SP25 CSE 5245 Interim Project Presentation

PanViS: Panoptic **Vi**deo **S**GG for Zero-shot Visual Comprehension

Mona Gandhi, Sriram Sai Ganesh, Abhinay Putta

What

We're focusing on enhancing video understanding by generating structured scene graphs from video content.

Our *scene graph from video generation system* helps *AI models and researchers* who want to understand changes in complex video content by extracting scene graphs and video-based downstream tasks like question answering.

Team 7: PanViS

Panoptic Video Scene Graph Generation for Zero-Shot Visual Comprehension

Team members:

Mona Gandhi Ram Sai Ganesh Abhinay Putta

Detailed Motivation

Interpretability is very low!

Relies on symbolic programs for QA

STAR BENCHMARK

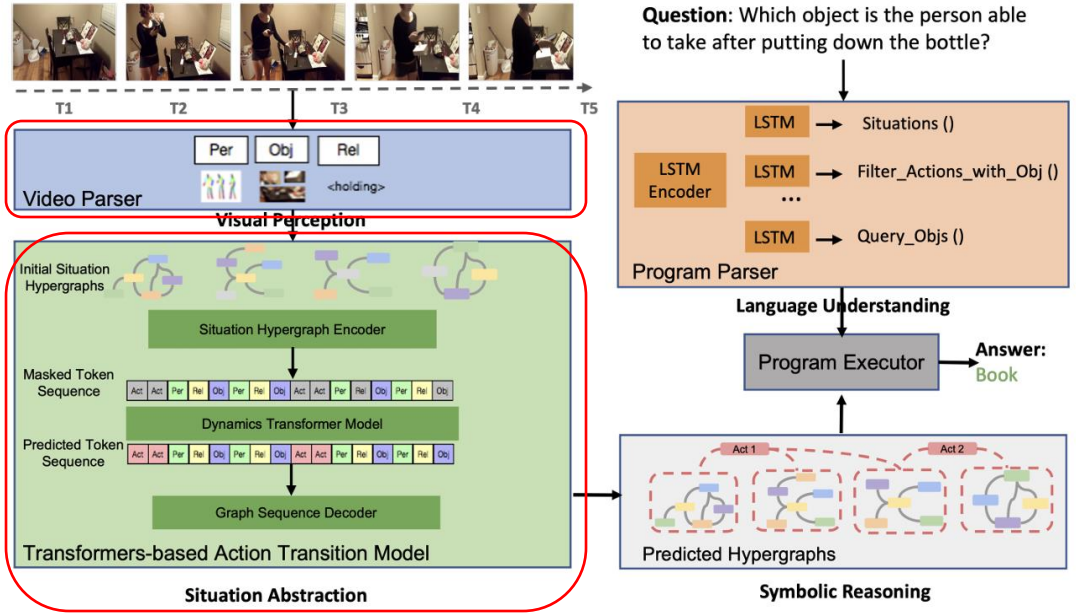


Figure 3: The architecture overview of NS-SR. It use a video parser to perceive entities, relationships and human-object interactions for visual situations. The present situation is sent to a transition model to learn complete situation abstraction and predict future situations in forms of a situation hypergraph. A program parser parses the question and options into a set of nested functions. The generated hypergraph fed to a symbolic program executor to get the answer. Best viewed in color.

