

# Video Recognition メタサーベイ

---

Video Recognition group

葛城 玲弥, 福沢 匠, 木林 佑太, 八木 拓真, 原 健翔

岡山大学  
名古屋工業大学  
東京大学  
産業総合研究所(AIST)

# サーベイ資料の構成

---

- イン트로ダクション
- 複雑な推論
- 少数サンプル学習と転移学習
- 長時間動画理解

# はじめに

---

- cvpaper.challengeにおける2024年のメタサーベイとしてvideorecogグループをはじめとしたメンバーにより作成
- **複雑化した動画認識のタスク**に焦点をあて、34本の論文を調査してまとめたサーベイ資料
  - 対象領域のすべての論文を網羅しているわけではありません
- 基本的な動画タスクやモデルについては以下の資料をご覧ください
  - <https://www.slideshare.net/cvpaperchallenge/ss-248333069>
  - <https://www.slideshare.net/slideshow/video-transformer/252751737>

# VideoRecognitionとは

- 動画処理・分析するタスク
- 画像タスクとは違い、**時間情報**も考慮

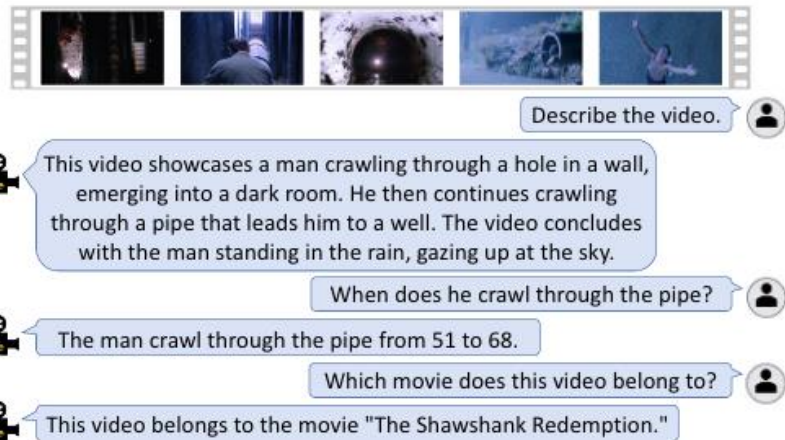


## 昨今のトレンド

- **タスクの複雑化** : Video Classification Task → Video Language Task



→ Riding a bike (ラベル)



\*"The Kinetics Human Action Video Dataset", Kay+, arXiv2017

\*\*[https://github.com/hueihan/Action\\_Recognition/blob/master/data/WIS/video/run/daria\\_run.avi](https://github.com/hueihan/Action_Recognition/blob/master/data/WIS/video/run/daria_run.avi)

# タスク複雑化の例：Video QA

- より難易度の高いデータセットが登場
- 簡単な質問
  - 一枚のフレームで答えが分かる
- 複雑な質問
  - ビデオ全体を理解しないと答えが分からない
    - いつ, どこで, 何回, なぜ, 未来の予測
  - マルチモーダルな情報が手掛かりとなる
    - 音声, 言語
  - 外部の知識を必要とする
    - 一般常識, web上の知識

**Factoid VideoQA:** directly asks about the visual fact.

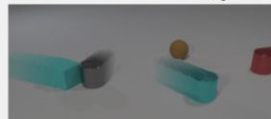


TGIF-QA (Jang et al. 2021)

**Factoid Question:**

Q: What is the color of the bulldog?  
A: Brown

**Inference VideoQA:** explore the logic reasoning ability.



CLEVRER (Yi et al. 2020)

**Inference Question:**  
Q: Which event will happen next?  
a: The cube collides with the red object ✓  
b: The cyan cylinder collides with the red object



NEXT-QA (Xiao et al. 2021)

**Inference Question:**  
Q: Why is the man in yellow and the man in black carrying a shirt-less man?  
A: throw him into water

**MM-VideoQA:** QA invokes visual, audio, subtitles, plots.



TVQA (Lei et al. 2018)

01:10.131 → 01:12.292  
(Meredith) I was swimming.  
01:12.367 → 01:14.358  
(Meredith) I was fighting.  
01:15.670 → 01:18.138  
(Meredith) And then I thought, just for a second.

Q: What was behind Meredith when she said she was swimming?  
0. A painting; 1. A couch;  
2. A metal shelf ✓  
3. A file cabinet; 4. A car

**KB-VideoQA:** QAs invokes external information outside video.



Penny was angry at Leonard in this episode.



KnowIT (Garcia et al. 2020)

(Howard) Grab a napkin, homey, you just got served.  
(Leonard) It's fine. You win.  
(Howard) What's his problem?  
(Sheldon) His imaginary girlfriend broke up with him.

Q: What girlfriend is Sheldon talk about?  
0. Priya; 1. Amy;  
2. Bernadette  
3. Penny ✓

“Video Question Answering: Datasets, Algorithms and Challenges”, Zhong+, EMNLP 2022

# 本資料の主なサーベイ内容

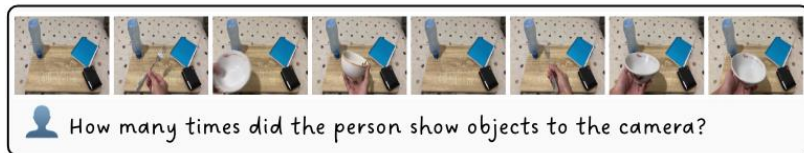
---

動画のコンテキストの理解，効率的な表現学習の観点から3つの項目に注目

- 階層的な特徴（スライドp10～）
  - 動画中のイベントの理解
  - 人物・物体の存在とその相互関係を表現する
- 少数サンプル学習と転移学習（スライドp38～）
  - 新しい概念や対象を少ないサンプルで学習
  - 既存の知識を新しいドメインに転移
- 長時間の動画理解（スライドp58～）
  - 数分から数時間に及ぶ長時間の動画
  - 動画ならではの時間的な文脈，ストーリーの展開，長期的な依存関係の認識が必要

# 長時間動画理解とは

- **目的** 数分から数時間にわたるVideoClip(EPIC-KITCHENSは約4秒)に対して、一部の時間帯ではなく、長時間の文脈や関係性を捉えた応答を行う
  - 長時間のcontextを捉えないと回答できないBenchmarkの作成(p○○,○○)
- **タスクの例**
  - Count: 動画中で何回actionや動きがあったかを数える
    - 動画の一部だけではなく、全体を見ないと回答不可
  - Action Prediction: 現在までの行動に基づいて将来の行動を予測する
    - 直前の行動だけではなく、今までの行動を捉えないと回答不可



引用 : <https://arxiv.org/abs/2311.17005>

# 少数サンプル学習・転移学習とは

- Video Recognitionの課題
  - データ収集+アノテーション → 動画のデータセット作成のコストは高い。少数サンプルからタスクを解けると、データ収集コストを削減できる。
- 少数サンプル（小規模データ）：数千規模のデータセット
- 少数サンプルの学習における懸念点
  - 「モデルの大規模化」による過学習
  - 「事前学習のバイアス」による性能低下

## 動画データ収集の限界

- 撮影のコスト
  - 数秒以上、撮影をし続ける必要がある
- アノテーションの人的コスト
- 不良サンプルが少ないタスク
  - 例：異常検知，医療診断



# 階層的な特徴

- VideoQAなどのタスクでは高度な動画理解が求められる
  - イベントの時間的關係
  - 物体・人物間の關係 などを捉える必要
- シーンの階層構造をモデル化した手法が提案されている



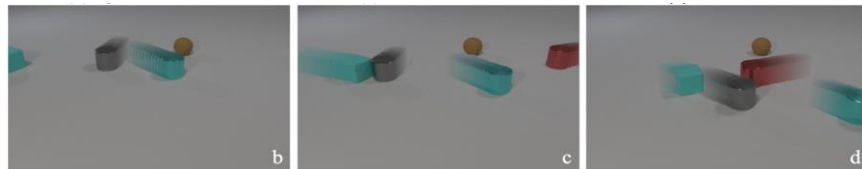
Q. **How** did the man in light blue jacket show that he wanted a drink?

A. point his finger [1]



Q. **Why** did the boy pick up one present and move to the sofa?

A. unwrap it [3]



Q. Which event **will happen next**

a). The cube collides with the red object. ✓

b). The cyan cylinder collides with the red object [2]

[1] "IntentQA: Context-aware Video Intent Reasoning", Li+, ICCV2023

[2] "CLEVRER: Collision Events for Video Representation and Reasoning", Yi+, ICLR2020

[3] "NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions", Xiao+, CVPR2021

# 階層的な特徴

# 背景

---

動画における

- 発生したイベント
- 人物・物体の相互関係 を表現する能力が求められている

課題

動画が持つ複雑な視覚情報を上手く表現することが難しい



シーンの階層的な特徴を反映し  
簡単なものの組み合わせで、動画表現を構成

# タスク例

## VideoQA

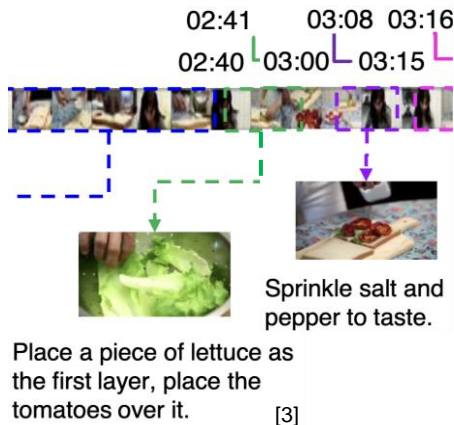


Q. **How** did the man in light blue jacket show that he wanted a drink? [1]



Q. **Why** did the boy pick up one present and move to the sofa? [2]

## Video Captioning



## Video Moment Retrieval



Query: Old man in white finishes filming then leaves.



Query: A man holding a camera comes into view. [4]

動画中の動作，物体の関係，意味，順序などを細かく理解したい

[1] "IntentQA: Context-aware Video Intent Reasoning", Li+, ICCV2023

[2] "NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions", Xiao+, CVPR2021

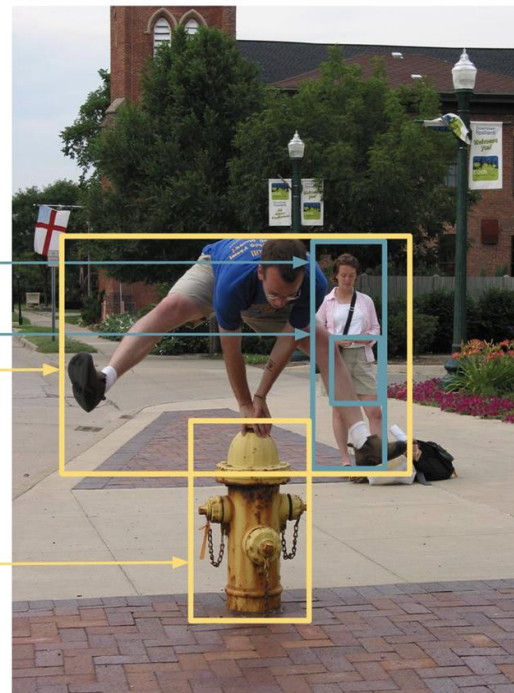
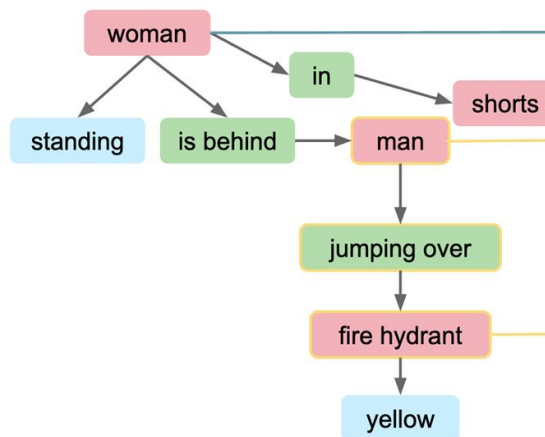
[3] Towards Automatic Learning of Procedures from Web Instructional Videos, Zhou+, AAAI 2018

[4] Weakly Supervised Video Moment Retrieval From Text Queries, Mithun+, CVPR2019

# 階層的な特徴

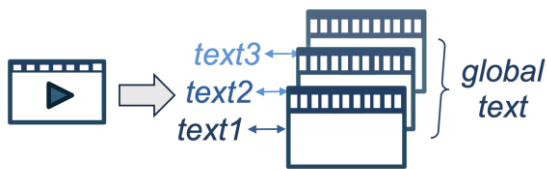
シーンの意味的な構造を反映してモデル化

- わかりやすい特徴の組み合わせで複雑な視覚情報を表現する



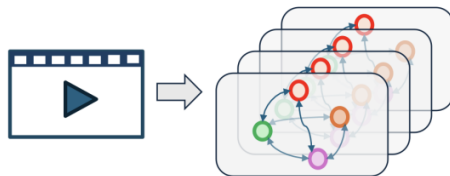
# 動画の階層的特徴の分類

## 時間的な階層性



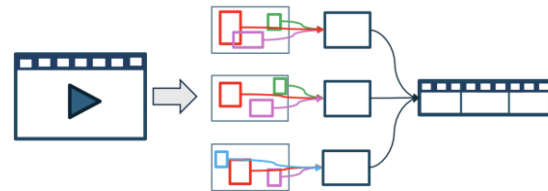
- 時間方向のイベントの階層性をモデル化
- 手順動作など時間的粒度の異なる動作の関係を学習

## 空間的な階層性



- シーンの物体・人物の関係をモデル化
- 物体間の細かな相互関係を捉えられる

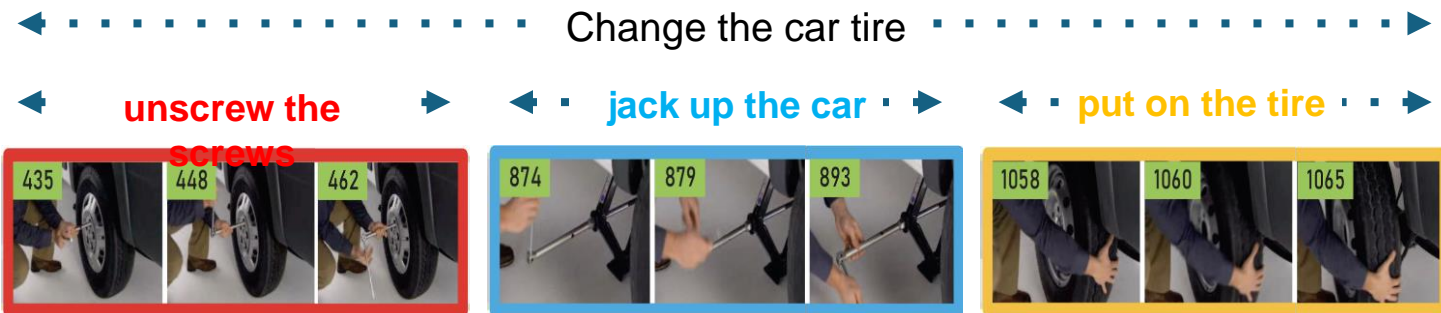
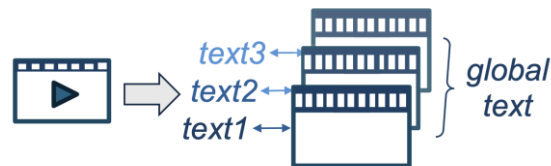
## 時空間的な階層性



- 局所的領域から段階的に動画全体の特徴を構成
- 時空間的に異なるレベルの視覚情報を統合できる

# 階層的な特徴 | 時間的

- 目的：長時間な動作の順序や階層性を学習する
  - 下流タスクとしてVQAなどに適用可能
- 動画をクリップに分割
  - クリップ単位で一致するテキストと対応付け

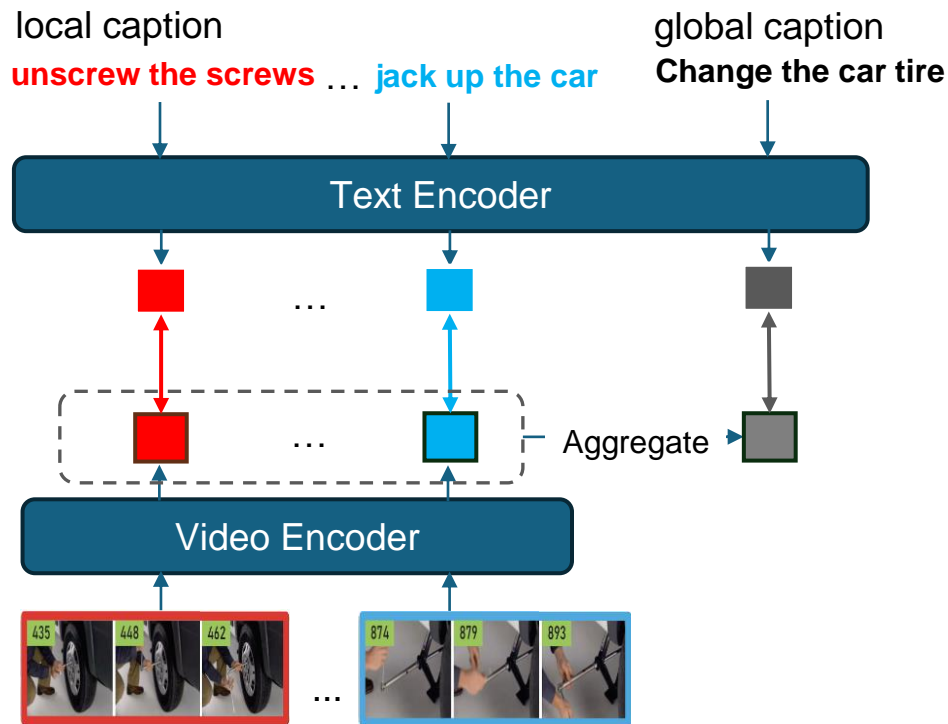


COINデータセットの動画とキャプションの例

“COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis”, Tang+,  
CVPR2019

# 階層的な特徴 | 時間的

- モデル構造の例
  - 2段階なV-L特徴の対応付け
    - 短い区間の動作と細かなテキスト
    - ビデオ全体の動作と要約テキスト
- 特徴
  - 良い点
    - 作業・料理などの手順の理解に有効
    - 動作間の関係を学習できる
  - 課題
    - 時間的に密なアノテーションが必要
    - 手順データセットはドメインが限定的





