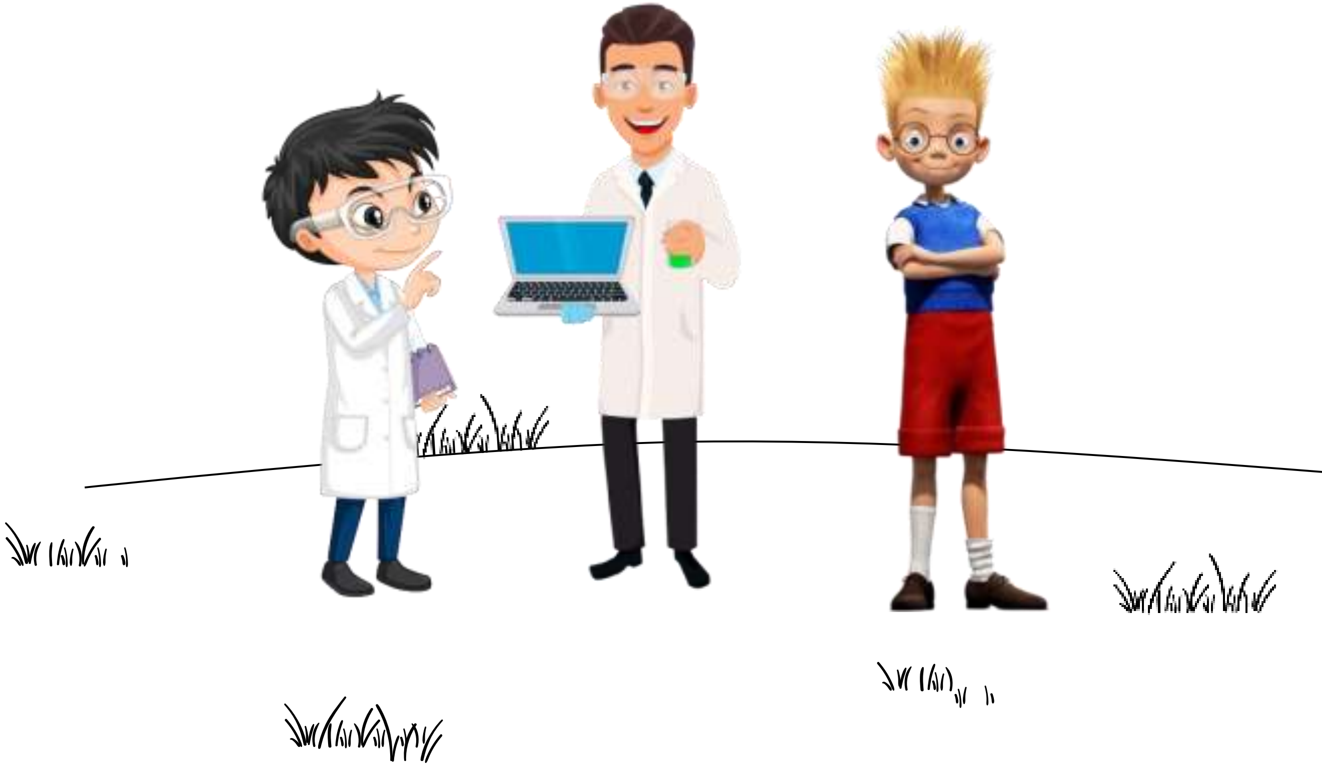
S-GEAR: Semantically Guided Representation Learning For Action Anticipation

Authors: Anxhelo Diko, Danilo Avola*, Bardh Prenkaj*, Federico Fontana, Luigi Cinque
Contact email: diko@di.uniroma1.it, bardh.prenkaj@tum.de, f.fontana@di.uniroma1.it



EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO
2024

SAPIENZA UNIVERSITÀ DI ROMA

TUM

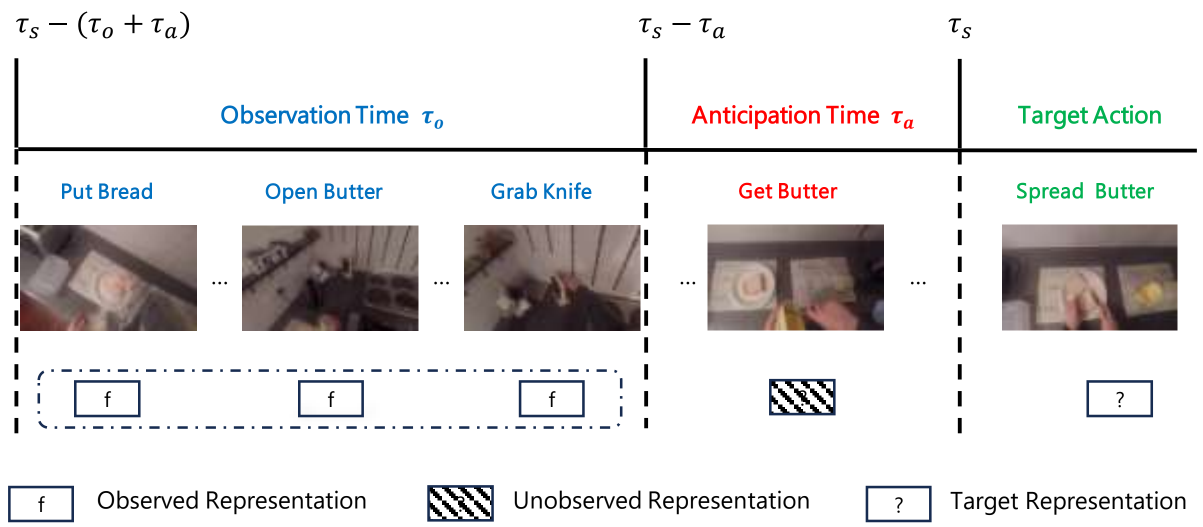
Content

Introduction

Action anticipation involves predicting an action category for an event starting at time τ_s requiring analysis of a sequence of events from the interval $[\tau_s - (\tau_o + \tau_a); \tau_s - \tau_a]$, where τ_s, τ_o and τ_a denote the starting, observation and anticipation periods.

Important applications:

1. Autonomous driving
2. Wearable assistants



Problem

Action anticipation, as an extension of action recognition, is prone to future uncertainty and the difficulty of reasoning upon interconnected actions.

Current approaches:

1. Action recognition methods

2. Temporal modeling through LSTMs or Causal Transformers

Important aspects that are not addressed:

1. Action semantic connectivity and co-occurrence

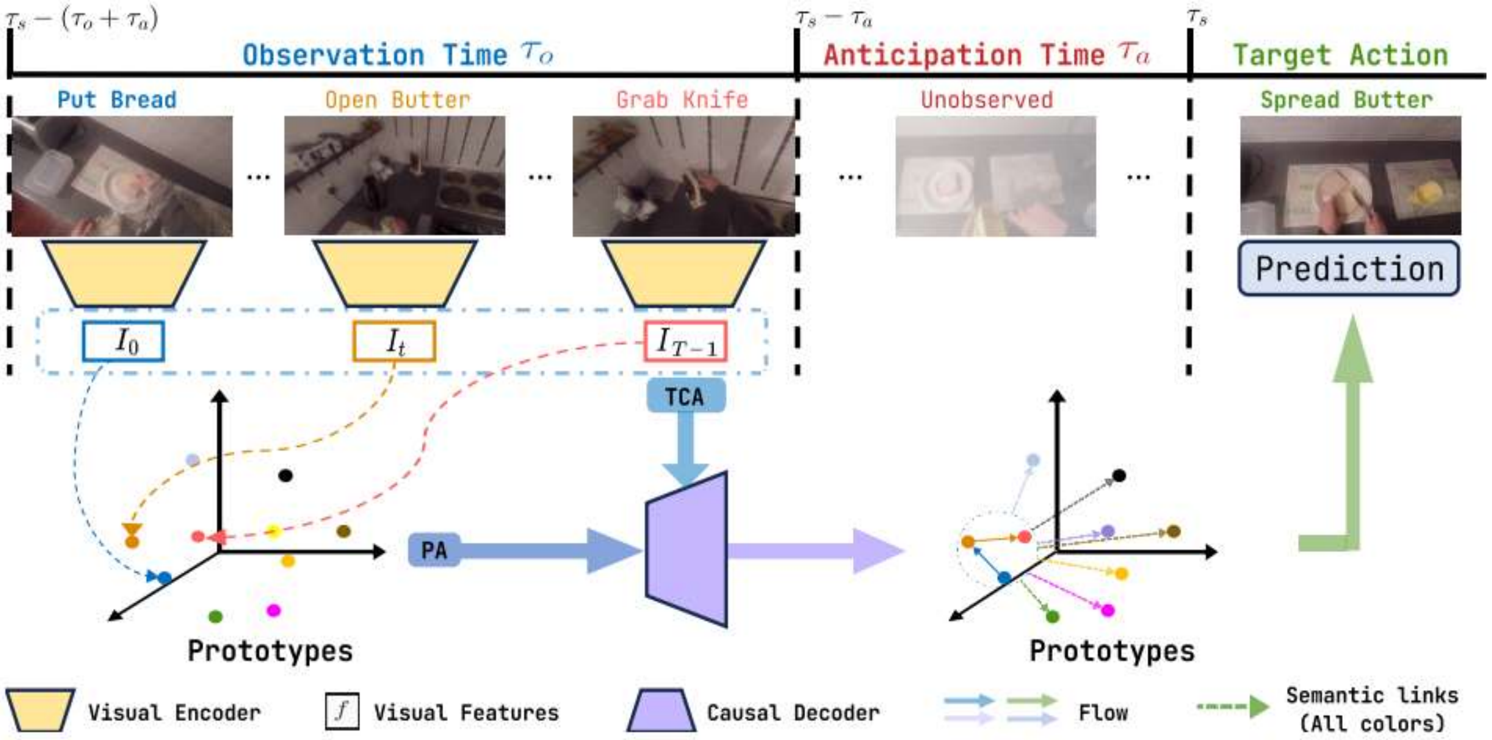
How do we deal with it?

