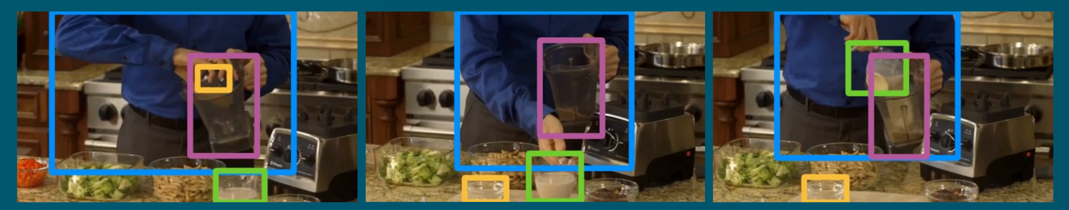
Large-scale Pre-training for Grounded Video Caption Generation

Evangelos Kazakos[1], Cordelia Schmid[2], Josef Sivic[1]
[1]Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University in Prague
[2]Inria, École normale supérieure, CNRS, PSL Research University

ICCV HONOLULU HAWAII OCT 19-23, 2025

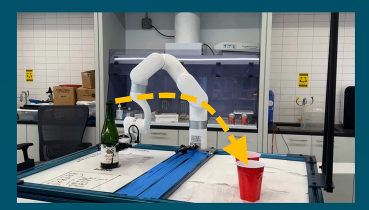
OVERVIEW

Goal: Simultaneous captioning & grounding of nouns to bounding boxes

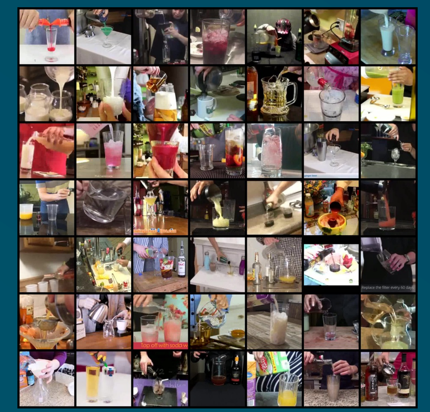


Caption: A man is adding ingredients from a bowl and cup into a jar to mix them well

Motivation



Task: pour wine into glass



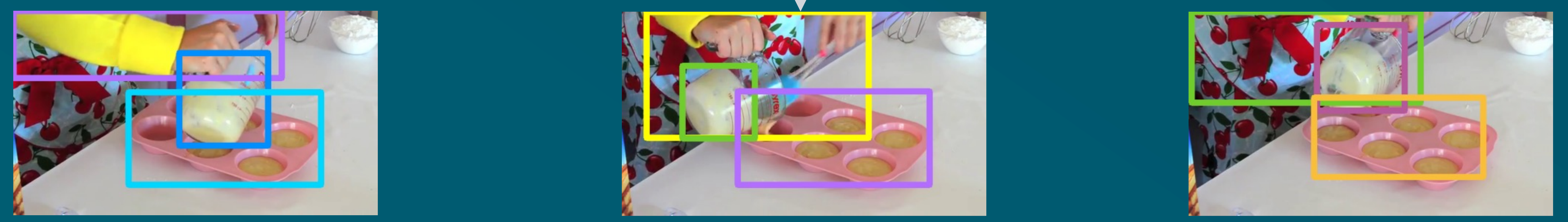
Learn by demonstration

Contributions

- ➤ Automatic annotation method
- ➤ HowToGround1M dataset
- ➤ Manually annotated iGround dataset
- ➤ The GROVE grounded video caption generation model
- ➤ SOTA on 5 grounding datasets

