Harnessing Object Grounding for Time-Sensitive Video Understanding

Tz-Ying Wu Sharath Nittur Sridhar Subarna Tripathi

Intel

WACV 2026
Tucson, Arizona • Mar 6 - 10, 2026

Time-Sensitive Video Understanding (TSV)

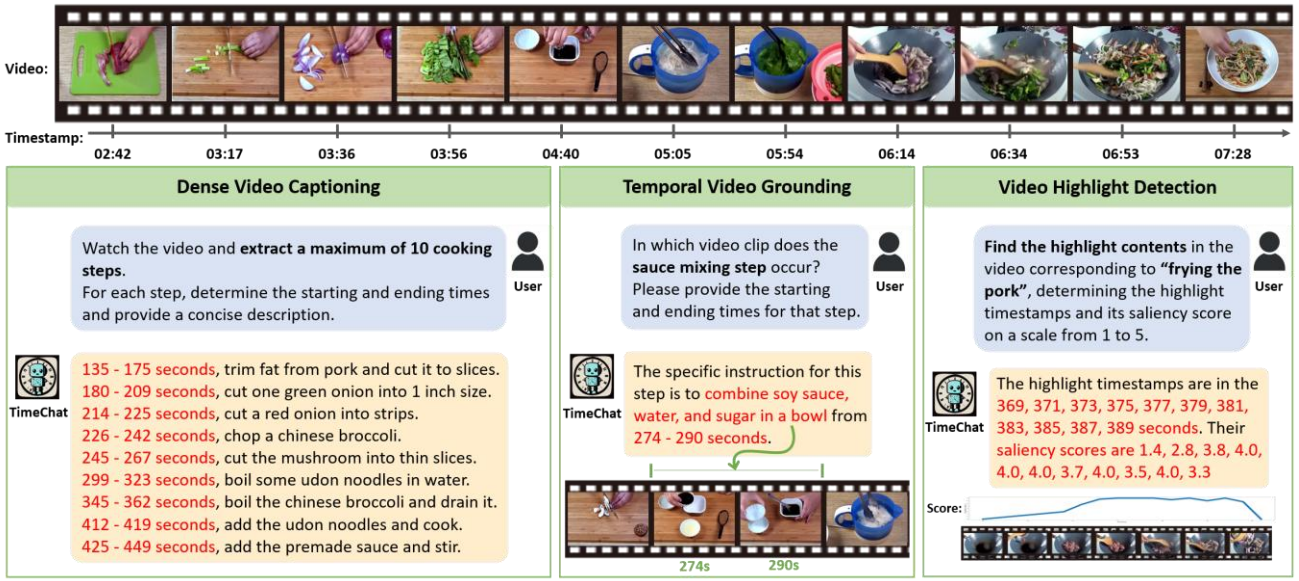
- TSV tasks are challenging as they require **fine-grained** scene comprehension with temporal reasoning.

Reasoning Temporal Localization (RTL)

Q: when does the woman's dance become the most **energetic** in the video?

A: Her dance is the most energetic between 7.37s and 12.81s when she does a **handspring**, which is more energetic than **standing** and **lying** on the floor.

[1] Huang et al., "LITA: Language Instructed Temporal-Localization Assistant". In ECCV 2024.



[2] Ren et al., "TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding". In CVPR 2024.

Motivation

- Video-LLMs compress frames into **highly condensed video tokens.**
- This works for global understanding but is **insufficient for TSV**.
- Objects provide key temporal cues yet are not modeled explicitly.
- We ask: *Can **grounded object (GO)** information improve Video-LLMs on TSV tasks?*

Probing Video-LLMs with GO Information

(i) Can providing GO information in Video-LLMs enhance TSV?

(ii) What levels of GO information is needed?

As a preliminary experiment, we probe existing Video-LLMs with object cues by injecting the GO information into the text prompt:

- **Class:** *"<Obj> Objects in this video are: man, window, ...</Obj>".*
- **Class+Time:** *"<Obj> Each object is provided with its timestamp and class label in the format of <time, class label>. Here are the objects: <91.2 second, man>, ...</Obj>".*
- **Class+Time+Bbox:** *"<Obj> Each object bounding box is provided with its timestamp and class label in the format of <time, (x1,y1,x2,y2), class label>. Here are the objects: <91.2 second, (0.0001, 0.1715, 0.0806, 0.3784), man>, ...</Obj>".*

Probing Video-LLMs with GO Information

- **Object cues help**, and more detailed cues yield larger gains.
- However, this naïve approach:
  - Is **not robust** to object grounding errors.
  - Introduces a **large** number of **extra tokens** (~42 tokens per object).