S-GEAR Framework

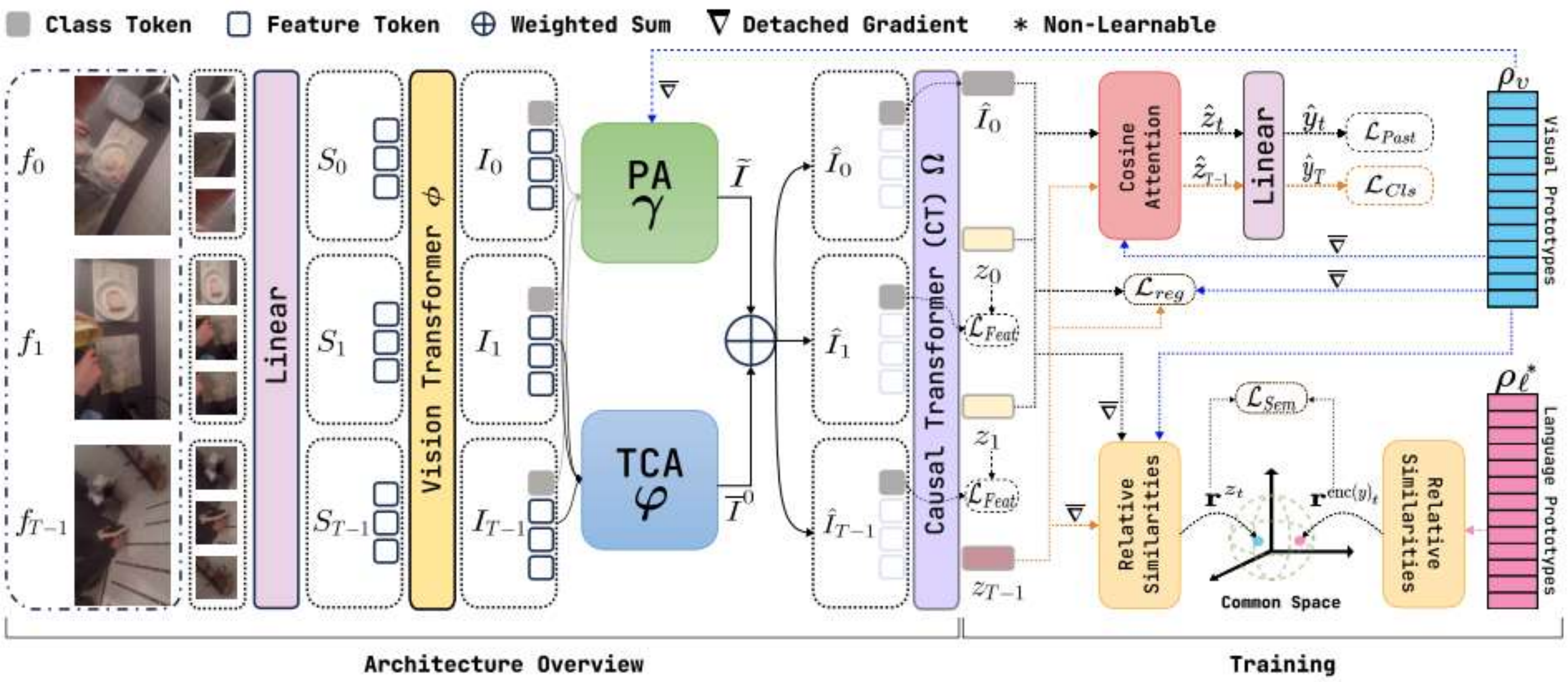
We present Semantically Guided Representation Learning Framework (S-GEAR) for action anticipation.

