frames
- Multiple objects per frame



Experimental Results

Method	METEOR	CIDER	AP50	Recall
GLaMM [21]	11.9	29.9	26.6	22.0
Pseudolabelling	13.8	40.0	27.1	20.4
GROC (ours)	**14.2**	**46.8**	**33.7**	**24.6**

Comparison with baselines

- Pseudo-labeling improves over an image-based approach
- Training our model on the pseudo-labels improves performance

Experimental Results

Pre-train	Fine-tune	METEOR	CIDEr	AP50	Recall
✓	✗	14.2	46.8	33.7	24.6
✗	✓	**20.7**	**78.0**	10.2	10.7
✓	✓	<u>20.6</u>	<u>72.7</u>	**36.2**	**26.8**

Evaluating fine-tuning on the GROC dataset

- Fine-tuning improves captioning significantly
- Our pre-training is necessary; without it fine-tuning fails

Multimodal data for generating automatic training data

- Large-scale weakly supervised data
  - HowTo100M dataset with 100M video-ASR pairs [HowTo100M. A. Miech et al., ICCV’19]

  - WebVid10M dataset with 10M video-text pairs [Frozen In Time, M. Bain et al., ICCV’21]

“Billiards, concentrated young woman playing in club”

“Female cop talking on walkie-talkie, responding emergency call, crime prevention”

Multimodal data for generating automatic training data

- Cross-modal supervision
  - Levering text model for annotating clips with question/answers
- Data Mining
  - Semi-automatic pipeline for generating a long video understanding dataset

Cross-model supervision: JustAsk

- Learning zero-shot video question answering with cross-modal supervision

Question: What type of animal do we see?

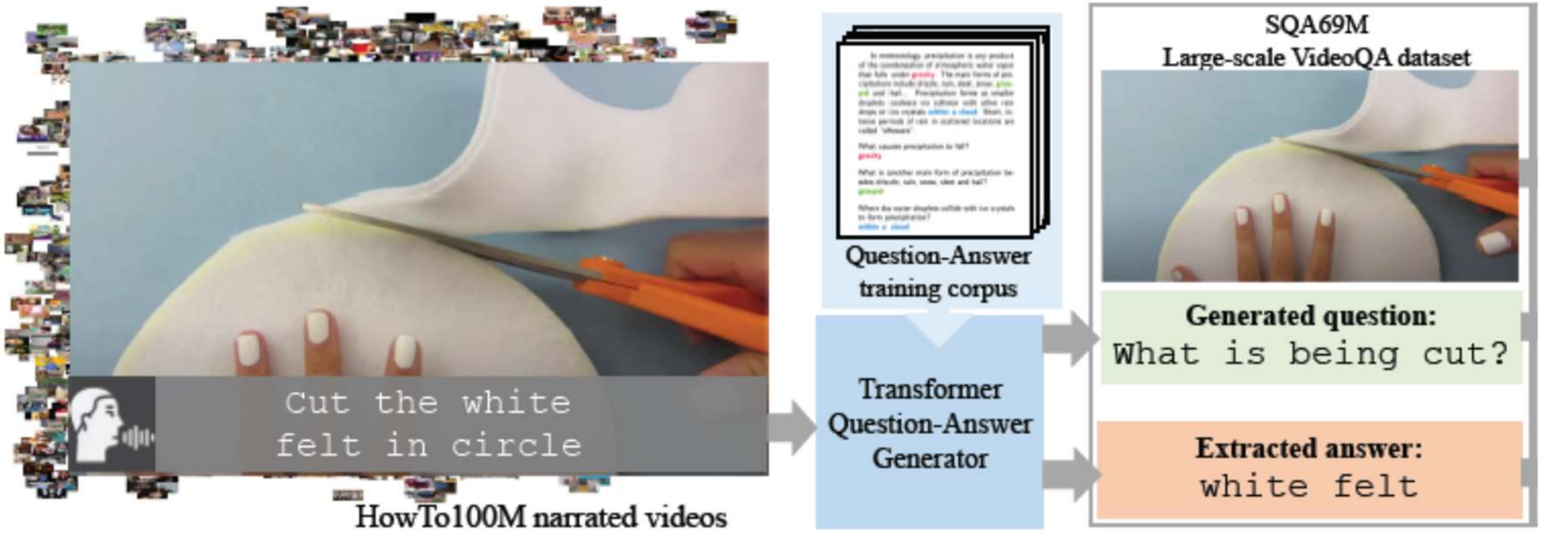
Our answer: Fish.

- Generate a large-scale video question answering dataset automatically (HowToVQA69M)

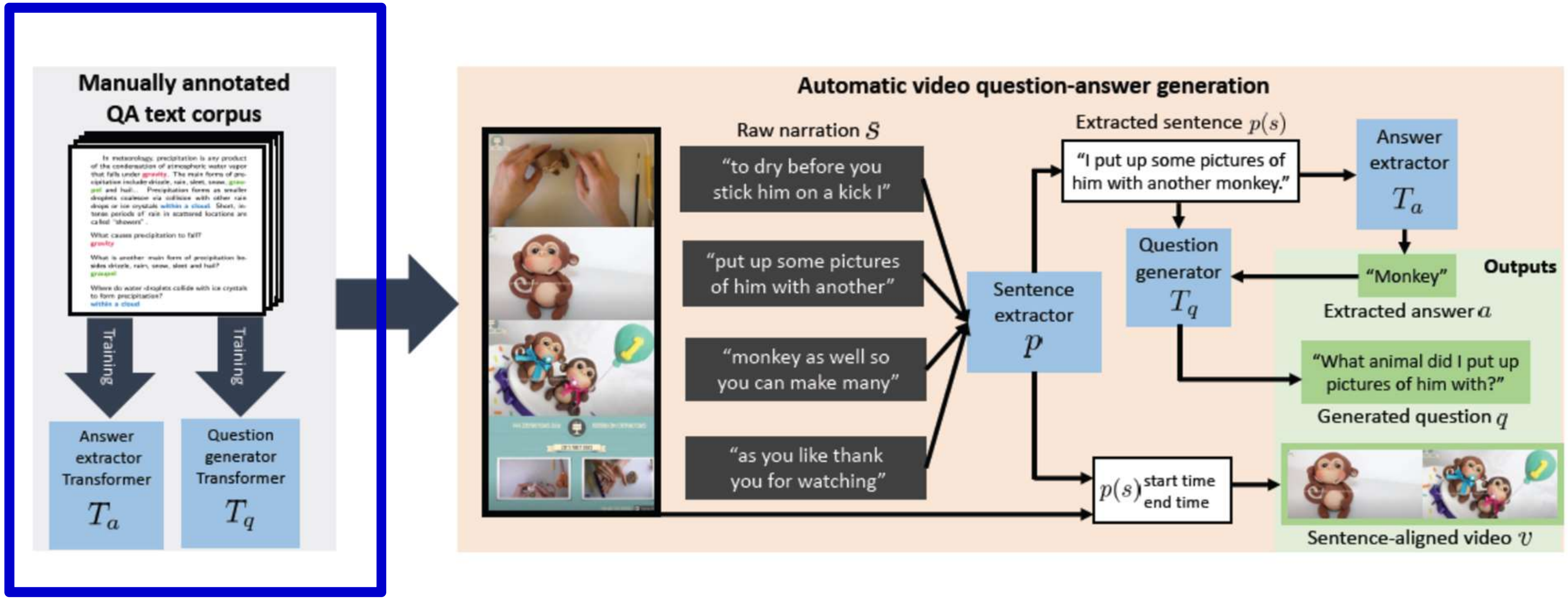
[JustAsk, A. Yang et al., ICCV'21]

Cross-modal supervision: JustAsk

- HowTo100M dataset with ASR captions
- Textual question-answer training corpus + transformer model
- Transformer extracts answer + question from ASR caption

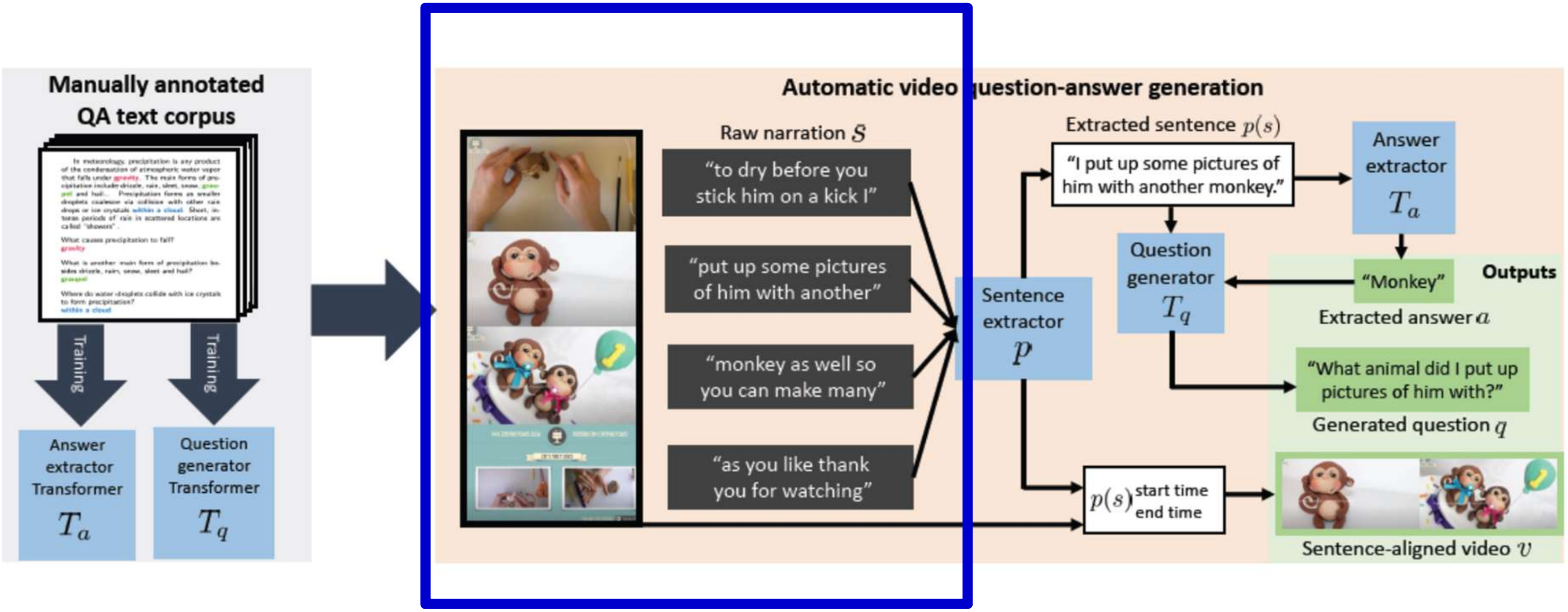


Cross-model training



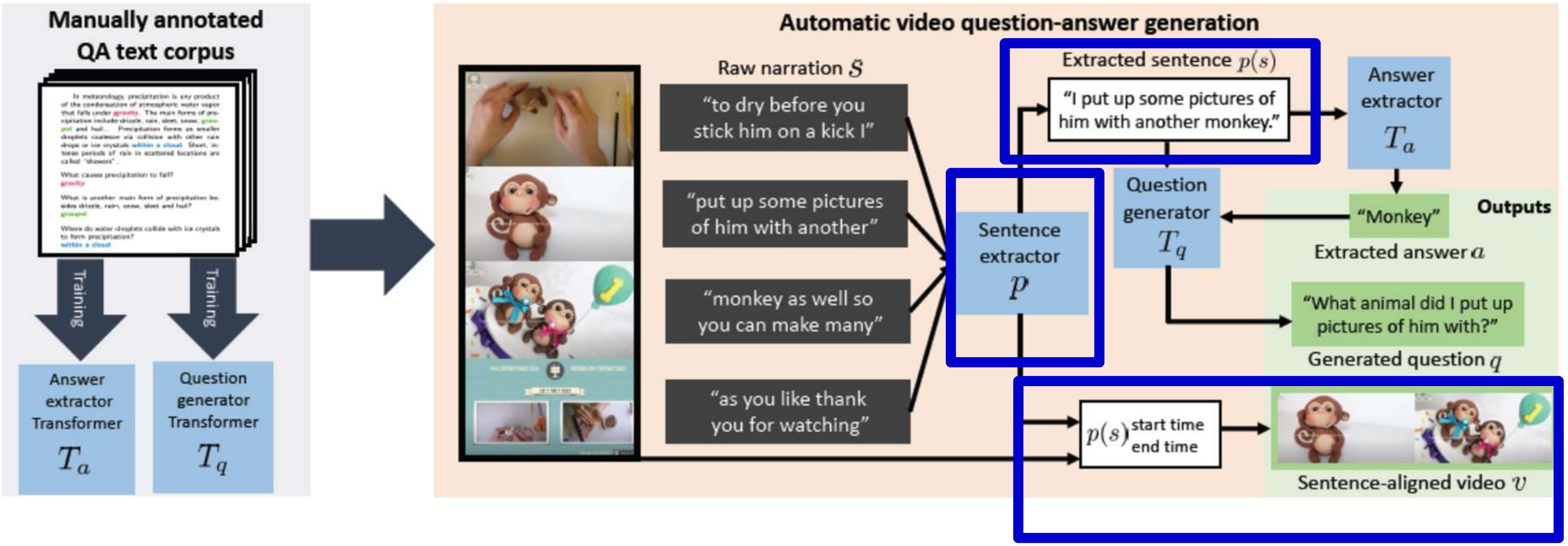
- Manually annotated QA text corpus: SQuADv1
  - 100k question-answer pairs for paragraphs from Wikipedia articles
- Transformers Ta and Tq are trained for answer extraction and answer-aware question extraction on SquADv1

Cross-model training



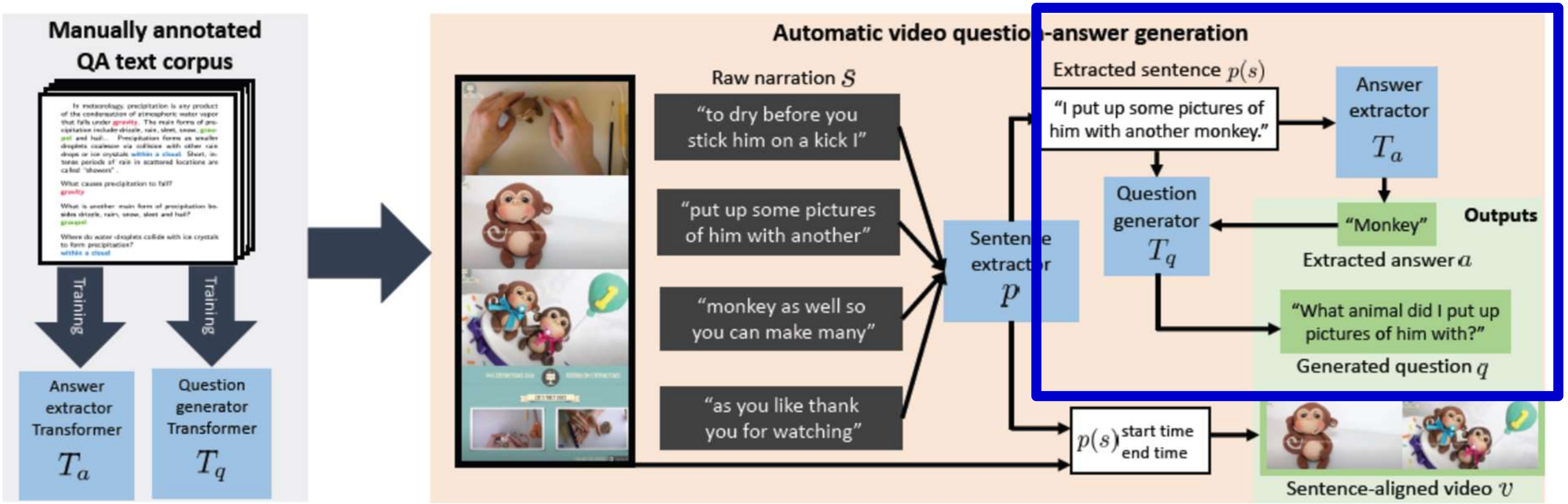
- HowTo100M clips + speech transcribed with ASR

Cross-model training



- HowTo100M clips + speech transcribed with ASR
- Sentence / punctuation extraction with recurrent network
  - Sentence aligned video

Cross-model training



- HowTo100M clips + speech transcribed with ASR
- Sentence / punctuation extraction with recurrent network
  - Sentence aligned video
- Answer + Question extraction with Ta and Tq

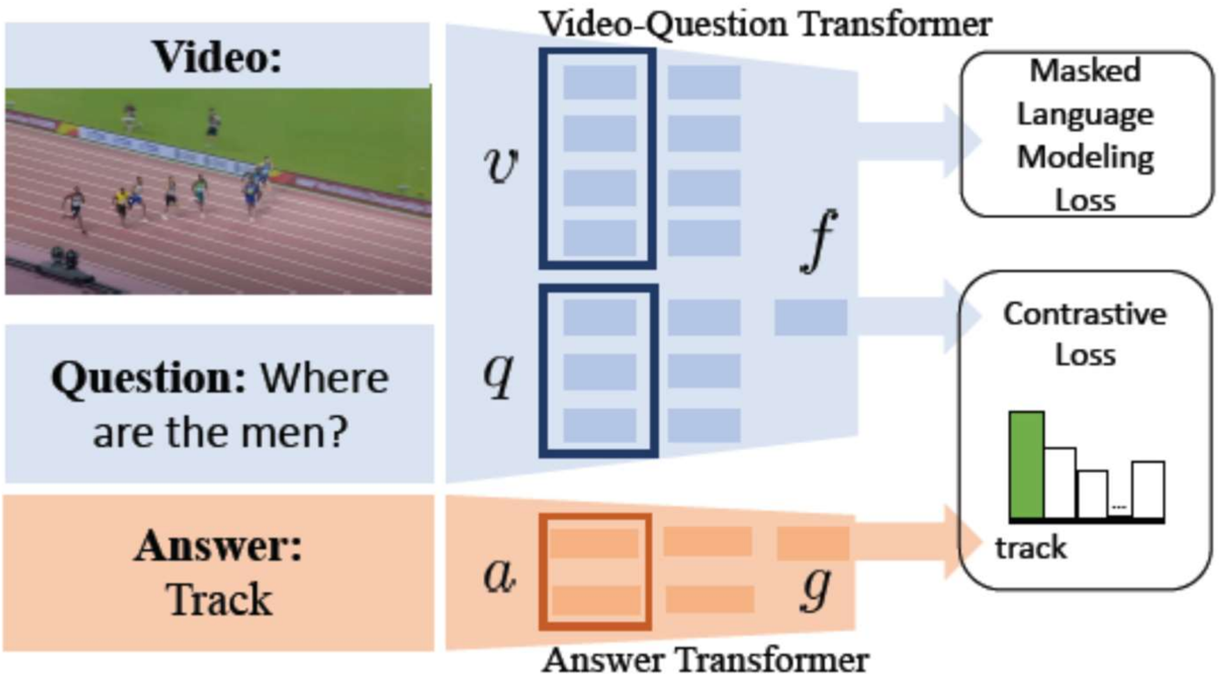
Example of generated question-answer

ASR: Add some of your favorite sprinkles give it a mix.