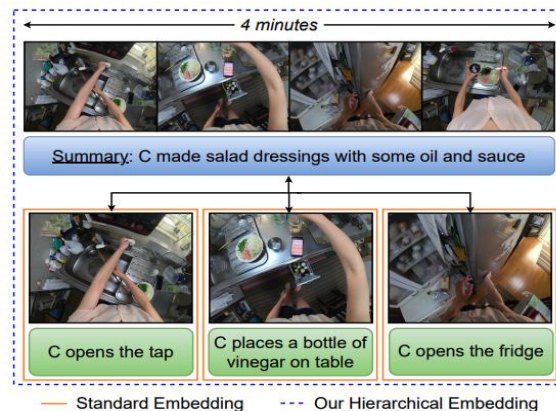
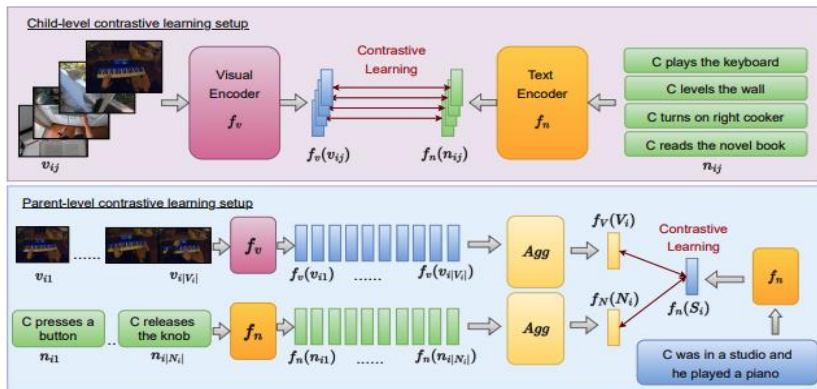
# HierVL: Learning Hierarchical Video-Language Embeddings, Ashutosh+, CVPR2023

## 長期的・短期的な関連性を同時に考慮するV-L埋め込み

- V-L間だけでなく、長期テキストと集約した短期テキストの間でも対象学習を行う
- Ego-4Dの長期・短期アノテーションで事前学習

結果：

- 長期ビデオモデリングで優れた結果を達成し、EPIC-KITCHENS-100、Charades-Ego、HowTo100Mなどの下流タスクで高い性能
- より広い動作の理解（動作の意図や理由）を掴むことができる



# Weakly Supervised Video Representation Learning with Unaligned Text for Sequential Videos, Dong+ , CVPR2023

## 正確なタイムスタンプがキャプションに提供されていない状況に対処

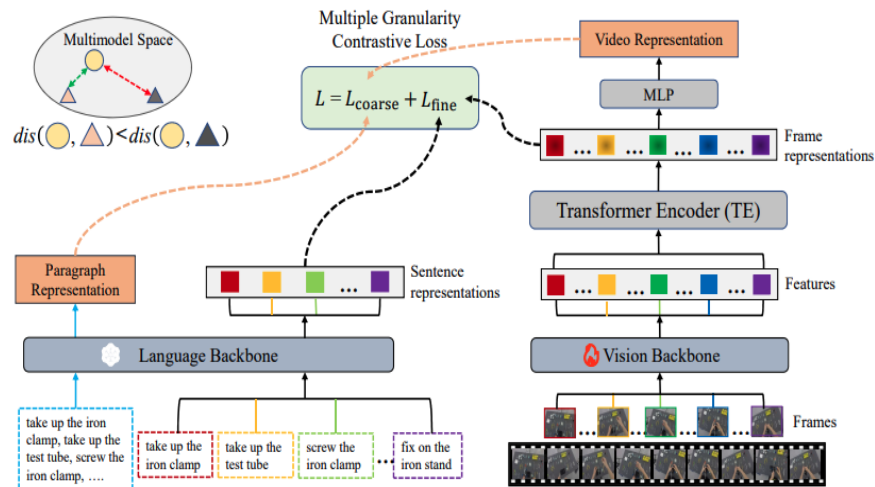
- 動作が順次発生することから、疑似フレームと文の対応を生成  
疑似ラベルを用いることで複数の粒度で損失を計算

### 結果

- COIN-SV、Diving-SV、CSVのデータセットで有効性が確認

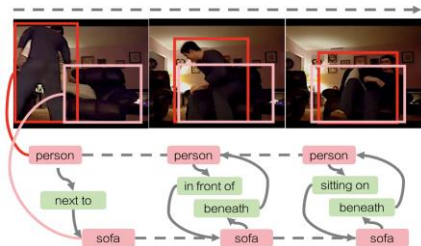
### 課題

- サブアクションが繰り返される場合に  
モデルが誤った疑似ラベルを生成する可能性



# 階層的な特徴 | 空間的

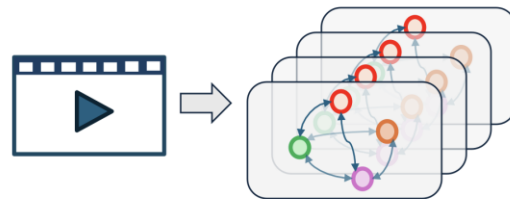
- 目的：物体間の細かな相互関係を捉える
- 物体間の関係をグラフで構成
  - グラフ：ノードとエッジの集合



## トップダウン型

### シーングラフ

- ノード：物体カテゴリ
- エッジ：関係（動作や位置関係）



## ボトムアップ型

### 特徴量ベースなグラフ

- ノード：ROI領域特徴
- エッジ：ノード間の類似度

# 階層的な特徴 | 空間的

- モデル構造の例

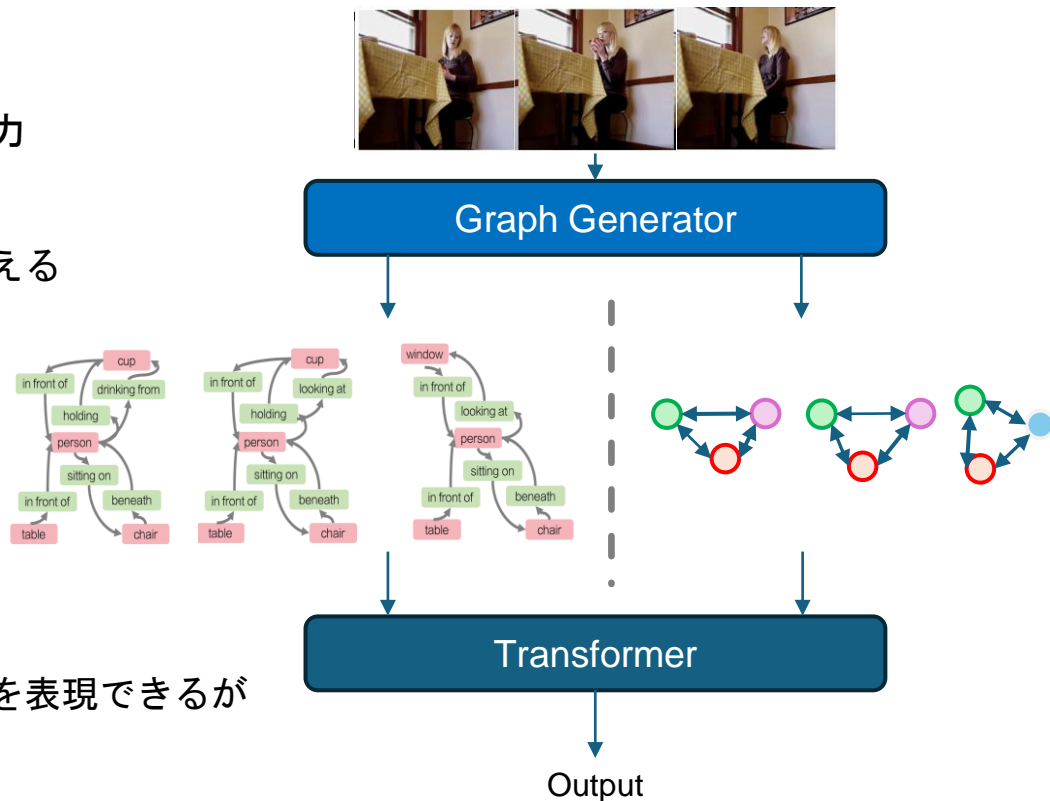
- 動画からグラフを生成しモデルに入力
- VQAの場合：
  - 各モデルの入力に質問テキストを与える

- 良い点

- 物体間の時空間的な相互作用を捉えられる

- 課題

- グラフの精度が結果に大きく影響
- トップダウン型はより意味的な関係を表現できるがシーングラフアノテーションが必要



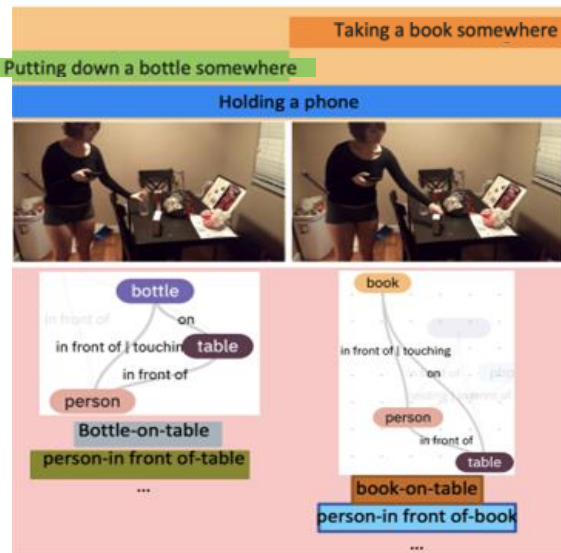
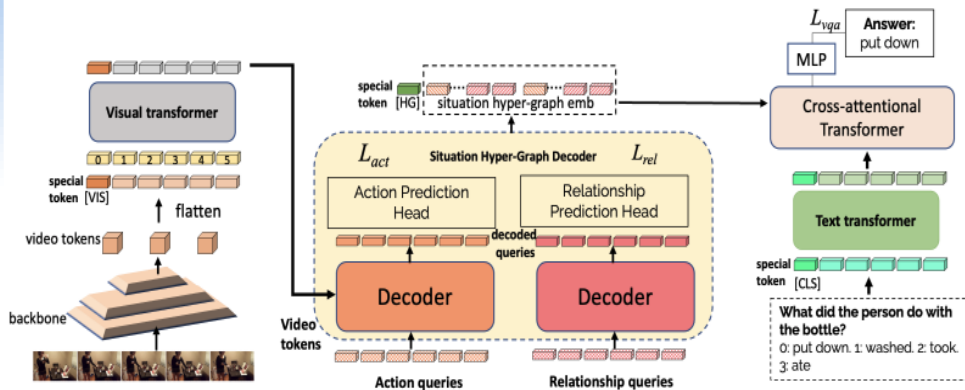
# Learning Situation Hyper-Graphs for Video Question Answering, Khan+, CVPR2023

動画からシーングラフを直接予測し，回答するVideoQAアーキテクチャ

- オブジェクト検出や他の事前知識を用いず，状況ハイパーグラフ（SHG）を直接学習する
- アクションとオブジェクト/ヒューマン-オブジェクト関係をデコードする点。

結果:

- AGQAとSTARでVQA精度を大幅に向上
- 人物，物体とそれらの関係だけでなく，関係の経時的な変化も捉える



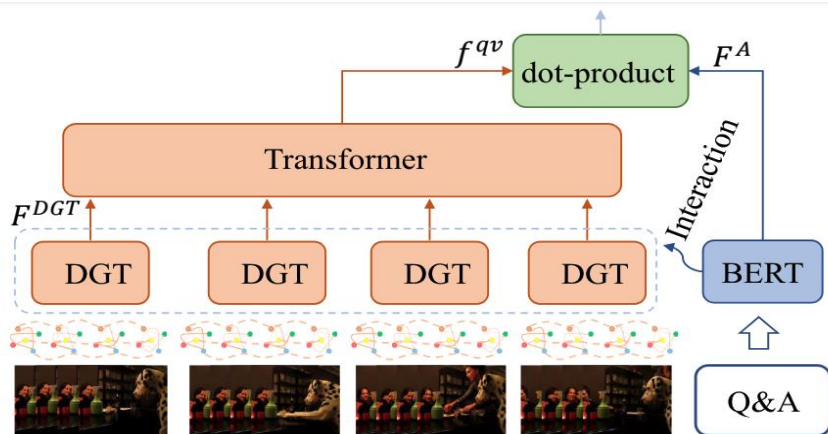
# Video Graph Transformer for Video Question Answering, Xiao+, ECCV2022

## Graph Transformerを用いてビデオ特徴を獲得する

- Dynamic Graph Transformer (DGT) を用いて視覚的な関係性の動的な変化を明示的に捉える
- 大規模な外部データを使用しない, データ効率の良い自己教師あり事前学習

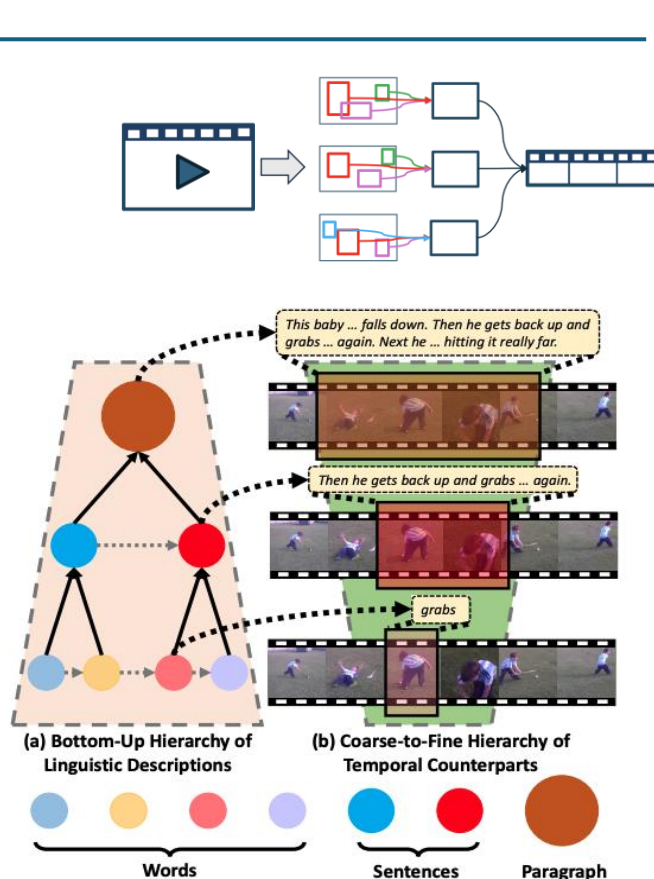
結果 :

- NExT-QA, TGIF-QAで高い性能
- 物体間の時空間的な相互作用, 関係の推論で高い性能



# 階層的な特徴 | 時空間的

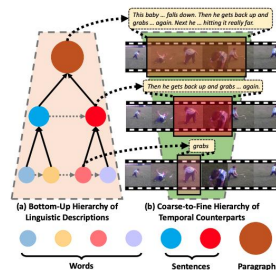
- 局所領域から全体への段階的な特徴の構成
  - 例：物体特徴→フレーム特徴→動画特徴
- 同レベルなテキストと視覚情報に対応付ける
- 良い点
  - 長い文章でも、細かなV-L相互関係を学習できる
- 課題
  - モデル構造が複雑
  - カスケードエラーの可能性
    - 下位レベルのエラーが上位に影響