S-GEAR, inspired from fundamentals of semantic connectivity, uses prototypical learning to:

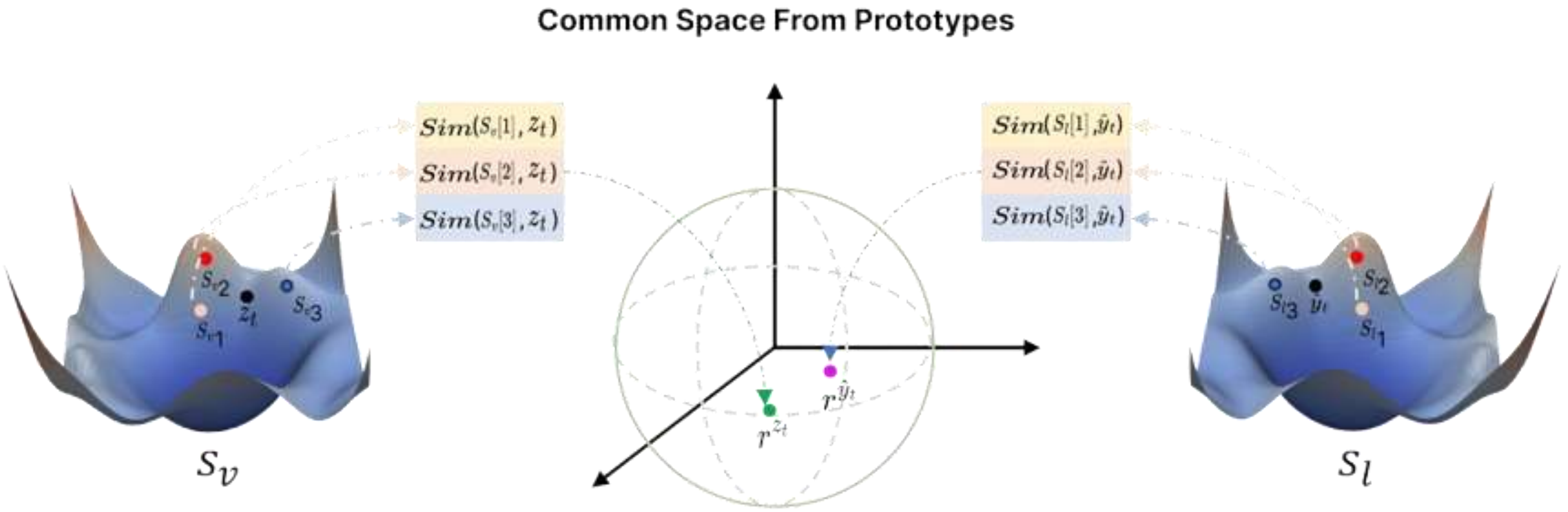
1. Model typical action patterns
2. Semantic relationships based on co-occurrence.