Generated question: What can you add to the mix?

Generated answer: Sprinkles.

VideoQA architecture



$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{e^{f(v_i,q_i)^\top g(a_i)}}{e^{f(v_i,q_i)^\top g(a_i)} + \sum_{(v',q',a')\sim\mathcal{N}_i} e^{f(v',q')^\top g(a')}} \right)$$

- Multi-modal transformer
- Contrastive loss with positive and negative answers
  - Can deal with large-scale data, here 16M different answers

Zero-shot VQA

- No use of any annotated examples for training
- Results on state-of-the-art datasets, use of test data only

Pretraining	iVQA Top 1	iVQA Top10	MSVD-QA Top 1	MSVD-QA Top 10
Random	0.09	0.9	0.05	0.5
HowToVQA69M	12.2	43.3	7.5	22.4

Zero-shot results

Question: What is the largest object at the right of the man?

Our answer: Wheelbarrow.

[Text only: Statue.]

Impact of training data

- Results on state-of-the-art dataset with training data

Pretraining	iVQA Top 1	iVQA Top10	MSVD-QA Top 1	MSVD-QA Top 10
Zero-shot HowToVQA69M	12.2	43.3	7.5	22.4
Training w/o pretraining	23.0		41.2	
Training with pretraining HowTOVQA69M	35.4		46.3	

Impact of pretraining data size

Pretraining data size	Zero-shot		Finetune	
	iVQA	MSVD-QA	iVQA	MSVD-QA
0%	—	—	23.0	41.2
1%	4.5	3.6	24.2	42.8
10%	9.1	6.2	29.2	44.4
20%	9.5	6.8	31.3	44.8
50%	11.3	7.3	32.8	45.5
100%	**12.2**	**7.5**	**35.4**	**46.3**

- Amount of pretraining data impacts performance
- Not yet saturated

Neptune: Benchmarking Long Video Understanding

What was the direct cause of Ottawa Fury FC's victory?

Ottawa Fury FC's victory was directly caused by Valfoul's successful **penalty kick** in the 91st minute.

1. Ottawa Fury FC's victory was directly caused by their **superior skill** and tactics.
2. Ottawa Fury FC's victory was directly caused by Tampa Bay Rowdies' **poor performance**.
3. Ottawa Fury FC's victory was directly caused by the **referee's decision to award a penalty kick**.
4. Ottawa Fury FC's victory was directly caused by **the crowd's support**.

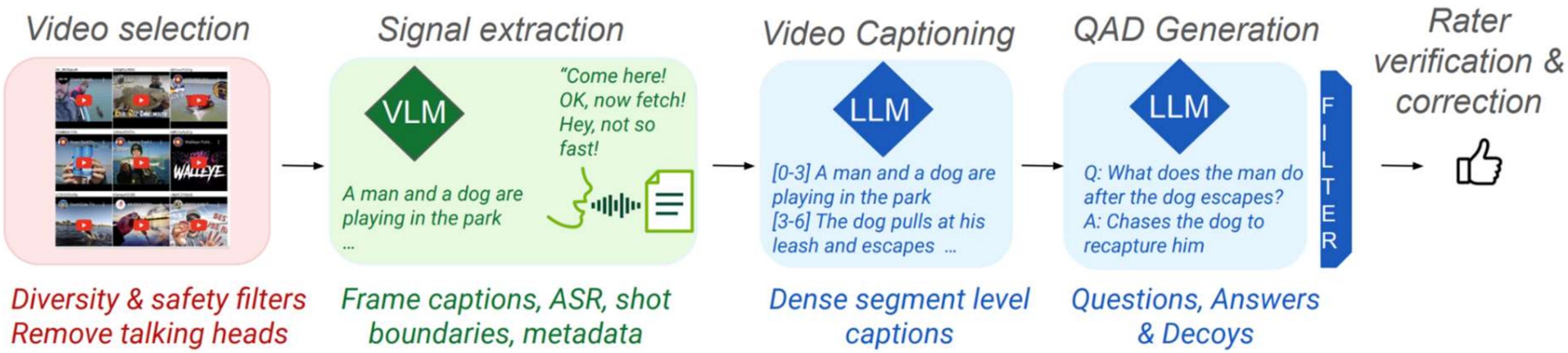
VideoQA is a great way to measure video understanding

- **Goal:** Answer questions about events, people, their motivations, understand temporal activities reason about cause and effect, people's relationships
- **Task:** video question answering
- **Project page:** https://github.com/google-deepmind/neptune

[A. Nagrani et al. Neptune: Benchmarking Long Video Understanding, 2024]

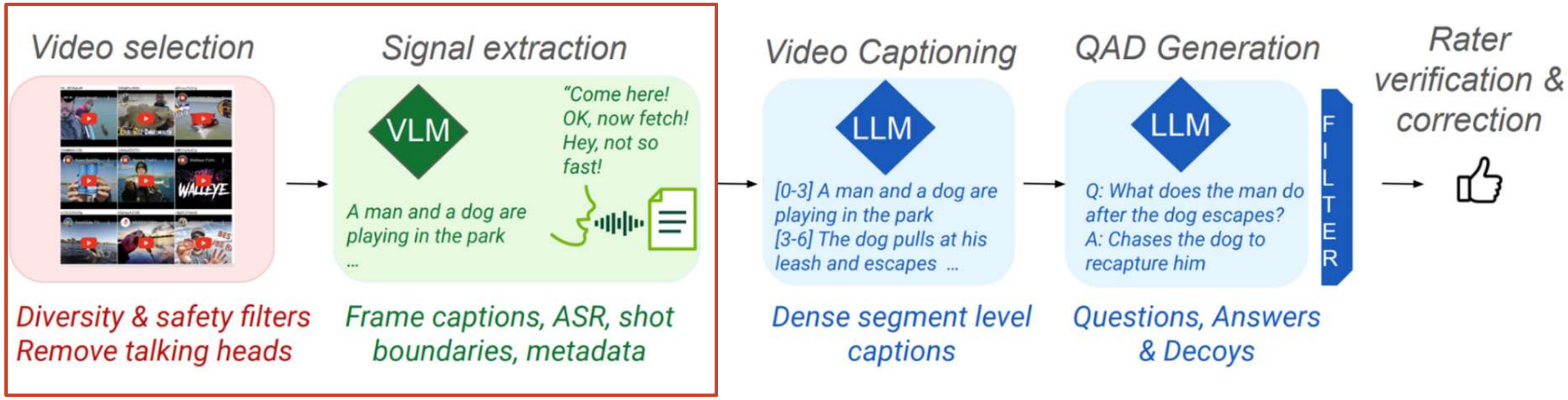
Semi-automatic pipeline

- *Annotation pipeline leveraging tools (YouTube filters, Gemini, VLMs) to reduce manual effort and achieve scale. Four automatic stages, followed by one manual rater stage.*



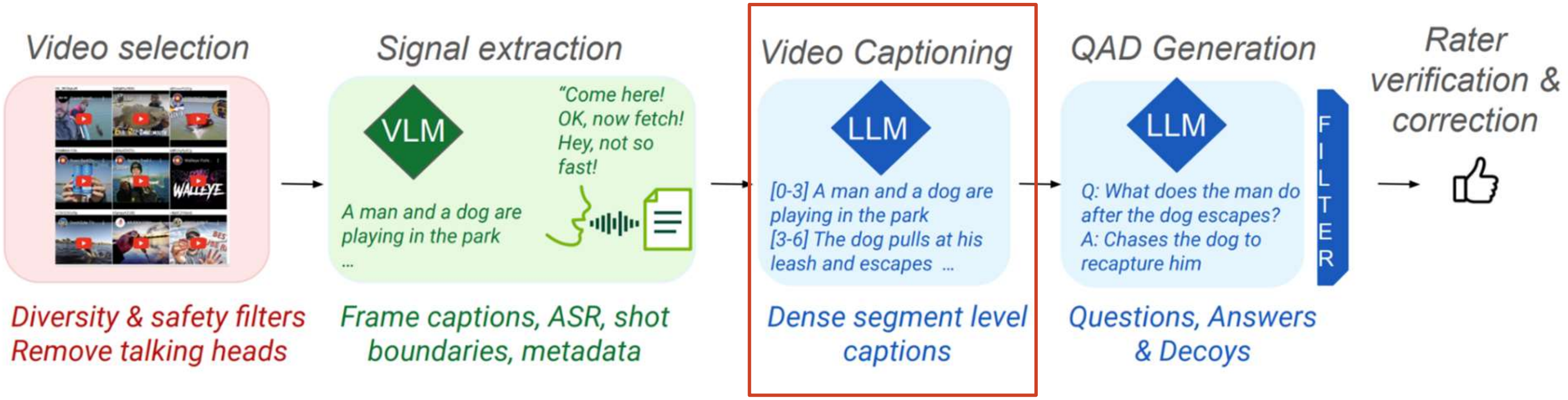
Video selection + extraction

- Filter suitable videos from the YT-Temporal-1Bn set
- Extract metadata



Video captioning

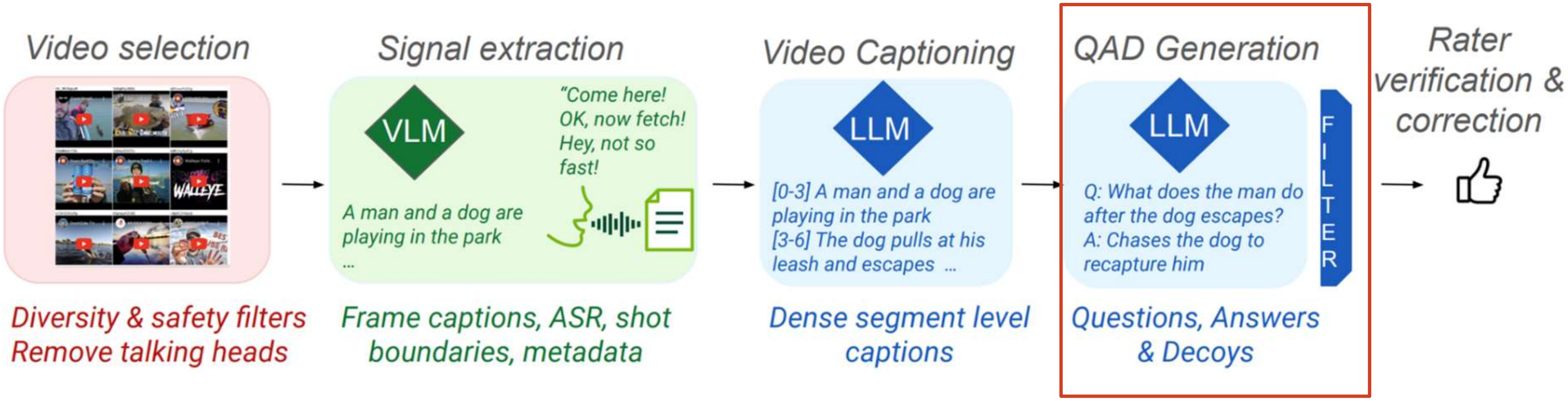
- Combine frame level captions into dense segment level captions automatically using Gemini



- This stage allows us to apply the pipeline to ANY video on YouTube (EgoSchema relies on manually generated captions)

QAD generation

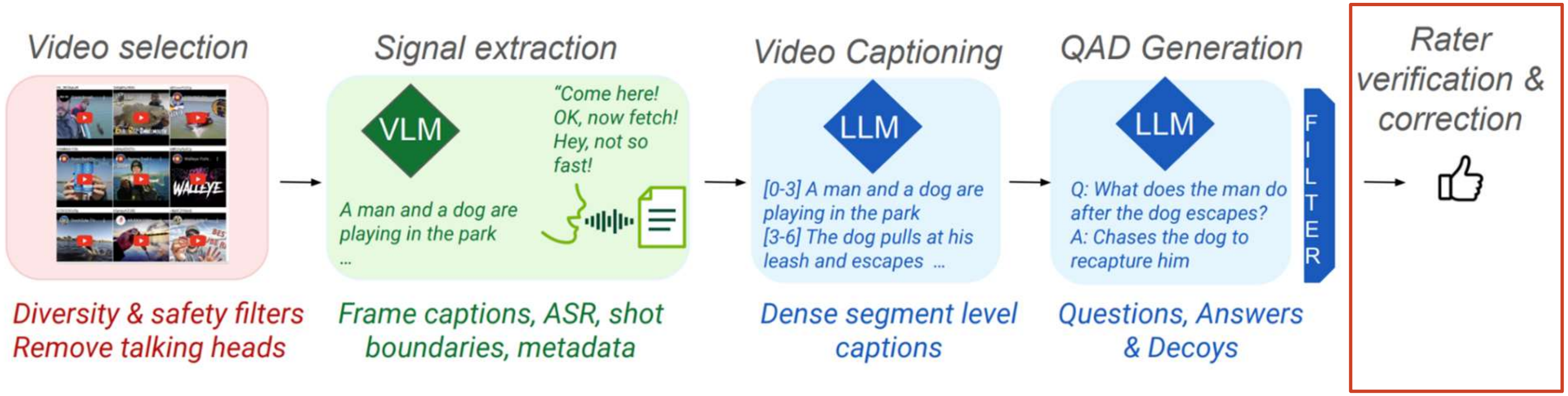
- Generate QADs in two stages:
  - (i) Given video captions from the previous step, first generate questions and answers;
  - (ii) generate six decoys given the questions and answers from the previous stage.



- Done using careful prompting of Gemini with in-context examples

Manual rater verification

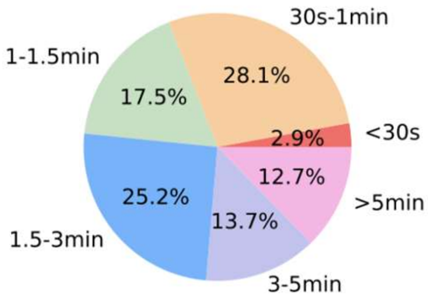
- Two rounds of manual rater verification to ensure quality
- Multiple raters per question (replication)
- Raters were trained with many feedback rounds



Neptune dataset - Statistics

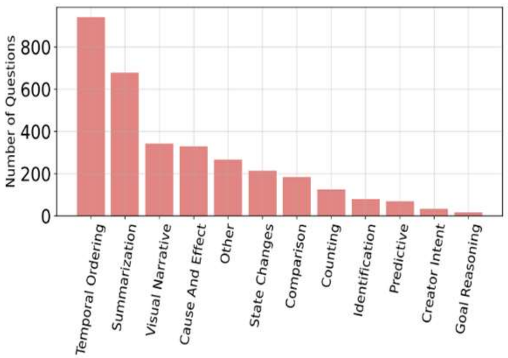
>3,200 questions **>2,400** videos **~100** hours of video.

Videos from **16** seconds to **15** minutes.

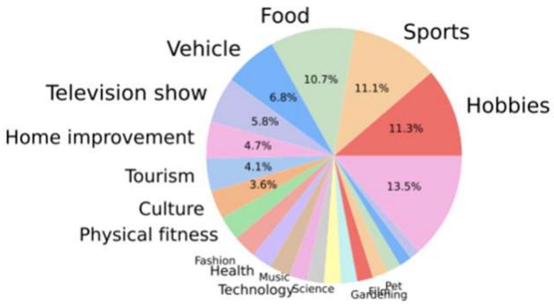


Video lengths

Different question types



Multiple Domains from YouTube



Domains

https://github.com/google-deepmind/neptune

Neptune dataset – Examples

Summarization

What are the key ingredients used in Vonn's recipe for smoked collard greens without meat?

Liquid aminos, smoked paprika, green peppers, garlic, and red peppers

1. Liquid aminos, smoked paprika, onions, garlic, and red peppers.
2. Liquid aminos, smoked paprika, green peppers, garlic, and yellow peppers.
3. Liquid aminos, smoked paprika, green peppers, garlic, and tomatoes.
4. Liquid aminos, smoked paprika, green peppers, garlic, and mushrooms.

Cause and Effect

Ottawa Fury FC's victory was directly caused by Valfoul's successful penalty kick in the 91st minute.

1. Ottawa Fury FC's victory was directly caused by their superior skill and tactics.
2. Ottawa Fury FC's victory was directly caused by Tampa Bay Rowdies' poor performance.
3. Ottawa Fury FC's victory was directly caused by the referee's decision to award a penalty kick.
4. Ottawa Fury FC's victory was directly caused by the crowd's support.