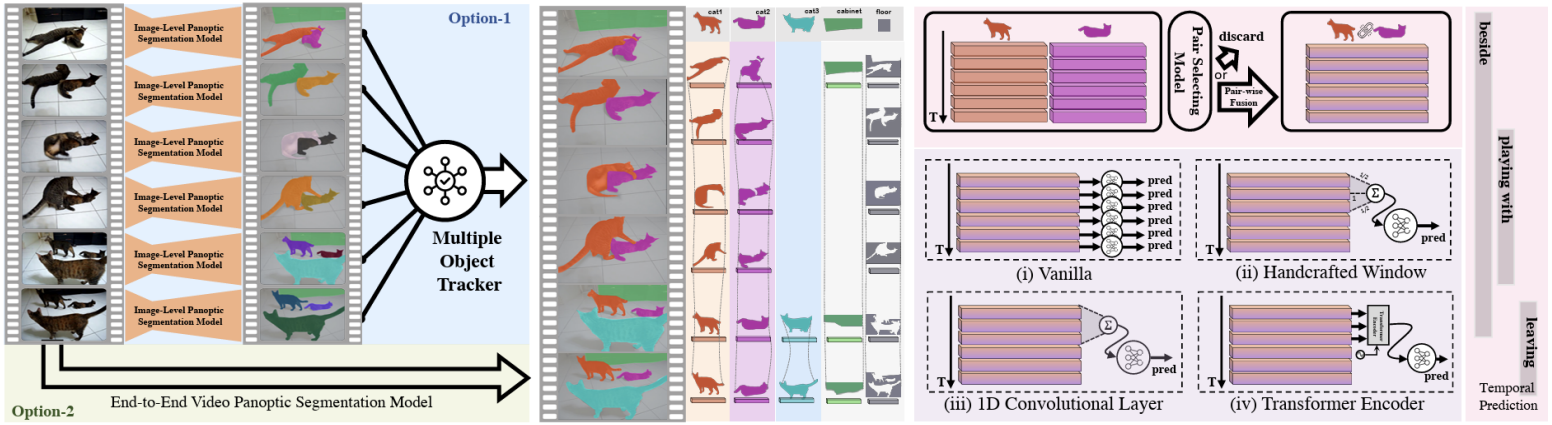
Detailed Motivation

Better Interpretability

A lot of annotations required for training!

Generalizability is very low!

Only handles subjects, that is ignore relations between background objects



(a) Stage-1: For Feature Tube and Mask Tube Output

(b) Stage 2: Relation Prediction

PVSG

Figure 5. **The two-stage framework to solve the PVSG task.** The goal of the first stage is to obtain the video panoptic segmentation mask for each object, as well as its corresponding video-length feature tube. Two options are provided to achieve the goal. The second stage predicts pairwise relations based on all the feature tubes from the first stage. Four options are provided for a comprehensive comparison.

Detailed Motivation

We aim to achieve the following through our approach to create a scene graph:

- Interpretability
  - Distributing each step to generate useful intermediate outputs
- Generalizability
  - Zero-shot Approach
  - Work with all type of video datasets
- Better Aligned with Natural Language
  - Make sure the relations between objects are plausible

Overall: Create richer scene graphs that can be used for various downstream tasks.

Deliverables

A pipeline that processes video input to generate structured scene graphs with

- ✅ Detected entities
- ✅ Entity video segmentation
- ✅ Relationships
- ⏳ Temporal consistency

To enable **improved Video–QA** performance.