GO at Inference	mIOU	P@0.5
n/a	23.04	18.78
Text *(Class)*	23.65	18.08
Text *(Class+Time)*	26.47	20.47
Text *(Class+Time+Bbox)*	27.63	23.71

Table 1: *Probing LITA-13B on the RTL task with different levels of grounded object information.*

Flip Ratio	mIOU	P@0.5
10%	27.07	23.25
20%	25.89	22.63
50%	24.63	20.13

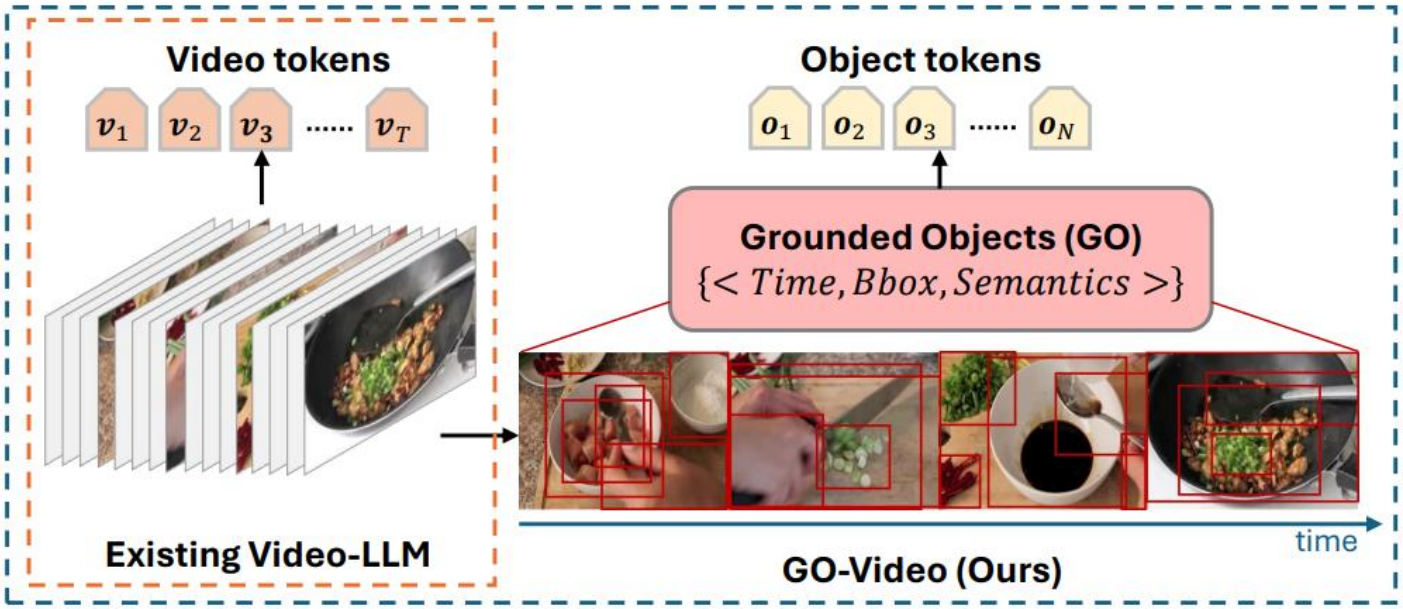
Table 2: *Flipping a fraction of object class labels with random classes.*

Shift	mIOU	P@0.5
$0.01 \times (H, W)$	27.37	21.13
$0.02 \times (H, W)$	26.42	19.01
$0.05 \times (H, W)$	24.71	17.14

Table 3: *Jittering bounding box locations wrt. the image height (H) and width (W) in pixels.*

