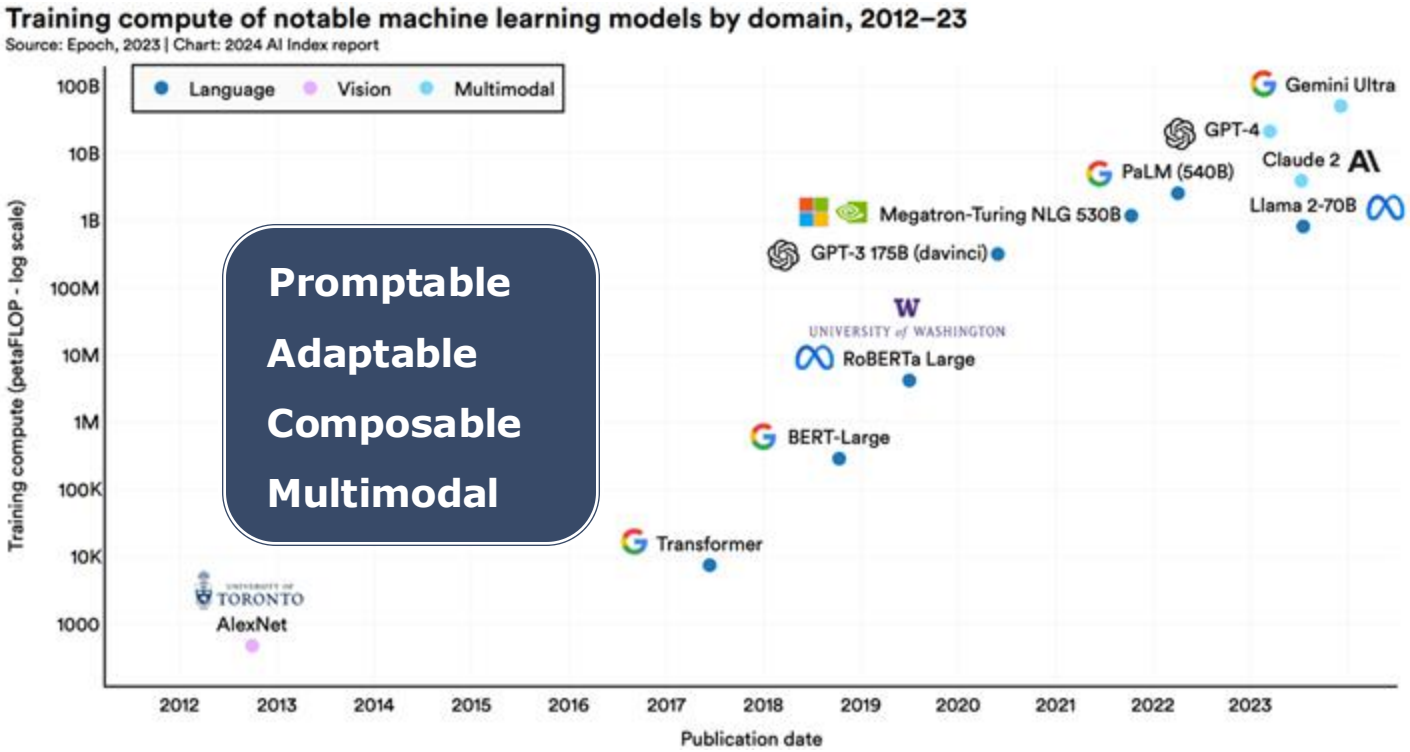
LASER: A Neuro-Symbolic Framework for Learning Spatio-Temporal Scene Graphs with Weak Supervision

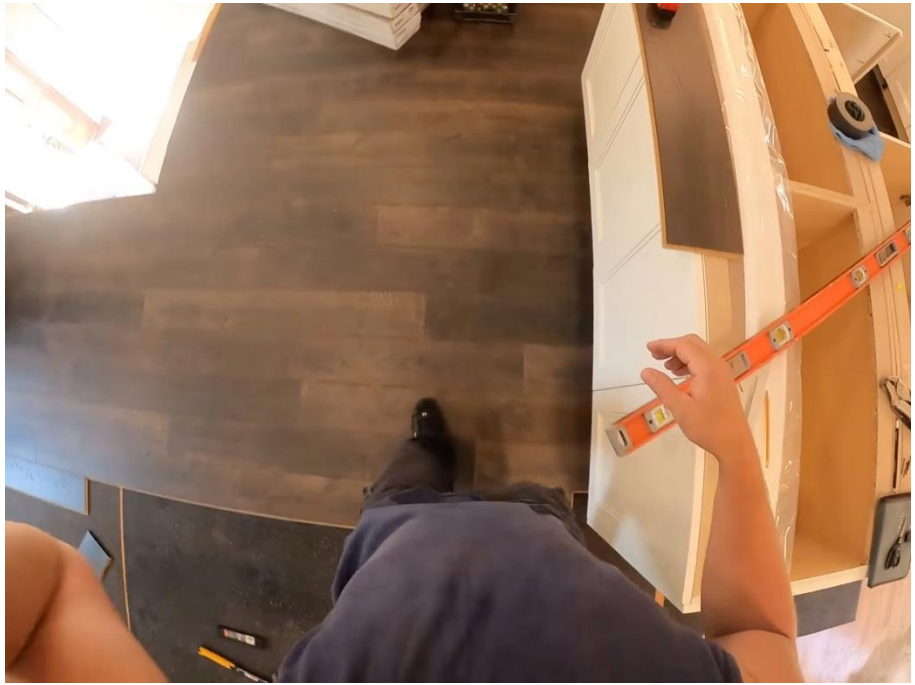
Jiani Huang
University of Pennsylvania

The Landscape of Foundation Models



Perrault, Ray, and Jack Clark. "Artificial intelligence index report 2024." (2024).

