SP25 CSE 5245 Presentation

Panoptic Video Scene Graph Generation

Jingkang Yang et al.

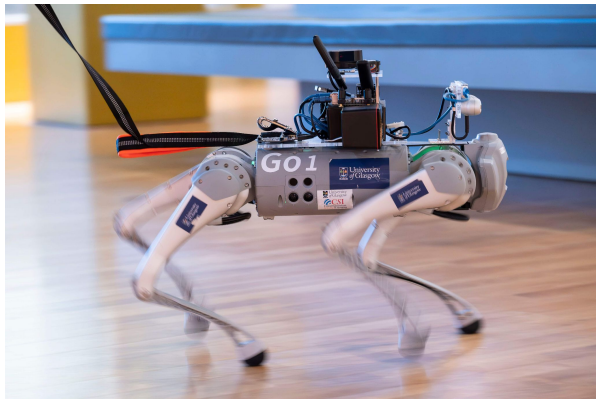
CVPR 2023

Paper: arxiv.org/abs/2311.17058

Presented by Abhinay, Mona & Ram

Introduction Utility

- AI needs to understand what's happening in a video, not just see what's in it
- Panoptic Video Scene Graphs provide more accurate object masks and relations between objects



Guide Robot for Blind
"Person in front of staircase"

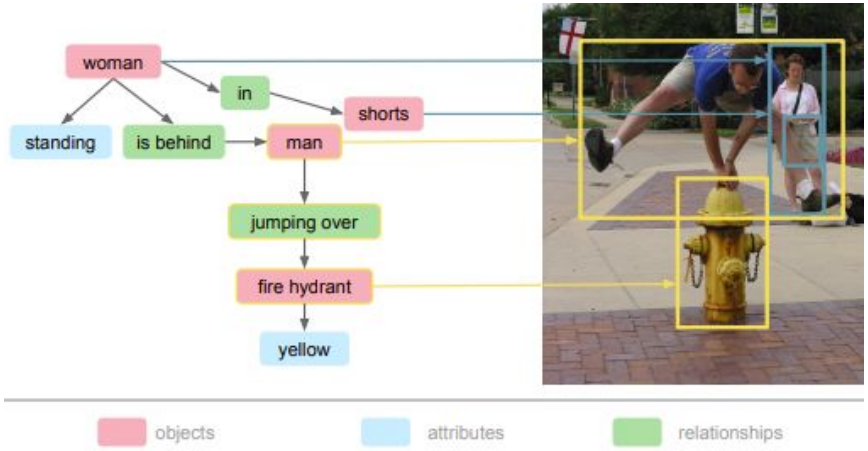
Theft
"Person grabbing bag"

Autonomous Dro
"Person trapped in car"

Background

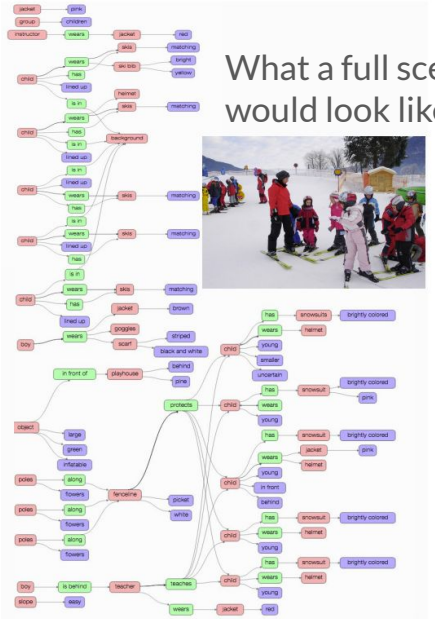
Panoptic Video Scene Graphs

- Scene graphs take the objects as nodes and have edges as the relations between objects
- Give you richer information about the relationships between objects for downstream tasks (VQA, reasoning, etc) information abou