Temporal Ordering

In what order do the following appear in the video?
(a) shot of customer service desk
(b) aerial view of the dealership
(c) interview with man and woman
(d) interview with woman only

(b) aerial view of the dealership
(d) interview with woman only
(a) shot of customer service desk
(c) interview with man and woman

(different orderings of the correct answer)

https://github.com/google-deepmind/neptune

Evaluation Metrics

Neptune supports two evaluation protocols

- **MCQ (5-way multiple choice questions)**
  - Accuracy as the metric
- **Open-Ended**
  - Answers are long (unlike existing datasets that often have one-word or closed set answers)
  - Accuracy is not sufficient!
  - Captioning metrics are either **rule-based** (eg. CIDEr or ROUGE-L) or **LLM-based** (Using ChatGPT or Gemini)
  - Introduction of GEM, an LLM-based open-source model, trained on an answer equivalence dataset, evaluated on a dev set

Evaluation of open-ended metrics on the GEM answer equivalence dev set

Traditional metrics are far from Gemini-1.5-pro

Gemma-9B fine-tuned on BEM gets close!

GEM (Gemma Equivalence Metric)

Metric	Fine-tuning data	F1-Score
CIDEr	None	56.4
ROUGE-L	None	62.2
BEM BERT model	BEM	61.5
Gemma-2B-IT	None	56.3
Gemma-7B-IT	None	65.2
Gemma-9B-IT	None	70.3
Gemma-9B-IT (GEM)	**BEM**	**<u>71.2</u>**
Gemini-1.5-pro	None	**72.8**

Benchmarking & Insights

Method	Modalities	MCQ Acc.	GEM
Random	-	20.00	
BLIP2	RGB (center frame)	28.10	8.50
Video-LLaVA	RGB (8 frames)	24.00	5.48
VideoLlaMA2	RGB (8 frames)	39.89	11.11
VLM captions + LLM (JCEF)	VLM captions (8 VLM captions)	56.45	11.50
MA-LMM	RGB (120 frames)	19.51	5.04
MiniGPT4-Video	RGB (45 frames)	22.89	6.19
MovieChat	RGB (150 frames)	30.30	1.01
Gemini-1.5-pro	QAD only	41.84	11.50
Gemini-1.5-pro	QAD+ASR only	65.76	41.59
Gemini-1.5-pro	RGB (all frames + ASR)	**75.32**	**43.36**

Single frame

Open source short-context MLLMs

Open source long-context MLLMs

Closed source long-context MLLMs

Open-source short context models actually do better than long-context ones!

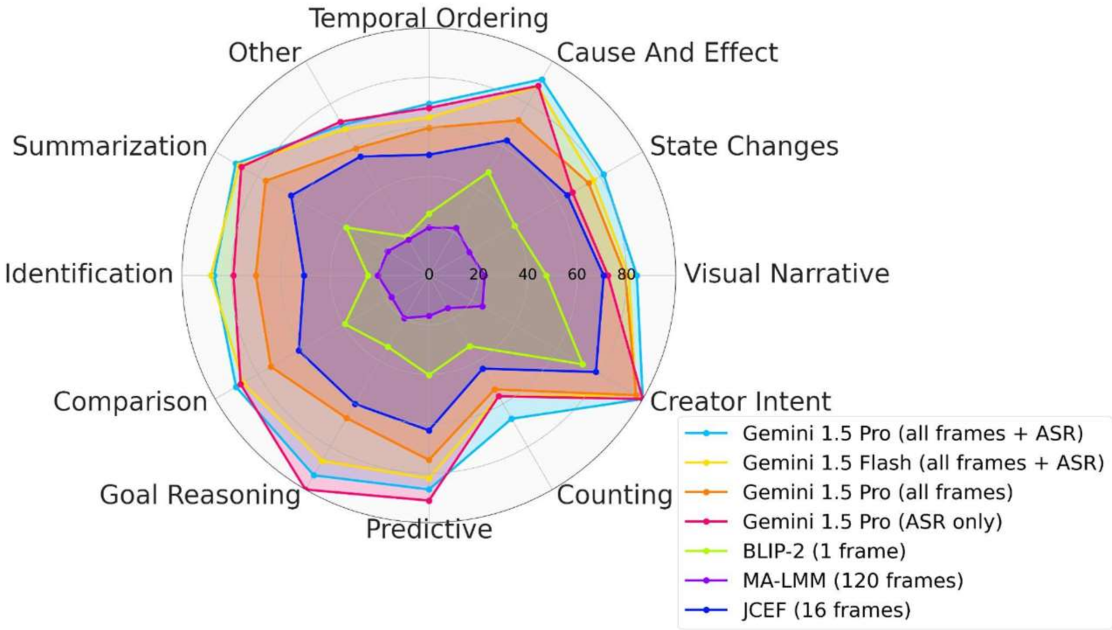
Big gap between open-source models and Gemini-1.5-pro

ASR and RGB are complementary

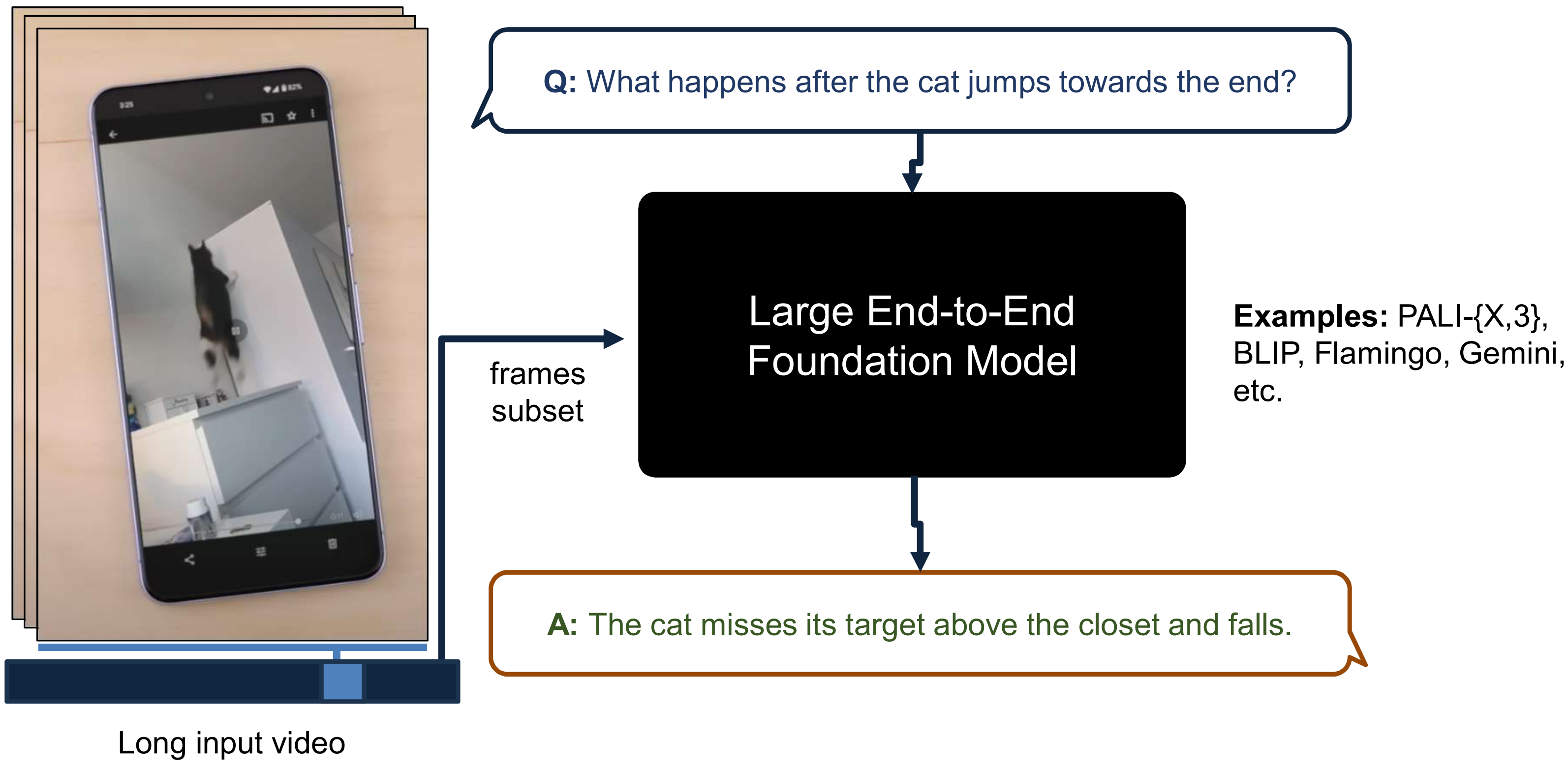
Results by Question Type

“**Counting**”, “**Temporal Ordering**” and “**State Change**” questions are the most challenging for all models

= **areas of focus for future video models**



Visual reasoning - Motivation



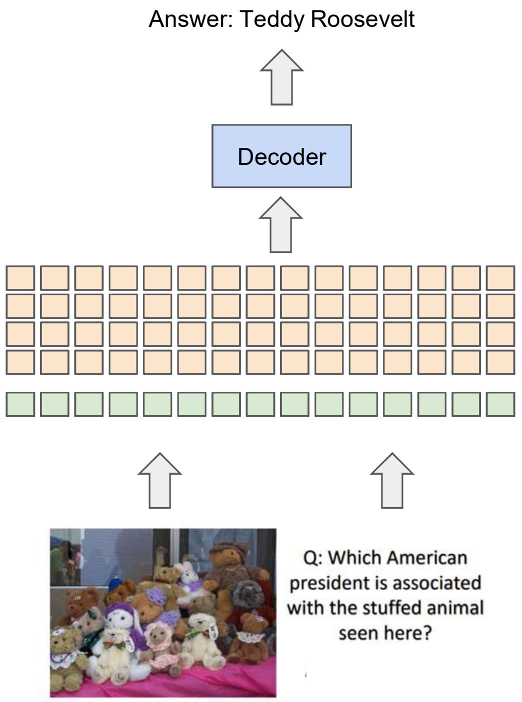
End-to-end trainable models are not interpretable, don't reason and can not use additional information

Visual reasoning

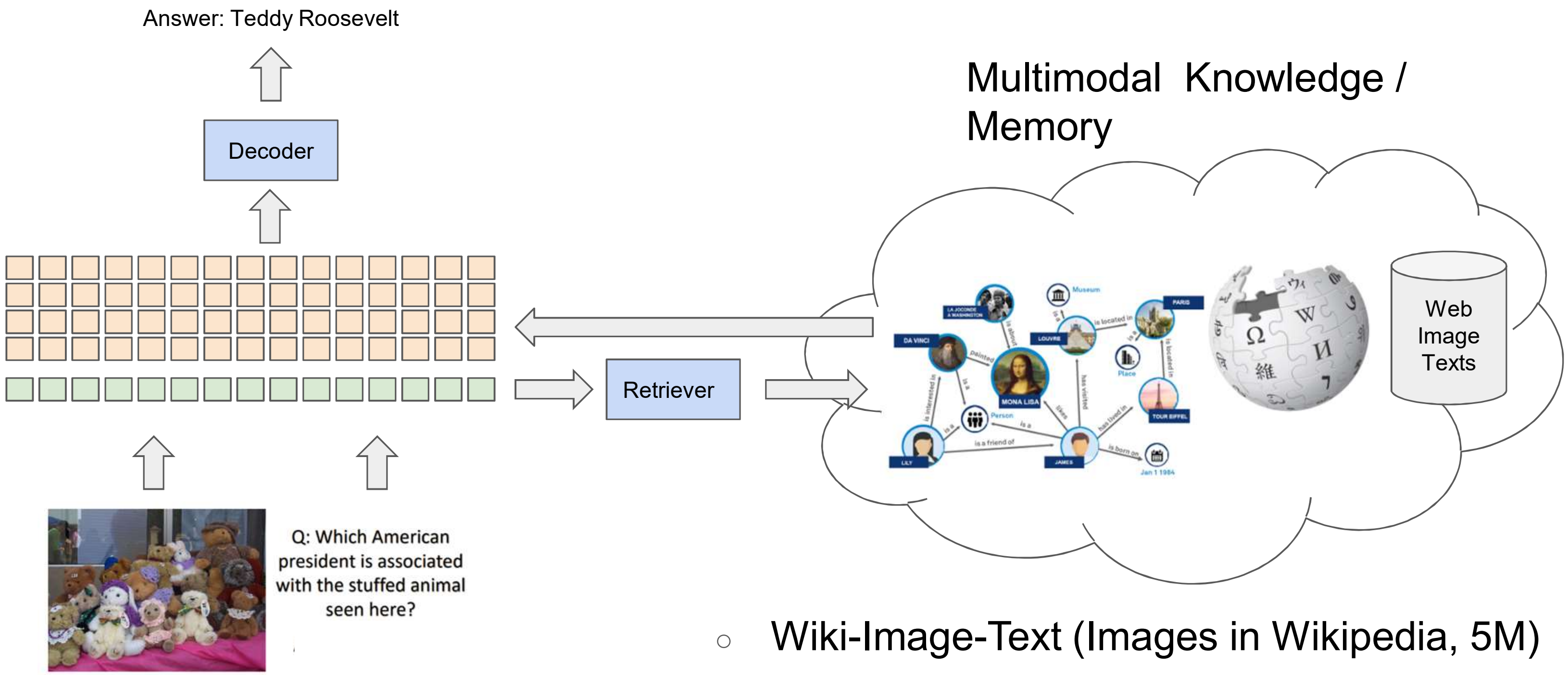
Different type of approaches

- Use of external memory (RAG - Retrieval-Augmented Generation)
  - Augment transformers with retrieved information
- Visual program generation with call of tools
  - Plan then execute paradigm
- Chain of reasoning with external tools
  - LLM-powered Agent (e.g., WebGPT, ReAct, etc.)

LLM with outside knowledge



LLM with outside knowledge



- Wiki-Image-Text (Images in Wikipedia, 5M)
- Wikidata (Knowledge Graph for Wikipedia entities, 12B triplets)

Why memory / knowledge?

- More accurate models: LLM are dedicated to high-level reasoning and memory to fine-grained and rare classes

- Disentangling knowledge from reasoning, use existing knowledge

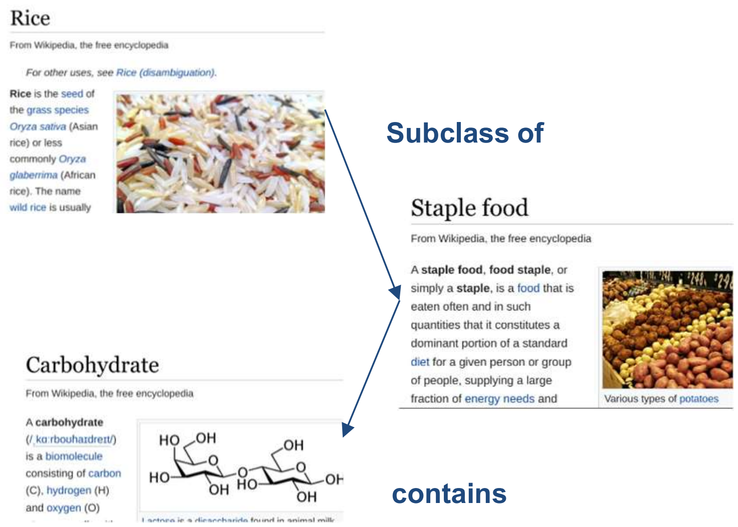
- Retrieved memory / knowledge can be used to interpret model decisions

- Incremental learning w/o catastrophic forgetting: memory update without requiring to update the model

Why memory/ knowledge for VQA?

Answering the question requires additional information

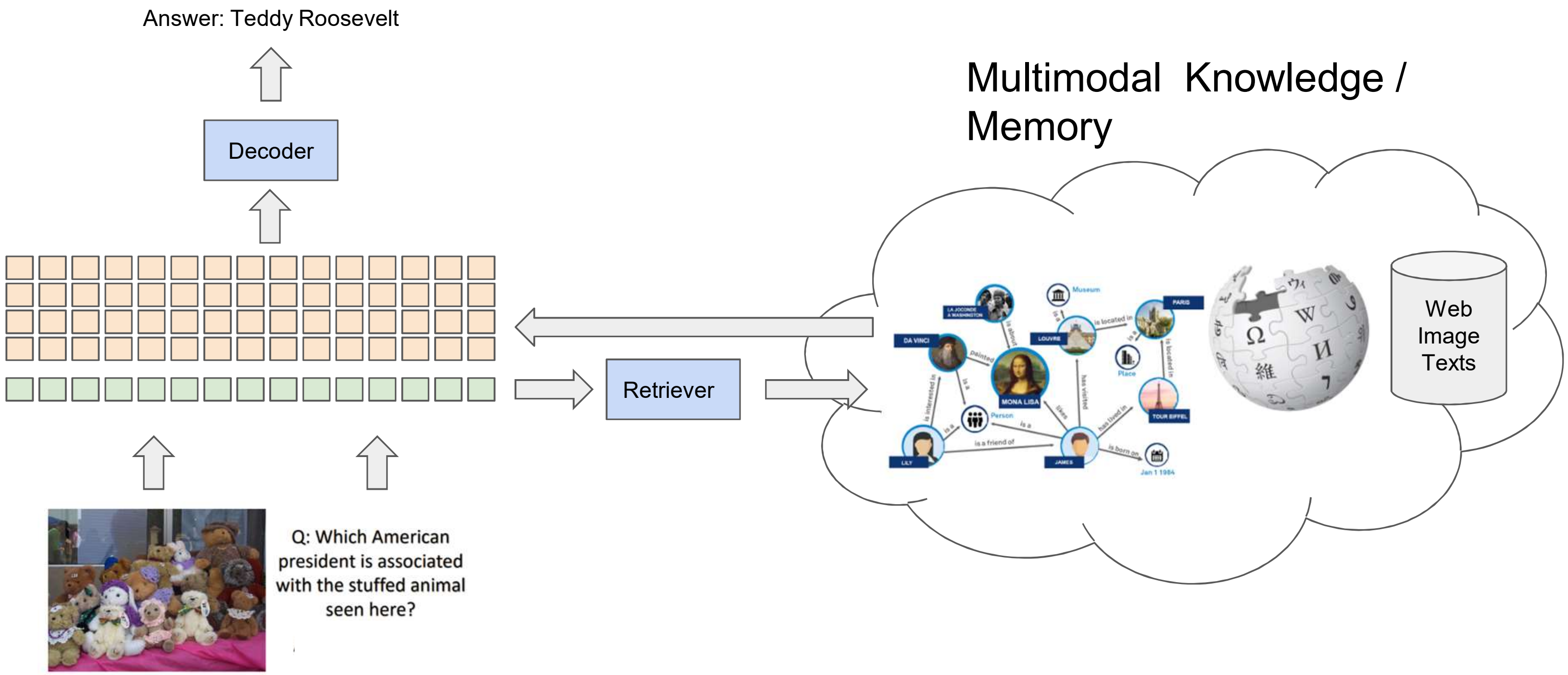
Question : Which part of this meal has the most carbohydrates?