A Coarse-Grained Video Task: Captioning



GPT-4

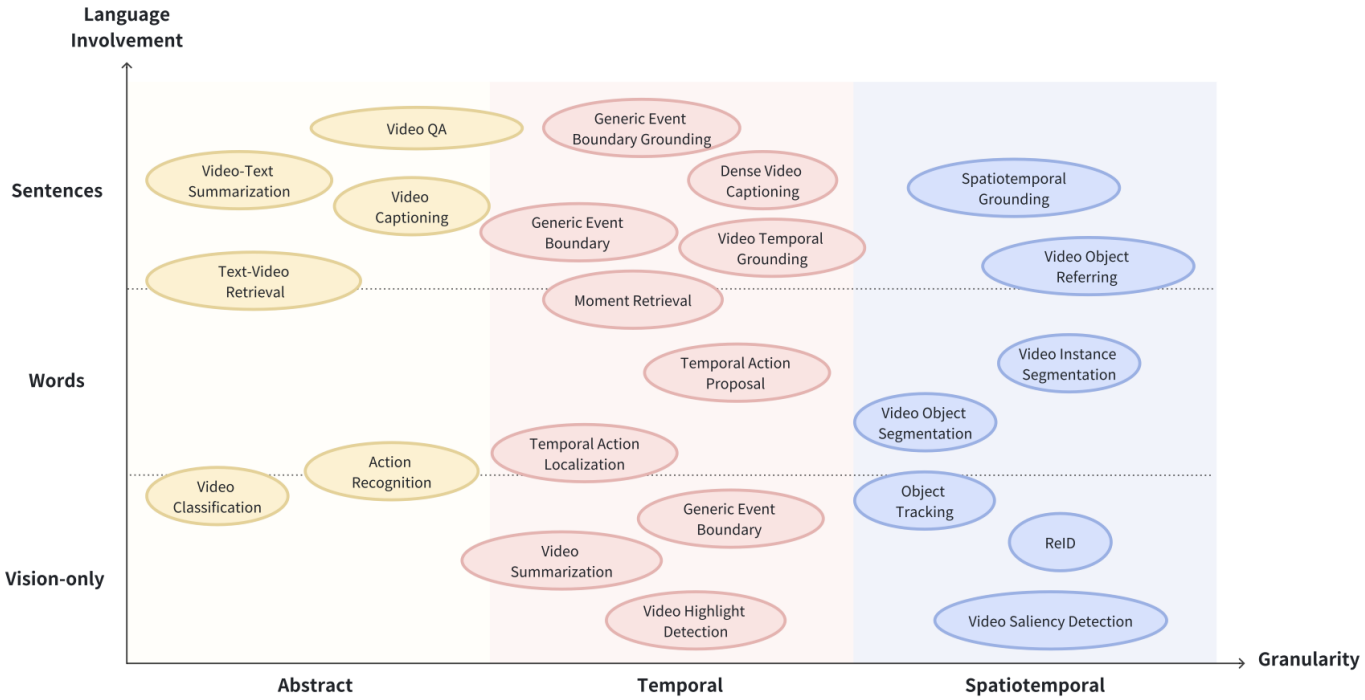
"A person is using a measuring tape to take measurements of a room."

A Finer-Grained Video Task: Object Detection

Where is the measuring tape?

