Compositional Video Understanding with Spatiotemporal Structure-based Transformers

CVPR JUNE 17-21, 2024 SEATTLE, WA

Hoyeoung Yun[1*], Jinwoo Ahn[1*], Minseo Kim[1] and Eun-Sol Kim[1†]

1 Hanyang University, Seoul, Korea

Project Page

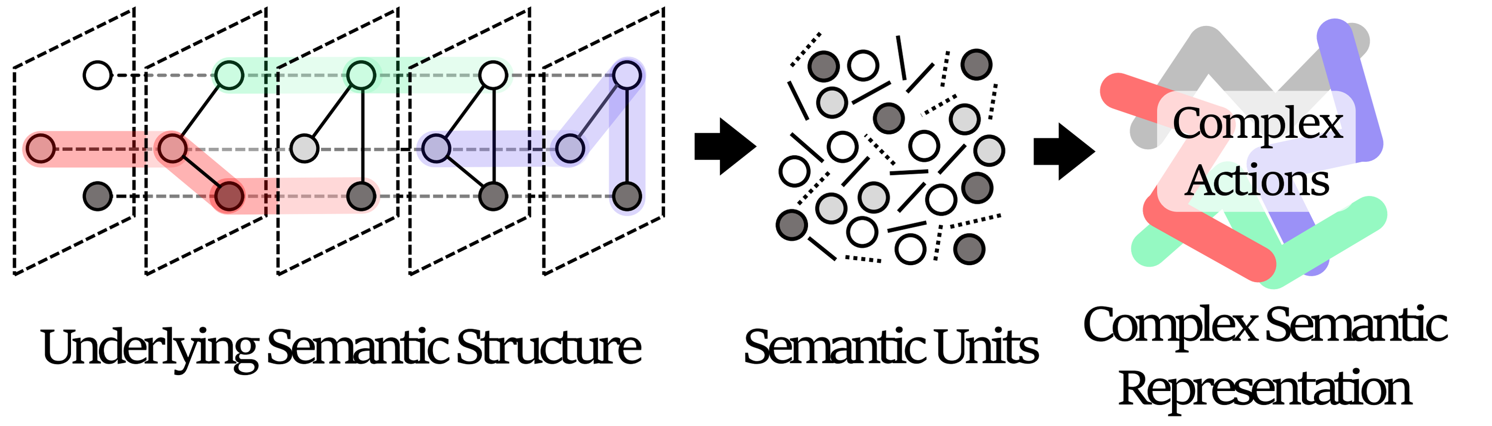
Introduction

Research Problem

- How to understand **multi-granular semantic structures** encompassing objects, scenes and video-wide contexts?
- How to understand video composed of **complex semantic structures**?

Key Insights

- **Spatiotemporal symbolic graph** as an input representation of a given video
- **Tokenizing** spatiotemporal graph and getting **semantic unit representation**
- Semantic units are then **combined to represent more complex semantics**



Compositional Generalization Ability

- Ability to **understand conceptual combinations unseen in the training process.**

Human Understanding Process (Train)

Compositional Generalization Ability (Inference)

Contributions

- A novel **object-oriented video understanding method** to learn the **multi-granular semantic structure of long videos** is suggested
- A novel **data split for compositional generalization test** of video understanding algorithms is proposed
- In experiments with both synthetic and real-world videos, we achieve new state-of-the-art performances