# Hierarchical Semantic Correspondence Networks for Video Paragraph Grounding, Tan+, CVPR2023

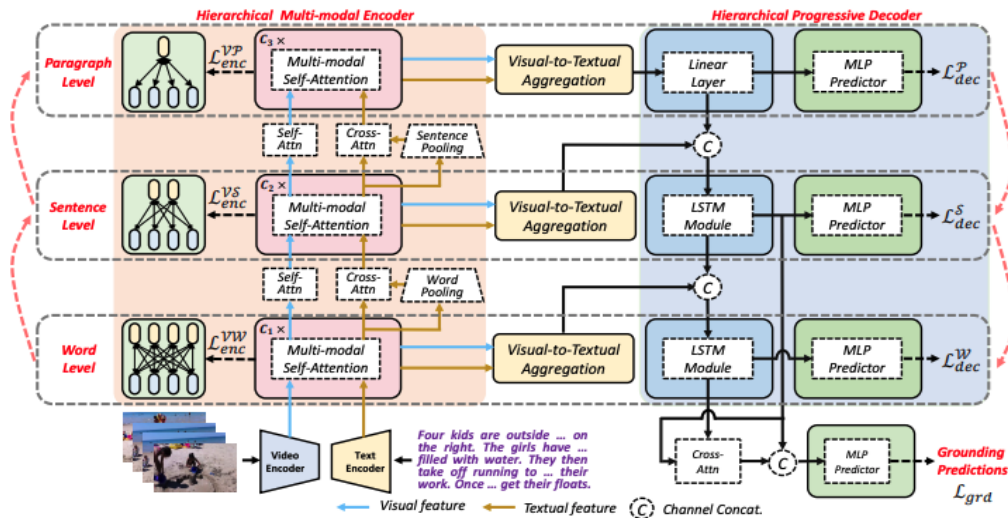
## Video Paragraph Groundingタスクに焦点

- 既存手法は単一レベルの意味対応のみに着目
- 視覚・言語特徴のための階層的なエンコーダー・デコーダー
- 上位レベルから下位レベルへとクエリを段階的にグラウンディング。



結果:

- ActivityNet-Captions, TACoSで既存手法を大きく上回る性能
- 特に長いビデオと複雑な段落を含むTACoSで有効。





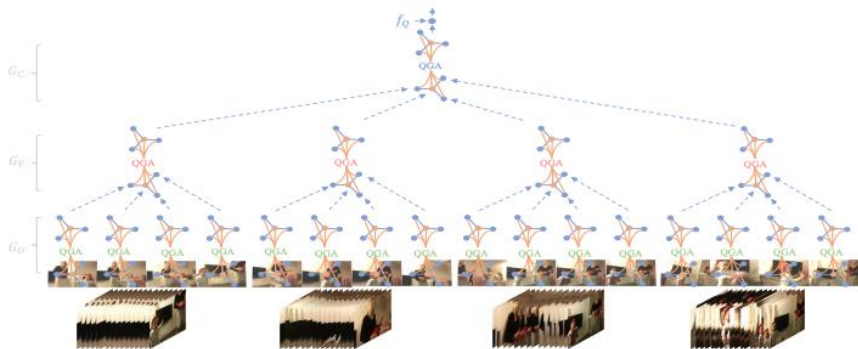
# Video as Conditional Graph Hierarchy for Multi-Granular Question Answerin, Xiao+, AAI2022

異なる粒度の視覚・言語の条件付きグラフ階層を用いたVideoQA

- 物体レベル, フレームレベル, クリップレベル, ビデオレベルと段階的に特徴を構成
- 言語クエリに基づいて視覚要素を識別し、質問に答えるための適切な情報を選択
- 各段階で視覚情報はテキストクエリとともにグラフアテンションネットワークに入力

結果:

- NEXt-QA, TGIF-QA, MSRv, MSVDで高い性能
- 予測の解釈可能性が向上



What does the girl in white do after bending down in the middle?

- 1 Grab her
- 2 Feed horse with grass** ✓
- 3 Run towards the camera
- 4 Umbrella
- 5 Put her arms up

$G_C$   $G_F$

# 階層的アノテーションの有無

---

正解ラベルに加えて階層的なアノテーションも利用可能か

## 教師あり学習

正解ラベル＋階層的アノテーション

- 学習時に出力だけでなく階層的特徴も評価可能
- 階層的な表現に対応するアノテーションはコストが高い

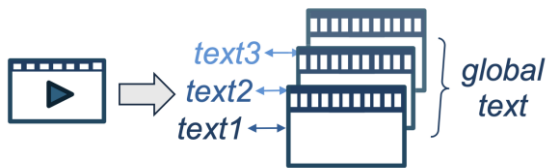
## 弱教師あり学習

正解ラベルのみ

- アノテーションが不要
- 学習時に階層的特徴を評価することが困難
- そのため正解しても正しい根拠を持っている保証がない

# 階層的アノテーションの有無

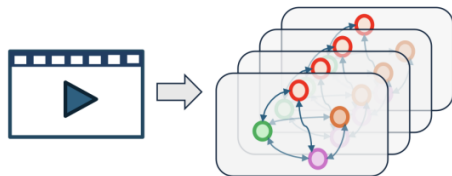
## 時間的な階層性



### • 教師あり学習

- 各セグメントとテキストの対応関係を学習
- 密なキャプションが必要

## 空間的な階層性



### • トップダウン型

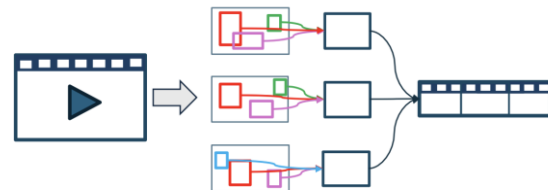
#### • 教師あり学習

- シーングラフ生成のためのアノテーションが必要
- グラフアノテーションを持つVQAデータセット [1,2]

### • ボトムアップ型

#### • 教師なし学習

## 時空間的な階層性



### • 教師なし学習

- ボトムアップ型であり階層構造に対応するアノテーションは不要

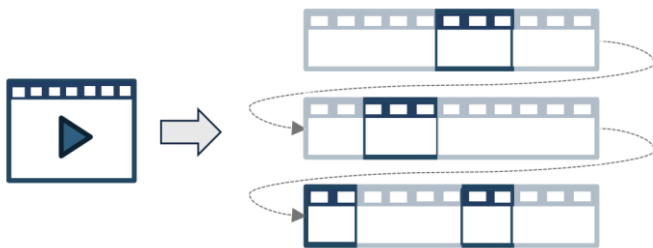
[1] "AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning", Grunde-McLaughlin+, CVPR2021

[2] "A Benchmark for Situated Reasoning in Real-World Videos", Wu+, NeurIPS2024

# その他

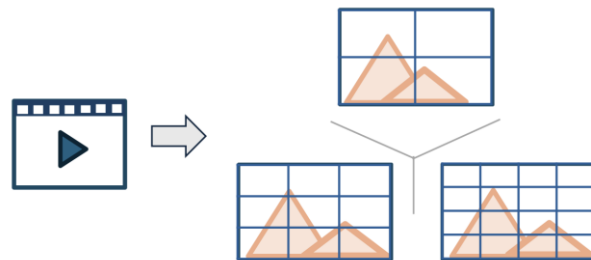
- 階層的な特徴ではないが  
時空間的な情報の局所性を捉えるための手法

重要な領域の選択



予測に関連する視覚情報を  
抽出する能力を獲得

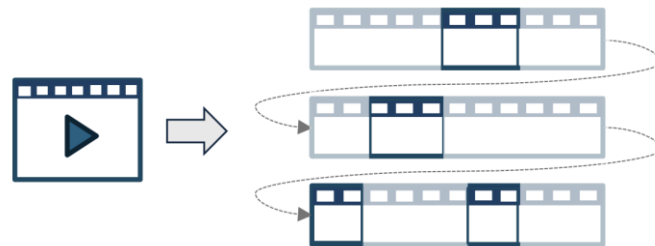
異なる解像度の入力



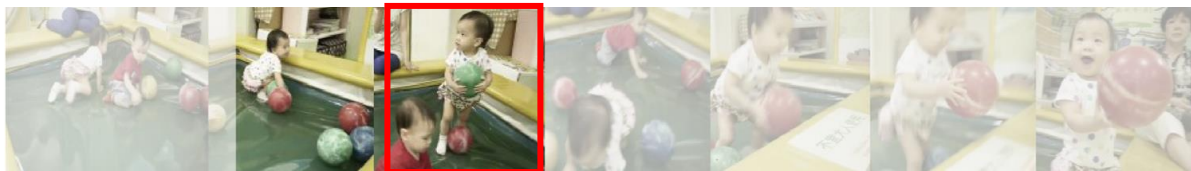
異なる粒度で視覚情報  
を捉える

# 重要な領域の選択

- 目的：VQAにおいて重要な視覚情報に着目
- 動画から回答に対して重要な瞬間を選択したのち，回答を予測する



Q: What did the girl do after she took the green ball?



A: Stand up

「質問」ではなく，「回答」に関連する瞬間を選ぶ

# 重要な領域の選択

- モデル構造の例

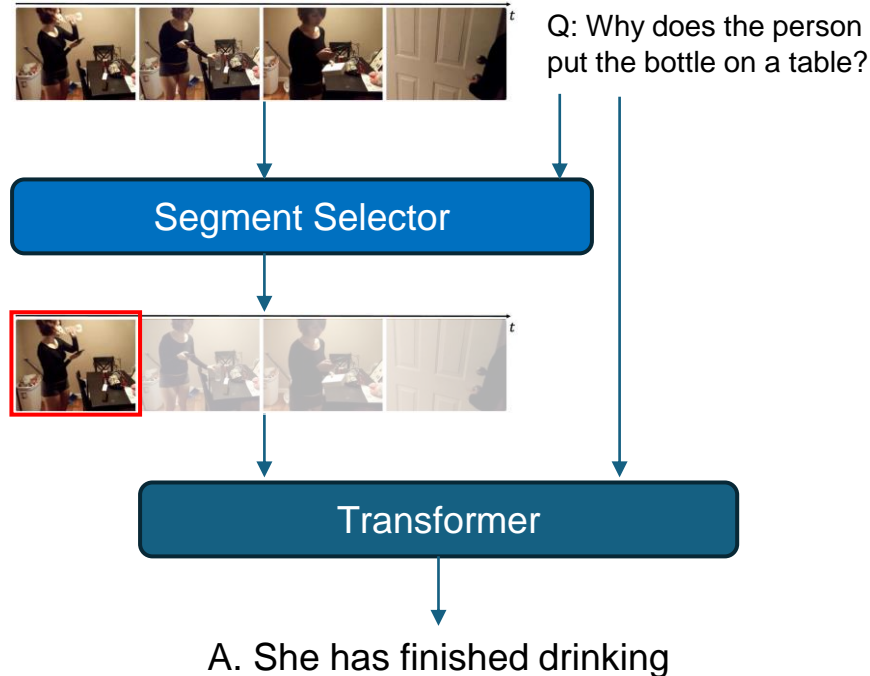
- 回答に重要なセグメントを選択
- 選択したセグメントと質問で回答を予測
- アノテーションは不要

- 良い点

- 回答の根拠となる視覚情報に注目する
- セグメント入力で計算量を抑えられる
- マルチイベントな質問に有効

- 課題

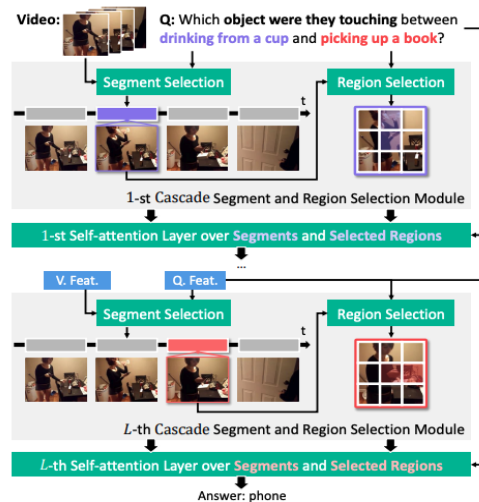
- セグメントの選択が性能に大きく影響
- 正解した場合でも、根拠として正しいセグメントを選択している保証がない



# MIST : Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering, Gao+, CVPR2023

## セグメント選択とパッチ選択の2段階による視覚情報の選択

- 長編ビデオの質問応答（VideoQA）システムで、
- 質問に関連するフレームと画像領域を選択することで、長編ビデオの効率的な処理を実現
- 結果:
  - MISTはAGQA、NExT-QA、STAR、Env-QAなど広い範囲のVideoQAデータセットで高い性能
  - 複雑なイベントから答えを見つける能力が向上
- 課題:
  - フレーム選択するため高頻度の動きを捉えることが難しい
  - 細かく複雑な物体レベルの推論には限界がある



# Discovering Spatio-Temporal Rationales for Video Question Answering, Li+, ICCV2023

## 質問に重要なセグメントとオブジェクトの選択

- 質問に関連する重要な瞬間とオブジェクトを選択する新しい手法を提案  
動画とテキストの相互作用を通じて、適応的に質問に重要なフレームとオブジェクトを選択
- 結果:
  - NExT-QAやCausal-VidQAなどのデータセットで有効性を示す
  - 複数のオブジェクトを含む長いビデオのQAにおいて既存手法より特に優れた性能

Question: What does the second person do after stopping her bike?



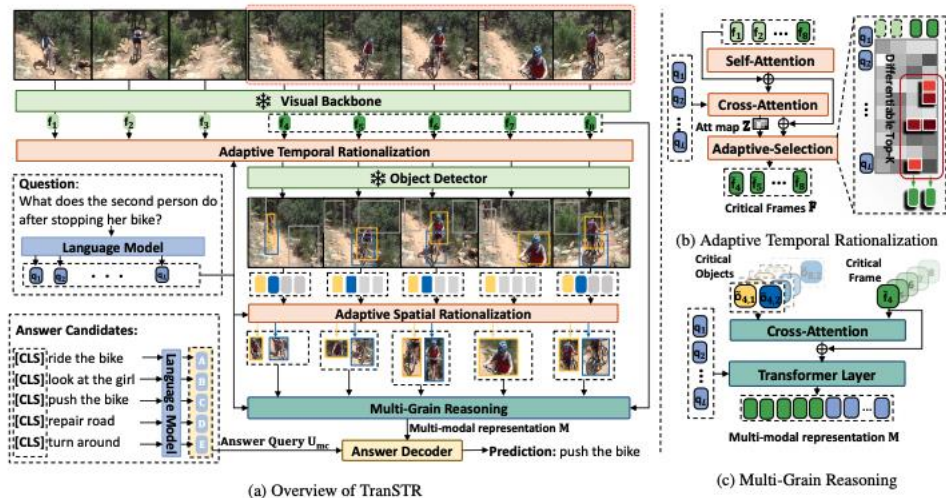
Answer Candidates:

- (A) ride the bike
- (B) look at the girl
- (C) push the bike
- (D) repair road
- (E) turn around



stopping

Answer: push the bike





# Can I Trust Your Answer? Visually Grounded Video Question Answering, Xiao+, CVPR2024

## 重要な瞬間をガウシアンマスクで予測

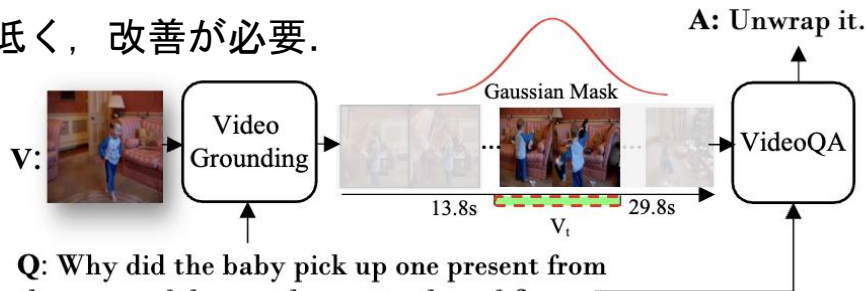
- VideoQAにおいてモデルが根拠として正しい視覚的区間に注目しているかを調査した論文
- セグメント選択ではなく、ガウシアンマスクを予測することで選択の難易度を下げる
- 他データの動画・テキストを負例として利用

結果：

- NExT-QAにおけるGrounding VideoQAの精度が向上し、既存手法より正しい根拠をもつことが確認された
- ただし、人と比べるとまだGrounding精度は低く、改善が必要。



What did the girl do after she took the green ball? Stand up.

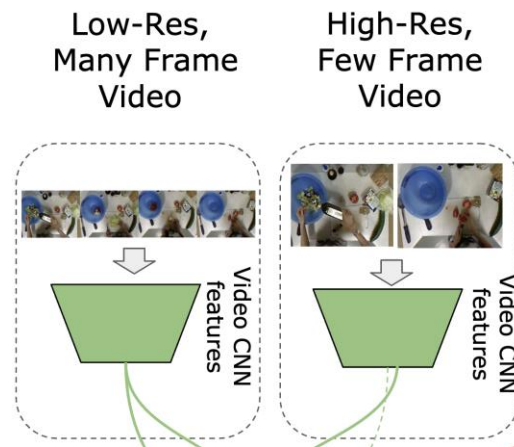
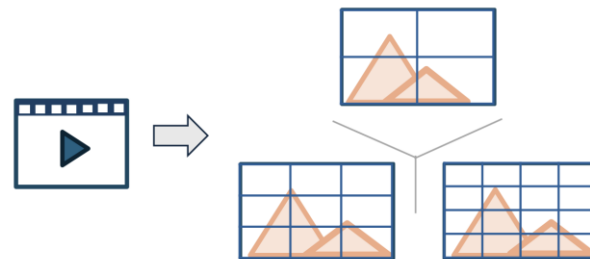
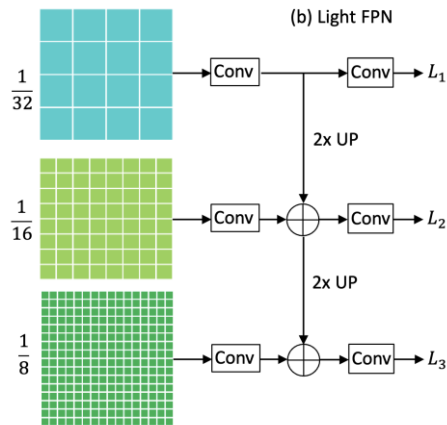


Q: Why did the baby pick up one present from the group of them and move to the sofa?