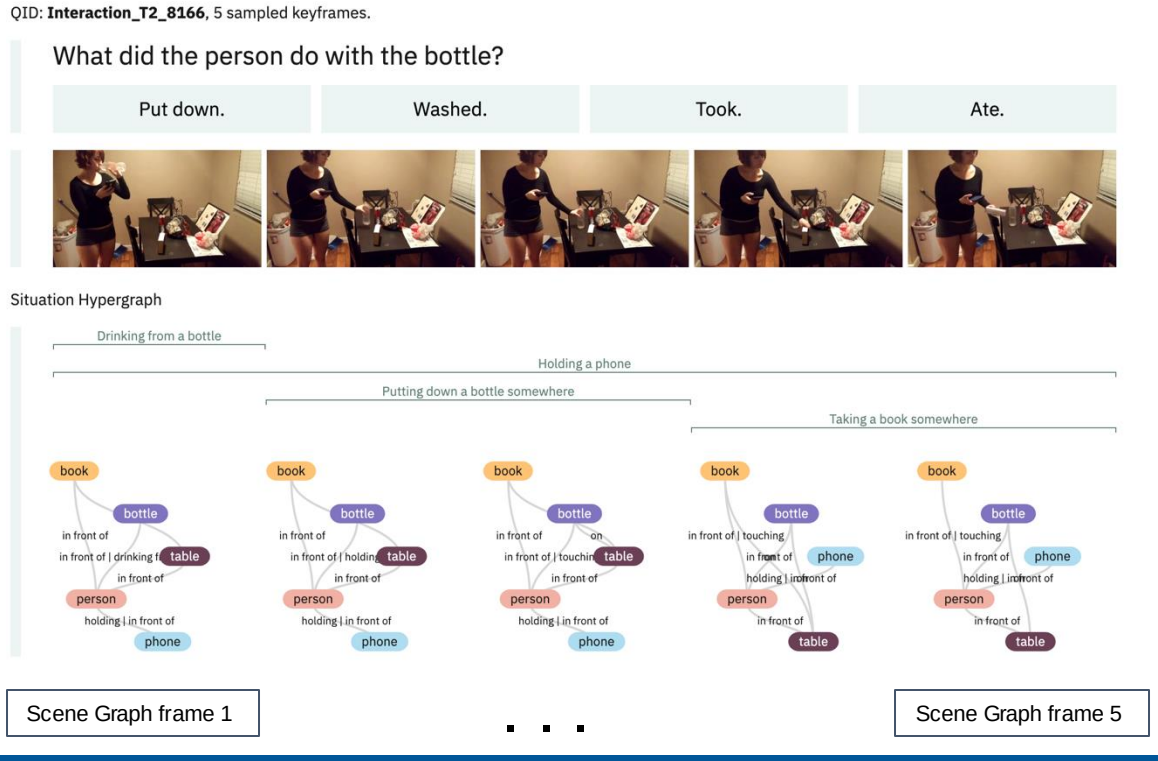
STAR Dataset

Question

Choices

Video

Actions across Frames



STAR Dataset

4 Question Types
- Interaction
- Sequence
- Predictive
- Feasibility

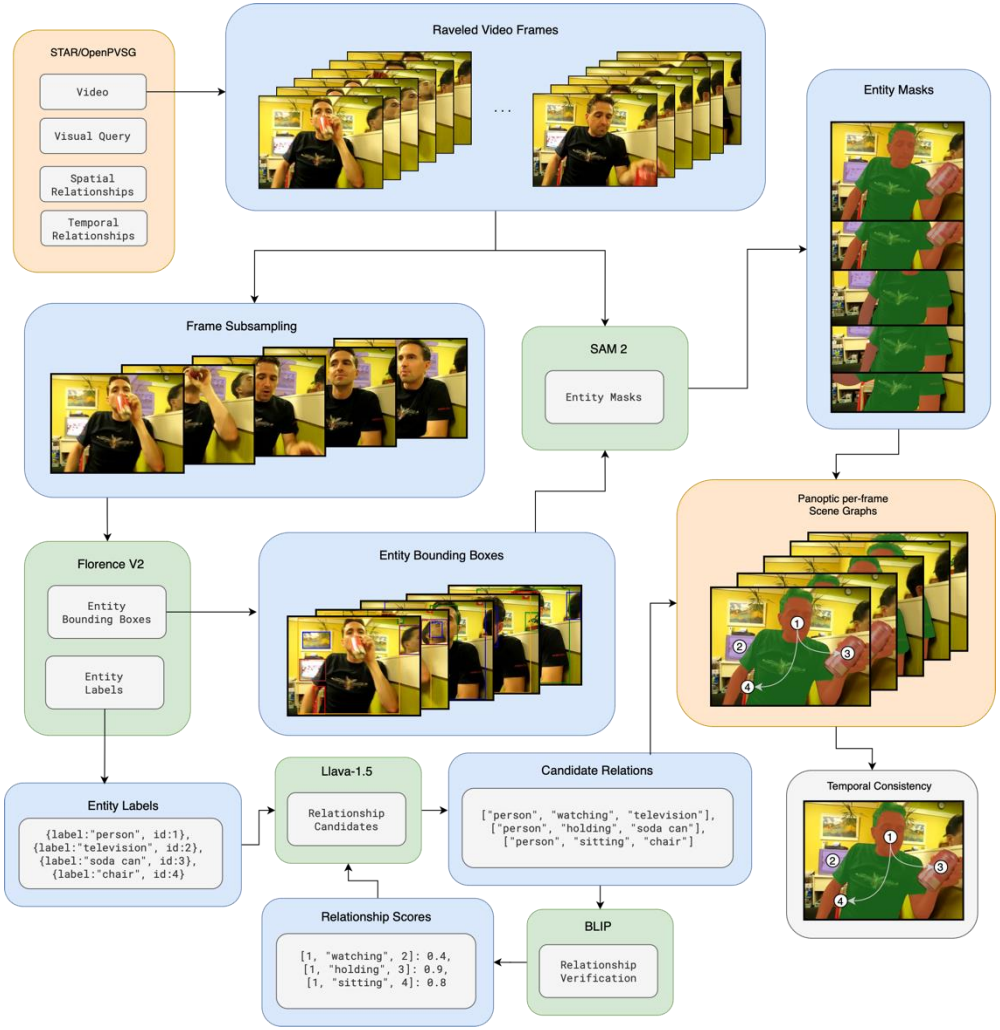
Annotation Statistics
- 111 action classes
- 37 entity classes
- 24 relationship classes

22K Situation Video Clips
60K Situated Questions
140K Situation Hypergraphs

Our Usage

- Only work with Validation data (currently) as we do not require any training.
  - 1k Situation Video Clips
  - 7k Situated Questions
- Annotations reduce our generalizability to other objects and actions.