AUTOMATIC ANNOTATION METHOD

Stage 1: Frame-level grounding with image VLM



A person is pouring a white liquid into a pink tray | A woman is pouring liquid into a muffin tin | Someone is pouring batter into a muffin cup

caption

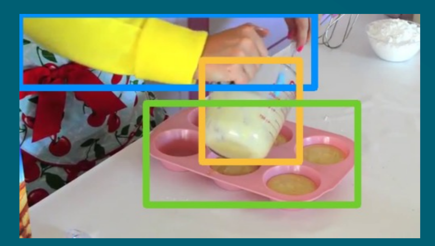
Stage 2: Caption aggregation with LLM

A person is pouring a liquid into a tray

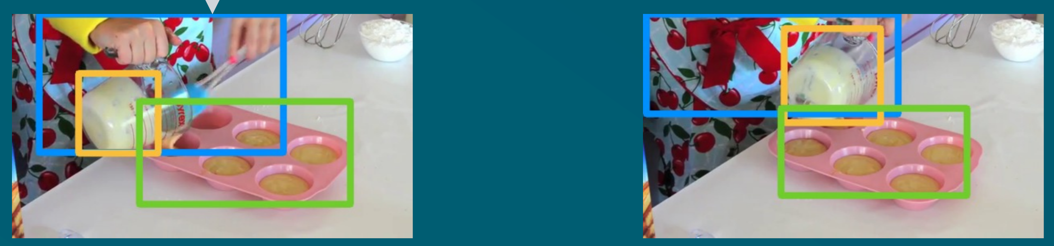
reference nouns

instance nouns

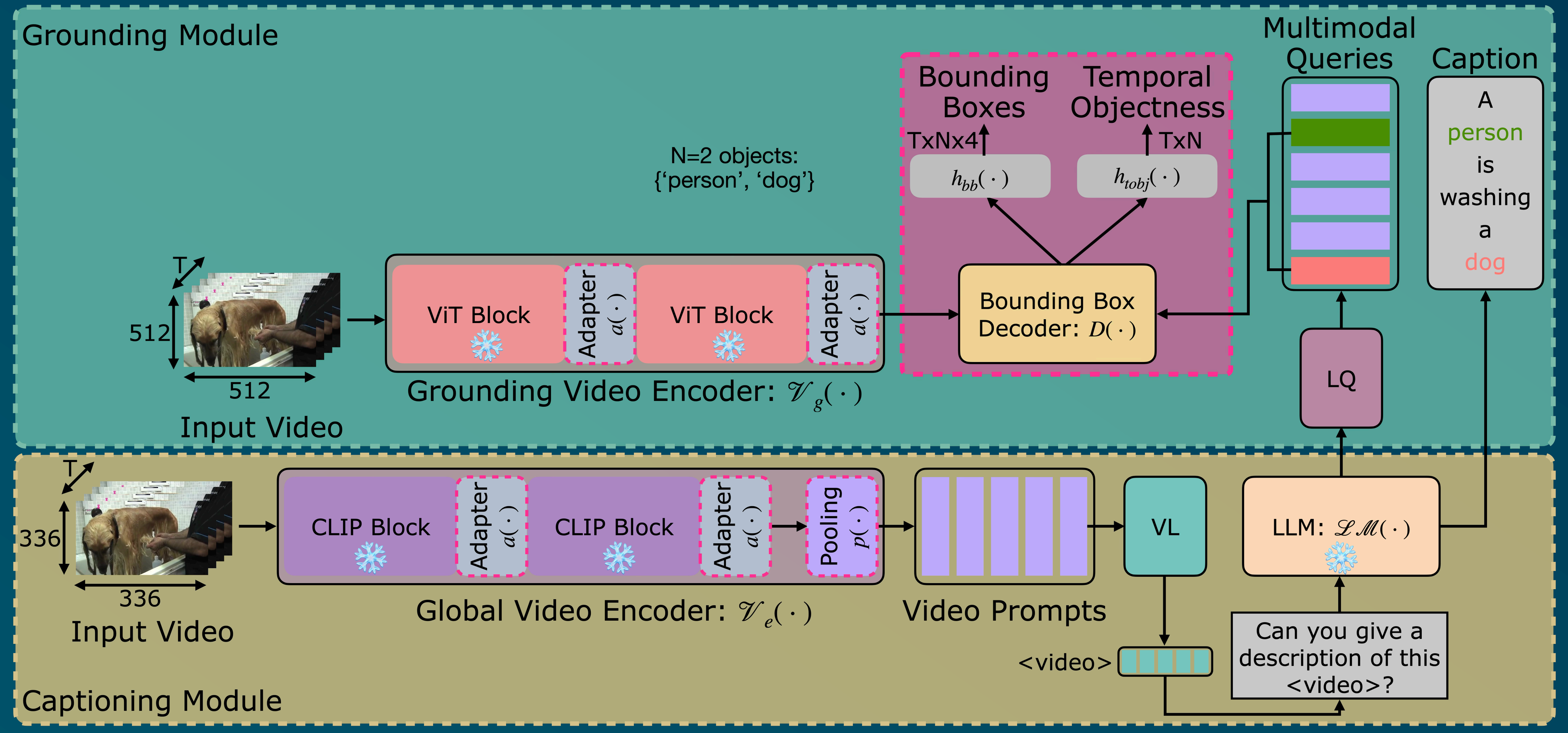
Stage 3: Connect boxes across frames with LLM