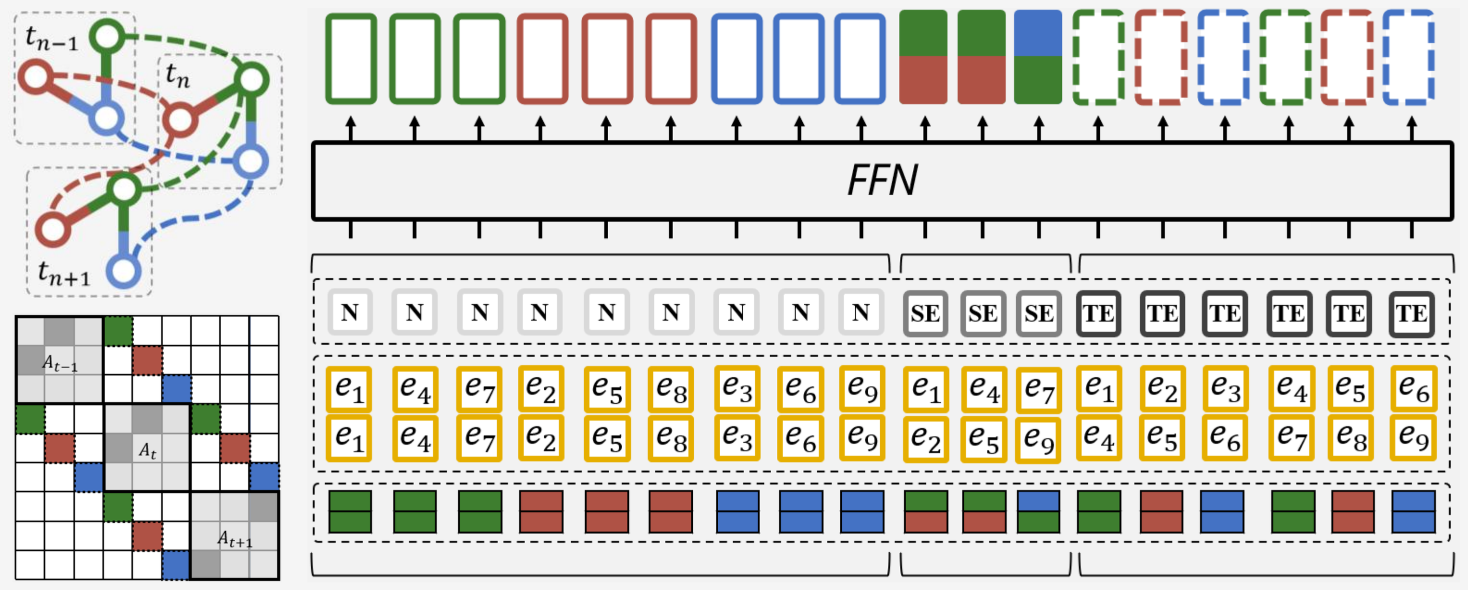
Proposed Method

Compositional Learning Framework

Spatiotemporal Graph Construction ➡ Spatiotemporal Graph Transformer ➡ Object-oriented Video Encoder ➡ Embedding Disentangling Module

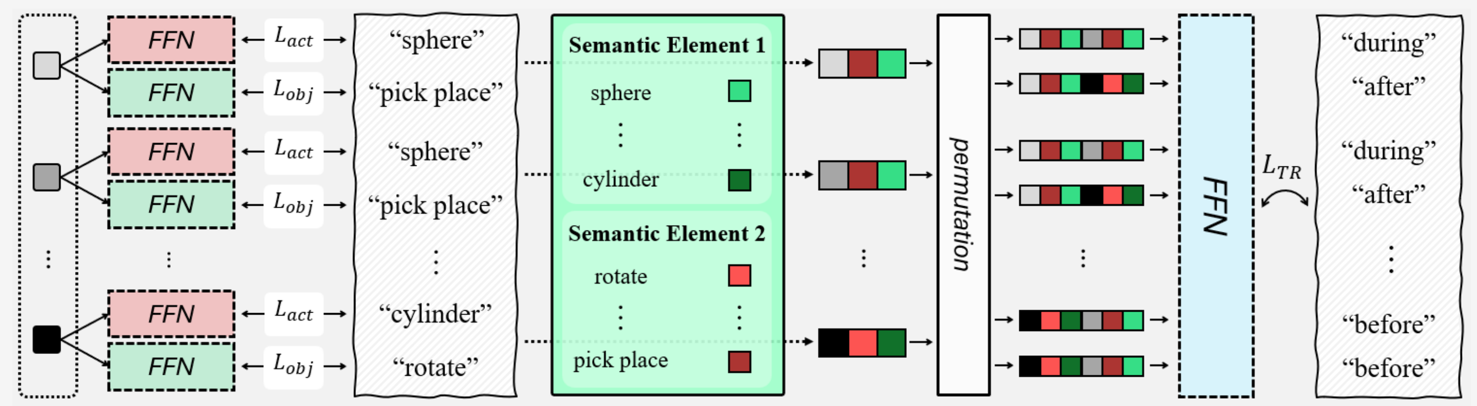
Spatiotemporal Graph Tokenizer & Graph Transformer

- ✓ All elements (**node, spatial edge, temporal edge**) are **tokenized** to be fed into the transformer as inputs
- ✓ We introduce **total adjacency matrix** which involves **temporal auxiliary edges** describing the temporal connection between two adjacent time
- ✓ As a positional embedding vector, **graph Laplacian of total adjacency matrix** is adapted