Answer: rice

Example from OK-VQA

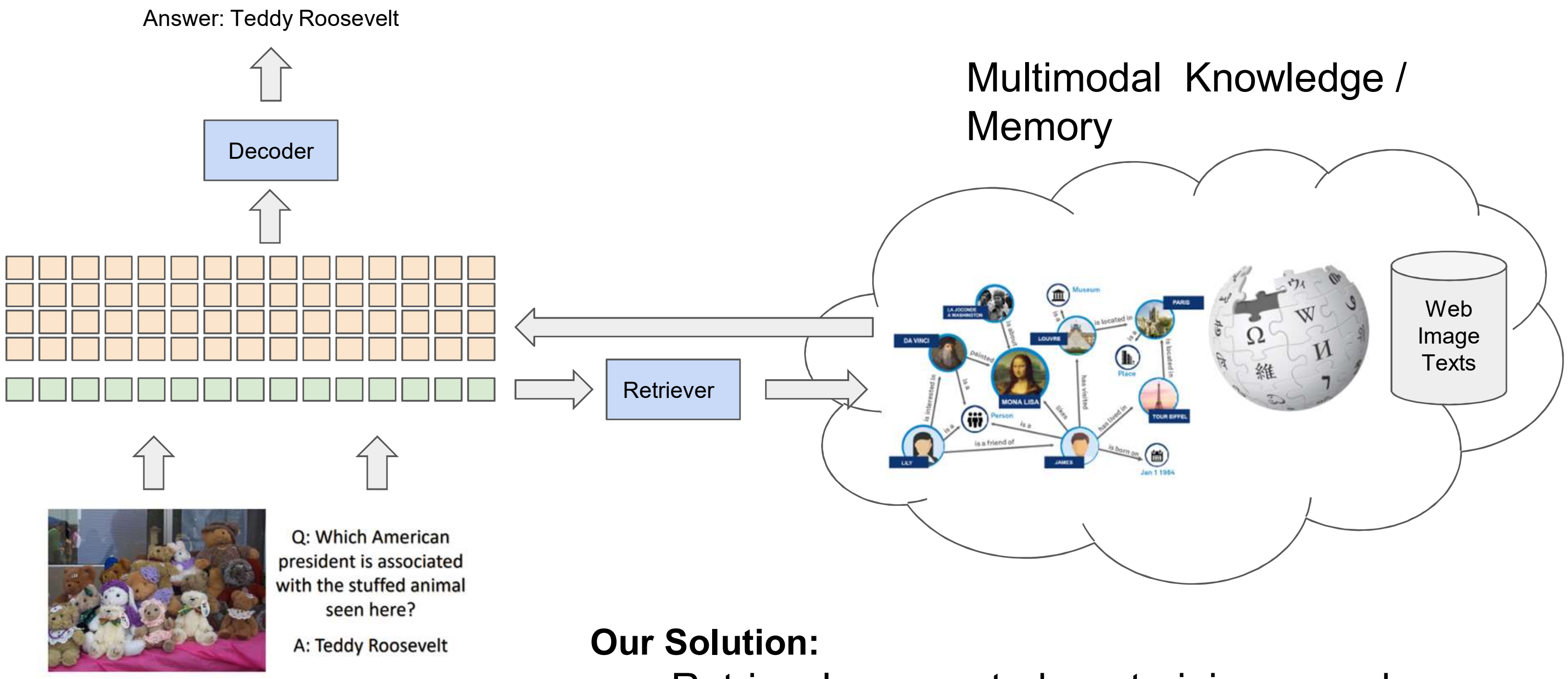
VLM with outside knowledge



Key Challenge:

- No direct supervision for retrieving relevant entries from knowledge base
- QA pairs are insufficient to train large model
  - OK-VQA (14055 pairs covering mostly factoid questions)
  - A-OK-VQA (24903 pairs covering world knowledge)

VLM with outside knowledge



Our Solution:

- Retrieval-augmented pre-training on web-scale image-caption datasets
  - Web Image Text (3B), Webli (10B)
- To generate captions, models are guided to retrieve relevant knowledge via end2end pre-training

Retrieval-augmented vision language model

Pretraining with image captioning

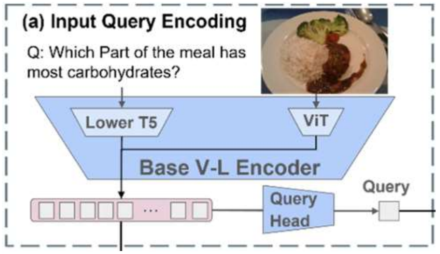
Multi-Source Multimodal Knowledge Memory

Knowledge Source	Corpus Size	Type of Text	Avg. Text Length
WIT [36]	5,233,186	Wikipedia Passage	258
CC12M [5]	10,009,901	Alt-Text Caption	37
VQA-V2 [11]	123,287	Question Answer	111
WikiData [39]	4,947,397	Linearlized Triplets	326

[REVEAL, Z. Hu et al, CVPR 2023]

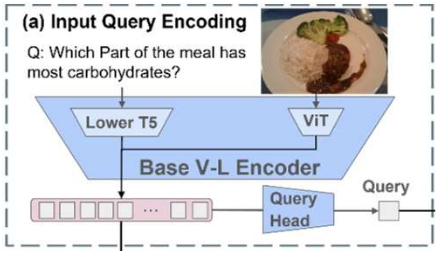
Retrieval-augmented vision language model

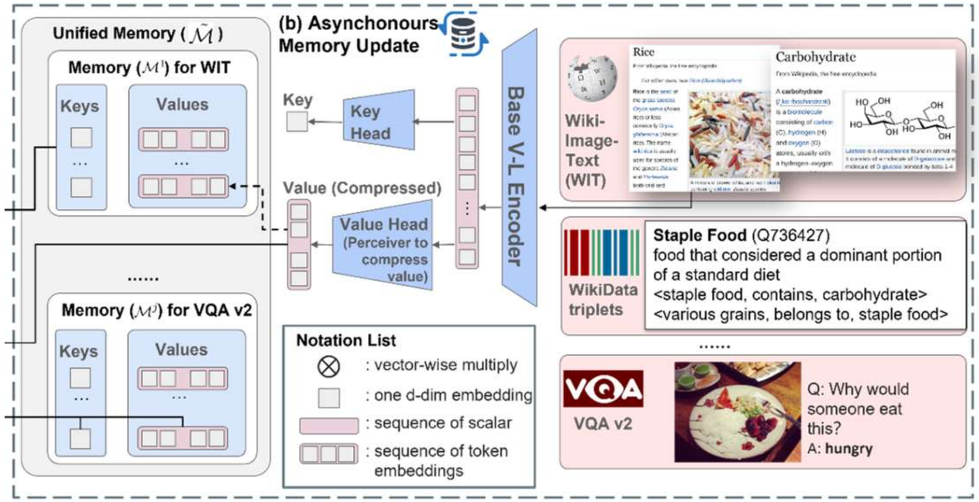
Model - Encoder



Retrieval-augmented vision language model

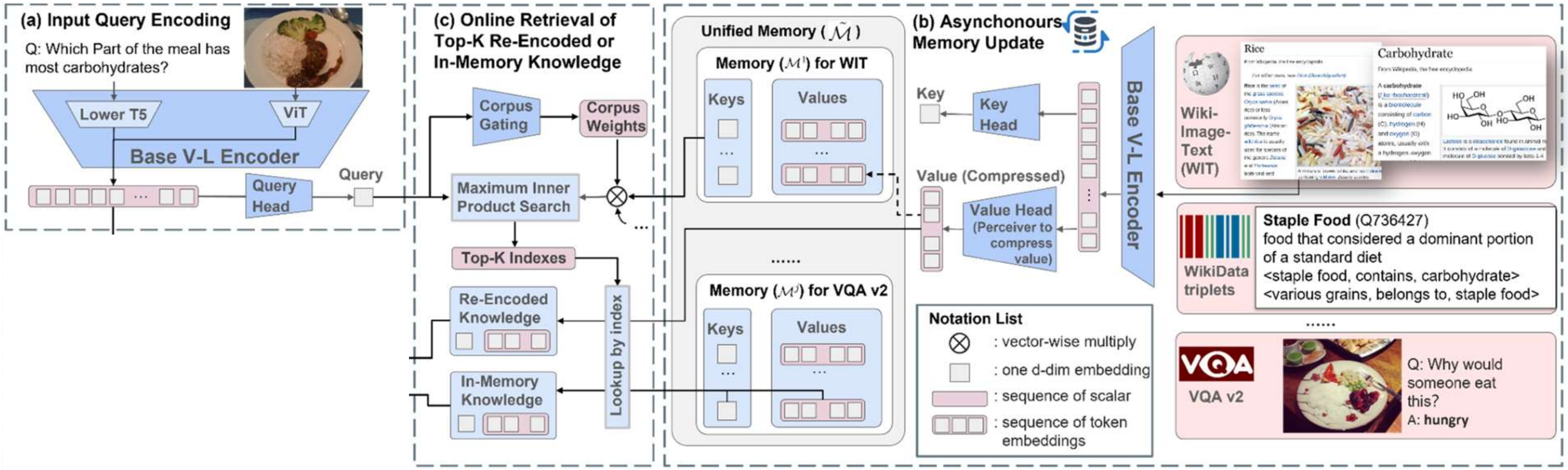
Model - Memory



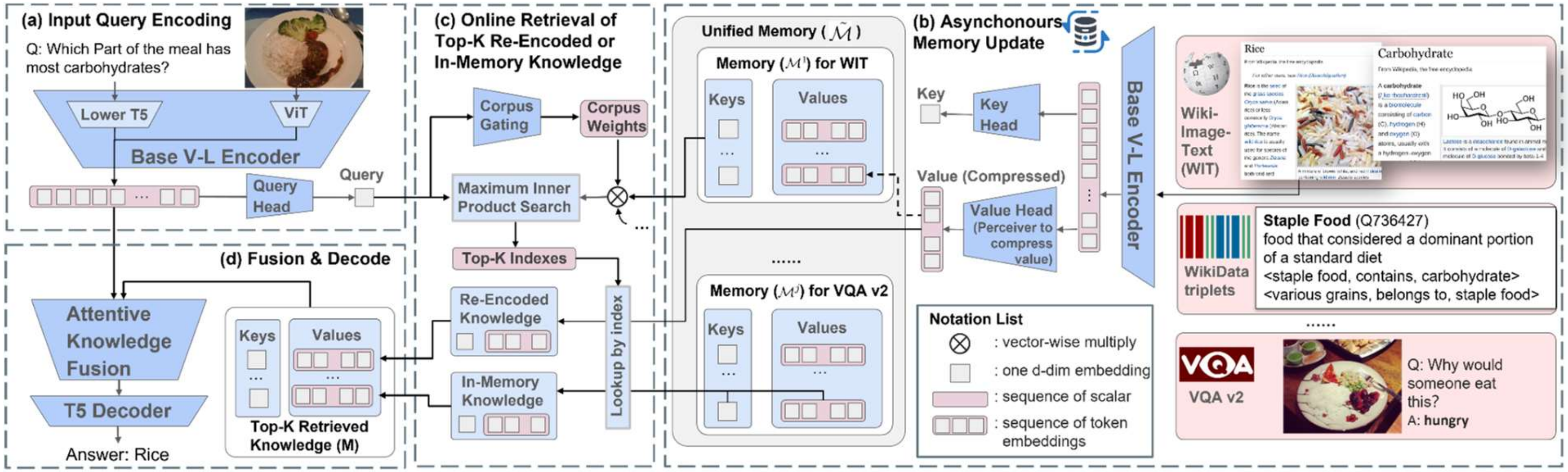


Retrieval-augmented vision language model

Model - Retriever



Retrieval-augmented vision language model



Model - Generator

Results on OK-VQA

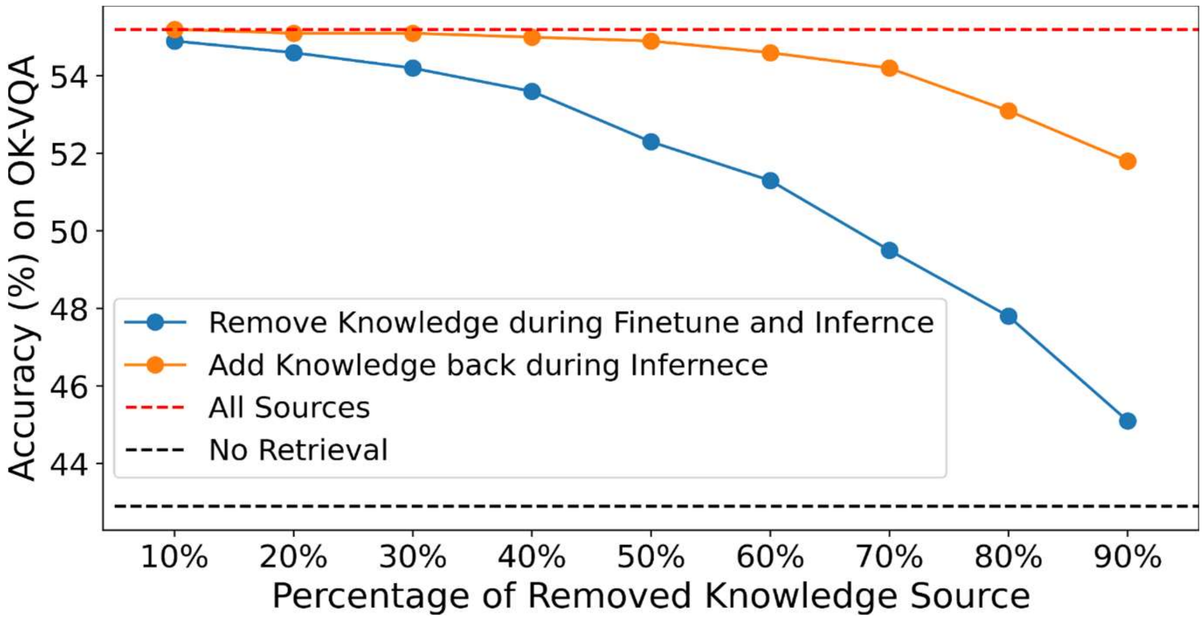
VQA Model Name	Knowledge Sources	Accuracy (%)	# params.
MUTAN+AN	Wikipedia + ConceptNet	27.8	-
ConceptBERT	Wikipedia	33.7	-
KRISP [27]	Wikipedia + ConceptNet	38.4	-
Visual Retriever-Reader	Google Search	39.2	-
MAVEx	Wikipedia+ConceptNet+Google Images	39.4	-
KAT-Explicit [12]	Wikidata	44.3	0.77B
PICa-Base [47]	Frozen GPT-3	43.3	(175B frozen)
PICa-Full [47]	Frozen GPT-3	48.0	(175B frozen)
KAT [12] (Single)	Wikidata + Frozen GPT-3	53.1	0.77B + (176B frozen)
KAT [12] (Ensemble)	Wikidata + Frozen GPT-3	54.4	2.31B + (176B frozen)
ReVIVE [23] (Single)	Wikidata + Frozen GPT-3	56.6	0.77B + (176.9B frozen)
ReVIVE [23] (Ensemble)	Wikidata+Frozen GPT-3	58.0	2.31B + (176.9B frozen)
REVEAL-Base	WIT + CC12M + Wikidata + VQA-2	55.2	0.4B
REVEAL-Large	WIT + CC12M + Wikidata + VQA-2	58.0	1.4B
REVEAL	WIT + CC12M + Wikidata + VQA-2	**59.1**	2.1B

Model Name	T5 Variant	Image Encoder	# params.	GFLOPs
REVEAL-Base	T5-Base	ViT-B/16	0.4B	120
REVEAL-Large	T5-Large	ViT-L/16	1.4B	528
REVEAL	T5-Large	ViT-g/14	2.1B	795

Table 2. Model configuration of different REVEAL variants.

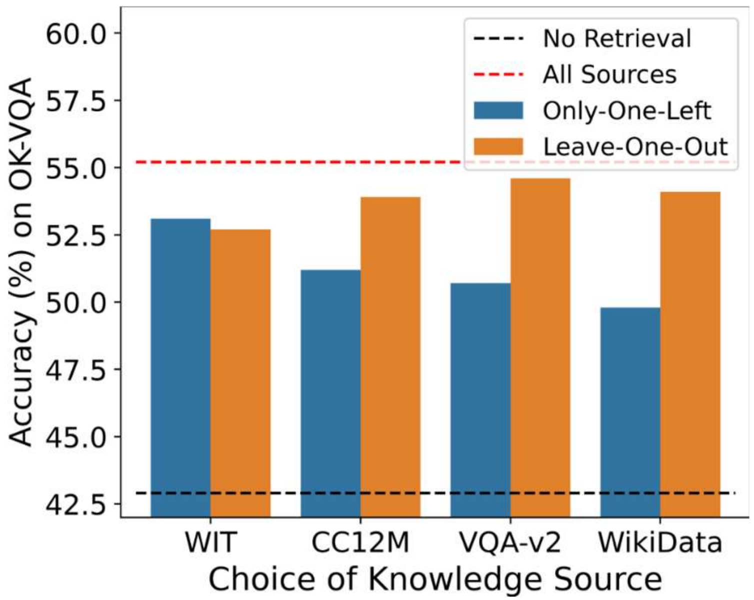
Example results

How useful is knowledge memory?