PanViS Architecture



Entity Detection

Microsoft's Florence V2

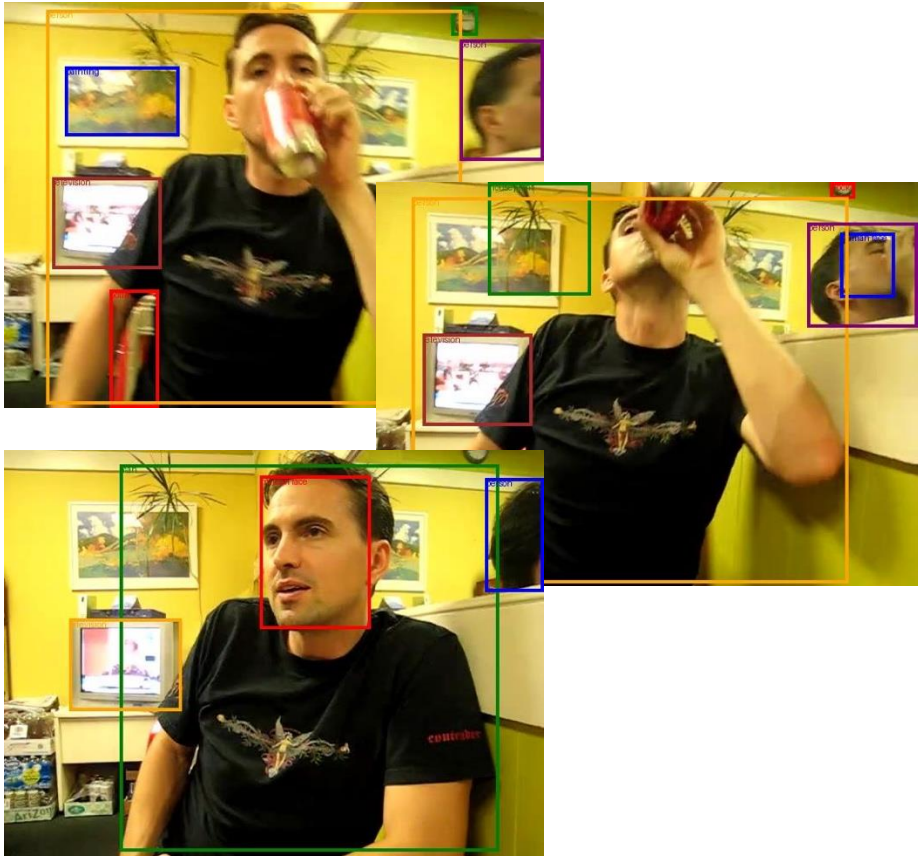
Zero-shot entity detection.

Input:

- Video frames (subsampled.)

Output:

- Labels & bounding boxes for entities.



Video Segmentation

Meta's SAM 2.1

Entity instance segmentation

across the video.

Input:

- Video frames &

  entity bounding boxes

Output:

- Entity binary masks (across video).

Relationship Proposal

Llava-1.5-7b-hf

Relationship proposal (pairs of entities)

Input

- Pairs of entity names:

  “<entity> and <entity>”.

Output

- Candidate relationships between entities: verbs or verb-phrases.

Candidate relationships between “**man**” and “**soda can**”:

```
["standing", "walking",
 "drinking", "throwing",
 "collecting"]
```

Relationship Verification

Salesforce's BLIP

Relationship verification.

Input:

- Image captions:

  "<entity> - <relation> - <entity>".

Output:

- Confidence scores in each caption.

"A man **standing** a soda can": 0.05
"A man **walking** a soda can": 0.04
"A man **drinking** a soda can": **0.41**
"A man **throwing** a soda can": 0.18
"A man **collecting** a soda can": 0.11

Evaluation

Scene graph Quality Evaluation:

Compare situated scene graphs from STAR (SG_{STAR}) to scene graphs generated using our pipeline (SG_{PanViS})

- **Missing_X** : Accuracy based metrics to verify all X from SG_{STAR} is present in SG_{PanVis}
- For each Frame:
    - Calculate Missing_Entities and Missing_Relations
- Note: We will have more objects and relationship in SG_{PanViS} than SG_{STAR}
- For overall video: calculate Missing_Actions (Temporal Relations)

Downstream QA Evaluation:

- Using SG_{PanViS}, how well can we answer questions from STAR dataset (currently)?
- Method:
    - Narrow the important frames based on objects in the question.
    - Get a subset of the scene graph involving those objects.
    - Prompt an LLM to use the subset of the scene graph to get the answer.

SP25 CSE 5245 Interim Project Presentation

PanViS: Panoptic **Vi**deo **S**GG for Zero-shot Visual Comprehension

Mona Gandhi, Sriram Sai Ganesh, Abhinay Putta

Thank you!
Questions?