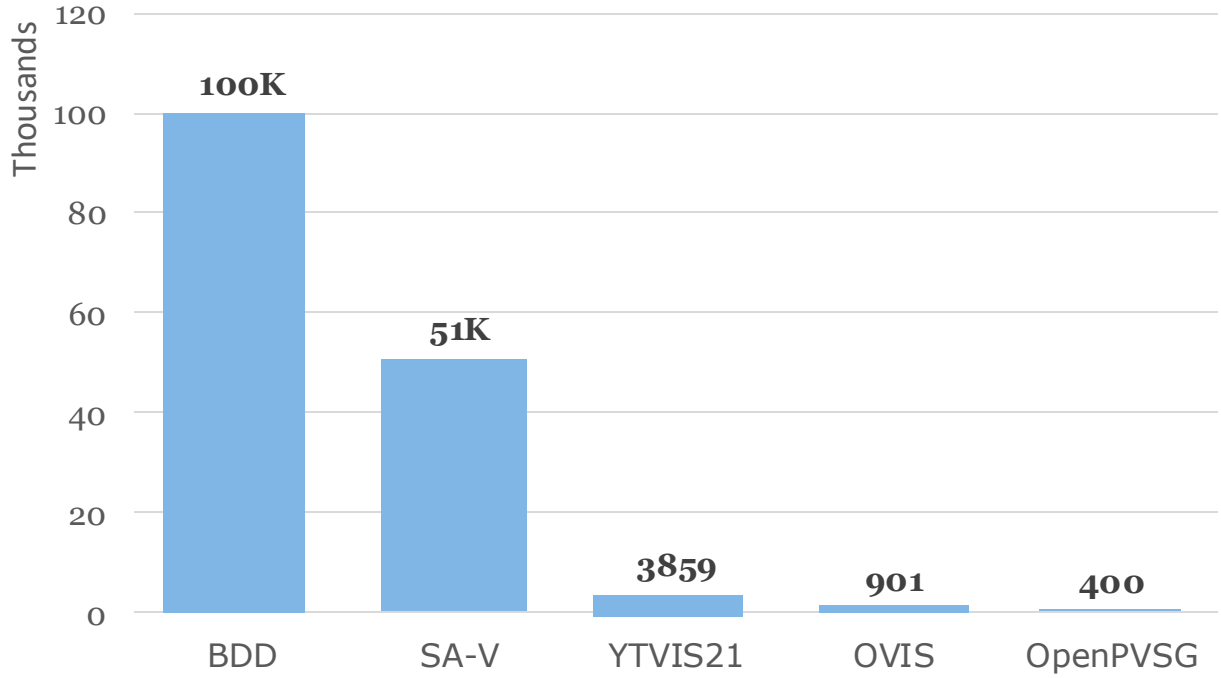
GPT-4

A Landscape of Video Understanding Tasks



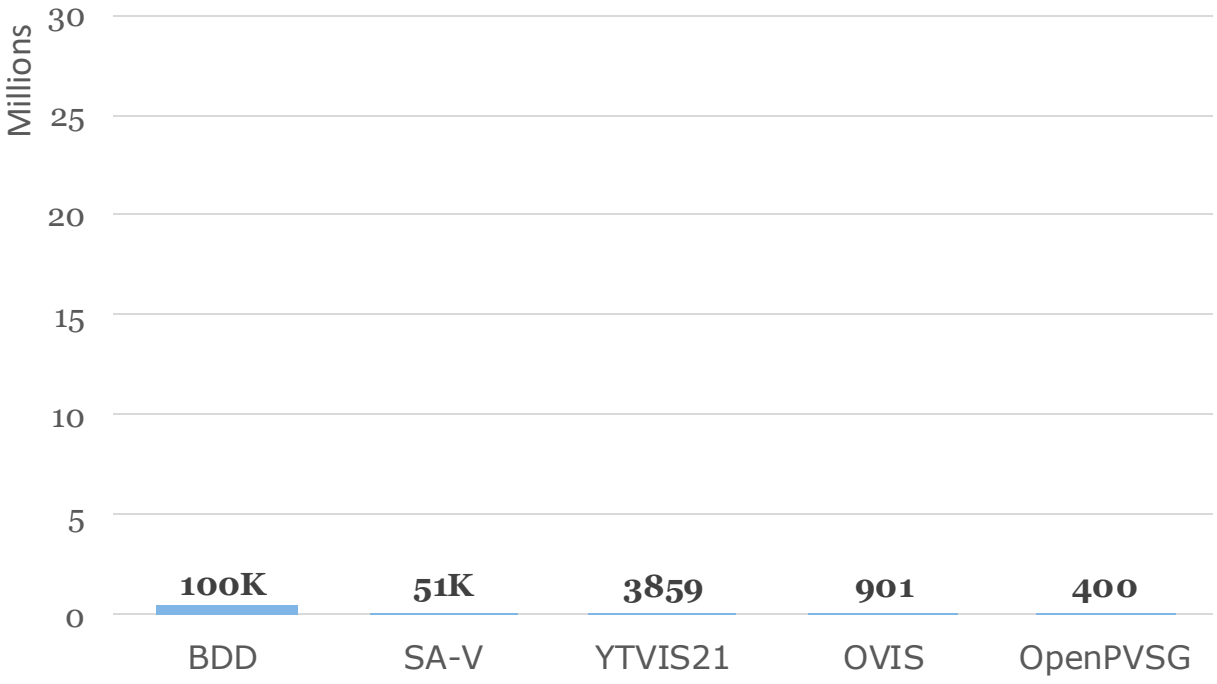
Tang, Yunlong, et al. "Video understanding with large language models: A survey." 2023.

Fine-grained Video Tasks Are Challenging



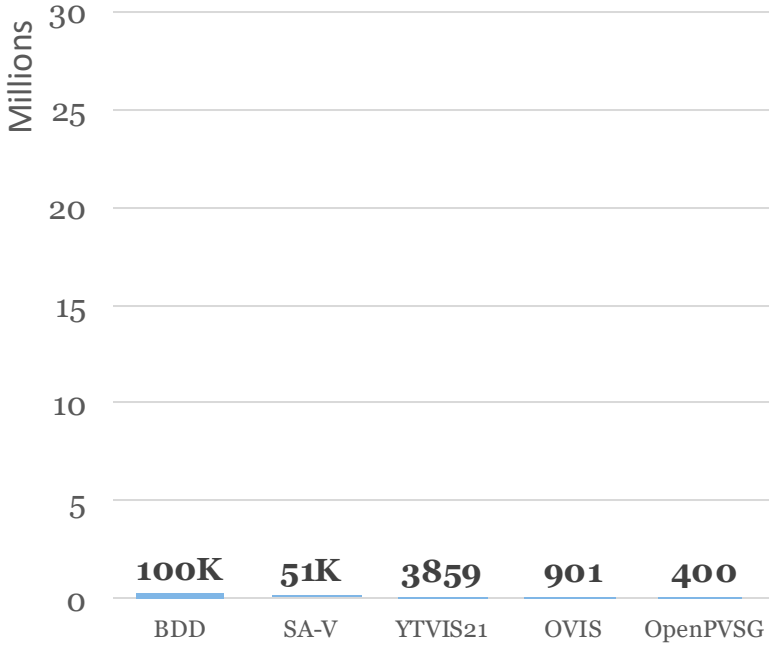
Video Datasets with Fine-grained Annotations, such as trajectories, entity classes, and relations