Adapted from: Krishna et al., Visual Genome, CVPR 2017

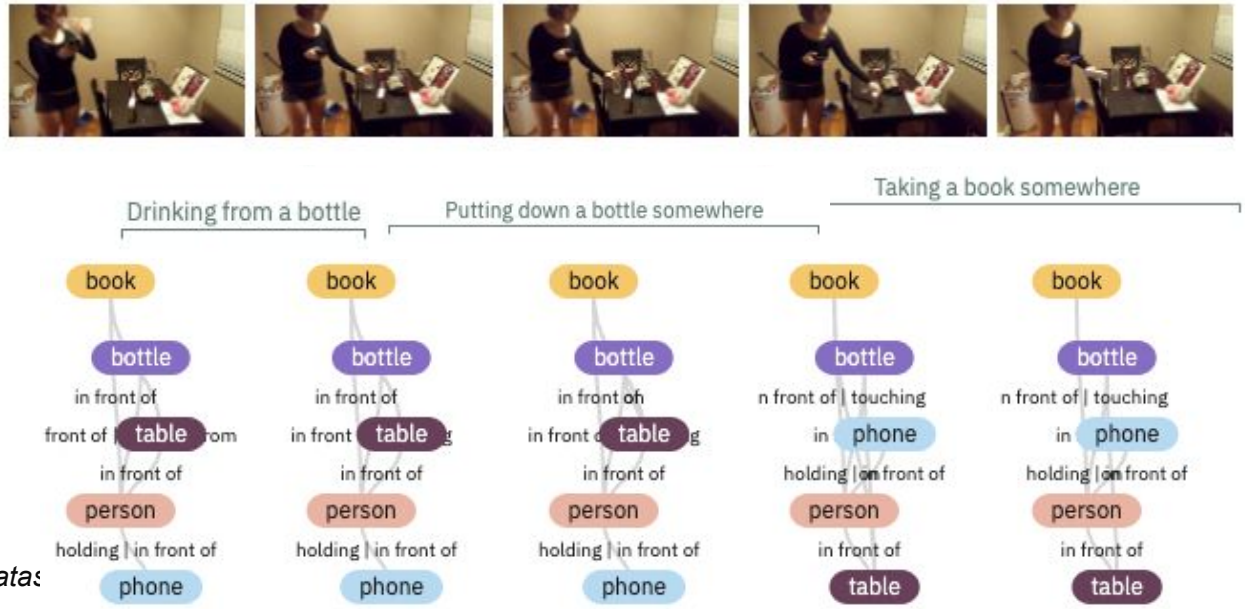
What a full scene graph would look like



Background

Panoptic Video Scene Graphs

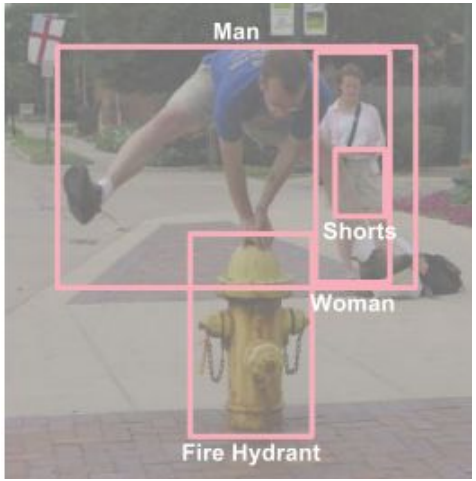
- Now imagine generating a graph like this for every number of interval between frames
- Capture the temporal element of when things happen and when things end



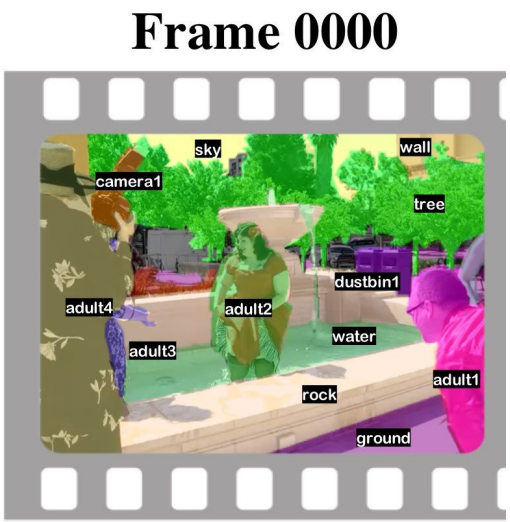
Adapted from: STAR datas

Background Panoptic Video Scene Graphs

- Bounding boxes can lead to loss of fine detail for video understanding
- Instead, panoptic segmentation masks use pixels to capture all objects and backgrounds
    - These masks cannot overlap