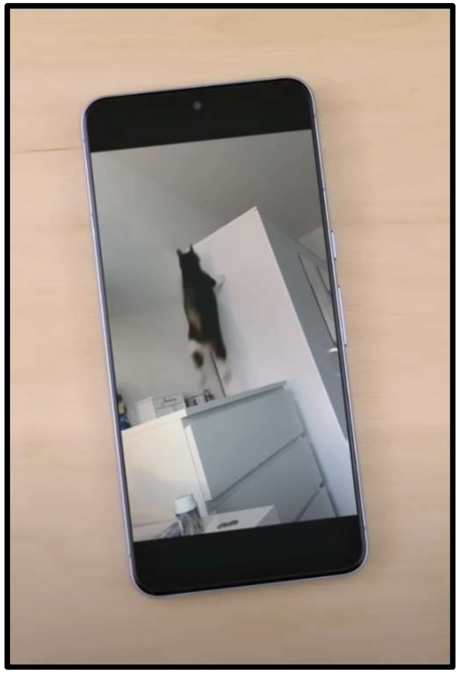
Blue curve: x% removed during fine-tuning and inference
Orange curve: x% removed during fine-tuning, but added during inference;
this simulates on-the-fly knowledge update

Contribution from each knowledge source



Only-One-Left: only use of a single knowledge source
Leave-One-Out: use all without this knowledge source

Video question answering: program generation



Long input video

Q: What happens after the cat jumps towards the end?

Prompt + APIs

Code Generation LLM

Program / Plan

A: The cat misses its target above the closet and falls.

Examples: ViperGPT, VisProg, CodeVQA, etc.

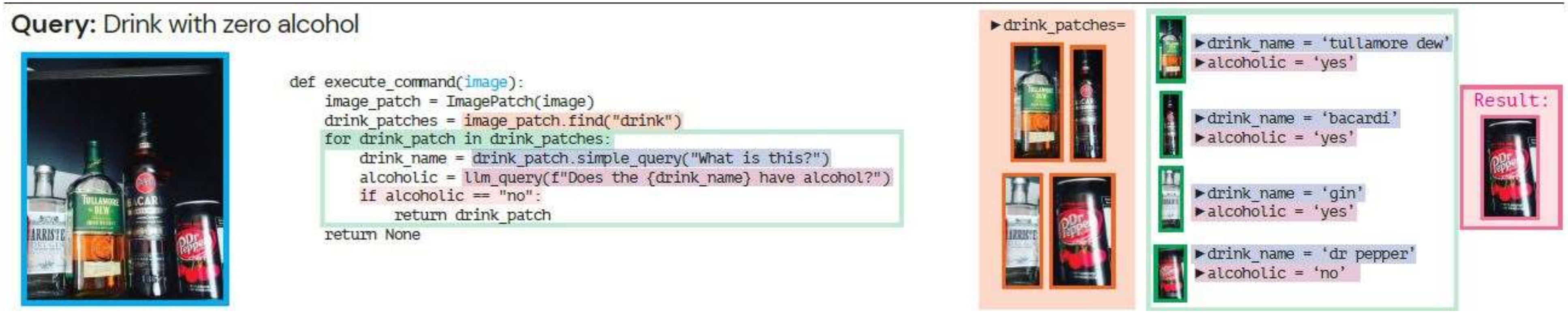
Modular design: uses vision/language base models as tools according to a generated program/plan

Models offer interpretability (+ show promise over E2E approaches in accuracy!)

Visual program generation: ViperGPT [Suris,Menon,Vonderick, ICCV'23]

Task: image question answering

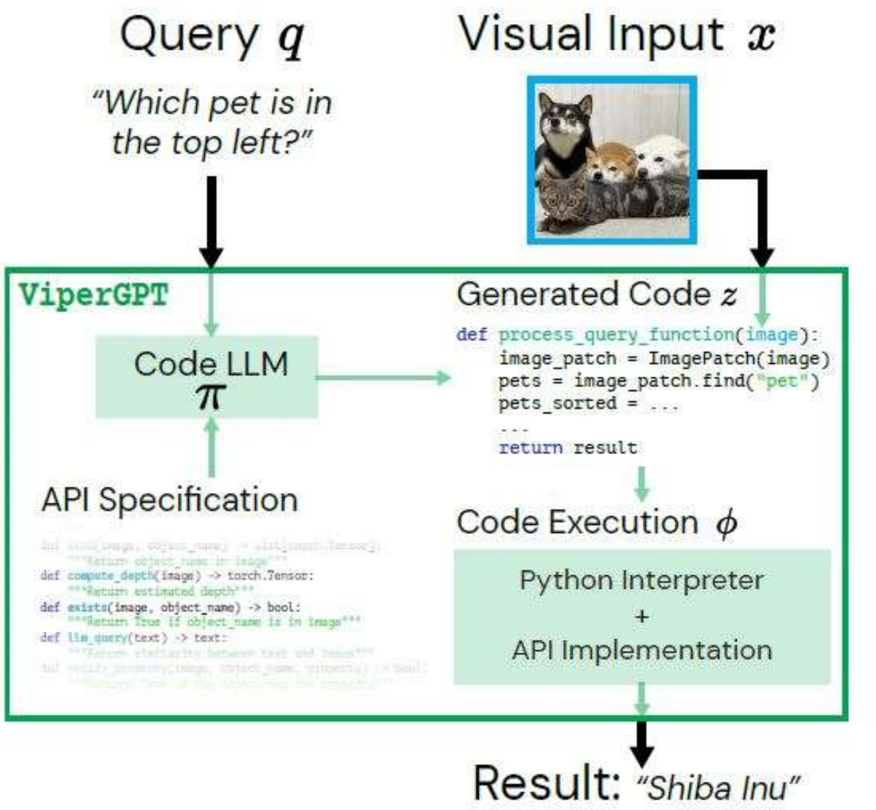
Visual program generation: ViperGPT [Suris,Menon,Vonderick, ICCV'23]



Task: image question answering

Visual program generation: ViperGPT [Suris,Menon,Vonderick, ICCV'23]

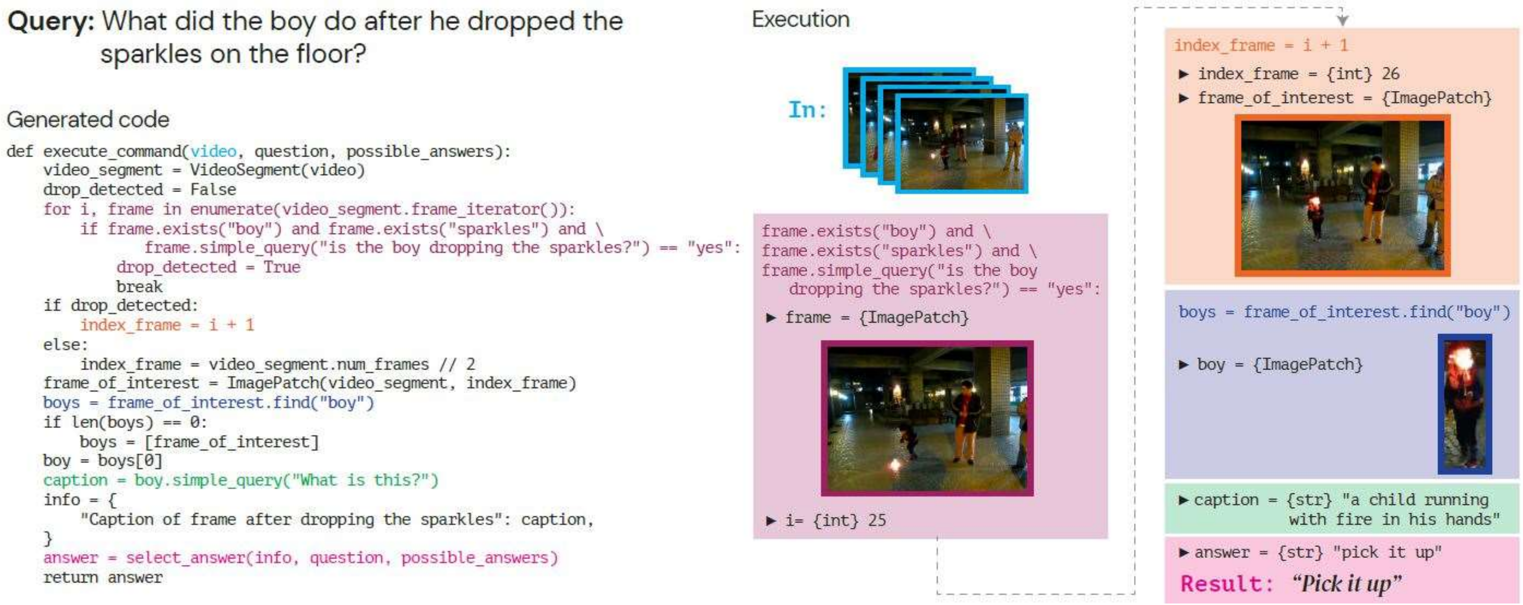
APIs:
- Detection/find
- Verification/presence
- Depth
- Frame selection



Overview Method

Code generation
- Codex (OpenAI) pretrained on internet data
- Prompt tuning with sample programs + API specifications

Visual program generation: ViperGPT [Suris,Menon,Vonderick, ICCV'23]



Task: video question answering on NeXT-QA

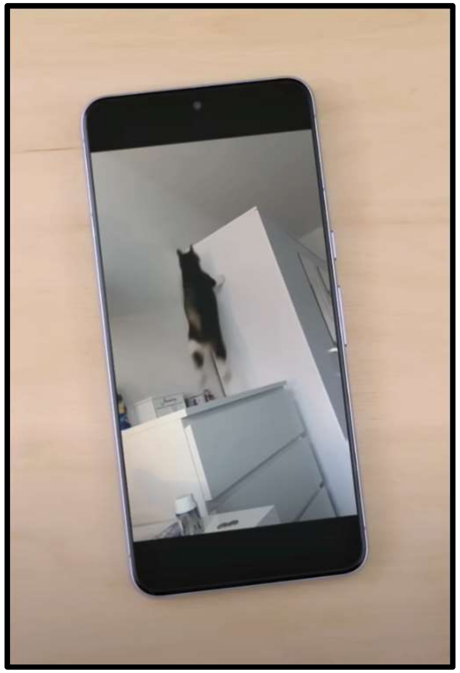
Visual program generation: ViperGPT [Suris,Menon,Vonderick, ICCV'23]

Table 4. **NExT-QA Results**. Our method gets overall state-of-the-art results (including *supervised* models) on the hard split. "T" and "C" stand for "temporal" and "causal" questions, respectively.

		Accuracy (%) ↑		
		Hard Split - T	Hard Split - C	Full Set
Sup.	ATP [7]	45.3	43.3	54.3
	VGT [58]	-	-	56.9
	HiTeA [61]	48.6	47.8	63.1
ZS	ViperGPT (ours)	**49.8**	**56.4**	60.0

Task: video question answering on NeXT-QA

Video question answering: program generation



Long input video

Q: What happens after the cat jumps towards the end?

Prompt + APIs

Code Generation LLM

Program / Plan

A: The cat misses its target above the closet and falls.

Examples: ViperGPT, VisProg, CodeVQA, etc.

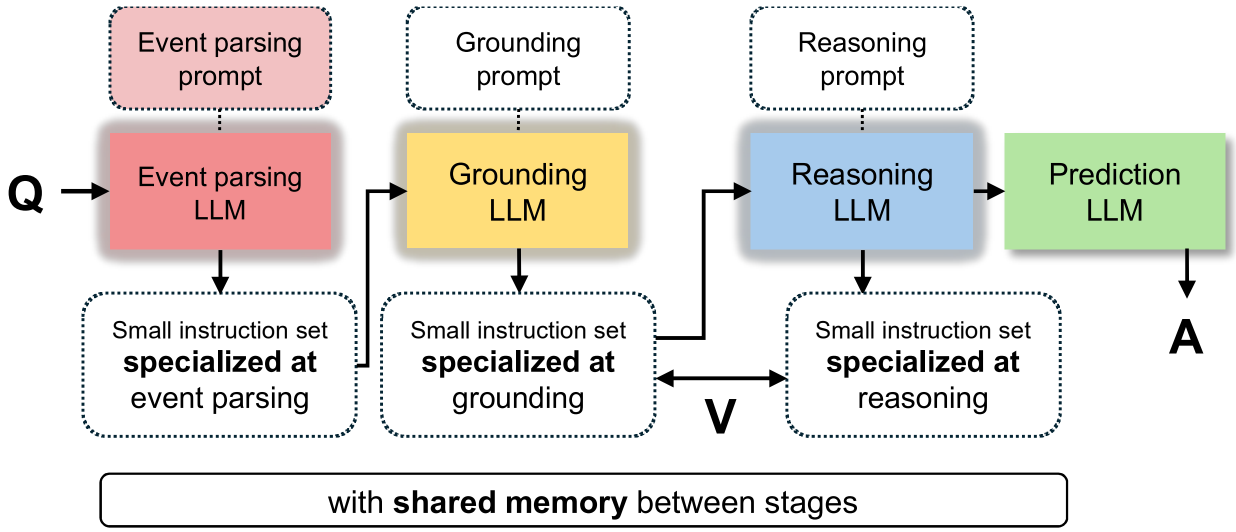
Modular design: uses vision/language base models as tools according to a generated program/plan

Models offer interpretability (+ show promise over E2E approaches in accuracy!)

Disadvantage: Single stage, no use of visual input for program generation

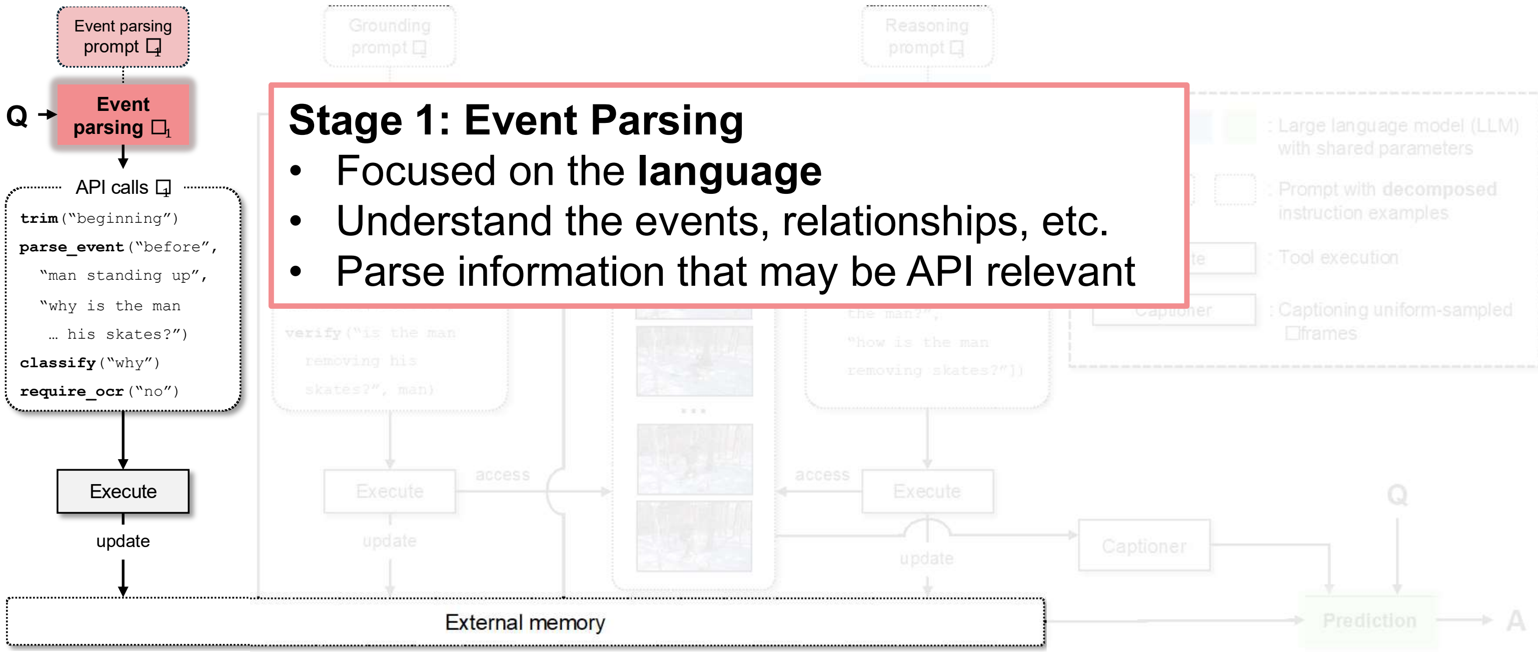
MoReVQA: Multistage, **Mo**dular **Re**asoning

➔ We introduce a new **multistage** <u>mo</u>dular <u>re</u>asoning <u>VQA</u> model (MoReVQA)



[MoReVQA, J. Min et al. CVPR'24]