Fine-grained Video Tasks Are Challenging

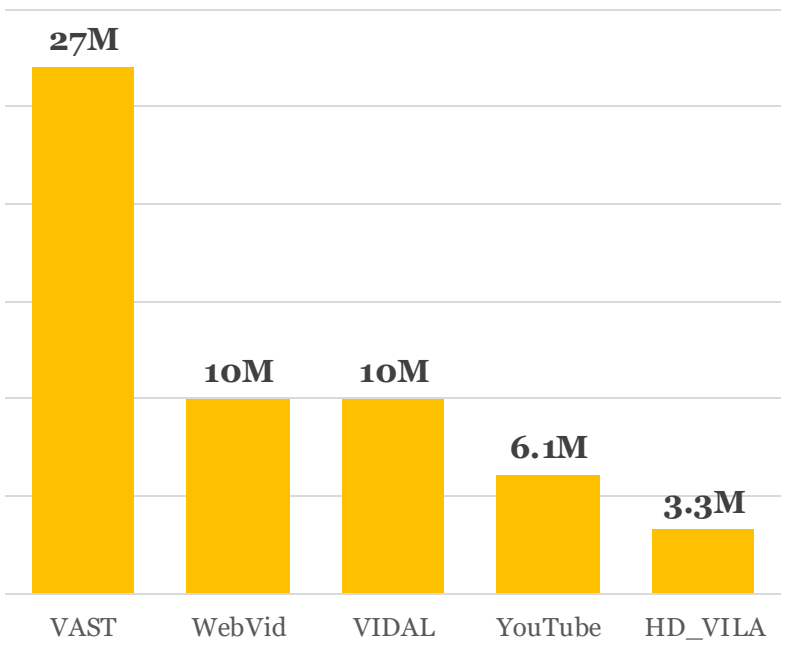


Video Datasets with Fine-grained Annotations

Fine-grained Video Tasks Are Challenging

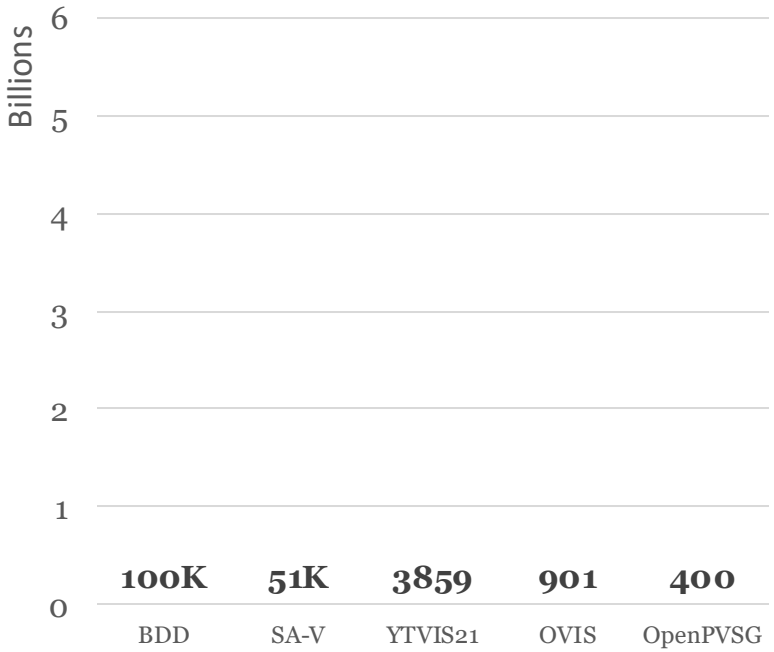


Video Datasets with Fine-grained Annotations

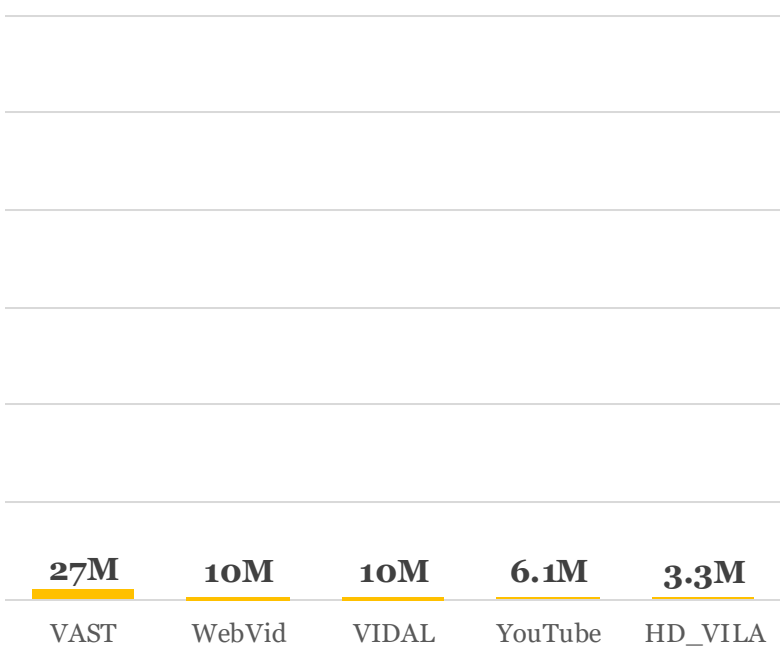


Video Datasets with Captions