Example: muffin tin → tray

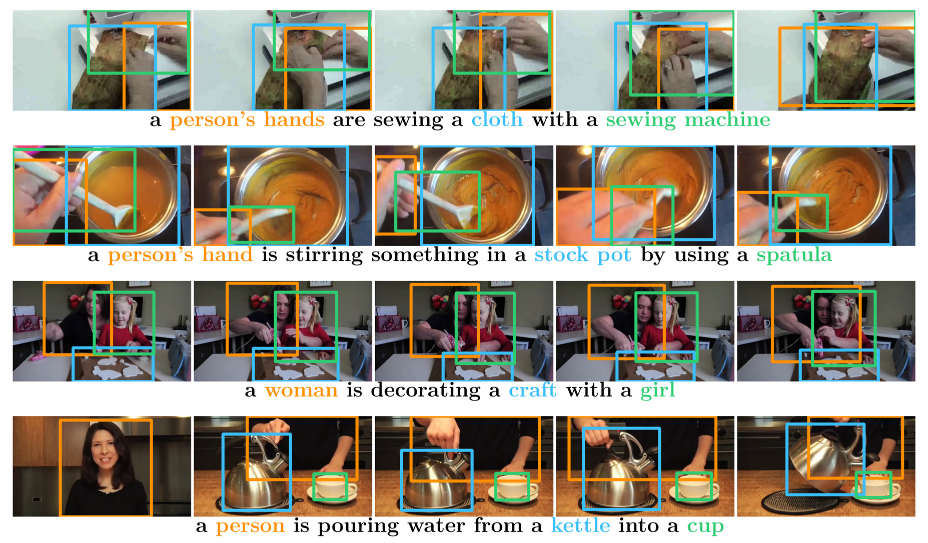


THE GROVE MODEL



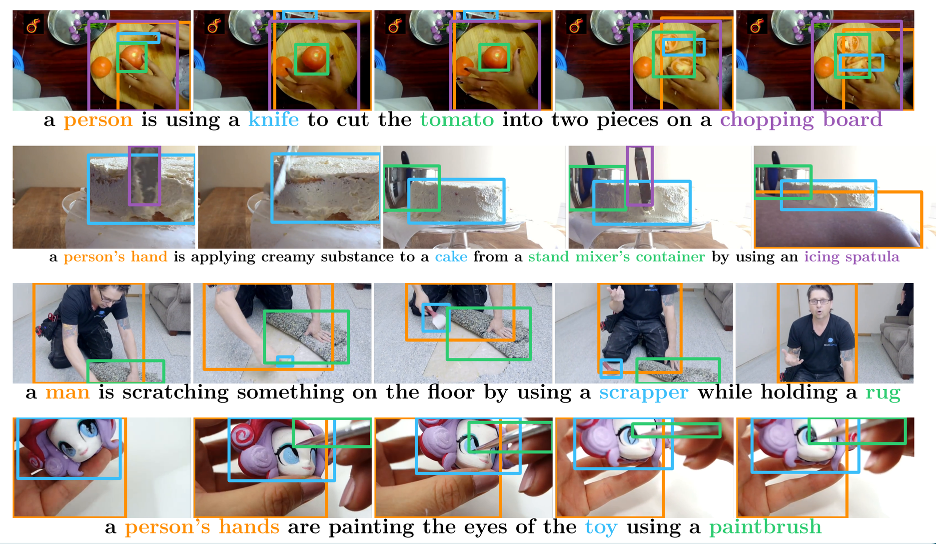
- ➤ 3D convolutional adapters for spatio-temporal modeling ($a(\cdot)$)
- ➤ Captioning module generates caption and tags noun phrases
- ➤ Cross-attention of embedded noun phrases with frame-level features ($D(\cdot)$)
- ➤ Temporal objectness head for predicting the presence of an object in a frame ($h_{tobj}(\cdot)$)

HowToGround1M DATASET



- ➤ Automatically annotated
- ➤ 1M videos
- ➤ 80.1M bounding boxes
- ➤ Ideal for pre-training

iGround DATASET



- ➤ Manually annotated
- ➤ 3500 videos
- ➤ Train/val/test: 2000/500/1000
- ➤ Ideal for fine-tuning and evaluation

RESULTS

TAKEAWAYS

- ➤ GROVE ≫ GLaMM in single-frame setup
- ➤ GROVE ≫ automatic annotation
- ➤ Pre-training is crucial for grounding
- ➤ Pre-training + fine-tuning 🔥

	Method	METEOR	CIDER	AP50	Recall
Center	GLaMM	11.9	29.9	20.8	19.3
Center	GROVE – PT (Ours)	14.3	50.6	27.0	22.5
Center	GROVE – PT + FT (Ours)	**21.4**	**83.5**	**31.7**	**26.2**
All	Automatic annotation	13.8	40.0	27.1	20.4
All	GROVE – PT (Ours)	14.3	50.6	33.6	24.3
All	GROVE – FT (Ours)	21.0	77.7	15.8	18.1
All	GROVE – PT + FT (Ours)	**21.4**	**83.5**	**40.0**	**28.7**

iGround test set

- ➤ SOTA in VidSTG
- ➤ Best performance without fine-tuning (interrogative)
- ➤ ++ SOTA in ANet-Entities, GroundingYT, YouCook-Inter. (👀 paper)