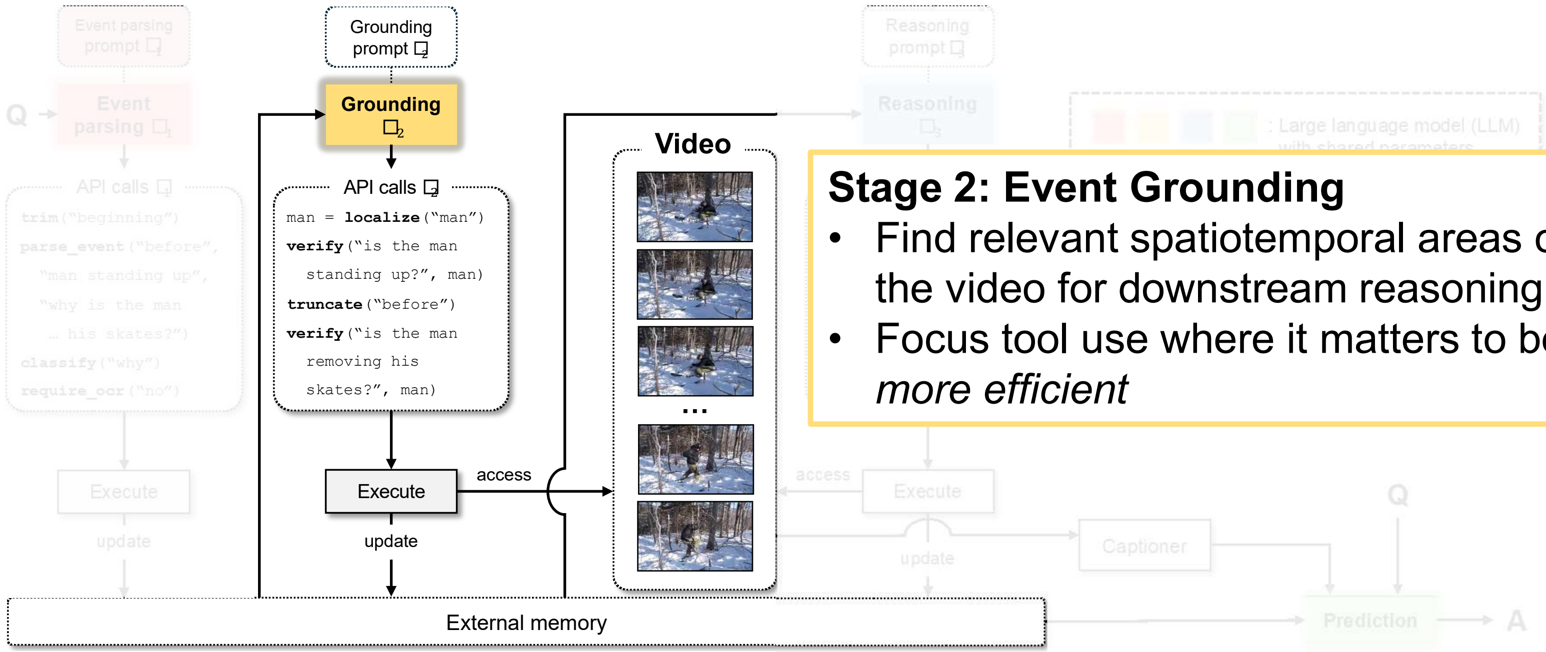
MoReVQA: Multistage, **Mo**dular **Re**asoning



Stage 1: Event Parsing

- Focused on the **language**
- Understand the events, relationships, etc.
- Parse information that may be API relevant

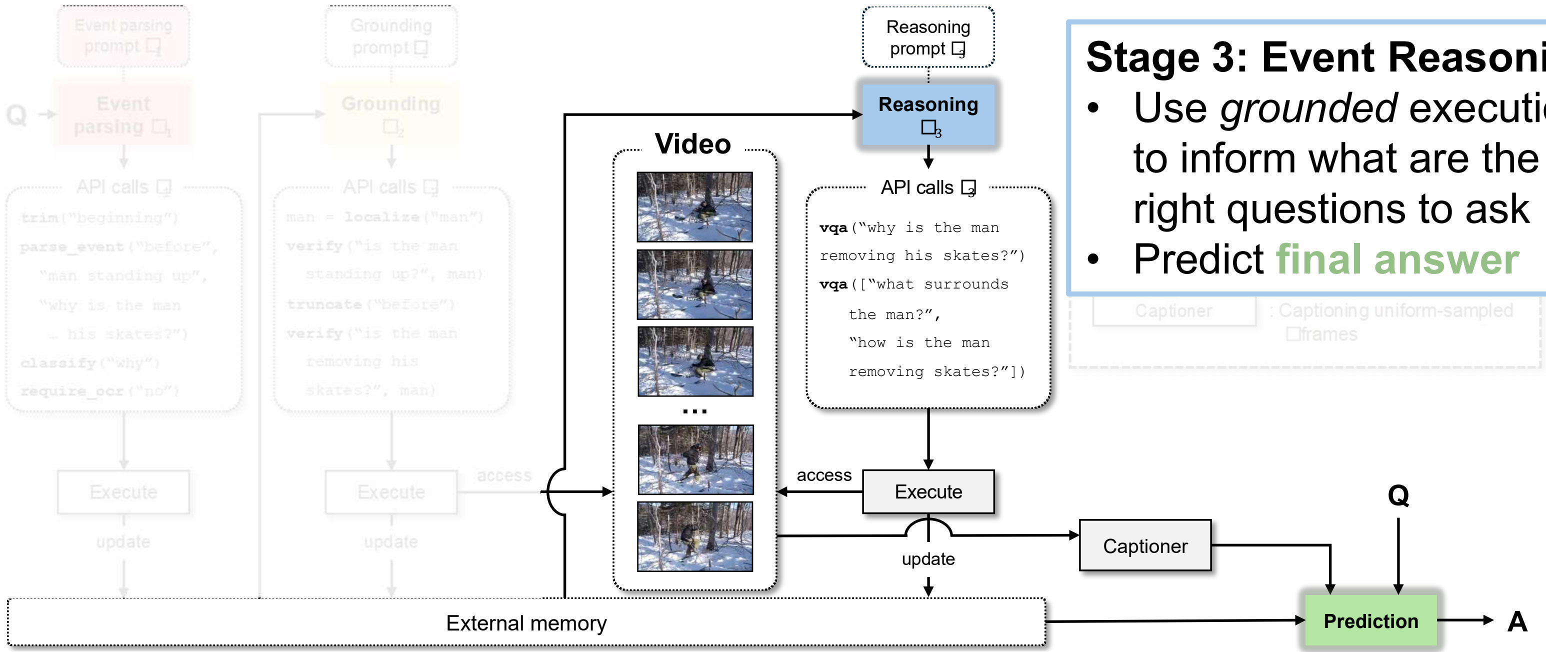
MoReVQA: Multistage, **Mo**dular **Re**asoning



Stage 2: Event Grounding

- Find relevant spatiotemporal areas of the video for downstream reasoning
- Focus tool use where it matters to be *more efficient*

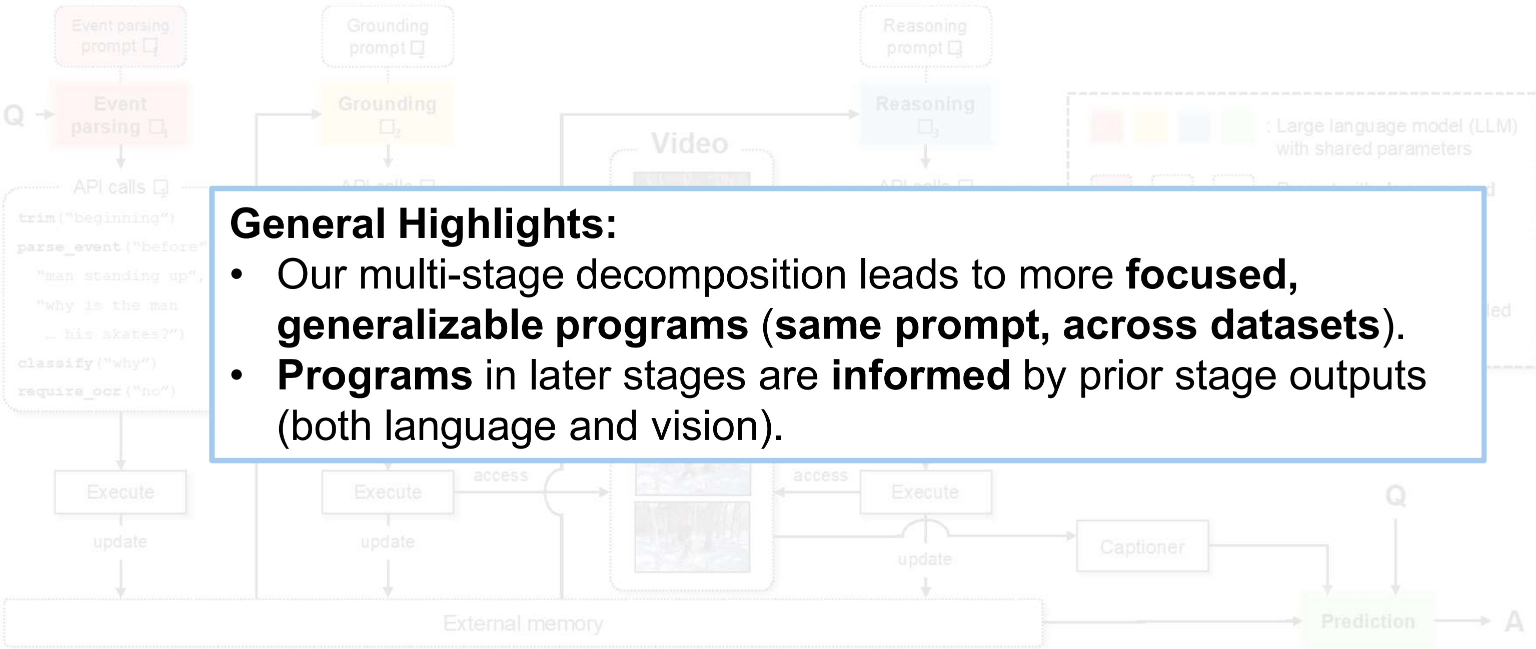
MoReVQA: Multistage, **Mo**dular **Re**asoning



Stage 3: Event Reasoning

- Use *grounded* execution to inform what are the right questions to ask
- Predict **final answer**

MoReVQA: Multistage, **Mo**dular **Re**asoning



General Highlights:

- Our multi-stage decomposition leads to more **focused, generalizable programs** (**same prompt, across datasets**).
- **Programs** in later stages are **informed** by prior stage outputs (both language and vision).

VideoQA Experiments

A diverse collection of video QA benchmarks:

- **NExT-QA:** Temporal/causal relationships
- **iVQA:** Instructional videos
- **EgoSchema:** Egocentric perspective
- **ActivityNet-QA:** General YouTube activities

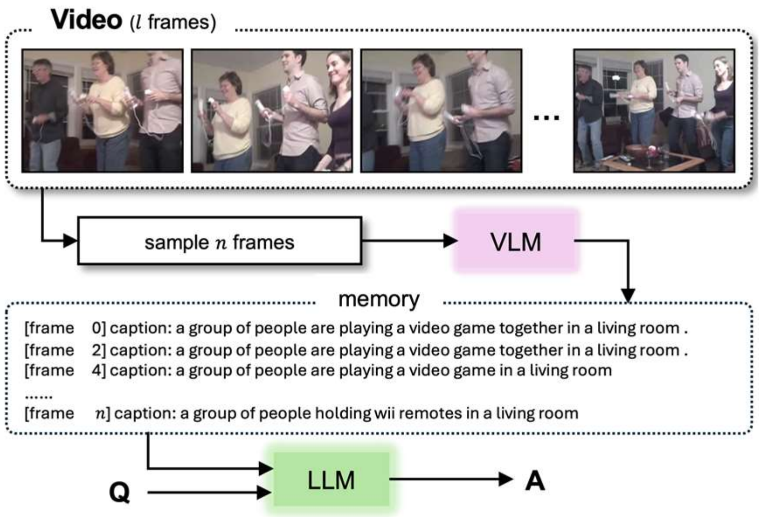
Open-ended

Long-videos

A New Simple Baseline: JCEF

Two of the core API modules (VLM and LLM) with a simple plan

"Just caption every frame" (JCEF)



We find this baseline **surprisingly effective**: *outperforms* ViperGPT!

Results: Overview

Method	Accuracy (%)			
	NExT-QA	iVQA	EgoSchema	ActivityNet-QA
Random	20.0	0.0	20.0	20.0
LLM-only [23]	48.5	15.0	41.0	-
ViperGPT [52]	60.0	-	-	-
ViperGPT+	64.0	46.6	49.3	37.1
JCEF	<u>66.7</u>	<u>56.9</u>	<u>49.9</u>	<u>43.3</u>

Our Just-Caption-Every-Frame (JCEF) baseline is **surprisingly strong**

Results: Overview

Method	Accuracy (%)			
	NExT-QA	iVQA	EgoSchema	ActivityNet-QA
Random	20.0	0.0	20.0	20.0
LLM-only [23]	48.5	15.0	41.0	-
ViperGPT [52]	60.0	-	-	-
ViperGPT+	64.0	46.6	49.3	37.1
JCEF	<u>66.7</u>	<u>56.9</u>	<u>49.9</u>	<u>43.3</u>
MoReVQA	**69.2**	**60.9**	**51.7**	**45.3**
	+5.2	**+14.3**	**+2.4**	**+8.2**

Our JCEF baseline is **surprisingly strong** relative to ViperGPT+

Our MoReVQA model **consistently improves** across all key datasets

Efficiency: 5x fewer "large model calls" with grounding!

Results: Ablation

Stages			NExT-QA	iVQA
Event parsing	Grounding	Reasoning	Val	Test
✗	✗	✗	66.65	56.89
✓	✗	✓	68.29	56.92
✓	✓	✗	68.71	57.53
✓	✓	✓	**69.22**	**60.88**

The MoReVQA stages are **complementary** + **synergistic**!

Results: State-of-the-art Comparisons

NExT-QA

Method	Val	FT
MIST-CLIP [21]	57.2	
HiTeA [68]	63.1	✓
SeViLa [71]	**73.8**	
ViperGPT [52]	60.0	
BLIP-2^{concat} [36]	62.4	
BLIP-2^{voting} [36]	62.7	✗
SeViLA [71]	63.6	
JCEF	66.7	
MoReVQA	**69.2**	

iVQA

Method	Test	FT
VideoCoCa [64]	39.0	
FrozenBiLM [66]	39.7	✓
Text+Text [39]	**40.2**	
FrozenBiLM [66]	27.3	
BLIP-2$_{(FlanT5XXL)}$ [36]	45.8	
InstructBLIP$_{(FlanT5XL)}$ [15]	53.1	✗
InstructBLIP$_{(FlanT5XXL)}$ [15]	53.8	
JCEF	56.9	
MoReVQA	**60.9**	

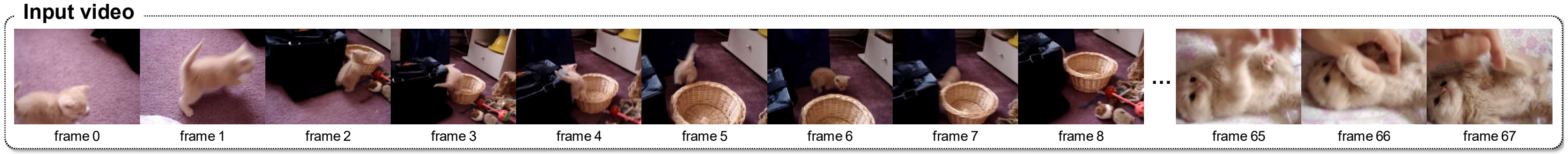
EgoSchema

Method	Test	FT
VIOLET [19]	19.9	
SeViLA [71]	22.7	
FrozenBiLM [66]	26.9	
mPLUG-Owl [69]	31.1	
InternVideo [57]	32.1	✗
*ShortViViT [46]	31.0	
*LongViViT [46]	33.3	
JCEF	50.0	
MoReVQA	**51.7**	

ActivityNet-QA

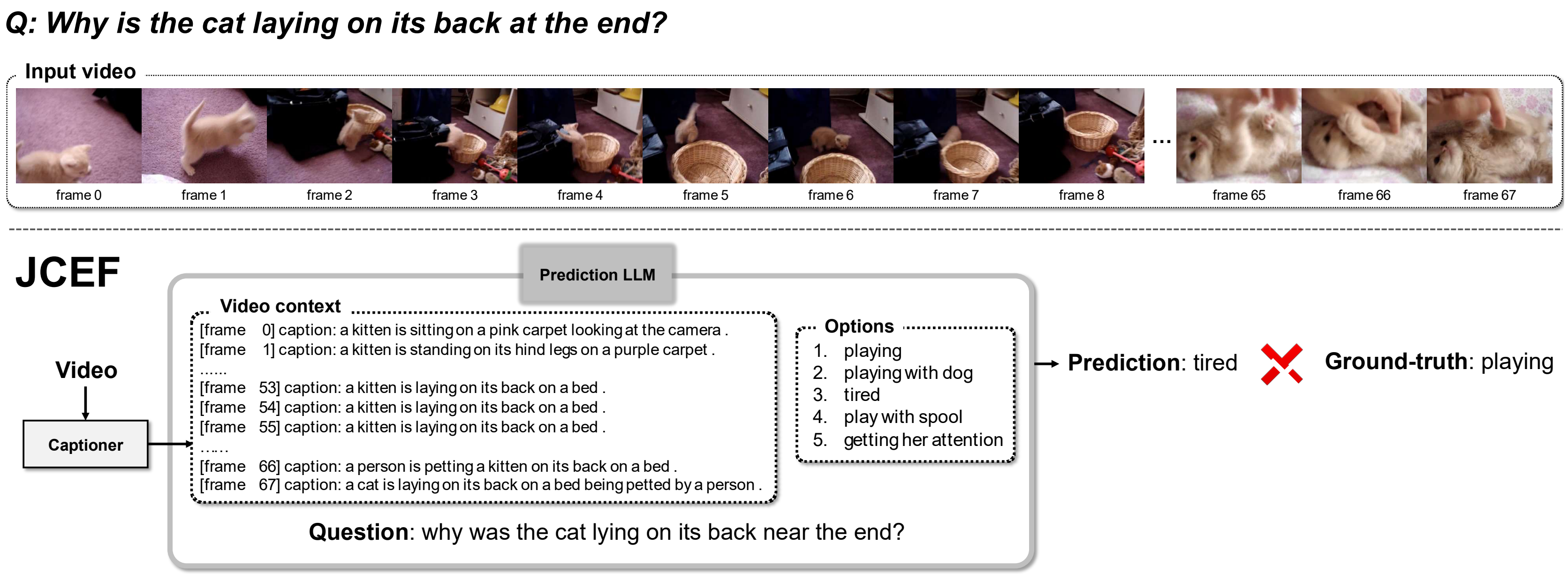
Method	Test	FT
Just Ask [65]	12.2	
VideoChat [37]	26.5	
*LLaMa adapter [77]	34.2	✗
*Video-ChatGPT [40]	35.2	
ViperGPT+	37.1	
JCEF	43.3	
MoReVQA	**45.3**	

Our **training-free** method outperforms prior work
(and even some finetuned + concurrent work!)