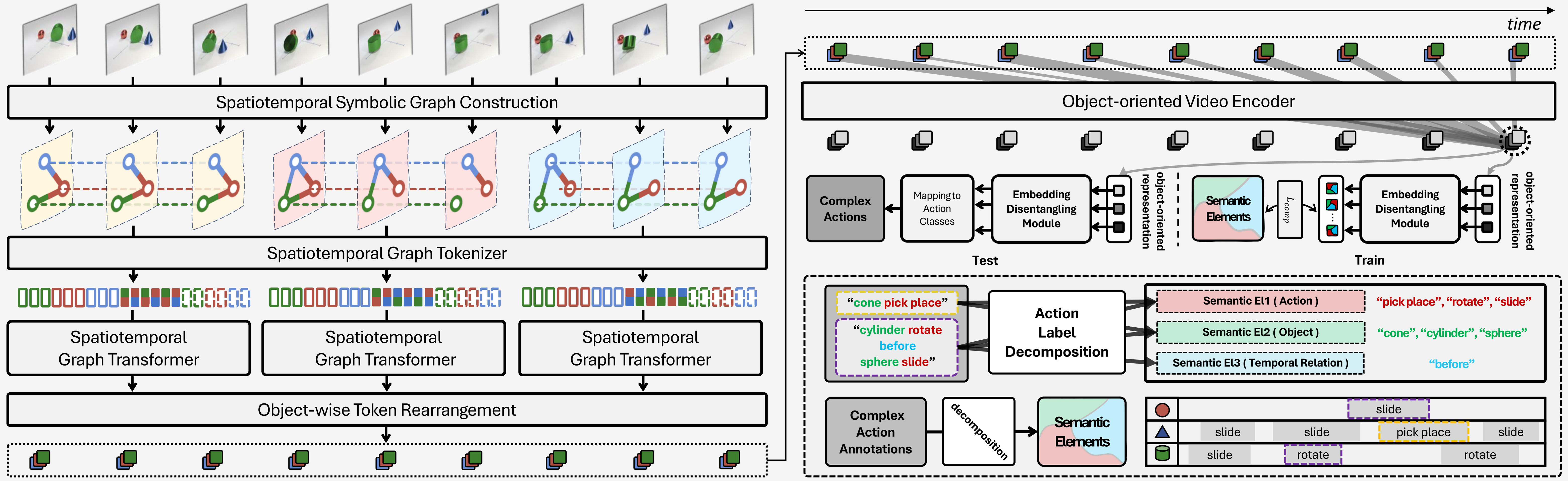


Embedding Disentangling Module

- ✓ To achieve compositional generalization ability, we introduce a method where **predictions are made for each semantic elements.** (e.g. object, action, temporal relation)
- ✓ To resolve this, **entangled feature embeddings are projected onto each independent subspace.** (implemented by learnable embedding)
- ✓ By concatenating each embedding, classification head is applied **to predict semantic labels that was previously decomposed.**



Overall Architecture



Experiments

Evaluation Methods

- Action Recognition with Synthetic Video Dataset (by CATER)
- Action Recognition with Real-world Video Dataset (by MOMA-LRG)
- **Compositional Generalization Test** (by newly suggested data splits)

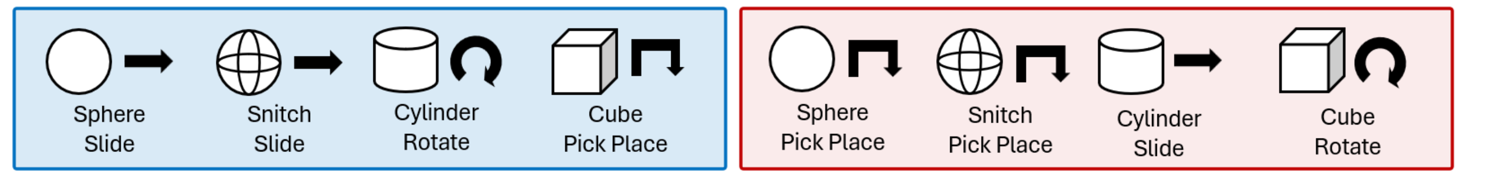
Compositional Generalization Test & Suggested Data Split

Train Phase

- Each of individual semantic elements is seen. **(ex. "sphere", "pick place" is seen)**
- But specific combinations of semantic elements is unseen. **(ex. "sphere pick place" is unseen)**

Test Phase

- Input video containing **"sphere pick place"** is given



Trainset & Testset Example of Object + Action Combination

Quantitative Results

Method	mAP
FasterRCNN [29]	25.45
R3D [4, 12]	44.2
R3D + NL [37]	45.9
R3D + LSTM [10]	53.4
R3D + NL + LSTM [10]	53.1
SCI3D + LSTM [29]	66.71
Single stream SCI3D + LSTM [29]	69.76
ViViT[1]	66.18
FROZEN[2]	66.64
Ours	**75.40**

Method	mAP
R3D [4, 12]	98.8
R3D + NL [37]	98.9
Ours	**99.88**

Method	mAP	# params	Inference time (ms)
ViViT [1]	65.09	99M	23
FROZEN [2]	69.36	115M	40
Ours	**72.83**	**7.3M**	**20**