Adapted from: Krishna et al., Visual Genome, CVPR 2017



Adapted from: Yang *et al.*, *Panoptic Video Scene Graph Generation*, arXiv:2311.17058, 2023.

Related Work

Scene Graph Generation

- Bounding-box-based models like MotifNet Graph R-CNN
- Mostly trained on Visual Genome
- Predict static relationships from one image

Limitations:

- No temporal info
- Not accurate by pixels in the box
- Can miss things like sky or grass

Video Scene Graph Generation

- Extends SGG to track relations over time
- Trained on datasets like VidOR, ImageNet-VidVRD
- Models like TRACE, MVSGG

Limitations:

- Uses bounding boxes
- Struggle with occlusion, overlapping objects, background

Related Work Video Panoptic Segmentation (VPS)

- VPS combines:
- Video Semantic Segmentation (segment background & uncountable materials/regions)
- Video Instance Segmentation (segment & track objects)
- Outputs non overlapping masks over time with IDs
- Models: VPSNet, VIP-Seg, VIP-Deeplab, K-Net, TubeFormer

Limitations:

- Segment and track objects but lacks info about the relationships between objects. This is where scene graphs come in!

PVSG The Problem

GOAL: Describe a given video with a **Dynamic Scene Graph**, with each node associated with an object and each edge associated with a relation in the temporal space.

Input: Video clip $V \in \mathbb{R}^{T \times H \times W \times 3}$ where T → #frames and H x W → frame size
Output: Dynamic Scene Graph G

$$\Pr(\mathbf{G} \mid \mathbf{V}) = \Pr(\mathbf{M}, \mathbf{O}, \mathbf{R} \mid \mathbf{V})$$

G comprises of :

- Binary mask tubes $M = \{m_1, ..., m_n\}$
- object labels $O = \{o_1, ..., o_n\}$
- Relations set $R = \{r_1, ..., r_l\}$

} Corresponds to n objects in the video

For $object_i$, the mask tube $m_i \in \{0,1\}^{T \times H \times W}$ collects all its tracked masks in each frame

PVSG Metric

Output: Predict a set of triplets
Example: relation r_i from t_1 to t_2 with subject with o_s and mask $m_s^{(t1,t2)}$, and an object with o_o and $m_o^{(t1,t2)}$. $m^{(t1,t2)}$denotes the mask tube m span across the period of t_1 to t_2.

Metric:
R@K → Triplet recall given the top K triplets
mR@K → Mean triplet recall given the top K triplets

Successful Recall:
1. The correct category labels of the subject, object, and predicate.
2. the predicted mask tubes the ground-truth tubes should have volume IOU over 0.5.

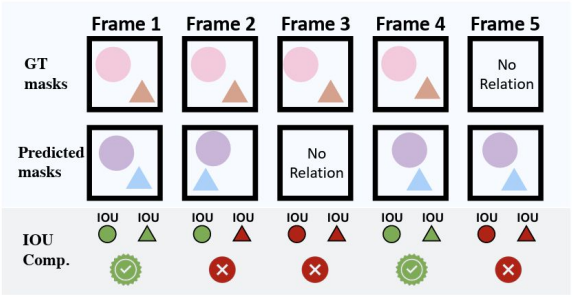


Figure 3. **Illustration of the PVSG Metric.** Assuming the classification of the triplet is correct, to further match the ground truth (GT) frame-wise, the predicted mask pair must have both subject and object masks with a mask IOU above 0.5. In this case, only Frames 1 and 4 satisfy this condition, yielding an intersection count of 2 and a union count of 5. Thus, the volume IOU is calculated as 0.4. As this value falls short of the 0.5 threshold, it is not considered a successful recall.