Architecture



Common Communication Space

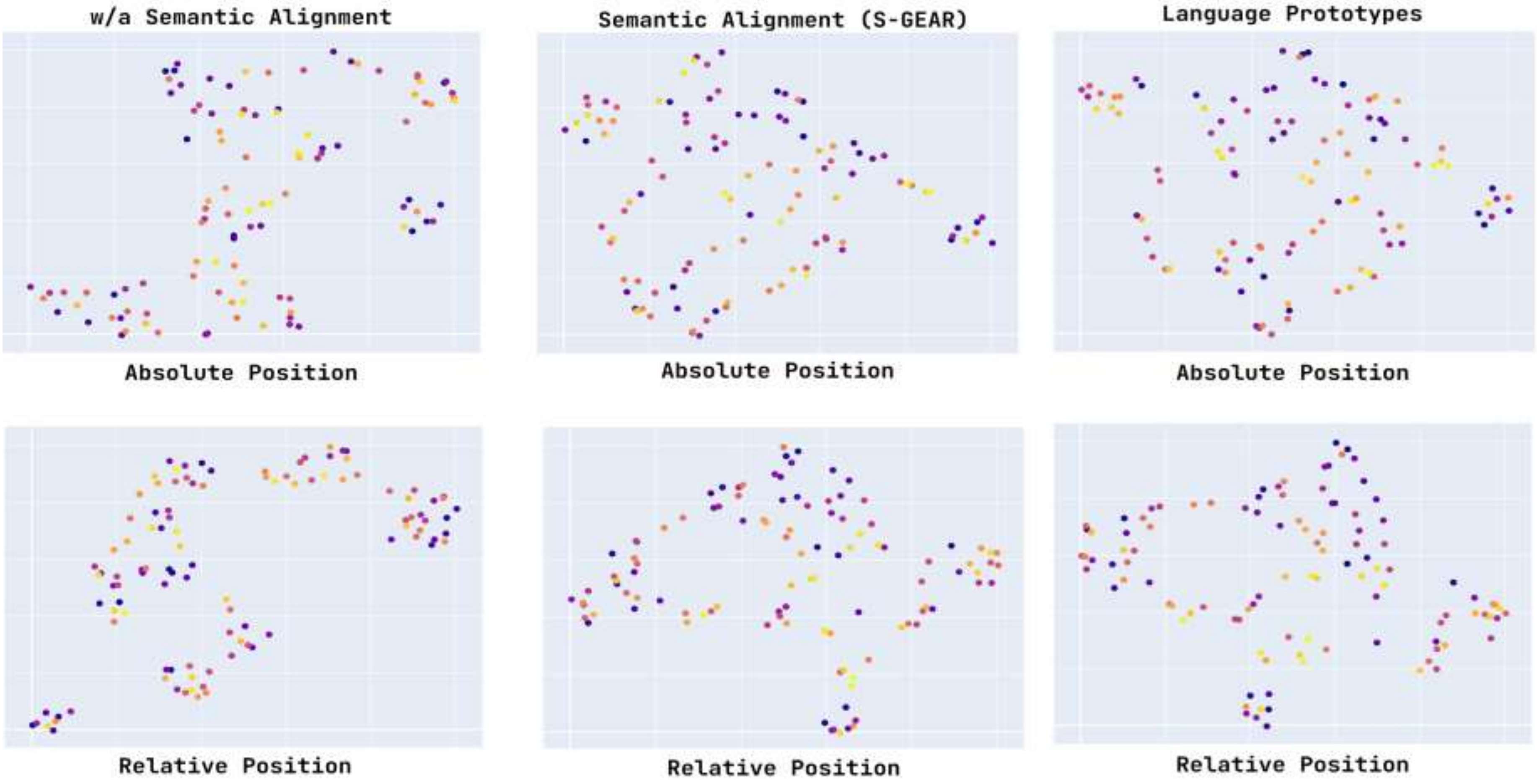


S_v - Visual Space
S_l - Language Space

r^{z_t} - Relative Representation of Visual Encoding z_t
$r^{\hat{y}_t}$ - Relative Representation of Language Encoding $\hat{y}_t$

Refer to main paper for more details.

Qualitative Results - Learned Representations



Refer to main paper for quantitative results and comparison with previous SOTA

Conclusions

Conclusions from our study:

- Modelling actions co-occurrence is critical for the task of action anticipation.
- Using common spaces to align modalities, allows S-GEAR to learn co-occurrence from language representation while preserving the visual information.
- S-GEAR lacks the ability to model the order of co-occurrence between action.

Thank you slide