(a) Framework of weakly-grounded VideoQA.

# 異なる解像度の入力

- 目的：異なる粒度で視覚情報を捉える
- 多様な解像度で動画入力  
周波数の異なる特徴に注目



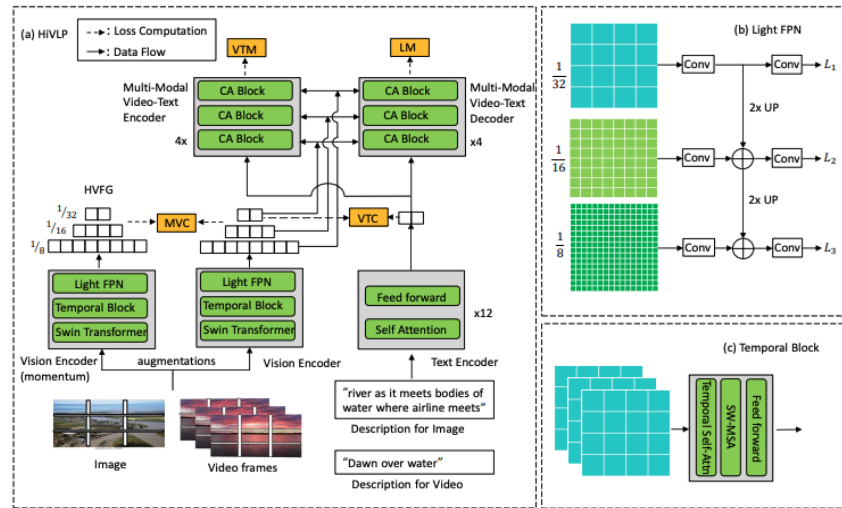
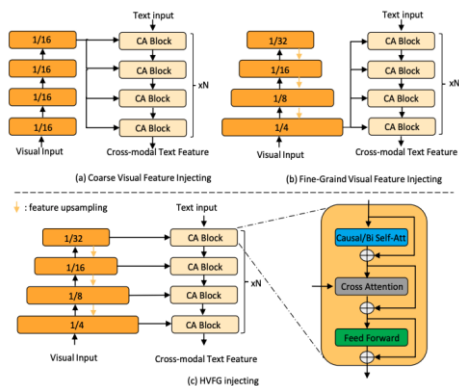
# HiVLP: Hierarchical Interactive Video-Language Pre-Training, Shao+, ICCV2023

## 解像度を段階的に変化させてモデルに入力

- 従来手法は単一スケールの視覚特徴を使用し、細かい相互作用が不足
- 高解像度はきめ細かい特徴、低解像度はグローバルな特徴を表す

結果:

- テキストビデオ検索, ビデオテキスト検索, ビデオキャプションで最先端の性能
- ビデオとテキストのより良いアライメントを実現
- 視覚トークンを損なう可能性が課題



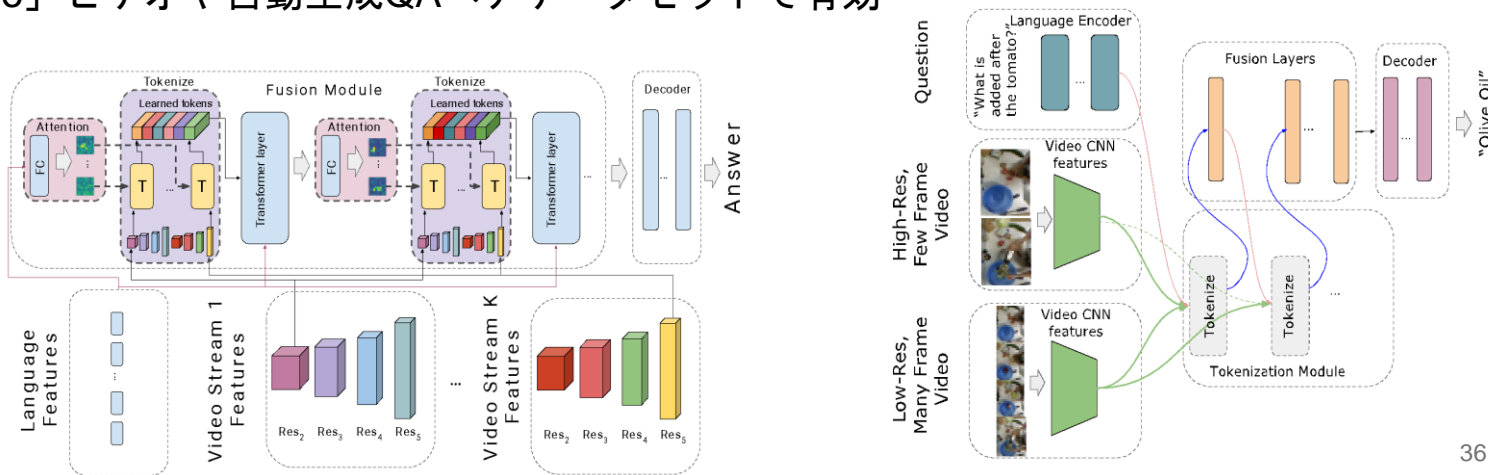
# Video Question Answering with Iterative Video-Text Co-Tokenization, Piergiovanni+, ECCV2022

## 2ストリームな解像度と、複数回の動画入力

- 動画とテキストの反復的な共同トークン化アプローチを新規に提案
- モダリティ間相互作用を重視するため、V-L表現の共同学習を行う

結果:

- MSRVTQ-QA、MSVD-QA、IVQAなど複数データセットで最先端手法を大きく上回る性能
- 「how-to」ビデオや自動生成QAペアデータセットで有効



# まとめと今後の展望

---

## 各階層的分類の課題

### 時間的な階層

- よりキャプションの時間的長さが多様でかつ、モダリティが広いデータセットが課題
- 生成AIを用いた多様なアノテーションの生成

### 空間的な階層

- トップダウン型
  - シーングラフを利用したVQAは発展途中
  - タスクごとに適したグラフの生成
- ボトムアップ型
  - 途中のグラフ要素と言語との対応付け
  - 予測のためにより有用なグラフ表現

### 時空間的な階層

- パッチやbboxと比較してより細かな局所領域の選択
- より効率的な階層的特徴の構成

階層性に限らず、動画中の複雑な視覚要素を効果的に表現する手法の検討は必要

# 少数サンプル学習・転移学習

# 少数サンプル学習・転移学習の重要性

- Video Recognitionの課題
  - データ収集+アノテーション → 動画のデータセット作成のコストは高い。少数サンプルからタスクを解けると、データ収集コストを削減できる。
- 少数サンプル（小規模データ）：数千規模のデータセット
- 少数サンプルの学習における懸念点
  - 「モデルの大規模化」による過学習
  - 「事前学習のバイアス」による性能低下

## 動画データ収集の限界

- 撮影のコスト
  - 数秒以上、撮影をし続ける必要がある
- アノテーションの人的コスト
- 不良サンプルが少ないタスク
  - 例：異常検知，医療診断

# モデルの大規模化

- 大規模化の要因
  - 動画の入力サイズ・特徴表現
  - スケーリング則\*
- パラメータ数が数十億規模の動画認識モデル\*\*
  - 事前学習のコスト
    - 学習時間：2週間以上
    - GPU (A 100)：64個
    - データ数：1.35M clips + 0.66 clips



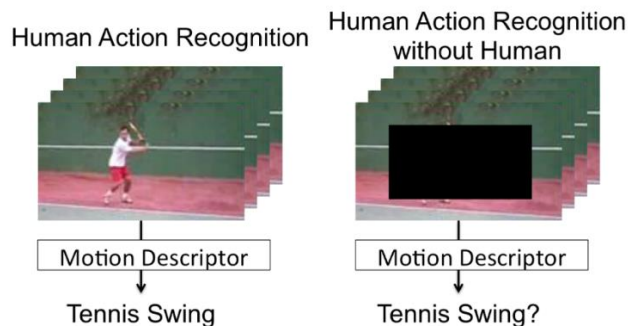
\*Zhai, Xiaohua, et al. "Scaling vision transformers." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

\*\*Wang, Limin, et al. "Videomae v2: Scaling video masked autoencoders with dual masking." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.



# 動画における空間バイアス

- 人（の動き）を見ず，背景のみでも分類できてしまう
  - 時間情報よりも空間情報が
- 事前学習に画像を使うと，空間情報に注目する傾向にある



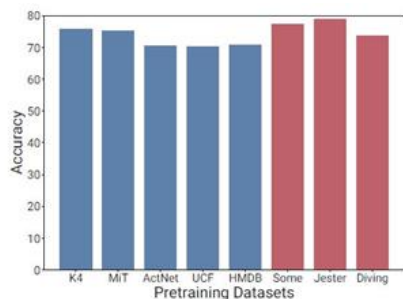
With or Without a Human	Stream	% on UCF101 (split 1)
With human	Spatial stream	51.26
	Temporal stream	40.50
	Two-stream	<b>56.91</b>
Without human	Spatial stream	45.33
	Temporal stream	26.80
	Two-stream	<b>47.42</b>

Table 2: Performance rate of human action recognition with or without a human

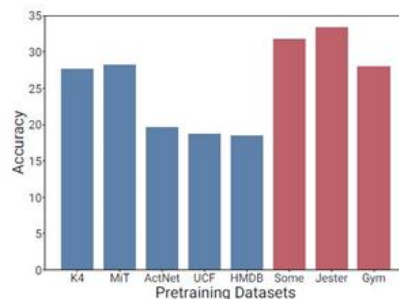
# 事前学習データの影響

- 一般行動認識・詳細行動認識それぞれのデータで事前学習
  - 青：一般行動認識のデータセットによる事前学習
  - 赤：詳細行動認識のデータセットによる事前学習
- ファインチューニングによる下流タスクへの影響
  - 一般行動認識：UCF-101
  - 詳細行動認識：Gym288, Diving48

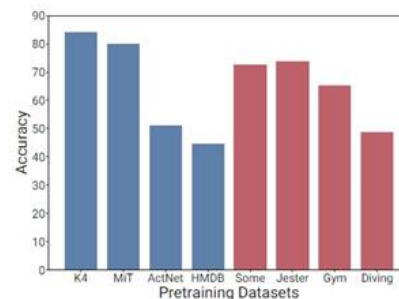
下流タスク（目的）に合わせた事前学習データセットが重要



(a) Gym288.



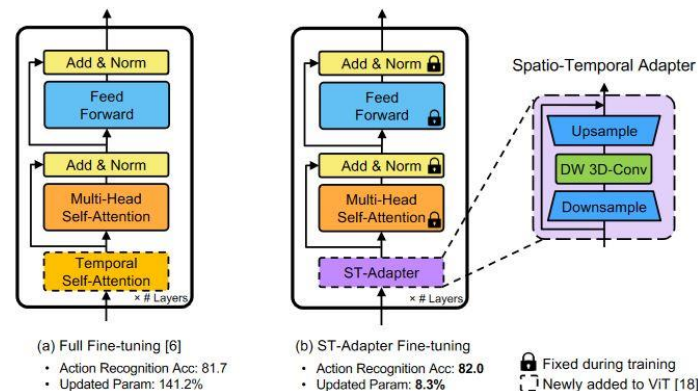
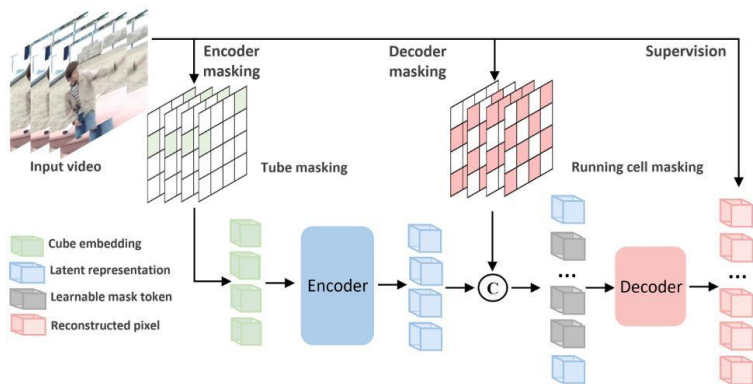
(b) Diving48.



(c) UCF-101.

# 問題点と解決策

- 問題点
  - 「モデルの大規模化」による過学習
  - 「事前学習のバイアス」による性能低下
- 解決策
  - Masked video modeling (MVM)
  - Parameter efficient transfer learning (PETL)



(a) Full Fine-tuning [6]  
• Action Recognition Acc: 81.7  
• Updated Param: 141.2%

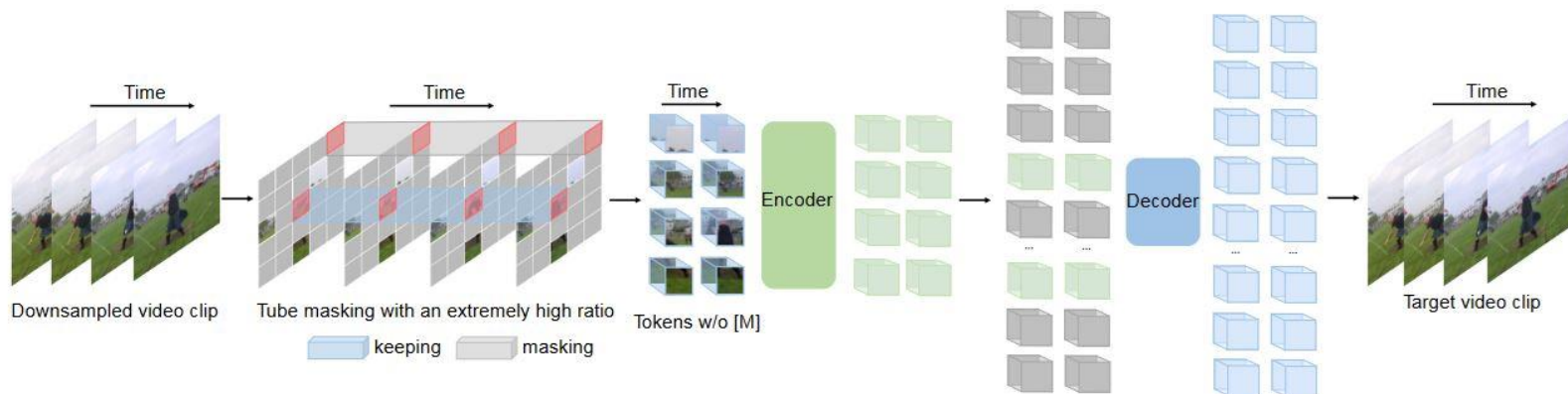
(b) ST-Adapter Fine-tuning  
• Action Recognition Acc: 82.0  
• Updated Param: 8.3%

\*Wang, Limin, et al. "Videomae v2: Scaling video masked autoencoders with dual masking." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

\*\*Pan, Junting, et al. "St-adapter: Parameter-efficient image-to-video transfer learning." *Advances in Neural Information Processing Systems* 35 (2022): 26462-26477.