Fine-grained Video Tasks Are Challenging

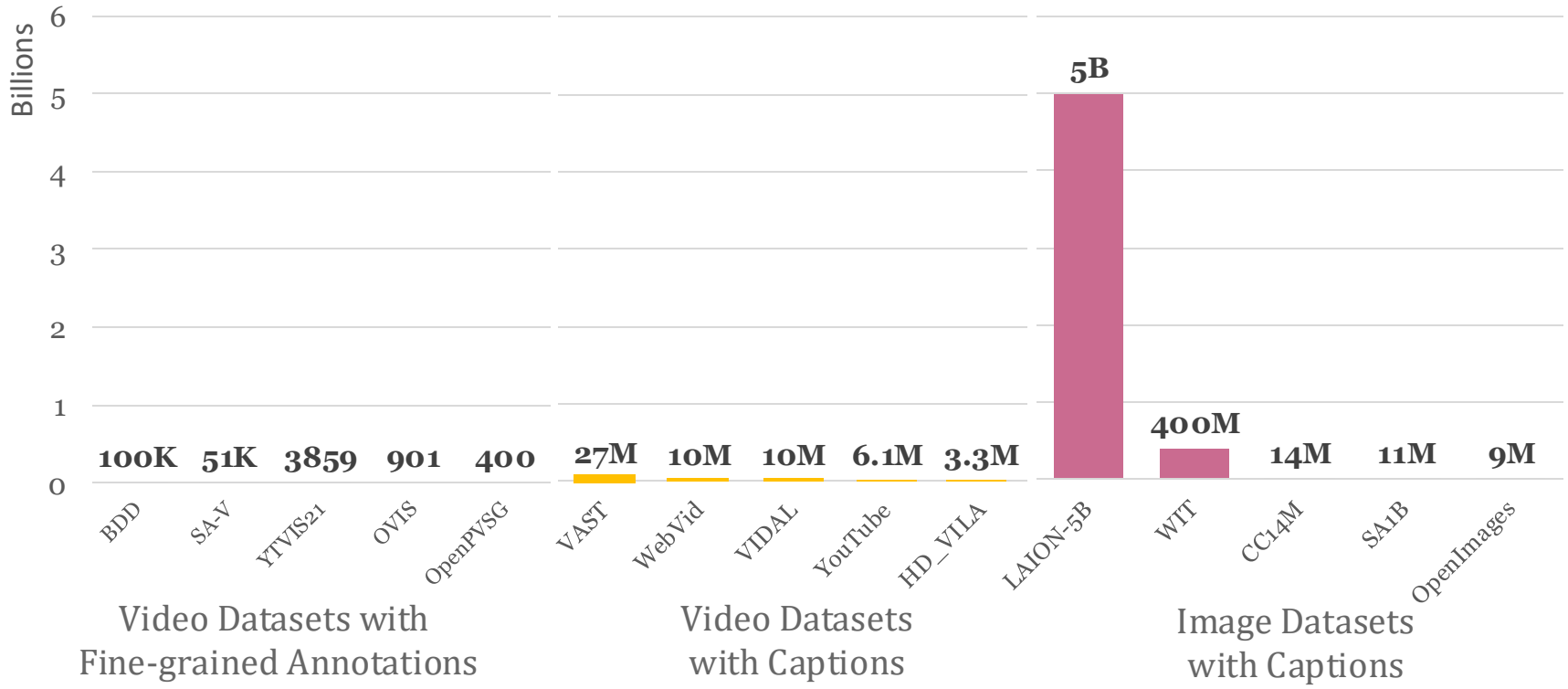


Video Datasets with Fine-grained Annotations

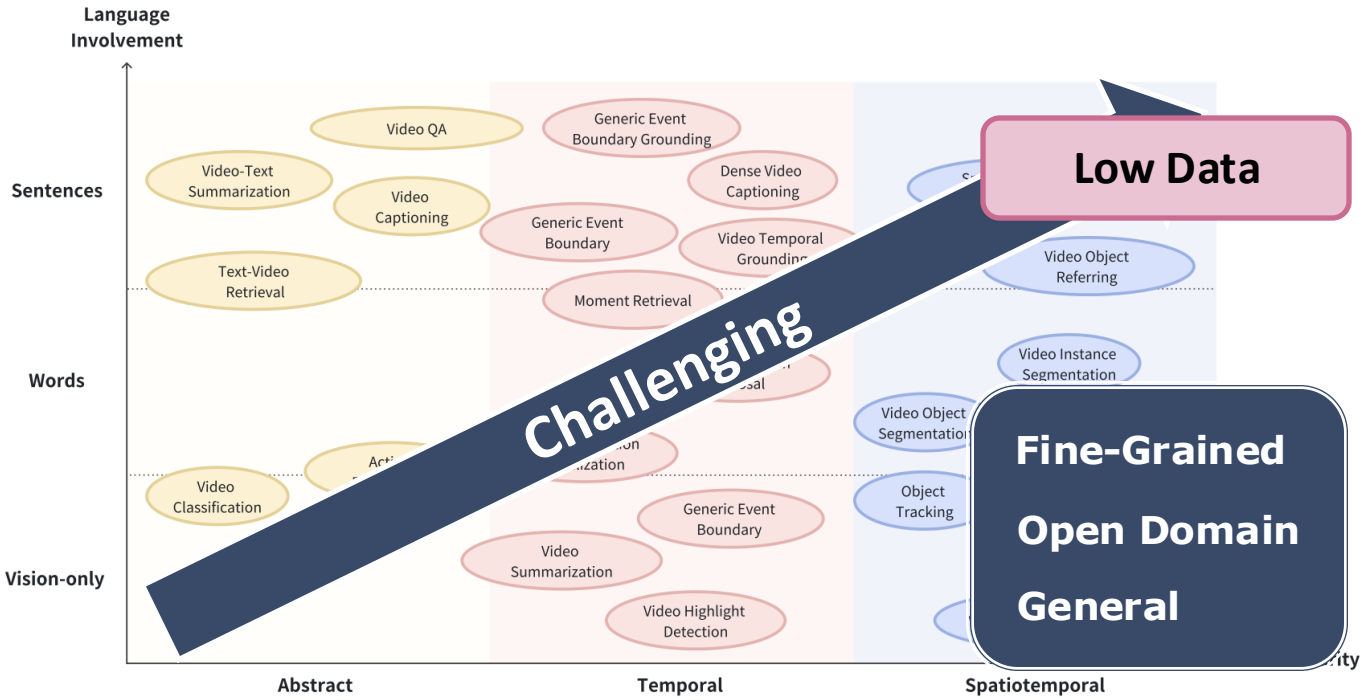


Video Datasets with Captions

Fine-grained Video Tasks Are Challenging

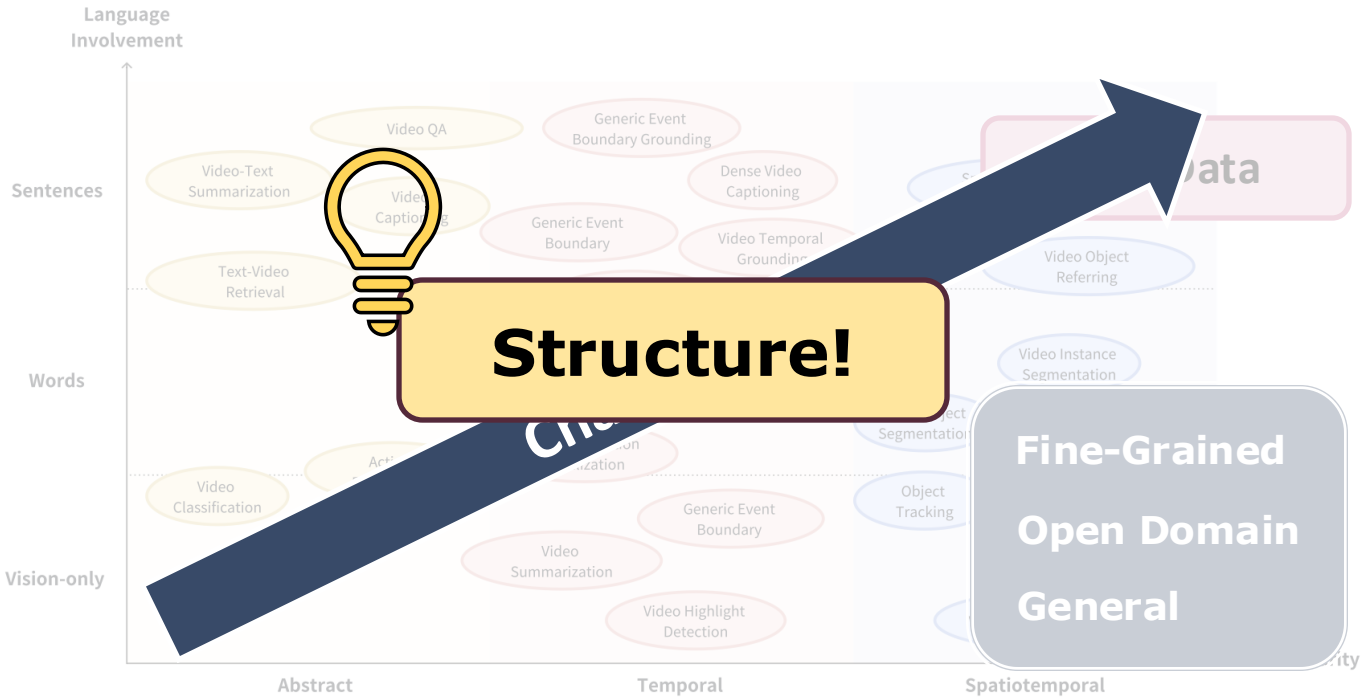


A Landscape of Video Understanding Tasks



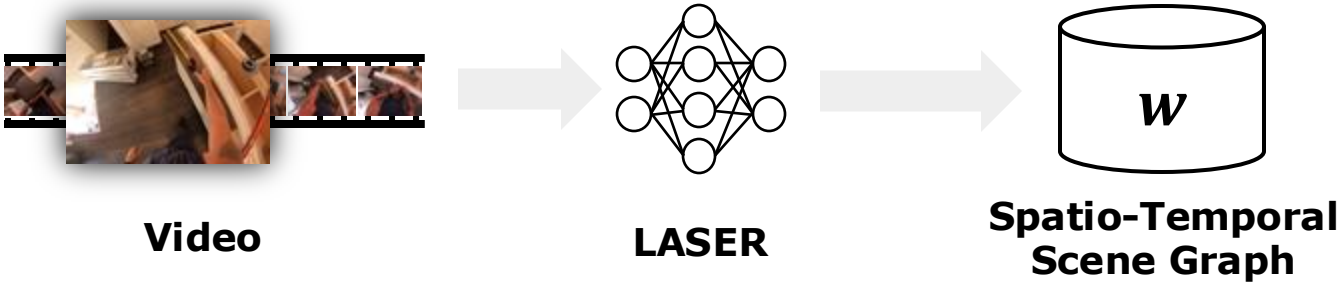
Tang, Yunlong, et al. "Video understanding with large language models: A survey." *arXiv preprint arXiv:2312.17432*(2023).

A Landscape of Video Understanding Tasks



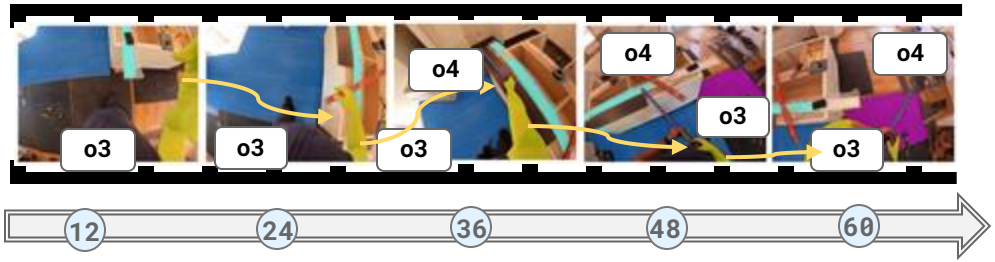
Tang, Yunlong, et al. "Video understanding with large language models: A survey." *arXiv preprint arXiv:2312.17432*(2023).