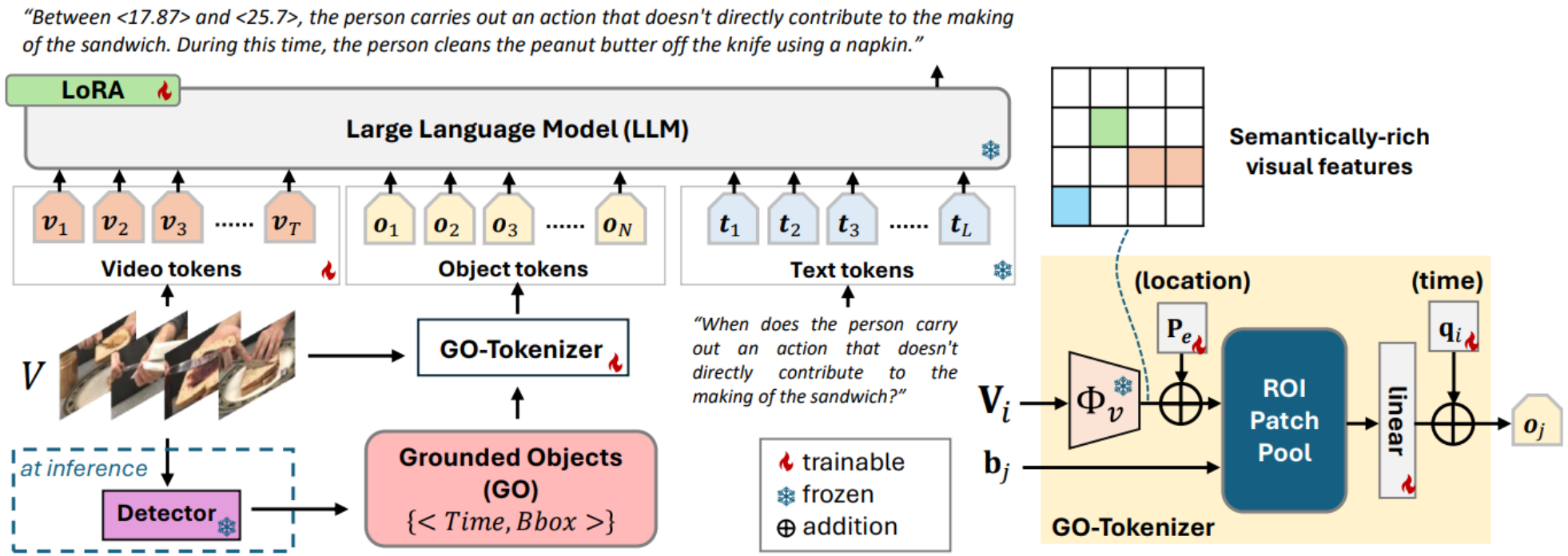
Augmenting Video-LLMs with Object Tokens

- We introduce **object tokens** to complement **video tokens**.
- They are **compact** representations encapsulating each object's *spatial*, *semantics*, and *temporal* information into a **single** token.



Grounded Object Tokenization

- To extract **object tokens**, we propose **Grounded Object Tokenizer (GO-Tokenizer)**, a lightweight add-on for existing Video-LLMs.



Results

- **Text** does not yield consistent gain when introducing GO in prompt.
- In contrast, **GO-(LITA/TimeChat)** consistently outperforms.

Model	GO Training	mIOU	P@0.5	Rel. Score
LITA	✗	28.28	26.40	54.7
LITA + Text (*Class+Time+Bbox*)	✗	28.07	25.83	54.3
Finetuned on ActivityNet-Captions & ActivityNet-RTL				
LITA	✗	28.72	24.53	54.7
LITA + Text (*Class+Time+Bbox*)	✗	28.82	25.10	55.6
GO-LITA	✓	**31.52**	**28.07**	**59.0**

Table 4: *Evaluating LITA-13B on the ActivityNet-RTL dataset with/without GO information.*

Model	GO Training	CIDEr	SODAc	F1
Valley [16]	✗	0.0	0.1	1.5
Video-LLaMA [27]	✗	0.0	0.0	0.1
TimeChat [22]	✗	3.4	1.2	12.6
Finetuned on ActivityNet-Captions				
TimeChat	✗	3.4	1.1	11.3
TimeChat + Text (*Class+Time+Bbox*)	✗	2.4	0.8	10.6
GO-TimeChat	✓	**3.9**	**1.4**	**18.5**

Table 5: *Zero-shot evaluation on Youcook2 dataset for dense video captioning.*

Results

- **Text** with **GO Training** still underperforms **GO-(LITA/TimeChat).**

Model	Go Training	Tokens/GO	CIDEr	SODAc	F1
TimeChat	✗	0	3.4	1.1	11.3
TimeChat + Text (*Class*)	✓	4	1.8	0.8	16.2
TimeChat + Text (*Class+Time*)	✓	12	1.8	0.9	16.8
TimeChat + Text (*Class+Time+Bbox*)	✓	42	1.6	0.8	13.0
GO-TimeChat	✓	1	**3.9**	**1.4**	**18.5**