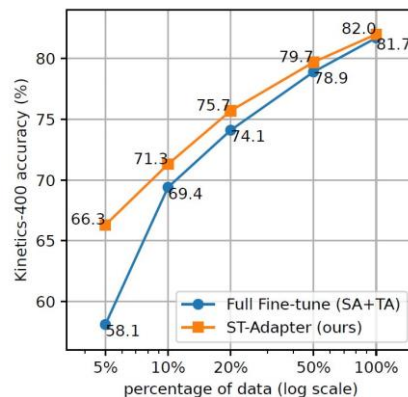
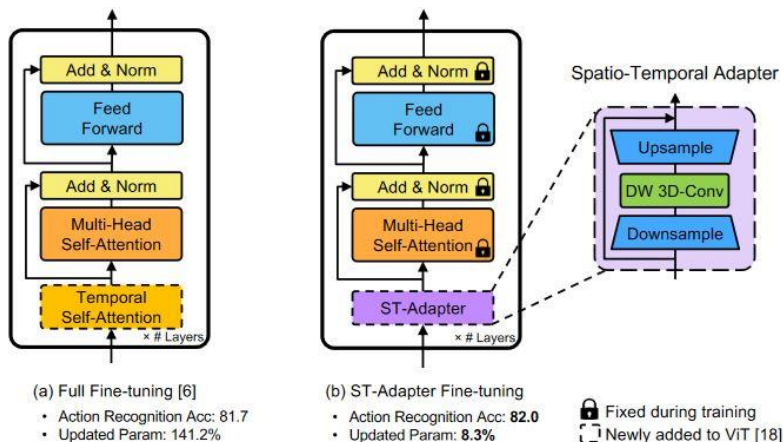
# Masked Video Modeling (MVM)

- 手法の概要
  - 自己教師あり学習による事前学習手法のひとつ
  - Encoder-decoderモデル
    - マスクされた部分を再構築するように学習
- 小規模データにおける利点
  - 単純な認識の学習ではないため、モデルの表現力が高い
  - マスクによるdata augmentation的な要素がある



# Parameter Efficient Transfer Learning (PETL)

- 手法の概要
  - 画像の基盤モデルなどをvideo recognitionタスクへ転移学習
  - Adapterなどを使い, パラメータ効率のよい学習
- 小規模データにおける利点
  - パラメータ数削減により, 小規模データでも学習可能に



# 今後の展開

- 事前学習の方法・データセットの変化
  - pre-training + post-pre-training
  - Video perlin noise (VPN) dataset\*による初期化
- アーキテクチャ
  - 数年はMVM系の手法が組み込まれる
  - 基盤モデルの活用

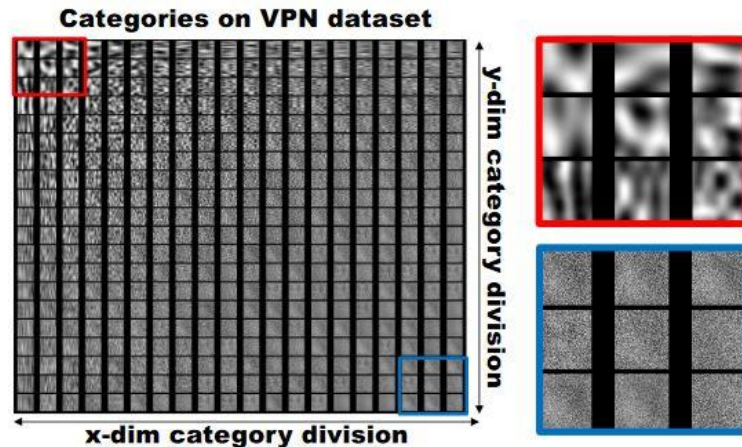


Figure 2. All categories in VPN dataset. The x and y coordinates correspond to the number of spatial divisions in  $(W, H)$ .

## 概要

- 行動認識において、3D CNNs (3D ResNet-50) が時間情報を活用しているか調査

## 手法

- 一般行動認識・詳細行動認識のデータセットで事前学習
- 時間情報を活用しているか確認するため、入力動画のフレームをシャッフル
- また、事前学習したモデルを別のデータセットでファインチューニングし比較

## 結果

- 時間情報の活用：一般行動認識の学習 < 詳細行動認識の学習
- 事前学習のデータセットが下流タスクの精度に影響
  - 詳細行動認識の下流タスク：詳細行動認識で事前学習した方が精度が高い傾向

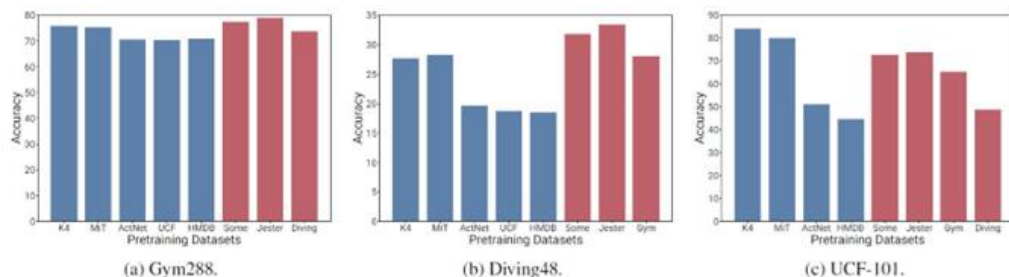


Figure 2: Finetuning 3D ResNet-50 on Gym288, Diving48, and UCF-101. Blue and red colors indicate generic and fine-grained action recognition datasets, respectively. K4, MiT, ActNet, UCF, HMDB, Some, Gym, and Diving means Kinetics-400, Moments in Time, ActivityNet, UCF-101, HMDB-51, Something-Something v2, Gym288, and Diving48, respectively.

Table 1: Training 3D ResNet-50 from scratch. *Something v2* indicates Something-Something v2. Larger *Log ratio*, which is the log ratio between the accuracies on original and shuffled inputs, indicates the smaller representation.

Dataset	Accuracy [%]		Log ratio
	Original inputs	Shuffled inputs	
Kinetics-400	64.3	32.7	0.68
Moments in Time	27.1	14.2	0.65
ActivityNet	33.4	29.6	0.12
UCF-101	45.3	32.0	0.35
HMDB-51	19.4	11.0	0.56
Something v2	44.3	2.3	2.95
Jester	92.0	20.1	1.52
Gym288	71.7	3.9	2.92
Diving48	15.1	4.2	1.29

## 概要

- ImageMAEをVideoに適用させるための研究
  - VideoMAE：動画の基盤モデル構築における自己教師あり事前学習器

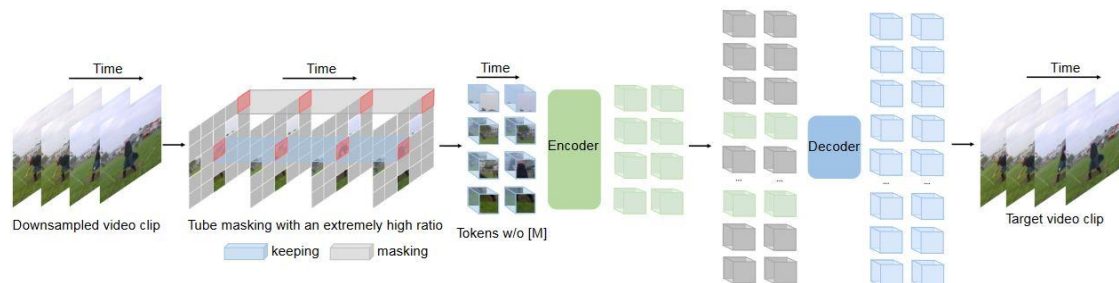
## 手法

- Tubeというマスク方法を利用
  - 全時間（フレーム）において同位置をマスク
  - 前後のフレームからの予測を避ける

## 結果

- マスク率90%~95%でも性能高：動画の冗長性の影響
- 小規模なデータ（3.5k~4.0k）でも学習性能向上

dataset	training data	from scratch	MoCo v3	VideoMAE
K400	240k	68.8	74.2	<b>80.0</b>
Sth-Sth V2	169k	32.6	54.2	<b>69.6</b>
UCF101	9.5k	51.4	81.7	<b>91.3</b>
HMDB51	3.5k	18.0	39.2	<b>62.6</b>





## 概要

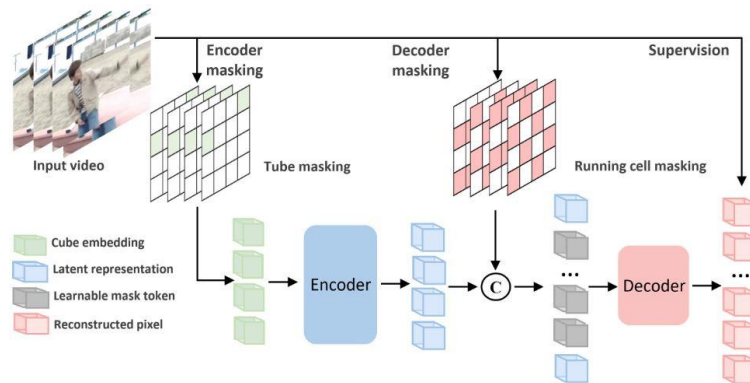
- VideoMAEの進化版
  - エンコーダ : information leakageの緩和
  - デコーダ : information complementの促進

## 手法

- デコーダ内のマスキング (cell masking : ビデオの情報全体をカバー)
- Intermediate fine-tuningを踏襲したprogressive training
  - 数百万規模のラベルなしビデオデータによるマスクありpre-training
  - ラベル付きハイブリッドデータ (1.35M clips) によるpost-pre-training
  - ターゲットデータセットによるspecific fine-tuning

## 結果

- デコーダのマスキングにより計算速度向上
- Progressive training : Kinetics400向上, Sth-Sth V2悪化



## 概要

- マスクされた特徴の予測を利用した二つの教師によるmasked video distillation (MVD) の提案

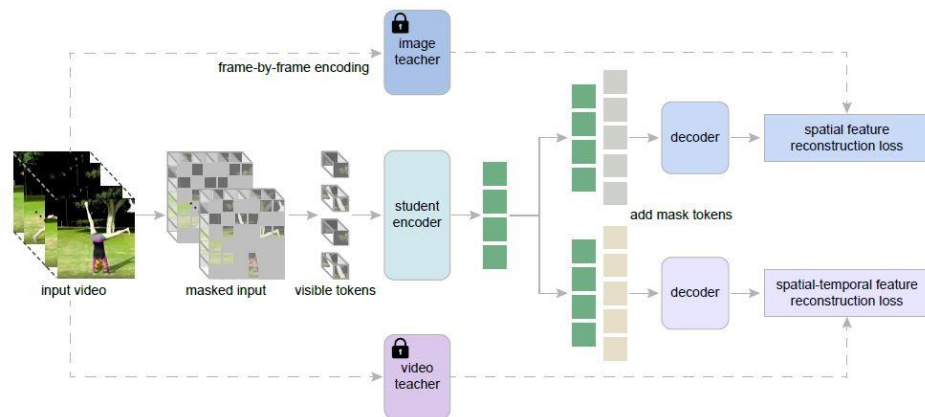
## 手法

- 低レベルではなく、高レベルの特徴を利用したマスク特徴モデリング
- 事前学習されたMIM, MVMモデルをマスク予測のターゲットに利用
  - Image teacher : MIMによる空間特徴の学習
  - Video teacher : MVMによる時空間特徴の学習

## 結果

- 画像・動画単一で使った場合と両方使った場合であまり変わらない (1%前後の差)
  - Kinetics, SSv2のtop1で評価
- VideoMAEより精度高

method	extra data	Param	UCF101	HMDB51
VideoMoCo R2+1D [46]	K400	15	78.7	49.2
MemDPC R2D3D [30]	K400	32	86.1	54.5
Vi <sup>2</sup> CLR S3D [11]	K400	9	89.1	55.7
CORP Slow-R50 [34]	K400	32	93.5	68.0
CVRL Slow-R50 [49]	K400	32	92.9	67.9
CVRL Slow-R152 [49]	K600	328	94.4	70.6
$\rho$ BYOL Slow-R50 [23]	K400	32	94.2	72.1
VIMPAC ViT-L [55]	HowTo100M	307	92.7	65.9
VideoMAE ViT-B [57]	K400	87	96.1	73.3
MVD-B (Teacher-B)	1N-1K+K400	87	97.0	76.4
MVD-B (Teacher-L)	1N-1K+K400	87	97.5	79.7



## 概要

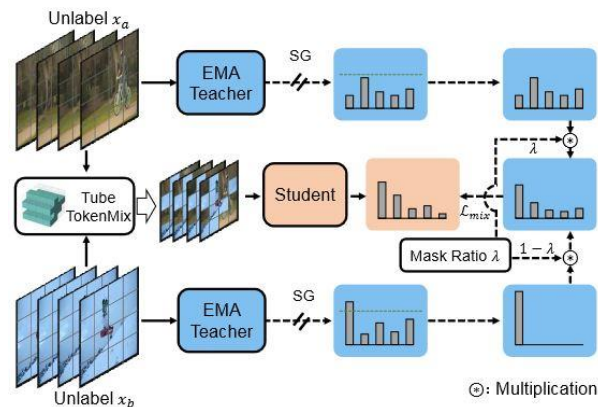
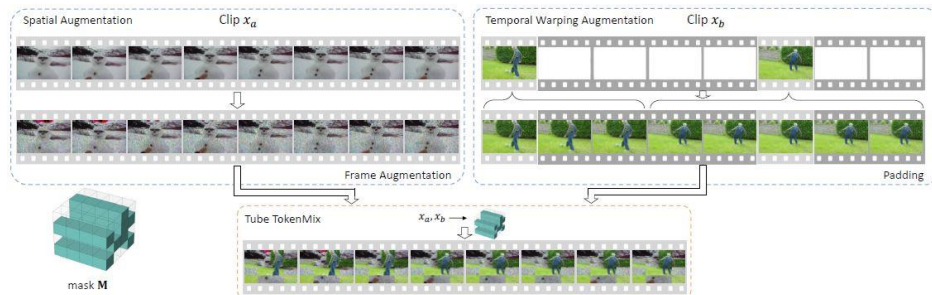
- Transformerベースの半教師あり学習方法SVFormerの提案
- トークンレベルのaugmentation方法 Tube TokenMix (TTMix) の提案

## 手法

- EMA Teacherによる疑似ラベルの生成
- Augmentation方法として, TTMix・Temporal Warping Augmentation (TWAug) の組み合わせ
  - TTMix: 時空間の相関をモデル化
  - TWAug: フレーム間の時間的変化をカバー

## 結果

- K400, UCF101, HMDB51でCNN系のSSLを上回る結果
- TokenMixでTubeが最も良いことを確認



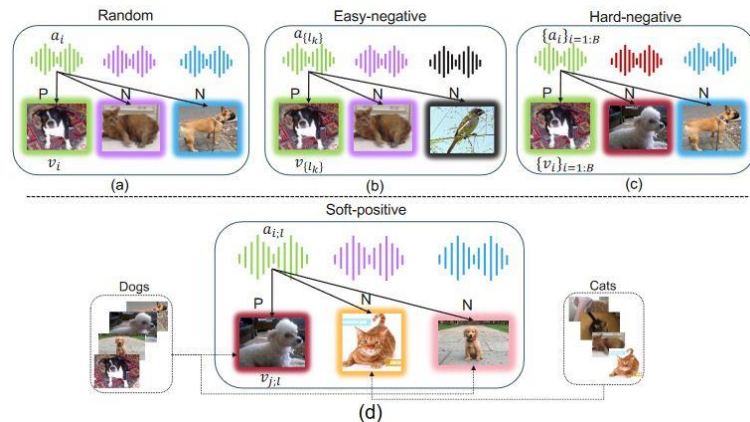
## 概要

- SSLにおけるAudio-Visual Correspondence (AVC) タスクにおけるサンプリング方法に関する研究
  - Soft-positive samplingの提案
- サンプリング方法の違いによる学習の比較
  - SSLにおいて, negative samplingだけでなくpositive samplingも重要では？

## 結果

- 下流タスクのデータが
  - 大きい場合：ランダムサンプリングでも特徴表現は失われない
  - 小さい場合：soft-positive samplingが効果的

- Pretextタスクのデータバランスは重要
  - Sampling strategyの効果が小さくなる



dが提案手法  
soft-positive  
sampling

## 概要

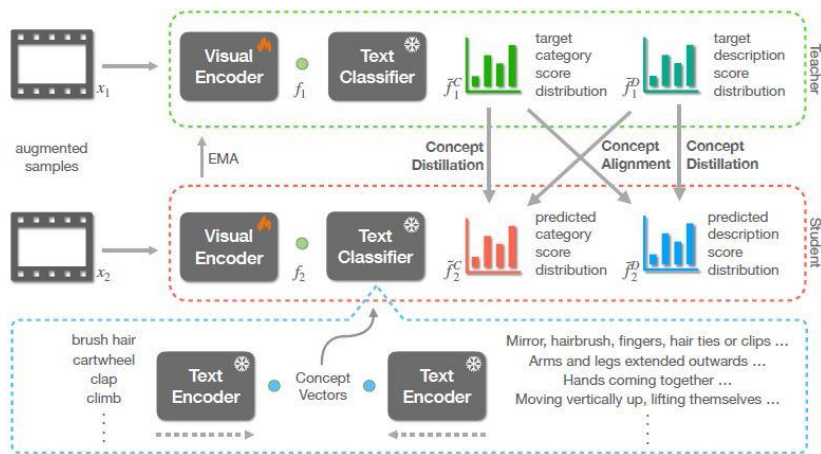
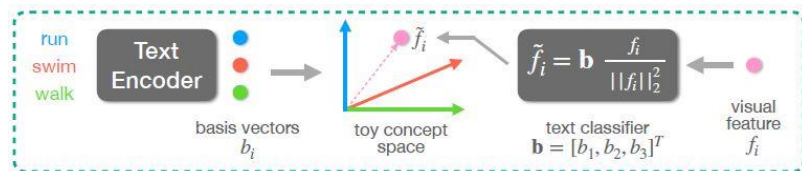
- 画像のCLIPモデルの動画ドメインに適用させる研究
- Concept distillationとconcept alignmentによる動画のSSL

## 手法

- Concept space : textの特徴表現とvideoの特徴表現から生成
  - Category concept space : 動画データカテゴリのラベルセットの抽出
  - Description concept space : LLMによる動画カテゴリの説明抽出
- Concept distillation : concept spaceの蒸留
- Concept alignment
  - Categoryとdescriptionそれぞれを付加情報に

## 結果

- K400, UCF101, HMDB51で検証
  - Linear probing・zero-shotで高精度
- オブジェクトレベルの動作認識は課題
  - Image CLIPの影響 (フレームの静的情報に焦点)