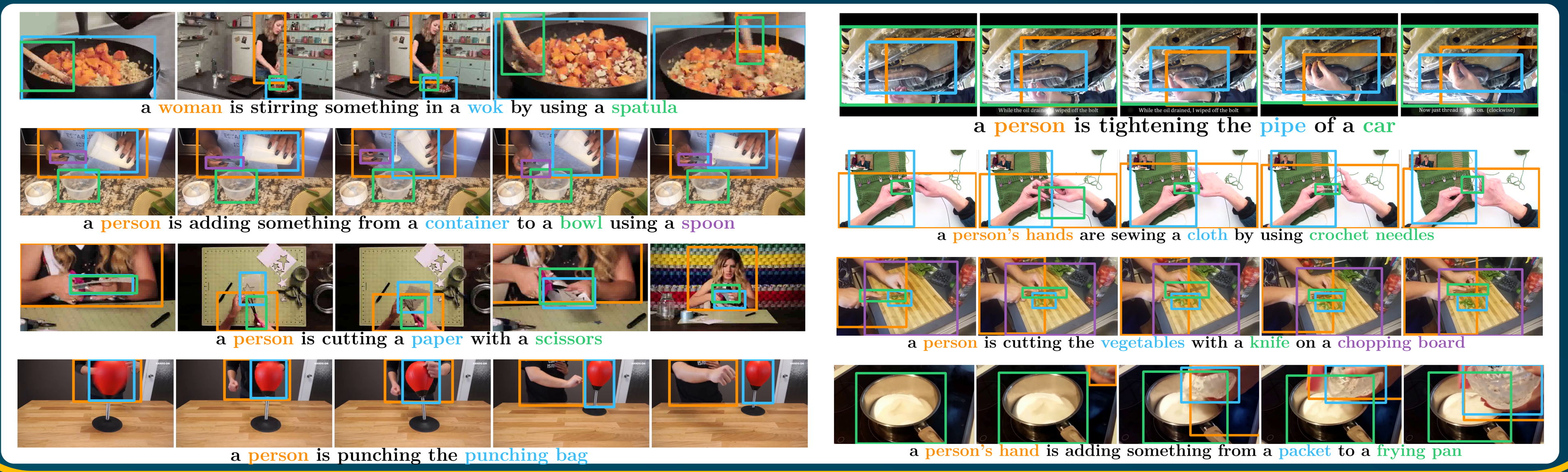
Method	m_{SIoU}
STVGBert	47.3
TubeDETR	59.0
STCAT	61.7
DenseVOC	61.9
GROVE – FT (Ours)	61.3
GROVE – PT+FT (Ours)	**63.7**

VidSTG (declar.)

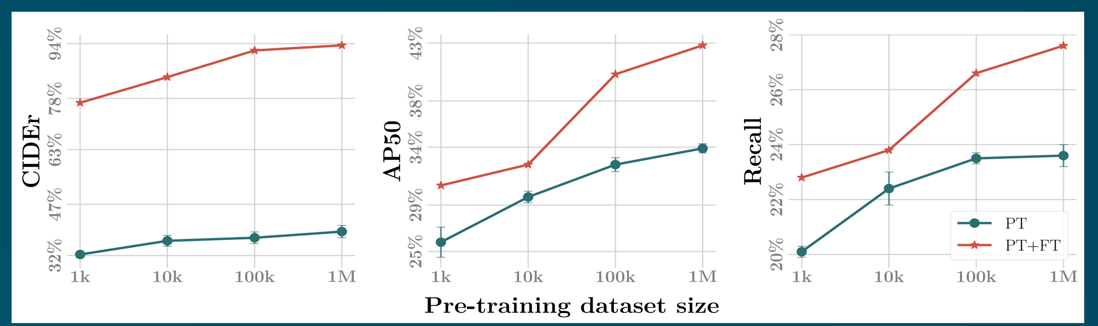
Method	FT	m_{SIoU}
PG-V-L (13B)	✗	35.1
GLaMM + SAM	✗	38.6
GROVE (Ours)	✗	**43.0**
VideoGLaMM	✓	39.7
GROVE (Ours)	✓	**55.0**

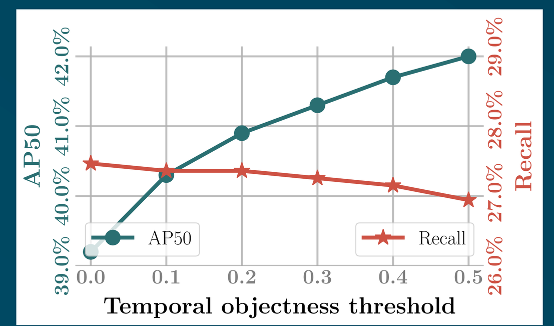
VidSTG (inter.)

▼ Qualitative results (iGround)



ABLATIONS





↓ 0.5 Recall
↑ 3.0 AP50

- ➤ **Scaling**: Performance scales with pre-training dataset size
- ➤ Fine-tuning benefits from scaling
- ➤ **Temporal objectness**: Little sacrifice in recall for 🚀 in AP50