LASER: Spatio-Temporal Scene Graph Generation



Probabilistic Spatio-Temporal Scene Graph (STSG)

Object Trajectories



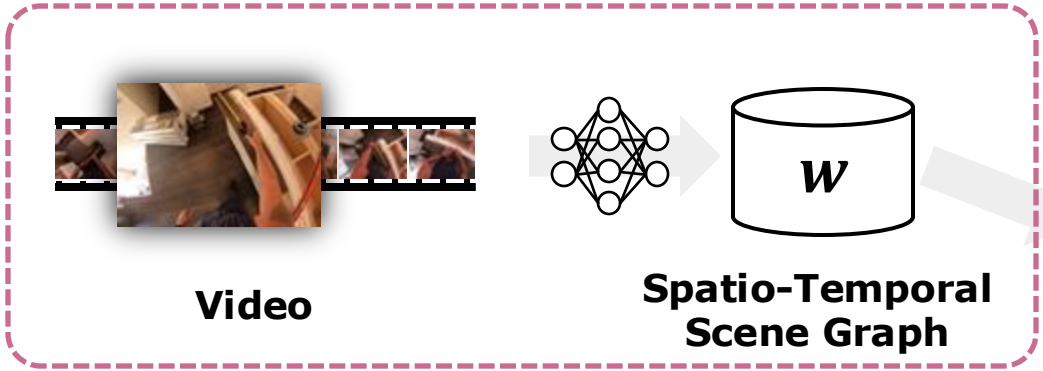
Probabilistic Object Names

prob	obj id	category
0.92	o3	adult
0.06	o3	clothes
0.82	o4	tape measure
...	...	...

Probabilistic Object Relations

prob	frame id	sub id	obj id	relationship
0.72	36	o3	o4	reaching for
0.16	36	o3	o4	holding
0.93	48	o3	o4	reaching for
...	...	...	...	...

An Illustration with Fine-Grained Video Understanding



Video

Spatio-Temporal Scene Graph

A person retrieves a tape measure to assess the length of a missing floorboard.

Caption

LLM

Spatio-Temporal Specification

Neuro-Symbolic Alignment Checker

Alignment Score

$\mathcal{L}$

Contrastive Loss

How to Obtain STSL Program from Caption?

Natural Language Video Caption

A person retrieves a tape measure to assess the length of a missing floorboard.

LLM

Structured Representations

Time Stamp 1

Description:
A person is reaching for a measuring tape.

Predicates:

```
name(v1, "person"),
name(v2, "tape measure")
relation("reaching for", v1, v2)))
```

Loc: early **Dur:** short

Time Stamp 2

Description:
The person is holding the tape measure.

Predicates:

```
relation("holding", v1, v2)))
```

Loc: early **Dur:** short

Time Stamp 3

Description:
The person is measuring the missing floorboard.

Predicates:

```
name(v3, "missing floorboard"),
relation("reaching for", v1, v3)))
```

Loc: early **Dur:** short

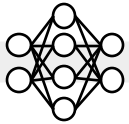
Code Synthesizer

Spatio-Temporal Specification

```
Globally(name(v1, "person"), name(v2, "tape measure"), name(v3, “missing floorboard”)) and
Until(Finally(relation("reaching for", v1, v2))), Until(Finally(relation('holding', v1, v2)),
          Finally(relation('measuring', v1, v3)))
```

An Illustration with Fine-Grained Video Understanding

Video



w

Spatio-Temporal Scene Graph

Neuro-Symbolic Alignment Checker

Alignment Score

$\mathcal{L}$

Contrastive Loss

A person retrieves a tape measure to assess the length of a missing floorboard.

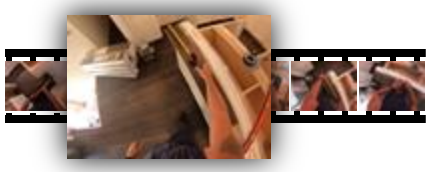
Caption

LLM

ψ

Spatio-Temporal Specification

Putting It All Together!



Video

A person retrieves a tape measure to assess the length of a missing floorboard.

Caption

LLM

backprop

w

Spatio-Temporal Scene Graph

ψ

Spatio-Temporal Specification

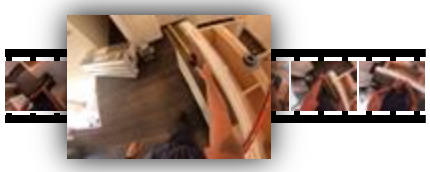
Neuro-Symbolic Alignment Checker

Similarity Score

$\mathcal{L}$

Contrastive Loss

Putting It All Together!



Video

A person retrieves a tape measure to assess the length of a missing floorboard.