## 概要 :

- 6つのVSSLアルゴリズムのdistribution shiftについて調査した最初の論文

## 結果 :

### Q1. 特徴表現の違い・Distributionシフトに対する頑健性について

- 結果はスライド右画像

### Q2. Finetuning・zero-shotへの影響

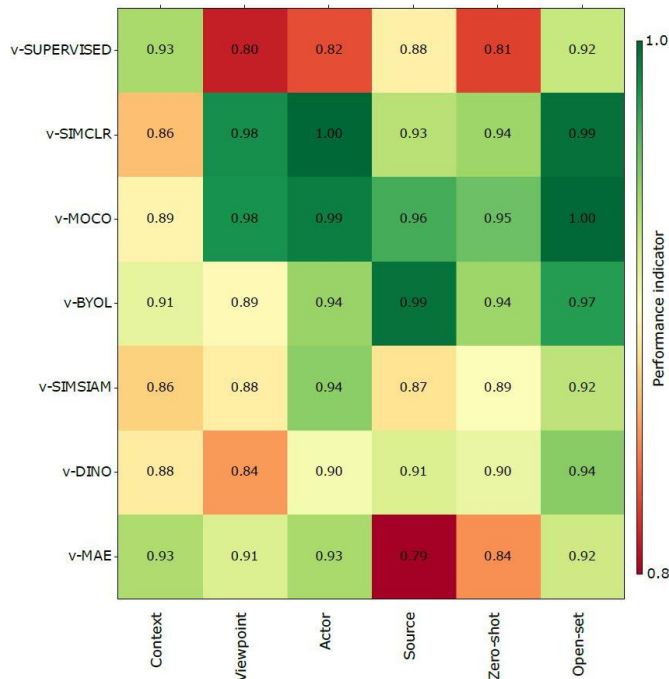
- Actor shift : 効果的, viewpoint shift : 効果的でない

### Q3. オープンセット問題におけるパフォーマンス

- Finetuning時のContrastive methodsが優位
- Closed-setとopen-setはトレードオフ
  - 事前学習モデルをfrozenして使った場合

### Q4. 意思決定の類似性 (同条件下での意思決定パターン)

- 予測の正確さではなく, 予測の類似性に注目
- Shiftタイプによって異なるが, actor shift以外は類似性低下



# ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning (NeurIPS'22)

## 概要：

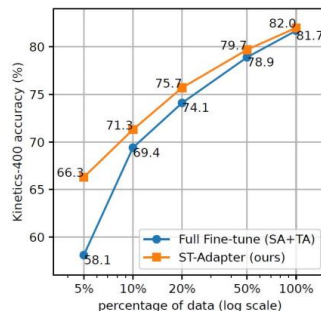
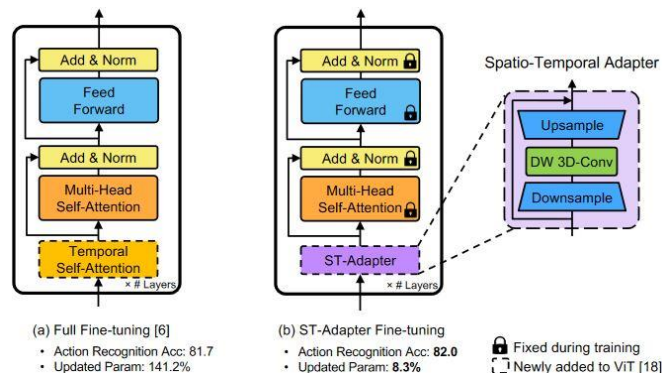
- 画像から動画へのクロスモダリティ転移学習
- パラメータ効率の良いファインチューニングのためのspatial-temporal adapter (ST-Adapter) の提案

## 手法：

- ST-Adapterにはdepth-wise-3D-convolution (DW 3D-Conv) を採用
  - パラメータサイズの抑制・開発のしやすさ

## 結果：

- 様々なfine-tuning方法と比較
  - 行動認識, 一人称視点動画で検証
  - CLIPの事前学習モデルでは, どの方法よりも精度高
- CLIPの事前学習モデルでfull fine-tuningと同精度
- 更新するパラメータ数を大きく削減 (7.2M vs. 121.6M)
- 小規模データでfull fine-tuningより有効



# Dual-path Adaptation from Image to Video Transformers (CVPR'23)

## 概要：

- 画像の基盤モデル→動画認識のクロスドメイン転移学習

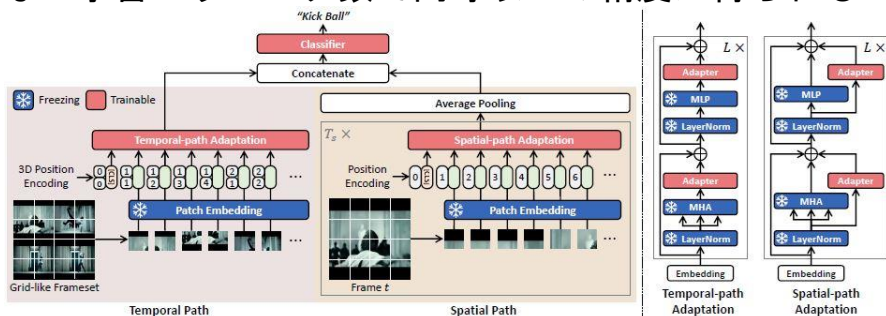
## 手法：

- Spatial adapter : MHA, MLPブロックと並列にadapterを接続
  - 画像基盤モデルの空間モデリング能力の活用
  - 並列adapterによる動画のための空間コンテキストの学習
- Temporal adapter : MHA, MLPブロックの上に接続
  - 3D位置エンコーディングによるパッチの時間順序・空間位置の考慮
  - Grid-likeなフレームセットへの変換によるフレーム間の関係を取得

## 結果：

- 定性的に時間コンテキストを学習できているか確認
- 四つのデータセットで少ない学習パラメータ数で同等以上の精度が得られることを確認

Method & Arch.	Classifier	Params	HMDB51
Full-tuning w/ ViT-B/16 [17]	Lin.	86M	59.4
Linear w/ ViT-B/16	Lin.	0.1M	61.2
VPT [35] w/ ViT-B/16	Trans.	7M	62.4
AdaptFormer [11] w/ ViT-B/16	Trans.	8M	63.7
Pro-tuning [49] w/ ViT-B/16	Trans.	9M	63.3
VideoPrompt [37] w/ ViT-B/16	Trans.	6M	66.4
ST-Adapter* [50] w/ ViT-B/16	Lin.	7M	65.9
<b>DUALPATH w/ ViT-B/16</b>	<b>MLPs.</b>	<b>10M</b>	<b>75.6</b>





# BEVT: BERT Pretraining of Video Transformers (CVPR'23)

## 概要：

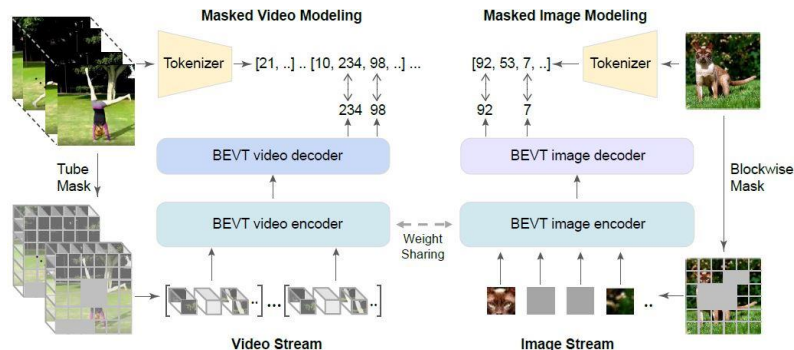
- BERTの学習方法をもとに設計されたBEVTの提案
  - 空間依存性・時間依存性それぞれに対応するtwo-stream型ネットワーク

## 手法：

- MIM, MVMによる空間表現の学習と時間ダイナミクスの学習の分離
  - 空間表現の学習：画像データのみ
  - 時間ダイナミクスの学習：動画データのみ
- エンコーダ間の重みの共有
  - 空間表現の学習：画像データのみ

## 結果：

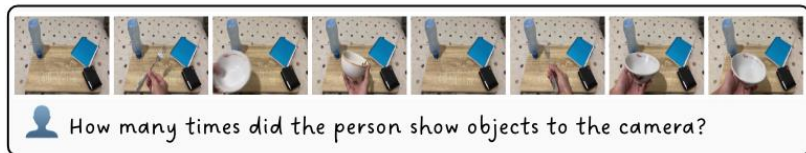
- K400, SSv2, Diving48による評価
  - 各データセットで比較手法と同等以上の精度
- シングルフレーム, フレームシャッフルによる検証
  - 空間情報が支配的なタスク場合：video streamの効果小
  - 時間情報が重要なタスクの場合：video streamの効果大
- 画像による事前学習の重要性も確認



# *Long-form Video Understanding*

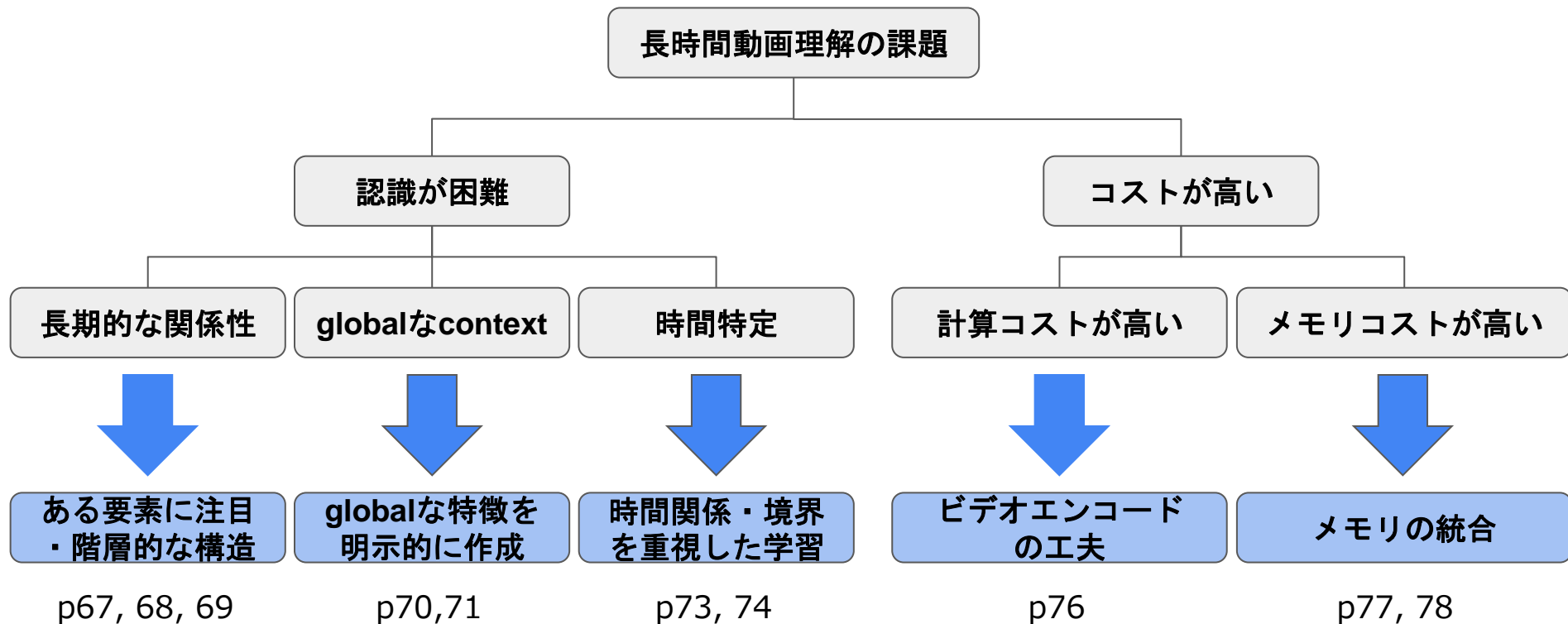
# 長時間動画理解とは（再掲）

- **目的** 数分から数時間にわたるデータ (EPIC-KITCHENSは約4秒) において、一部の時間帯ではなく、長時間のcontextを捉えた応答を行う  
→ 長時間のcontextを捉えないと回答できないBenchmarkの作成(p62,63)
- **タスクの例**
  - Count: 動画中で何回actionや動きがあったかを数える
    - 動画の一部だけではなく、全体を見ないと回答不可
  - Action Prediction: 現在までの行動に基づいて将来の行動を予測する
    - 直前の行動だけではなく、今までの行動を捉えないと回答不可



引用 : <https://arxiv.org/abs/2311.17005>

# 長時間動画理解の課題と解決策



# 1. 長時間動画理解のBenchmark

2. 認識の困難さに対応する研究

3. 認識のコストに対応する研究

# EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding[Mangalam+, NIPS2023]

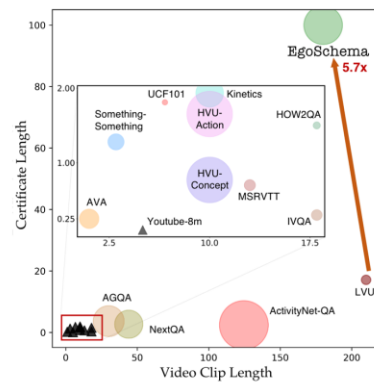
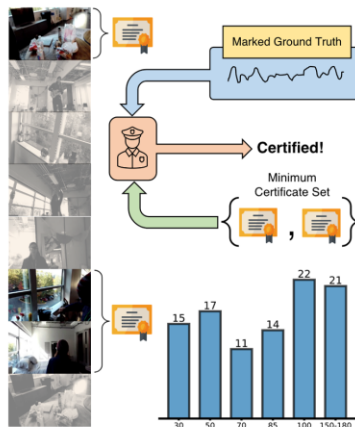
**概要:** 長時間のビデオを理解するためのQAベンチマークを作成した。Ego4Dデータセットから派生した多肢選択式のQAで、各質問は3分間のビデオクリップに基づいて、5つの選択肢から正しい回答を選ぶ

**新規性:** ビデオクリップの長さではなく、タスクを正しく理解するために必要十分なサブクリップの長さ (time certification) で動画の長さを評価し、他のデータセットに対し、2番目のデータセットに比べて5.7倍長いEgoSchemaを構築

**結果:** 人間が76%の精度をEgoschemaで達成するものの、最新のビデオモデルですら33%未満の精度となり、長時間映像理解の問題点を明らかにした

**Example 1:** Question: "What is the overarching behavior of C and the man in the video?"  
Options:  
1. C teaches the man game rules but the man seems distracted and is not paying attention  
2. The man teaches C how to play the card game while organizing the deck for future games  
3. C and the man are playing a card game while keeping track of it in a notebook  
4. C shows the man how to properly shuffle cards while the man plays them  
5. The man shows C a new card game while C takes notes for future reference

**Example 2:** Question: "Observe the video in terms of characters' actions and interactions. How do these shifts contribute to the overall narrative?"  
Options:  
1. The video displays a profound sense of conflict and tension arising between the characters  
2. The man is showing C the issues that need fixing in the apartment in a professional manner  
3. Both the characters display an increasingly urgent need to solve an issue in the apartment  
4. C and the man admire and interact with several objects in the apartment that look beautiful.  
5. Actions and interactions are causal and relaxed, reflecting a comfortable environment.