PVSG The Dataset

- Problems with a variety of Video Datasets:
  - *Action Genome, Charades:* curated scripts produce random action sequence, limited potential to explore contextual logic and reasoning.
  - *ImageNet-VidVRD, VidSGG:* limited size.
  - *VIP-Seg*: lacks temporal relations.
- Humans rely on unpolished videos to form essential understanding.
- Candidates:
  - *VidOR:* unedited, natural and diverse.
  - *Ego4D-STA:* good for exploring logical relationships, supports long and short term actions.
  - *Epic-Kitchens-100 (including VISOR)*: rich action data.

PVSG The Dataset

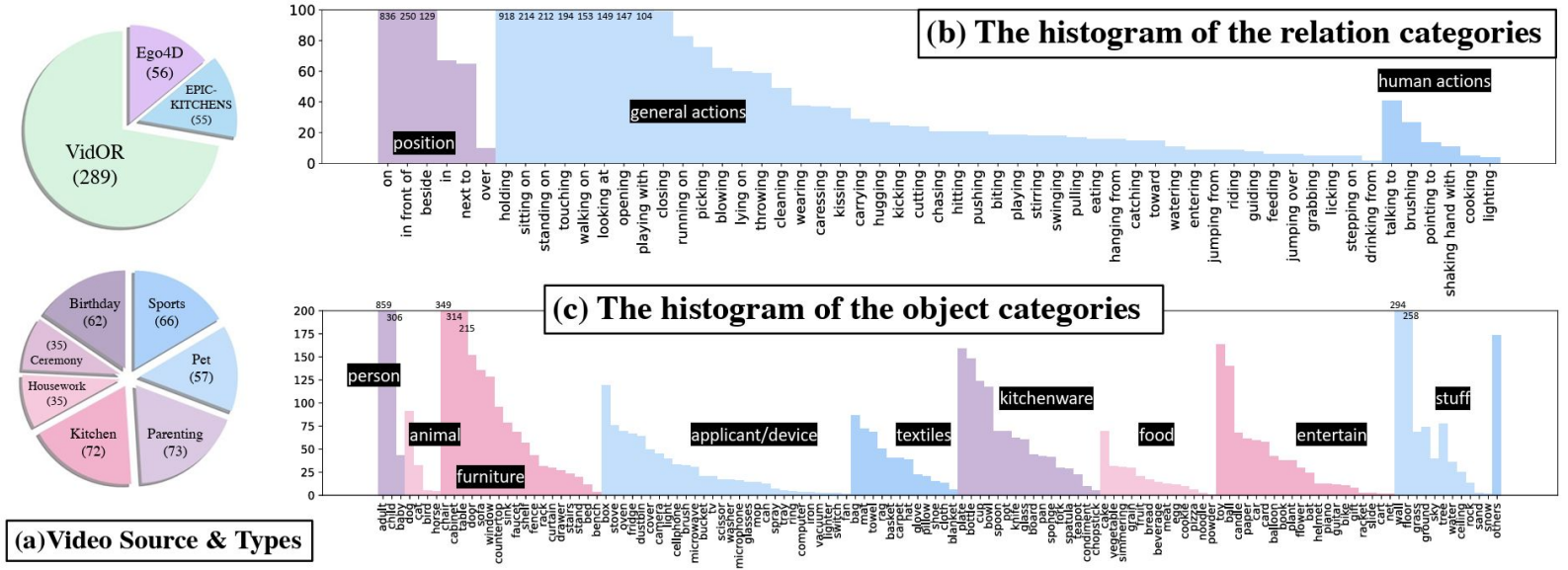


Figure 2. **The PVSG dataset statistics.** The PVSG dataset contains 400 third-person and ego-centric videos from diverse environments, as shown in (a). The statistics of object classes and relation classes are shown in (b) and (c).

PVSG

The Dataset Construction

Step 1: Video Clip Selection
Carefully select 300 long, daily, unedited videos with storyline from VidOR and 100 from Epic-Kitchens and Ego4D.