Token Types			Dataset	
N	SE	TE	CATER	MOMA-LRG
✓	-	-	72.49	49.76
✓	✓	-	74.49	66.47
✓	-	✓	74.48	66.78
✓	✓	✓	**75.40**	**72.83**

	Task 1		Task 2	
	val	test	val	test
L_{vid}	90.42	90.38	56.16	53.53
$L_{vid} + L_{comp}$	99.89	99.88	76.07	75.40

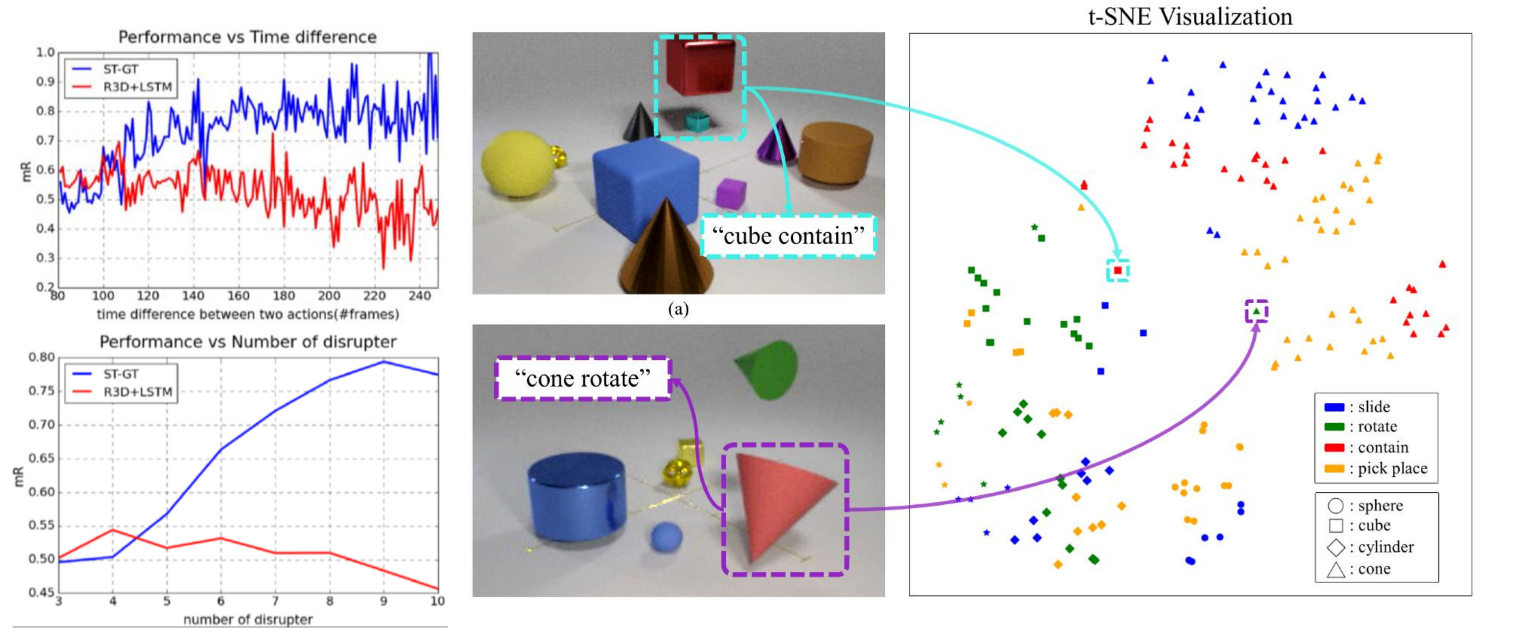
Action Recognition Performance & Ablation Study

	Validation		Test	
	Baseline	Ours	Baseline	Ours
$\mathcal{D}^1_{test}$	25.78	72.95	25.05	71.53
$\mathcal{D}^1_{train}$	93.95	98.9	93.04	99.1

Split		R3D	R3D+LSTM	Ours
Validation	$\mathcal{D}^2_{test}$	4.31	4.00	69.45
	$\mathcal{D}^2_{train}$	60.32	76.49	81.1
Test	$\mathcal{D}^2_{test}$	3.47	3.51	69.2
	$\mathcal{D}^2_{train}$	58.23	74.84	80.9

Compositional Generalization Result

Qualitative Results



Acknowledgement

This work is supported by IITP grant funded by MSIT (Grant No. 2022-0-00264/40%, 2022-0-00612/20%, 2022-0- 00951/20%,) and IITP Artificial Intelligence Graduate School Program for Hanyang University funded by MSIT (Grant No. RS-2020-II201373/20%).