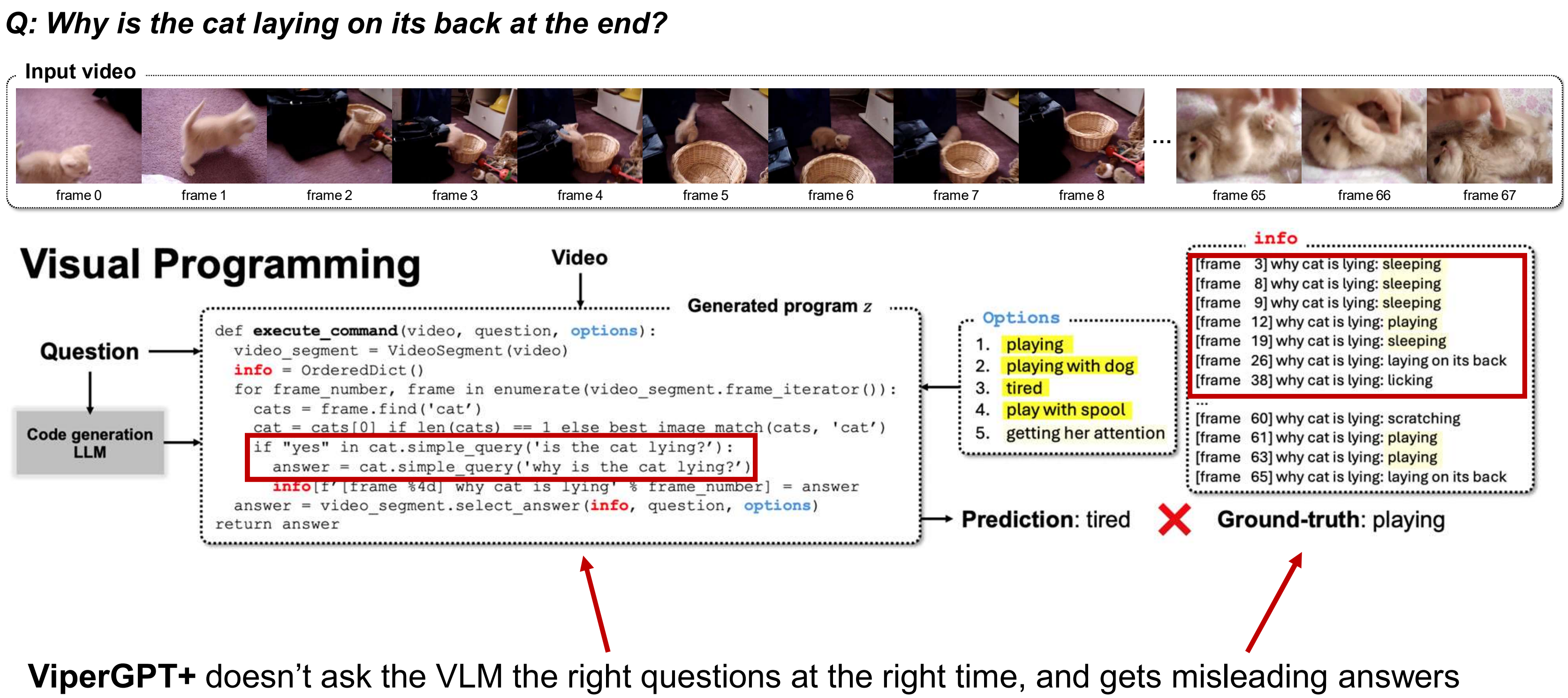


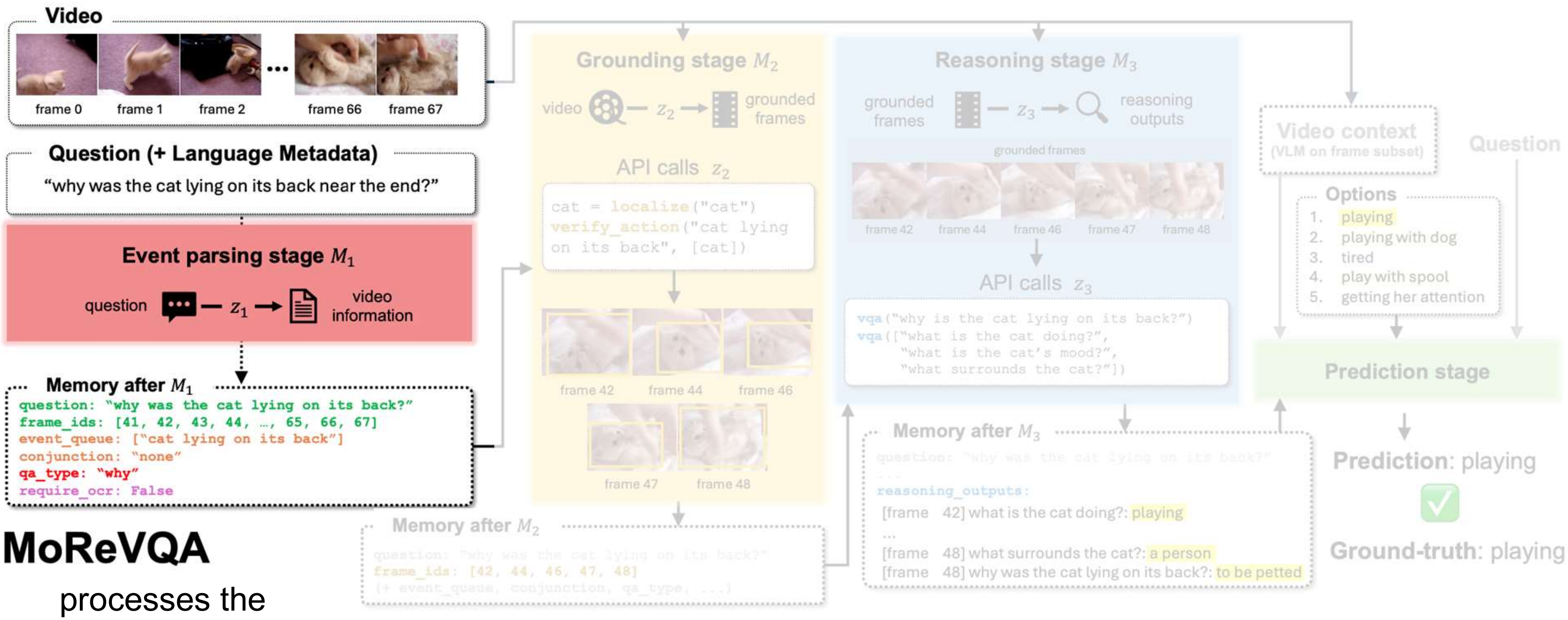
Let's take an input video of a cat:

Q: Why is the cat laying on its back at the end?



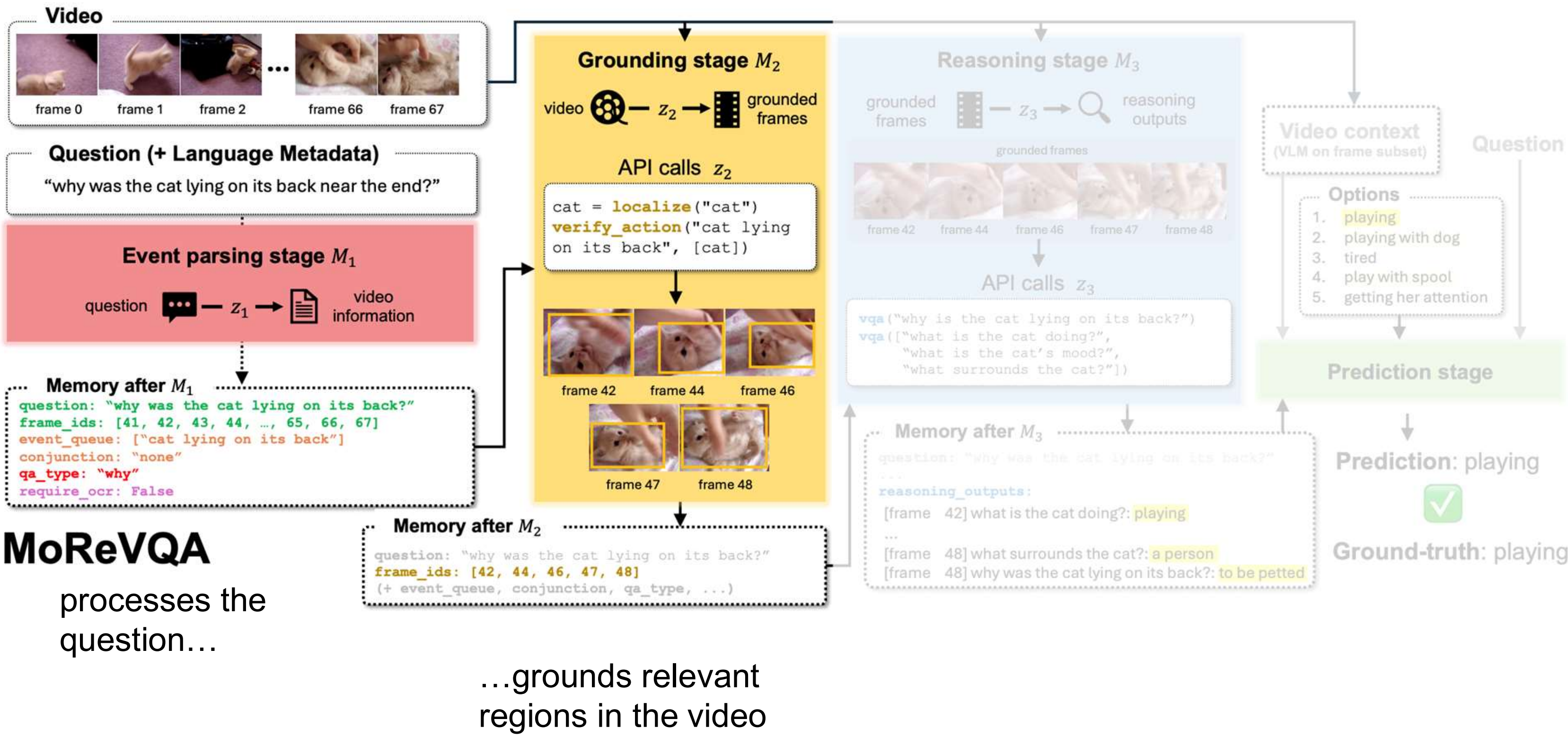
JCEF offers general captions, so misleading captions can impact the final prediction

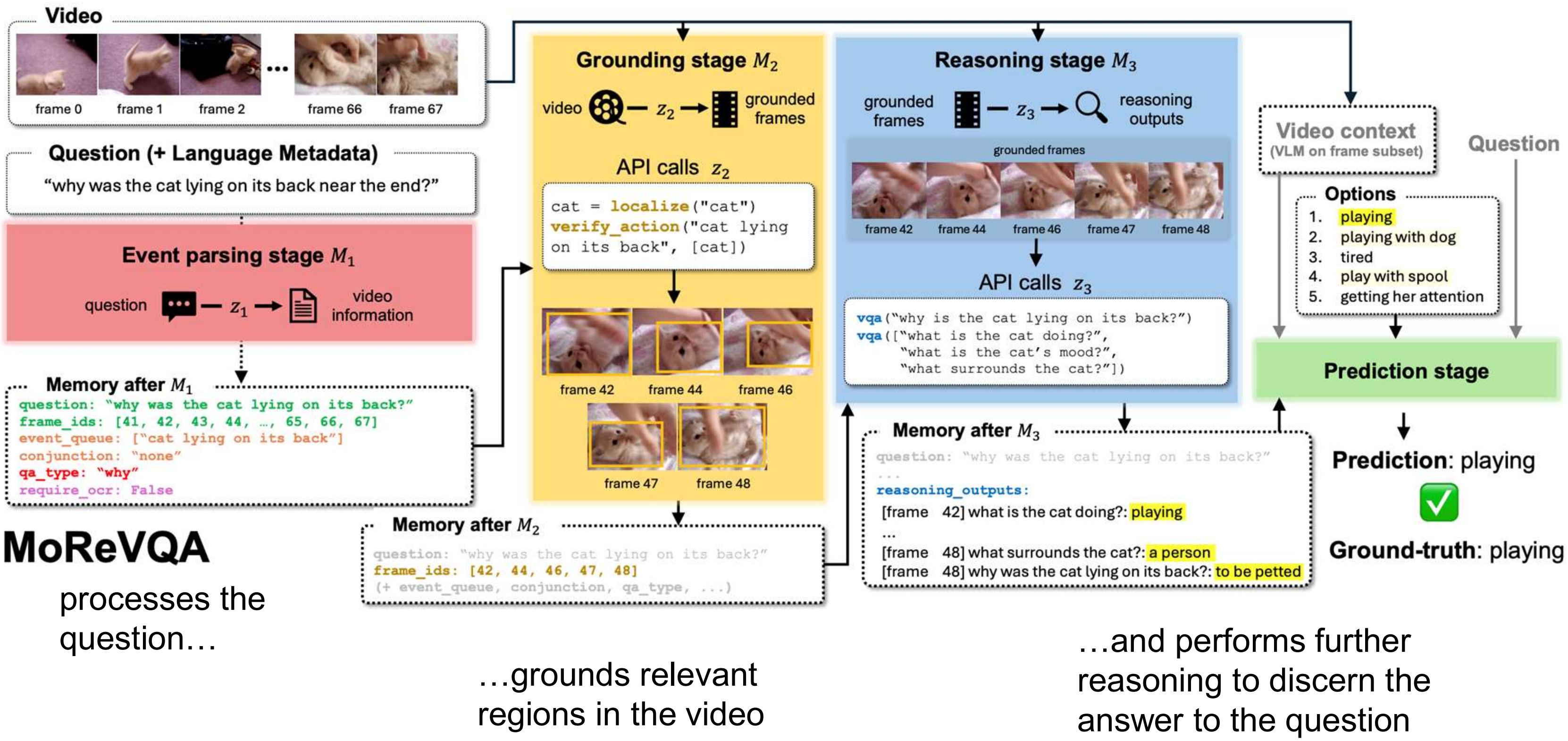

Q: Why is the cat laying on its back at the end?
Input video
frame 0
frame 1
frame 2
frame 3
frame 4
frame 5
frame 6
frame 7
frame 8
...
frame 65
frame 66
frame 67
Visual Programming
Video
Question
Code generation LLM
Generated program z
def execute_command(video, question, options):
video_segment = VideoSegment(video)
info = OrderedDict()
for frame_number, frame in enumerate(video_segment.frame_iterator()):
cats = frame.find('cat')
cat = cats[0] if len(cats) == 1 else best_image_match(cats, 'cat')
if "yes" in cat.simple_query('is the cat lying?'):
answer = cat.simple_query('why is the cat lying?')
info[f'[frame %4d] why cat is lying' % frame_number] = answer
answer = video_segment.select_answer(info, question, options)
return answer
Options
1. playing
2. playing with dog
3. tired
4. play with spool
5. getting her attention
info
[frame 3] why cat is lying: sleeping
[frame 8] why cat is lying: sleeping
[frame 9] why cat is lying: sleeping
[frame 12] why cat is lying: playing
[frame 19] why cat is lying: sleeping
[frame 26] why cat is lying: laying on its back
[frame 38] why cat is lying: licking
...
[frame 60] why cat is lying: scratching
[frame 61] why cat is lying: playing
[frame 63] why cat is lying: playing
[frame 65] why cat is lying: laying on its back
Prediction: tired
Ground-truth: playing
ViperGPT+ doesn't ask the VLM the right questions at the right time, and gets misleading answers



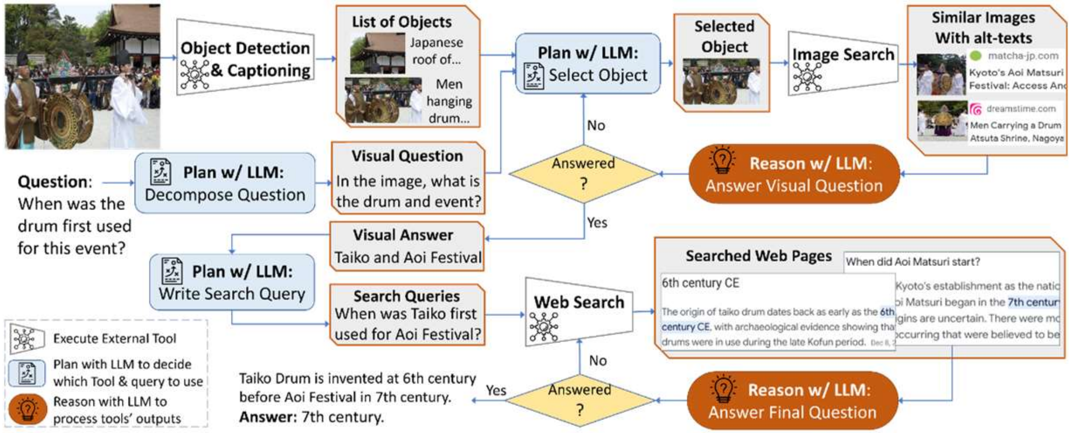
MoReVQA

processes the question…


Video
frame 0
frame 1
frame 2
frame 66
frame 67
Question (+ Language Metadata)
"why was the cat lying on its back near the end?"
Event parsing stage M_1
question — z_1 → video information
Memory after M_1
question: "why was the cat lying on its back?"
frame_ids: [41, 42, 43, 44, …, 65, 66, 67]
event_queue: ["cat lying on its back"]
conjunction: "none"
qa_type: "why"
require_ocr: False
MoReVQA
processes the question…
Grounding stage M_2
video — z_2 → grounded frames
API calls z_2
cat = localize("cat")
verify_action("cat lying on its back", [cat])
frame 42
frame 44
frame 46
frame 47
frame 48
Memory after M_2
question: "why was the cat lying on its back?"
frame_ids: [42, 44, 46, 47, 48]
(+ event_queue, conjunction, qa_type, ...)
…grounds relevant regions in the video
Reasoning stage M_3
grounded frames — z_3 → reasoning outputs
grounded frames
API calls z_3
vqa("why is the cat lying on its back?")
vqa(["what is the cat doing?", "what is the cat's mood?", "what surrounds the cat?"])
Memory after M_3
question: "why was the cat lying on its back?"
reasoning_outputs:
[frame 42] what is the cat doing?: playing
[frame 48] what surrounds the cat?: a person
[frame 48] why was the cat lying on its back?: to be petted
Video context
(VLM on frame subset)
Question
Options
1. playing
2. playing with dog
3. tired
4. play with spool
5. getting her attention
Prediction stage
Prediction: playing
Ground-truth: playing