# MVBench: A Comprehensive Multi-modal Video Understanding Benchmark

[Li+, CVPR2024]

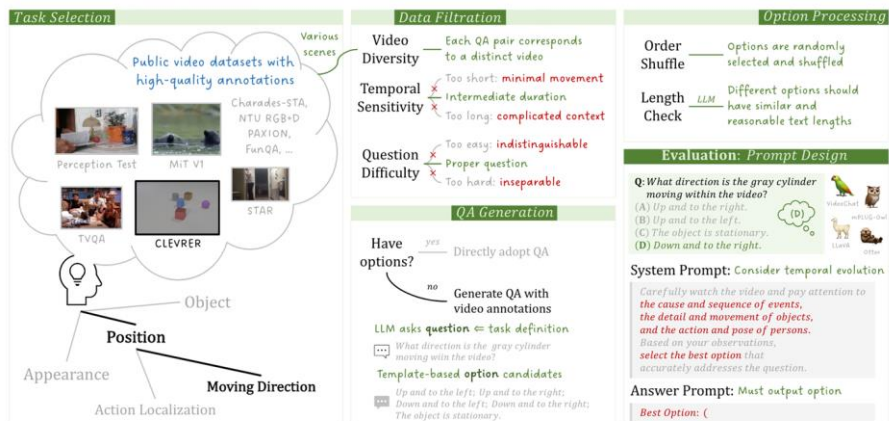
**概要**：動画中の時間経過や全体を考慮する性能を評価するベンチマークの作成

**新規性**：静的画像タスクを動的ビデオタスクに変換、複数の公的なビデオベンチマークからアノテーションを自動的に変換し、アノテーション負荷を削減しながらベンチマークを作成

**手法**：新ベンチマークの設計とそのベンチマークで高性能を実現するVLLMの提案

静的な画像タスクを動画タスクに変換(“Is the man on the stage?” → “What direction is the man moving?”), 自動アノテーション生成を行うことでアノテーションコストを削減

**結果**：提案モデルは最先端モデルと比較して、MVBenchの評価タスクで平均15%以上の精度向上を達成



Spatial	Temporal	Source	Example
Action Sequence	STAR	STAR	What happened after the person took the food? (A) Ate the medicine. (B) Tied up the blanket. (C) Put down the cup/glass/bottle. (D) Took the box.
Action Prediction	STAR	STAR	What will the person do next? (A) Put down the pillow. (B) Open the door. (C) Take the book. (D) Open the closet/cabinet.
Action Antonym	PAXION	PAXION	Which one of these descriptions correctly matches the actions in the video? (A) not sure (B) scattering something down (C) piling something up
Fine-grained Action	MIT V1	MIT V1	What is the action performed by the person in the video? (A) watering (B) leaking (C) pouring (D) planting
Unexpected Action	FunQA	FunQA	What unexpected event contributes to the humor in the video? (A) The man left without dancing. (B) Two women hugged each other at the end. (C) The man finally danced with the woman. (D) Two men hugged each other unexpectedly.
Object Existence	CLEVRER	CLEVRER	Are there any moving green objects when the video ends? (A) not sure (B) yes (C) no
Object Interaction	STAR	STAR	Which object was tidied up by the person? (A) broom (B) cabinet (C) blanket (D) table
Object Shuffle	Perception Test	Perception Test	Where is the hidden object at the end of the game from the person's point of view? (A) Under the first object from the left. (B) Under the third object from the left. (C) Under the second object from the left.
Moving Direction	CLEVRER	CLEVRER	What direction is the cyan sphere moving within the video? (A) The object is stationary. (B) Up and to the right. (C) Down and to the left. (D) Down and to the right.
Action Localization	Charades-STA	Charades-STA	During which part of the video does the action 'person sitting on a couch' occur? (A) In the middle of the video. (B) At the end of the video. (C) Throughout the entire video. (D) At the beginning of the video.
Scene Transition	MoVQA	MoVQA	What's the right option for how the scenes in the video change? (A) From the reception desk to the conference room. (B) From the kitchen to the dining area. (C) From the server room to the control center. (D) From the classroom to the library.
Count	Action Count	Perception Test	How many times did the person launch objects on the table? (A) 3 (B) 2 (C) 4
Moving Count	CLEVRER	CLEVRER	How many metal objects exit the scene? (A) 2 (B) 3 (C) 1 (D) 0
Attribute	Moving Attribute	CLEVRER	What shape is the moving object when the video begins? (A) cylinder (B) sphere (C) cube
State Change	Perception Test	Perception Test	Is the lighting device on at any point? (A) Yes (B) I don't know (C) no
Pose	Fine-grained Pose	NTU RGB-D	What is the pose performed by the person in the video? (A) pack up (B) sit down (C) drop (D) stand up
Character	Character Order	Perception Test	What letter did the person write first on the paper? (A) I (B) C (E)
Epocentric Navigation	VLN-CE	VLN-CE	For an agent following instruction: "Go left through the door." What is the next action it should take? (A) Turn left and move forward (B) Move forward (C) Stop (D) Turn right and move forward.
Who did	Episodic Reasoning	TVQA	Who did Emily dress like a fairy when he was speaking to Emily? (A) To get her to trust him. (B) He secretly loved fairies. (C) He lost a bet with Emily. (D) It was dressed like a fairy day at school. (E) Mrs Ruiz made him dress up.
Cognition	Counterfactual Inference	CLEVRER	Which of the following will happen if the cylinder is removed? (A) The cyan rubber object and the blue cube collide. (B) The brown cube collides with the metal cube. (C) The cyan rubber object and the metal cube collide. (D) The cyan rubber cube collides with the sphere.

Figure 2. Generation pipeline of MVBench. Within public annotations, data is carefully filtered and relevant multiple-choice QAs are auto-generated. The effective system prompt and efficient answer prompt are employed to guide LLMs toward precise outputs.

まとめBY: 木林

1. 長時間動画理解のBenchmark
- 2. 認識の困難さに対応する研究**
3. 認識のコストに対応する研究



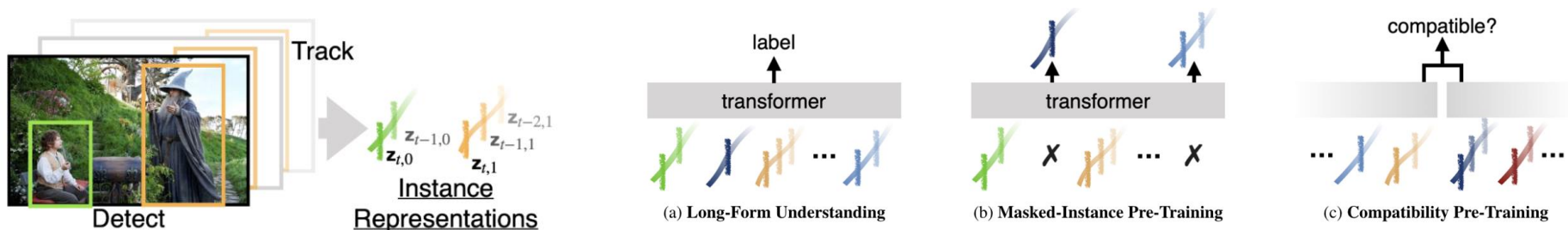
長期的な関係性やグローバルなcontextの活用

**概要：** オブジェクトに着目したトランスフォーマーベースのビデオ認識アーキテクチャと自己教師あり学習を導入することで、7つの多様なタスクでの性能向上を実現

**新規性：** ピクセル全体ではなく、トラッキングされたオブジェクトの軌跡について相互作用をモデリング。大規模データセットでの評価プロトコルを開発し、既存手法と比較して評価。

**手法：** トラッキングされたオブジェクトの軌跡をトランスフォーマーを用いて学習。オブジェクトをマスクし、マスクされた軌跡を予測したり、2つの表現が互換可能かを分類するpre-text taskを考案（BERTから）

**結果：** 9つの長期間ビデオタスクのほとんどで当時の最先端手法を上回る。SlowFastなどの短期的な動画に対するモデルと比較して、より優れたパフォーマンスを示し、オブジェクト間の相互作用のモデリングが重要であることを示唆



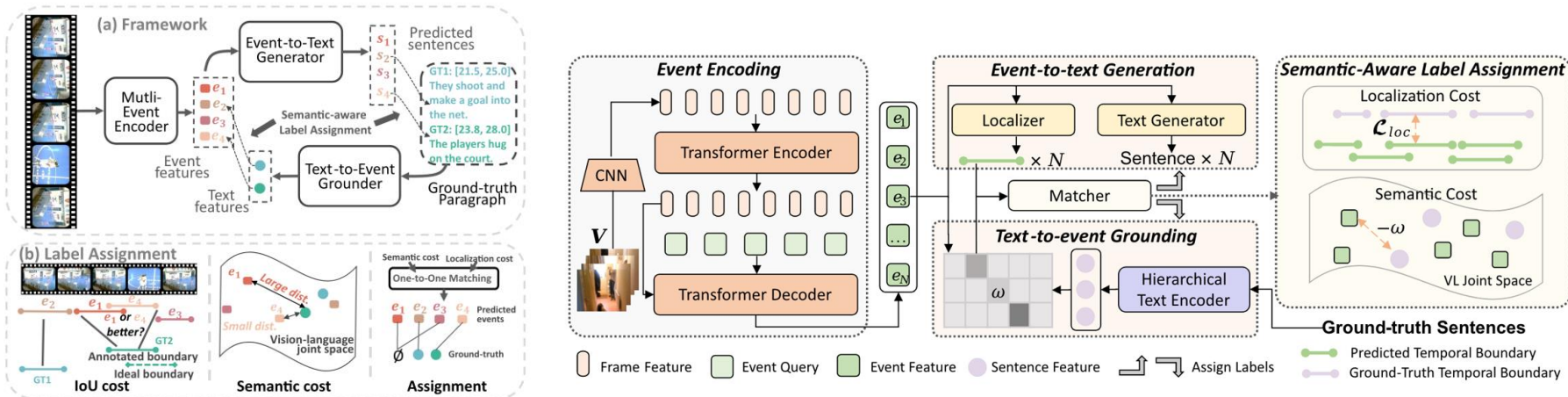
# Learning Grounded Vision-Language Representation for Versatile Understanding in Untrimmed Videos[Wang+, arxiv2023]

**概要：** 未編集の長時間動画から自動的に情報イベントを検出し、複数の文説明と対応するイベントセグメントとの間のアライメントを効果的に行い、Dense Video Captioningにて高性能を達成

**新規性：** 自動イベント検出を行い、イベントとテキストのマッチングを行う pretext task 曖昧な境界アノテーションによって引き起こされる最適でないマッチング結果の軽減

**手法：** ビデオからイベントレベルの特徴を抽出するエンコーダを設計し、文章クエリから対応するイベントを見つけるTEGタスクとイベントセットからテキストを生成するETGタスクを提案。イベント同士の対応はセマンティック空間での距離を用いて行うことで曖昧な時間境界アノテーションの影響を軽減

**結果：** キャプション生成、video groundingにおいて当時のSOTAを達成



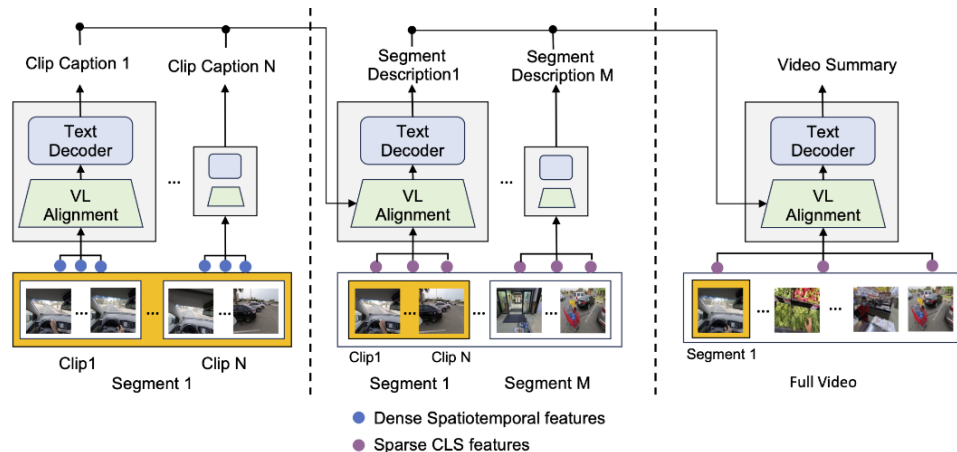
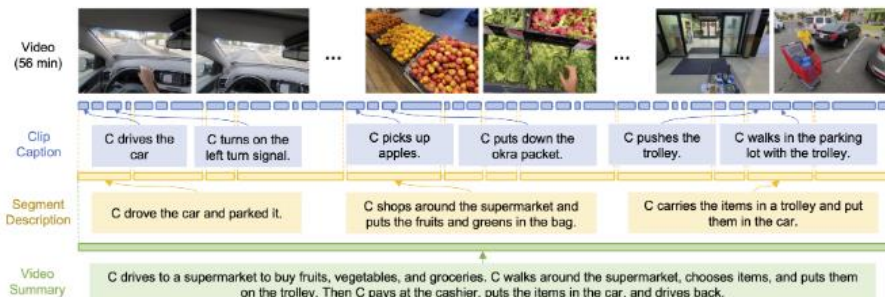
# Video Recap: Recursive Captioning of Hour-Long Videos[Islam+, CVPR2024]

**概要：** 現実世界のビデオが有する時間的粒度の階層構造を考慮した階層的なキャプションモデルを提案

**新規性：** LLMを使用して生成された擬似要約データを追加データとして使用、異なる階層レベルのキャプションを提供するデータセットEgo4D-HCapデータセットの導入、再帰的なキャプション生成

**手法：** 異なる階層レベルのキャプションを生成するための再帰的なアプローチを採用、低レベルのキャプションから始め、徐々に高レベルのキャプションへと学習

**結果：** すべての階層レベルでベースラインを大幅に上回る性能(CIDEr, ROUGE-L, METEOR)、LLMによる擬似注釈データの有効性も確認 生成された階層的キャプションは、EgoSchemaデータセット上での質問応答タスクにおいても優れた性能を発揮



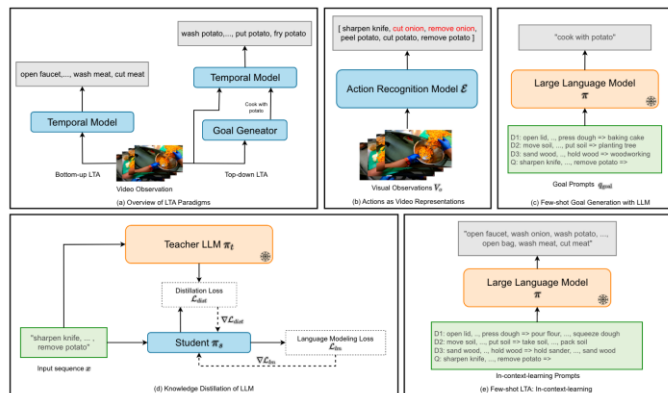
# AntGPT: CAN LARGE LANGUAGE MODELS HELP LONG-TERM ACTION ANTICIPATION FROM VIDEOS?[Zhao+, ICLR2024]

**概要：** 長期行動予測 (LTA) タスクにおいて、大規模言語モデル (LLM) の活用を提案。LLMを用いて時間的ダイナミクスをモデリング。

**新規性：** LLMを使用して、手順テキストデータから得られる事前知識を利用して推論を行う

**手法：** 動画セグメントをアクション認識モデルを使用してアクションラベルに変換し、LLMを用いてICLでgoalを生成。goalを条件として未来のアクションシーケンスをLLMで予測するTop-down LTAとgoalを用いないBottom-up LTAを組み合わせる。

**結果：** Ego4DLTAなどで従来の手法と比較して優れたパフォーマンスを示し、LLMの事前知識が長期行動予測に有用であることが確認された



The screenshot shows the interface with GT Observations and Future Actions. Below, it displays LLM prompts and answers for different tasks:

(a) Predict Goal  
Instruction: "What is the most possible goal according to given previous actions?"  
D1: take fork, wash fork, put fork, take knife, wash knife, take spoon, wash spoon, put spoon => clean tableware  
Q: put stick, put stick, pull string, paint wall, paint wall, paint paint, paint wall =>  
[Answer]: "painting a room"

(b) Bottom-up: Predict Future  
Instruction: "What will be the possible next 20 actions according to given previous actions? You should follow the following rules..."  
D1: lie leaf, carry leaf, put plant, adjust cloth, take plant, take sickle, take sickle => cut plant, put sickle, ..., cut plant, cut plant, hold plant  
Q: put stick, put stick, pull string, paint wall, paint wall, paint paint, paint wall =>  
[Answer]: "clean brush, dip brush, ..., paint wall, ..., paint wall, paint wall, clean brush"  
Verb\_ED = 0.55  
Noun\_ED = 0.95

(c) Top-down: Predict Goal and Future  
Instruction: "What will be the possible next 20 actions according to given previous actions? You should follow the following rules..."  
D1: lie leaf, carry leaf, put plant, adjust cloth, take plant, put plant, take sickle, take sickle Q1: What's the goal according to previous actions? Q2: What are the future 20 actions? => The goal is gardening. Future 20 actions are: cut plant, put sickle, ..., cut plant, cut plant, hold plant  
Q: put stick, put stick, pull string, paint wall, paint wall, paint paint, paint wall Q1: What's the goal according to previous actions? Q2: What are the future 20 actions? =>  
[Answer]: "The goal is painting a room. Future 20 actions are: dip brush, paint wall, ..., dip brush, ..., paint wall, dip brush, paint wall"  
Verb\_ED = 0.93  
Noun\_ED = 0.60