Caption

LLM

w

Spatio-Temporal Scene Graph

ψ

Spatio-Temporal Specification

backprop

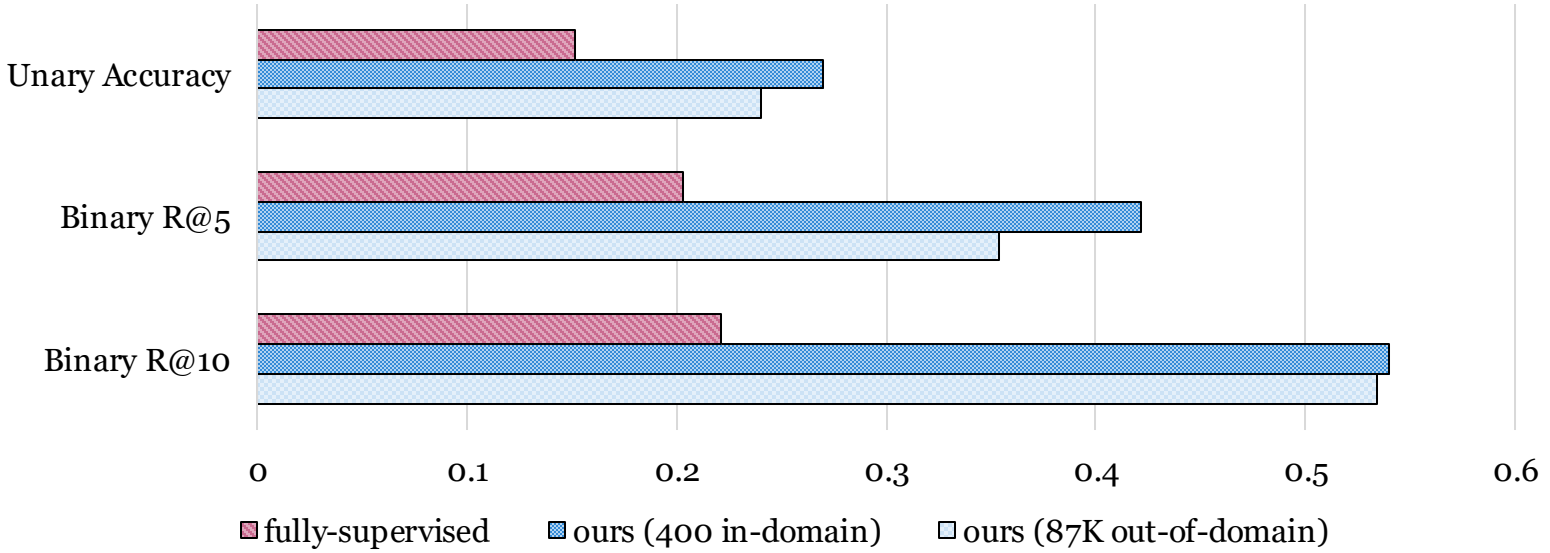
Neuro-Symbolic Alignment Checker

Similarity Score

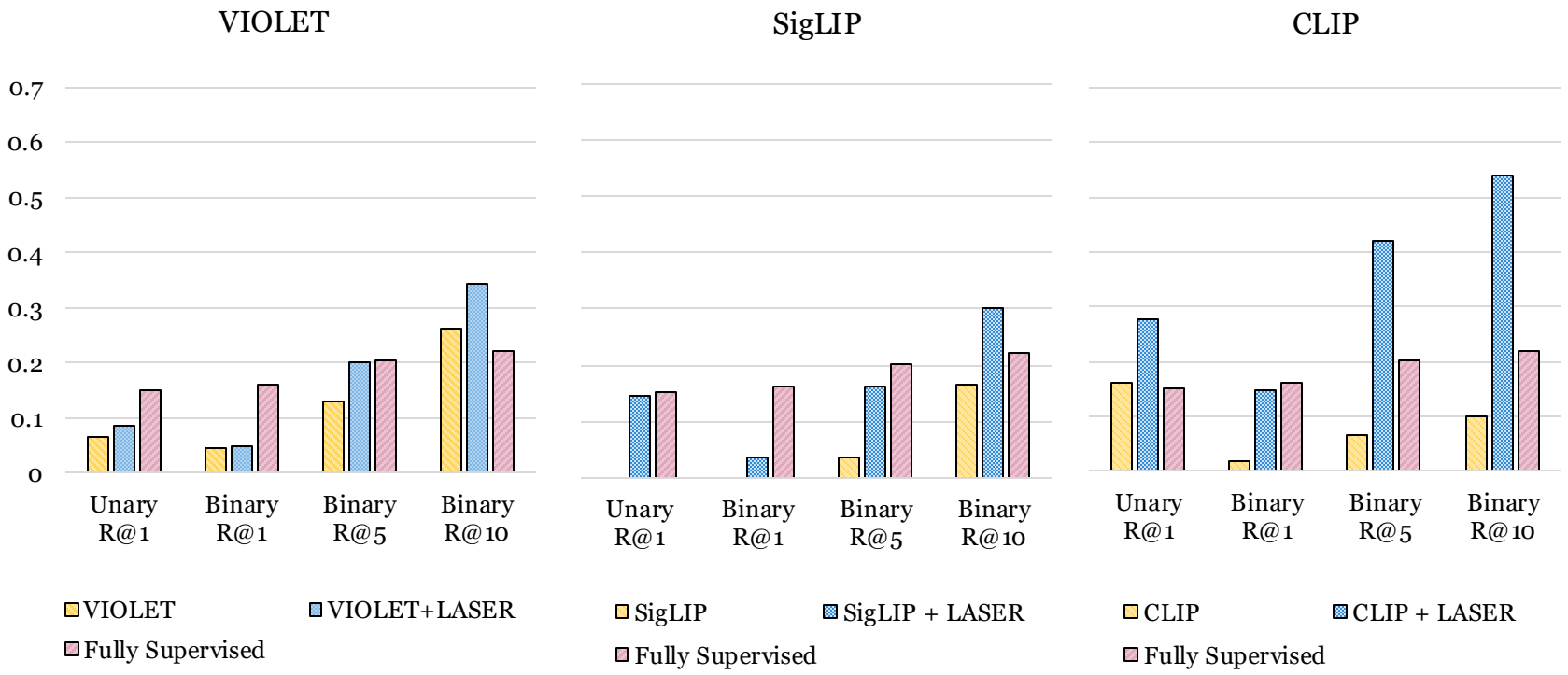
$\mathcal{L}$

Contrastive Loss

Video Spatio-Temporal Scene Graph Generation



Performance with Different Backbone Models



Thank you!