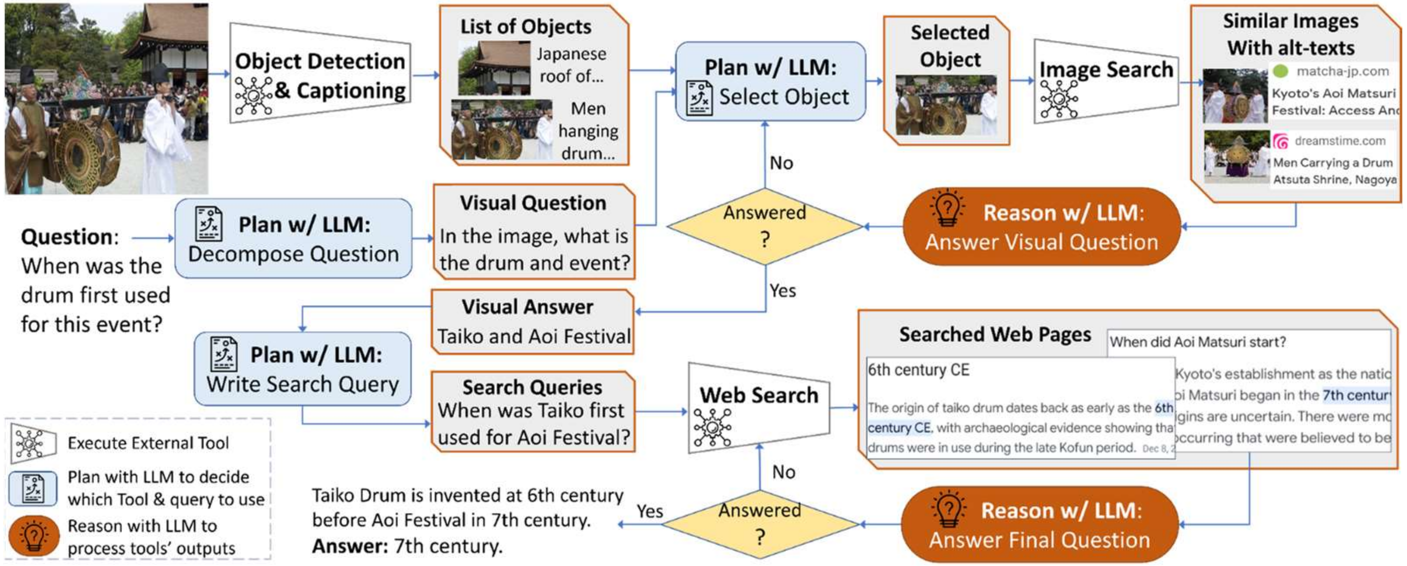

Video
frame 0
frame 1
frame 2
frame 66
frame 67
Question (+ Language Metadata)
"why was the cat lying on its back near the end?"
Event parsing stage M_1
question — z_1 → video information
Memory after M_1
question: "why was the cat lying on its back?"
frame_ids: [41, 42, 43, 44, …, 65, 66, 67]
event_queue: ["cat lying on its back"]
conjunction: "none"
qa_type: "why"
require_ocr: False
MoReVQA
processes the question…
Grounding stage M_2
video — z_2 → grounded frames
API calls z_2
cat = localize("cat")
verify_action("cat lying on its back", [cat])
frame 42
frame 44
frame 46
frame 47
frame 48
Memory after M_2
question: "why was the cat lying on its back?"
frame_ids: [42, 44, 46, 47, 48]
(+ event_queue, conjunction, qa_type, ...)
…grounds relevant regions in the video
Reasoning stage M_3
grounded frames — z_3 → reasoning outputs
grounded frames
frame 42
frame 44
frame 46
frame 47
frame 48
API calls z_3
vqa("why is the cat lying on its back?")
vqa(["what is the cat doing?", "what is the cat's mood?", "what surrounds the cat?"])
Memory after M_3
question: "why was the cat lying on its back?"
...
reasoning_outputs:
[frame 42] what is the cat doing?: playing
...
[frame 48] what surrounds the cat?: a person
[frame 48] why was the cat lying on its back?: to be petted
…and performs further reasoning to discern the answer to the question
Video context
(VLM on frame subset)
Question
Options
1. playing
2. playing with dog
3. tired
4. play with spool
5. getting her attention
Prediction stage
Prediction: playing
Ground-truth: playing

Visual Information Seeking with an LLM Agent (AVIS)



- Prompt-engineering is fragile → human traces
- LLM-based planner with external tools → addition of external knowledge
- LLM-based reasoner to process tool output

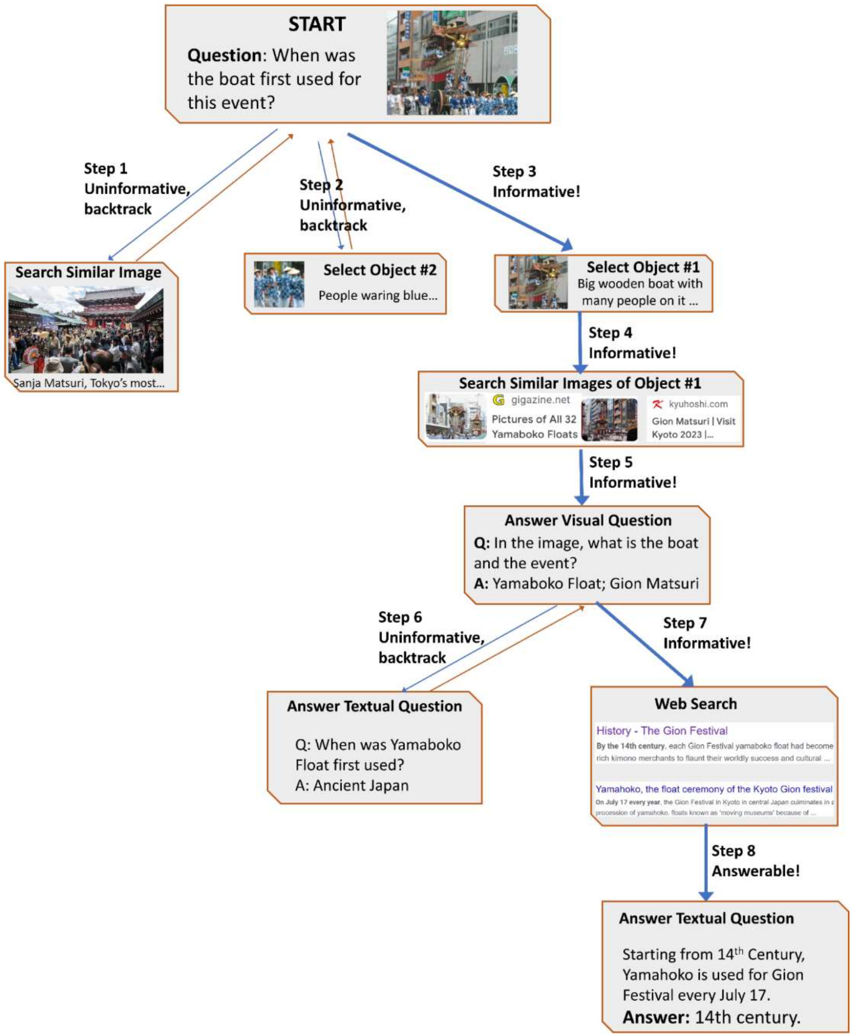
Visual Information Seeking with an LLM Agent (AVIS)



Example of generated workflow:

- LLM-based planner the dynamically selects the external tool
- LLM-based reasoner to process tool output
- Use of human behavior as guidance for decision making

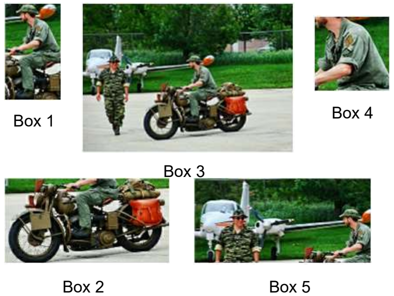
[AVIS: Autonomous Visual Information Seeking with Large Language Model Agent, Z. Hu et al. Neurips'23]



Example of AVIS generated workflow for VQA

Human Study

Question: In what year was this motorcycle built?



(a) Input visual question and detected objects

PALI Caption

PALI VQA Query:

PALI VQA

Search Query:

Search

PALM Query:

PALM

Show entity of box 1

Show caption of images similar to box 1

Show entity of box 2

Show related products to the object in box 2

Show caption of images similar to box 2

Show caption of identical images to box 2

Show entity of box 3

Show caption of images similar to box 3

Show related products to the object in box 4

Show caption of images similar to box 4

Show caption of images similar to box 5

(b) Tools shown to user

Outputs of “show entity of box2**”**

- Harley-Davidson
- Harley-Davidson, Inc. is an American motorcycle manufacturer headquartered in Milw manufacturers to survive the Great Depression along with its historical rival, Indian M arrangements, periods of poor economic health and product quality, and intense glob brand widely known for its loyal following. There are owner clubs and events worldwi
- Harley-Davidson is noted for a style of customization that gave rise to the chopper m motorcycles with engine displacements greater than 700 cc, but it has broadened its
- Harley-Davidson manufactures its motorcycles at factories in York, Pennsylvania; Mil
- Despatch rider
- A despatch rider is a military messenger, mounted on horse or motorcycle.
- In the UK 'despatch rider' is also a term used for a motorcycle courier.
- Despatch riders were used by armed forces to deliver urgent orders and messages be telecommunications were limited and insecure. They were also used to deliver carrie

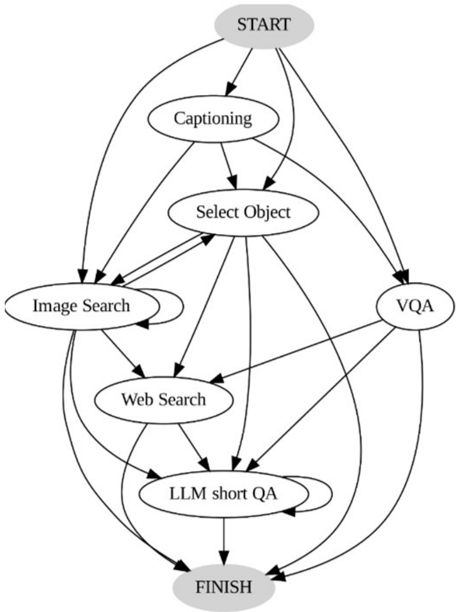
Useless API call

Success! Found the Answer!

Could't find the Answer!

(c) Tool Output

Human Study



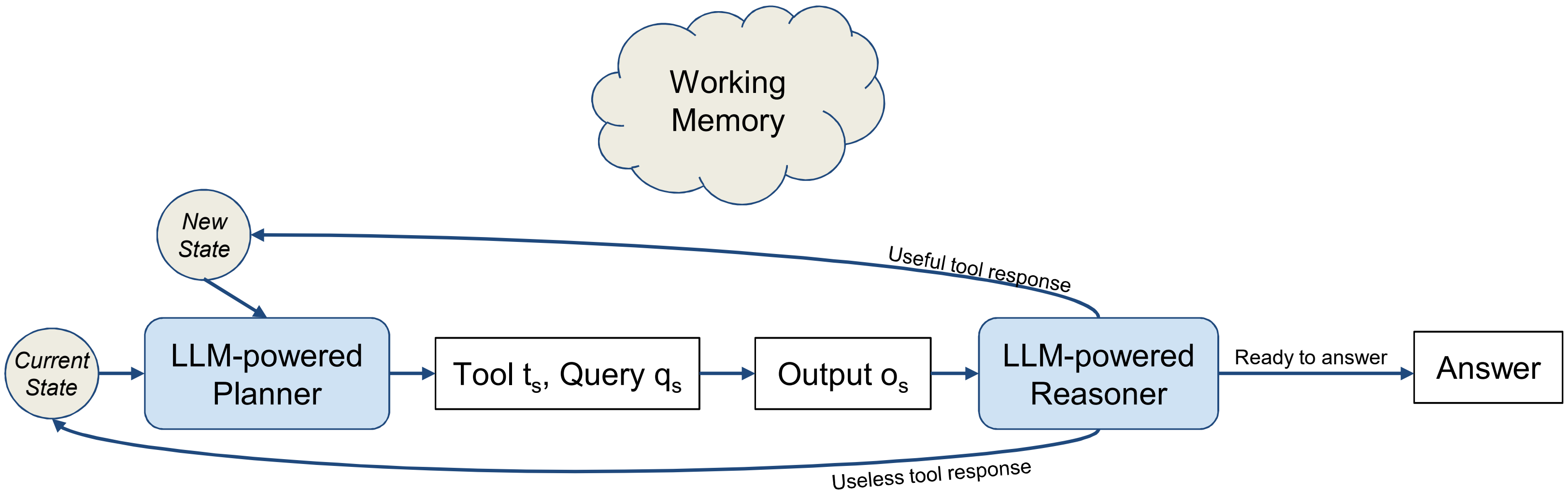

```
    Please think step by step. In the following, you will be given a "Query", a list of "Objects".

    Your task is to predict the object #ID that is mostly relevant to answer the querys. Please generate the
      detailed explanation why you select this object, and then output ID in "Object #ID".


Query: which city is this place?
Object #0 [
  a row of pillars sitting next to a dirt road . There is a building and this is plant. Here we can see
      pillars and a sky. (Caption, whole image)
  Query Suggestion: Palmyra Archaeology (score=100.0),
  Great Colonnade at Palmyra (ancient_roman_architecture_structure): The Great Colonnade at Palmyra was the
      main colonnaded avenue in the ancient city of Palmyra in the Syrian Desert (score=90.3),
]
Object #1 [
  a green plant sitting next to a brick wall . There is a plant and this is wall. And there is a sky.
      (Caption, center)
  Date palm (type_of_palm_trees): Phoenix dactylifera, commonly known as date palm, is a flowering plant
      species in the palm family, Arecaceae, cultivated for its edible sweet fruit called dates (score=81.7),
]
Object #2 [
  a wicker basket sitting on top of a rock . There is a blur image of a rock. (Caption, lower right)
]
Output: The query asks about the city of the place. Only Object #0 (whole image) mentions city name Palmyra,
      which is an acient city. Also, Object #0 contains Query Suggestion "Palmyra Archaeology".
Therefore, the predicted Object #ID is 0.
```

Transition Graph

Example from human decision making

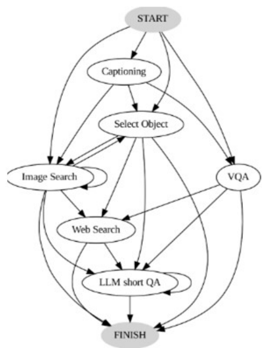
→ Guiding LLMs using human decision making examples

Our approach



Tools

- Web search
- Image search
- LLM
- VLM
- Object selection
- OCR
- Object entity
- Similar image captions
- Identical image captions



Transition Graph from User Study

Experimental results – InfoSeek Dataset

Q: What is this bridge named after?
A: George Washington

Q: What is the length of the wingspan in millimetre of this insect?
A: 33.0-45.0

Q: Who is the founder of the aircraft in the image?
A: Olive Ann Beech

Q: In which year was this equipment retired from operational service?
A: 2006

Experimental results – InfoSeek Dataset

Model	Unseen Entity	Unseen Question
PALM [9] (Q-only, few-shot)	3.7	5,1
OFA [22] (fine-tune)	9.7	14.8
PALI [6] (VQA, zero-shot)	1.8	2.2
PALI [6] (fine-tune)	16.0	<u>20.7</u>
PALM [9] w/ CLIP [32] (few-shot + external knowledge)	<u>21.9</u>	18.6
FiD [44] w/ CLIP [32] (fine-tune + external knowledge)	20.7	18.1
(—baselines without dynamic decision making, sequentially execute the tools—)		
baseline-PALM w/ (PALI*, few-shot)	12.8	14.9
baseline-PALM w/ (PALI* + Object, few-shot)	31.3	36.1
baseline-PALM w/ (PALI* + Object + Search, few-shot)	36.1	38.2
AVIS (ours, few-shot)	**50.7**	**56.4**
w/o PALI*	47.9	54.2
w/o Object	41.2	48.4
w/o Search	42.5	49.6

Internships

- Topics
  - Long video understanding + visual reasoning
  - Interpretability by reasoning and cross-modal information
  - Image and video generation
  - LLM for 3D understanding
  - Vision language for robotics

- Please contact me direct with a CV + the name of two referees
  - Email: Cordelia.Schmid@inria.fr