# Koala: Key frame-conditioned long video-LLM[Tan+, CVPR2024]

**概要:** 短時間のvLLMsを長時間のビデオに対応させるために

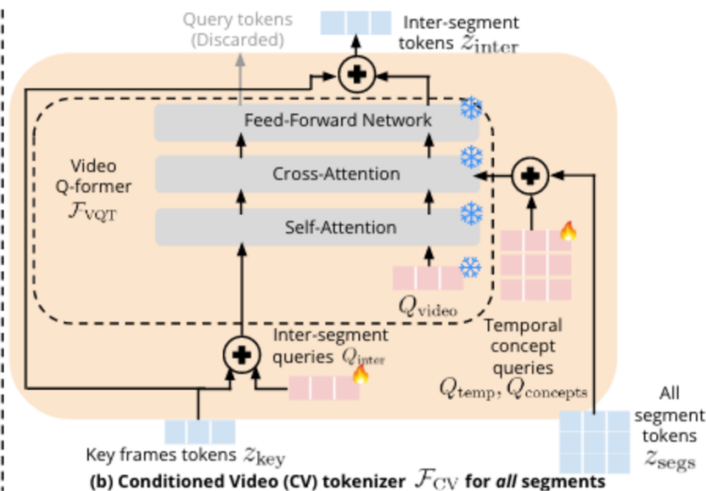
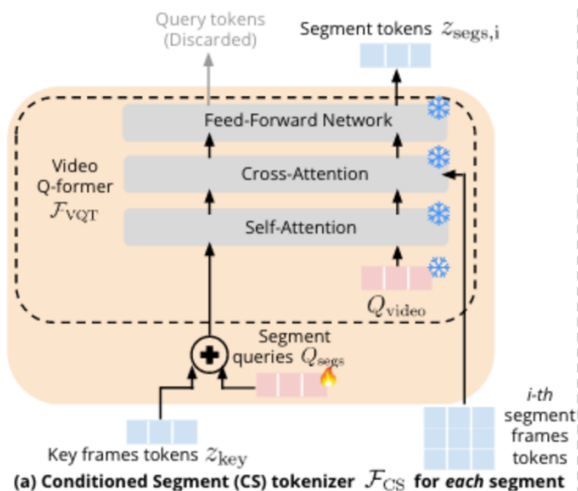
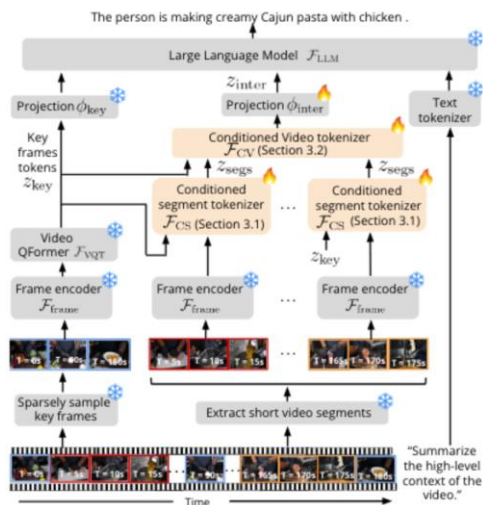
**新規性:** 疎にサンプリングされたキーフレームを使用し、グローバルコンテキストを捉え、vLLMsを条件付

**手法:** 2つの新しいトークナイザを導入、

Conditioned Segment (CS) トークナイザ: グローバルコンテキストに基づいて各フレームをエンコード

Conditioned Video (CV) トークナイザ: セグメント間の文脈関係を推論しトークン化

**結果:** 長時間VQAや手順理解、行動認識タスクにおいて既存のvLLMsを上回る性能



時間特定能力の向上

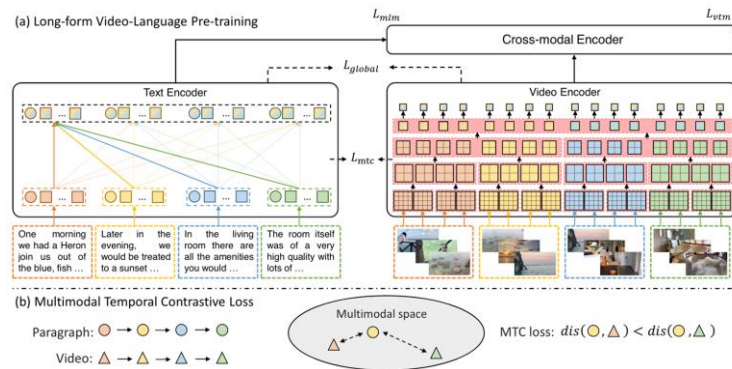
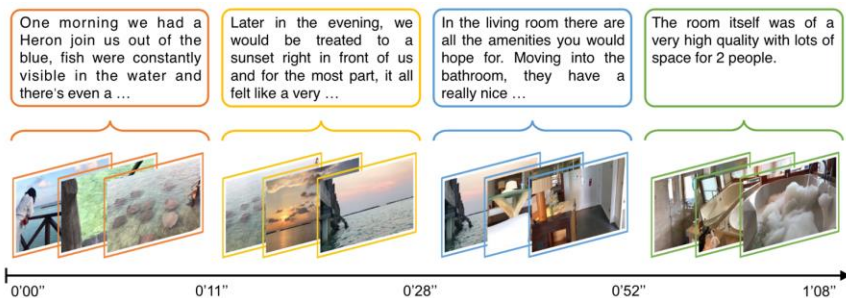
# Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning [Sun+, NIPS2022]

**概要:** ビデオとテキストを共同で学習させるモデル(LF-VILA)を提案

**新規性:** ビデオクリップとテキストの時間的関係を学習させるために、時間的に近い2つを射影空間上でも近くするContrastive Loss(MTC Loss)、計算コストの削減と離れた時間帯の関係性を学習するためにTransformerの時間窓を徐々に拡大するAttention(HTWA)

**手法:** 提案したLossとAttentionを用いてBERTを模したPre-trainingを行う

**結果:** 長時間動画から特定のParagraphに対応する部分を検索するタスクにおいて従来のSOTAモデルをR@1で大幅に更新。VQAやClassificationでも高性能。





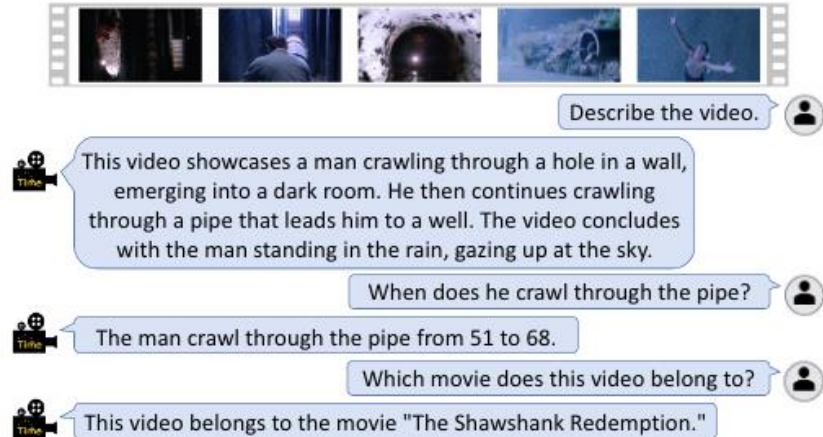
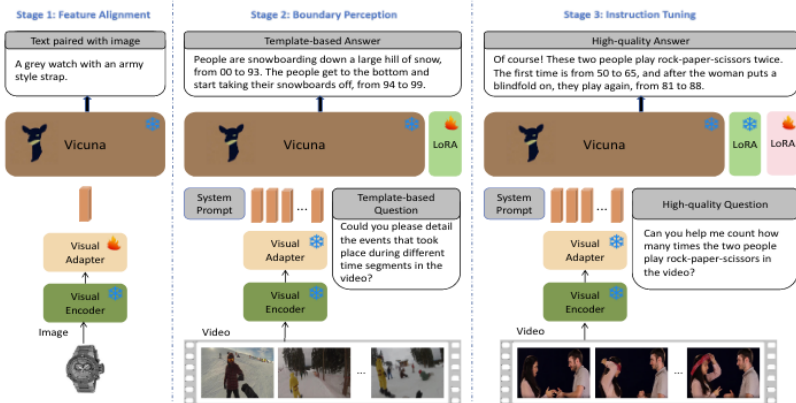
# VTimeLLM: Empower LLM to Grasp Video Moments[Huang+, CVPR2024]

**概要：** LLMの優れたテキスト理解能力を動画に対しても適用し、イベントの開始と終了の時間境界を正確に把握することに特化したvLLMを提案。詳細な時間境界に基づくイベント理解と推論を可能にした

**新規性：** Boundary-aware Trainingという時間境界の認識を強化する3stageの学習戦略を提案

**手法：** 画像特徴をLLMのセマンティック空間に画像-テキストペアを用いてalignment → video内のイベントをQA形式に変換し、LLMの時間境界認識と対応するイベントの理解を強化 → 高品質な対話データセットでLLMをInstruction Tuning

**結果：** 詳細な時間認識を必要とするタスク(Temporal Video Grounding, Dense Captioning)において優れた性能を確認



1. 長時間動画理解のBenchmark
2. 認識の困難さに対応する研究
3. **認識のコストに対応する研究**

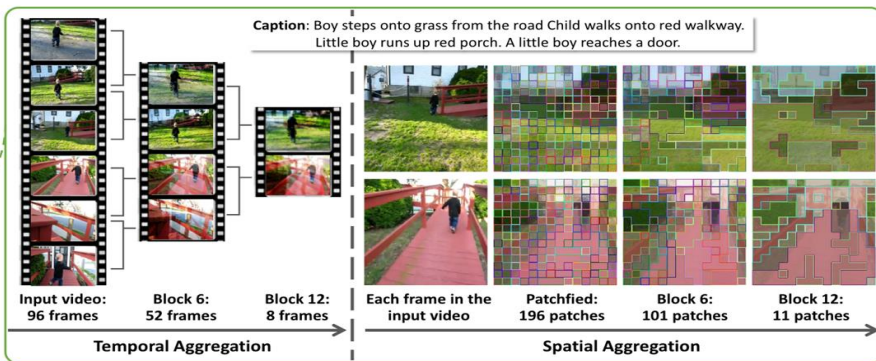
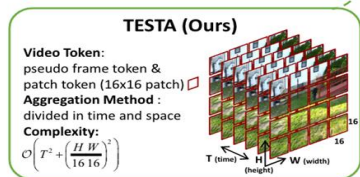
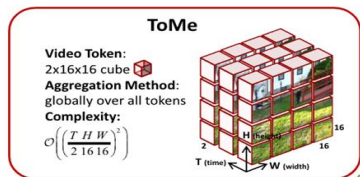
# TESTA: Temporal-Spatial Token Aggregation for Long-form Video-Language Understanding[Ren+, ACL2023]

**概要：** 長編動画の理解において効率的なビデオエンコード手法を提案

**新規性：** ビデオエンコーダブロック内で時間的（フレーム）および空間的（パッチ）にトークンを集約するモジュールを導入し、ビデオエンコードの計算負荷を削減

**手法：** フレームトークン（フレーム全体を代表する疑似トークン）とパッチトークン（16x16の2Dパッチ）を使用し、類似したトークンを見つけ出し、平均プーリングにより集約

**結果：** 同じフレーム数での入力において、計算効率が1.7倍向上。フレーム数を増やすことで性能の向上が確認された。複数のデータセットに対して高い性能を示し、特に長時間の理解タスクで優れた結果



S1: A girl in a pirate hat walks towards the camera.

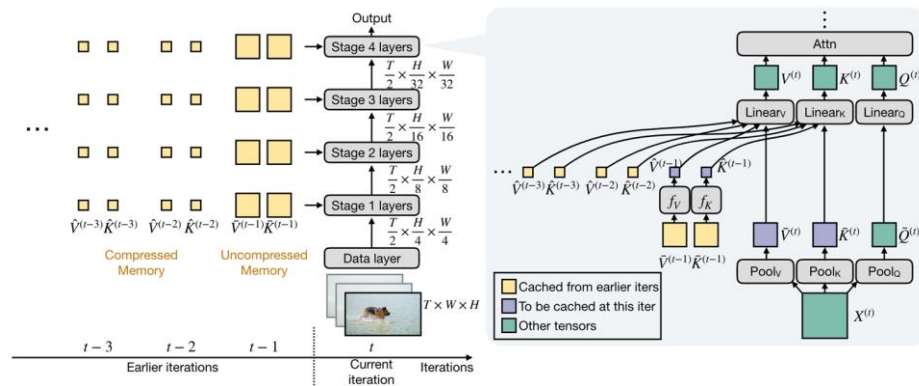
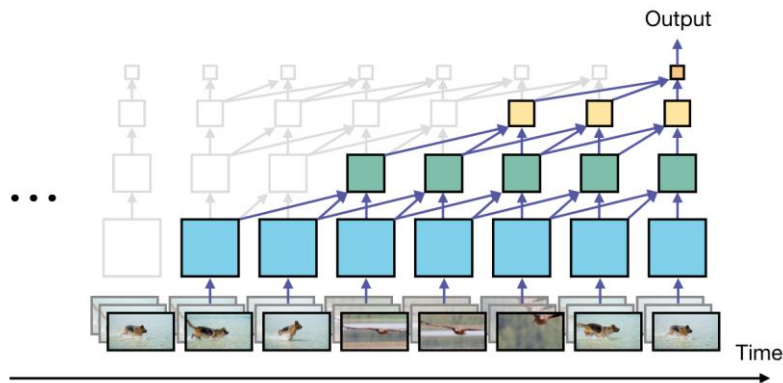
# MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition[Wu+, CVPR2022]

**概要:** ビデオをオンラインで処理し、各反復ごとにメモリをキャッシュすることにより4.5%の計算量で30倍の長さの時間をモデリング

**新規性:** メモリ（キャッシュ）ベースのアプローチを導入

**手法:** Multiscale Vision Transformer (MViT) をベース。ビデオを細かいクリップに分割して処理時に特徴量をキャッシュし、重要なキューのみを保持するように学習されたメモリ圧縮モジュールで圧縮。各層で過去の情報にアクセスすることで長期間のコンテキストを取得

**結果:** SpatioTemporalActionLocalizationやClassificationTaskにおいて性能改善が確認された



# MovieChat: From Dense Token to Sparse Memory for Long Video Understanding

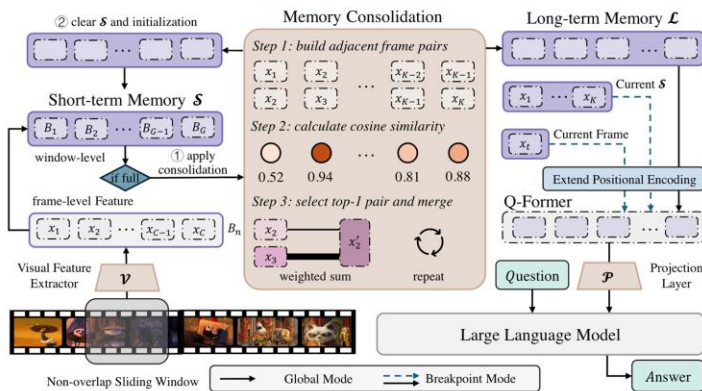
[Song+, CVPR2024]

**概要:** ビデオ基盤モデルと大規模言語モデル (LLMs) を統合することで、特定の視覚タスクに依存しない汎用的な動画理解を実現。

**新規性:** 手動アノテーション付きの長編動画理解ベンチマーク「MovieChat-1K」を作成し、評価専用のメモリ機能を導入し、計算複雑性とメモリコストを削減しながら、長期間のcontextを維持

**手法:** 短期メモリと長期メモリを用意。フレームごとに抽出した特徴を短期メモリに保存し、長期メモリで隣接する類似フレームを統合することで、密なトークンを疎なメモリに変換し、長期間の情報を保持。長期メモリのみ、または短期メモリも合わせて推論。

**結果:** MovieChat-1Kテストセットを用い、従来の手法 (Video Chat, Video LLaMA, Video-ChatGPT) に対して高い性能を確認。



The screenshot shows the MovieChat interface with a video player at the top and a Q&A section below. The Q&A section is divided into two modes: Breakpoint Mode and Global Mode.

**Question #1: What may happen next step?** (Breakpoint Mode)

- Video Chat:** Based on the video, it appears that the person is cooking meat and vegetables on a stove top using different utensils and ingredients such as onions, garlic, mustard, and olive oil. It is likely that the person will continue cooking the meat until it is fully cooked, and then serve it to their family or friends. Additionally, they may add other ingredients such as mushrooms or potatoes to the [redacted], it is possible that they will use different types of cooking methods such as sautéing or grilling to ensure the meat is cooked evenly and to perfection.
- Video LLaMA:** At frame 4250, the video shows the hand of a man cutting meat on a cutting board. It's not clear what happens next, but it is possible that the man continues cutting the meat.
- Video ChatGPT:** The next step in the video is to place the cooked meat on a plate and serve it. The person in the video is seen cutting the meat into smaller pieces and placing it on a plate. The person then takes a fork and begins to eat the meat.
- Movie Chat:** I think the man will prepare the food in the kitchen. It seems like he is going to cook it with spices and herbs. He has a large wooden board on the kitchen counter and a large white pot on the stove, and he will probably use some tools to prepare the food. He might also use his hands or a spatula to cook the meat.

**Question #2: Which step took the longest to complete this dish?** (Global Mode)

- Video Chat:** Based on the video, it appears that the longest step in making the dish is preparing the pan for cooking the steak. Specifically, the man takes the pan out of the oven and starts cleaning it by wiping it with a rag, then rubbing it with olive oil, and finally heating the pan to high heat. It is also mentioned that the man has to wait for the pan to heat up before putting the steak in the pan.
- Video LLaMA:** The dish that this video portrays took several minutes to complete. However, the most time-consuming task seems to have been preparing the ingredients on the white kitchen counter. The man in the video is seen chopping vegetables and using a sharp knife to prepare a meal. Finally, he will sit down to enjoy the meal, savoring each bite and reflecting on the process that led to this culinary creation.
- Video ChatGPT:** The longest step in preparing the dish was the time it took to cook the meat. The man in the video cooks the meat on a stove for a long time, which is a crucial step in the process. The video shows the man cooking the meat on a stove for a considerable amount of time, which indicates that it took a considerable amount of effort and patience to achieve the desired result.
- Movie Chat:** The longest step to complete this dish would be the preparation of the food on the cutting board. It requires chopping the meat and adding other ingredients like broccoli and carrots, and mixing them together before putting them in the pan. The chef had to carefully chop and arrange the ingredients before cooking the dish.

まとめBY: 木林