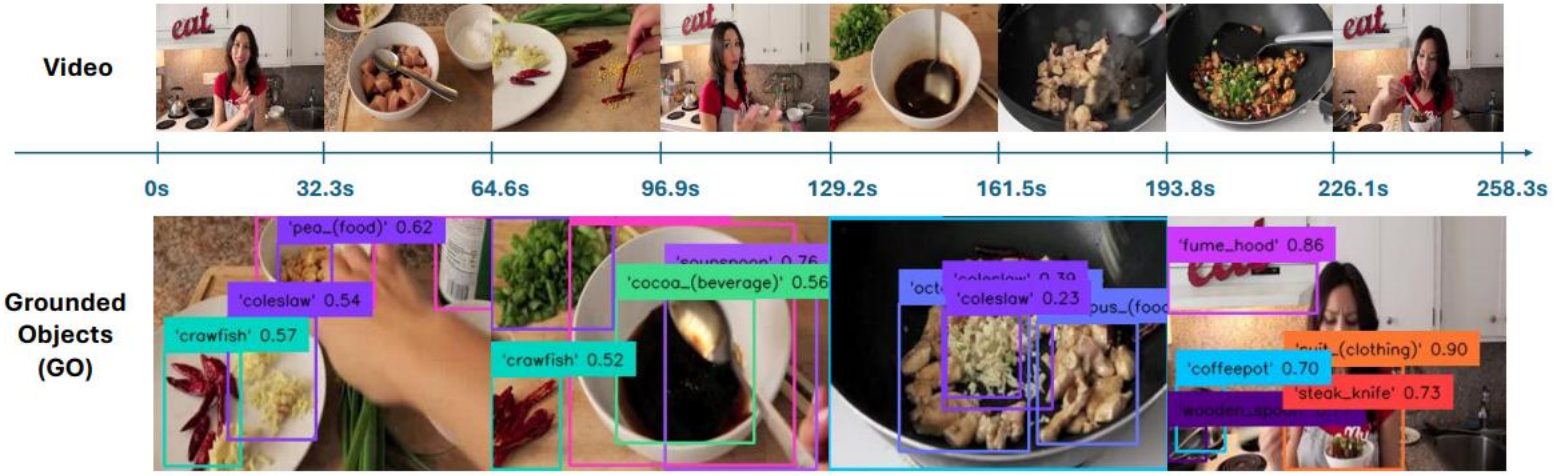
Table 6: *Comparison to Text with GO Training using TimeChat as the base model.*

Model	Go Training	mIOU	P@0.5	Rel. Score
LITA	✗	28.72	24.53	54.7
LITA + Text (*Class+Time+Bbox*)	✓	30.05	27.97	55.5
GO-LITA	✓	**31.52**	**28.07**	**59.0**

Table 7: *Comparison to Text with GO Training using LITA as the base model.*

Visualization

- Video-LLMs with **GO-Tokenizer** generates captions with higher quality compared to **Text**.



Text (Class +Time +Bbox)

20.6s - 44.8s: a person chicken is cooking in a white bowl on dining.
44.8s - 70.2s: the ingredients like soy sauce, sesame seeds, green chillies and chicken breast are being put on the dining table.
70.2s - 130.9s: the chicken breast is fried in a white bowl on dining table.
130.9s - 149.3s: the chicken is served in a white bowl with a spoon.
149.3s - 207.5s: the stove is on and a black handle is sticking out of the pot and there are two sauce pans in it.
207.5s - 245.7s: a toothbrush and pencil are being kept on a vase on dining table.

GO-Tokenizer (ours)

0.0s - 11.2s: a woman is talking while standing in a kitchen.
11.2s - 37.8s: she is adding various ingredients to a white bowl, including chicken and seasonings.
37.8s - 47.2s: she shakes the bowl to combine the ingredients.
47.2s - 65.5s: she then adds some liquid to the bowl and mixes it in.
65.5s - 84.9s: she then adds a little more seasoning to the bowl.
84.9s - 114.6s: she uses a spatula to stir the ingredients.
114.6s - 134.0s: she then adds a little more seasoning and mixes it in.
134.0s - 147.3s: she then adds a little more liquid to the bowl and mixes it in.
147.3s - 164.8s: she adds a little more seasoning and mixes it in.
164.8s - 185.1s: she then adds a little more seasoning to the bowl and mixes it in.
185.1s - 207.4s: she adds a little more liquid to the bowl and mixes it in.
207.4s - 225.7s: she then adds a little more seasoning to the bowl and mixes it in.
225.7s - 250.2s: she then uses two chopsticks to take a sample of the sauce and eats it.

Thank you!

Harnessing Object Grounding for Time-Sensitive Video Understanding

Tz-Ying Wu Sharath Nittur Sridhar Subarna Tripathi

Intel

WACV 2026
Tucson, Arizona • Mar 6 - 10, 2026