Step 2: VPS Annotation
Rely on off-the-shelf VOS (Video Object Segmentation) model called AOT for human-machine interactive annotation process.
Coarse VPS Annotation: Identify key objects to annotate and identify key frames. After annotating key objects in corresponding frames → use AOT to propagate the mask.
Fine VPS Annotation: Conduct several rounds (> 5) of human machine revision process for final annotation.

Step 3: Relation Annotation
Describe the video with several sentences → annotate relations accordingly.
Use unambiguous predicates like “sitting on” rather than “on”.

Methodology Stage 1: Video Panoptic Segmentation

Goal: Segment and track each pixel in a non-overlapping manner.

$$\{y_i\}_{i=1}^{N} = \{(\mathbf{m}_i, p_i(c))\}_{i=1}^{N}, \text{ where } \mathbf{m}_i \in \{0,1\}^{T \times H \times W}$$

Predicted Video mask

Prob of assigning class c to clip m

Methodology Stage 1: Video Panoptic Segmentation

Goal: Segment and track each pixel in a non-overlapping manner.

IPS+T: Image Panoptic Segmentation with Tracker:

- Mask2Former → Transformer encoder-decoder architecture with a set of object queries.
- Input → Mask2Former → object queries. Two MLPs are used to project queries into two embeddings → mask classification and mask prediction.
- Obtain panoptic segmentation of each frame and then using UniTrack obtain final N tracked video cubes.

VPS: Video Panoptic Segmentation Baseline:

- Replace the backbone and neck in Video K-Net with Mask2Former feature extractor.
- Use temporal contrastive loss to perform directly on the output queries from the decoder.
- During training, two nearby frames are sent to model to learn association embedding, use this learnt association during inference to perform instance-wise tracking to match masks frame by frame.

Methodology Stage 2: Relation Classification

Input: Object query (feature) tubes $\{Q_i\}^N$

Pair Selection:

Training Phase: Selected based on their match with ground truth.

Testing Phase: N x (N-1) pairs → impractically large number.

Solution: Trainable pairing component: Transformer encoder to cross-attend to all other object features within each frame. Uses max-pooling to condense query → calculate pairwise similarity → optimizes towards ground truth pairing using multi-label loss.

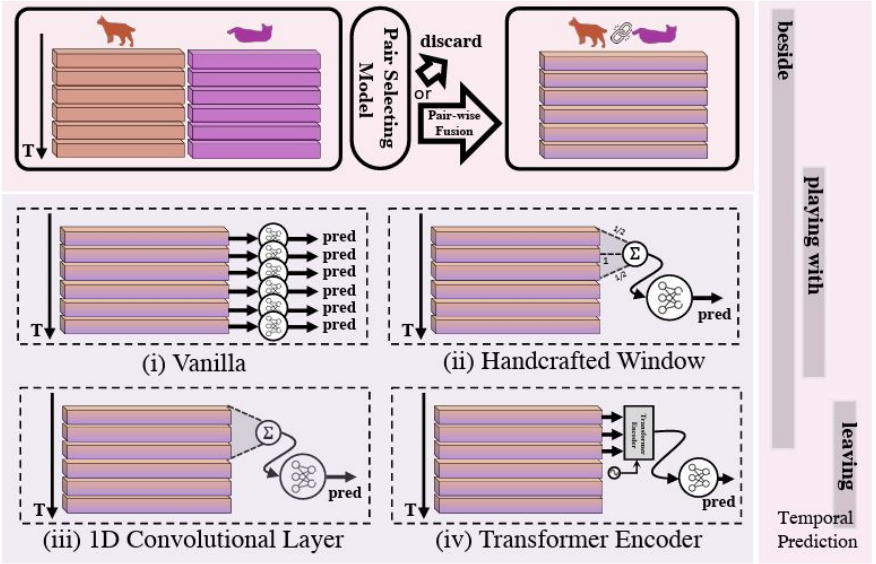
Methodology Stage 2: Relation Classification

Input: Object query (feature) tubes $\{Q_i\}^N$

Pair Selection: Gives pairs of objects

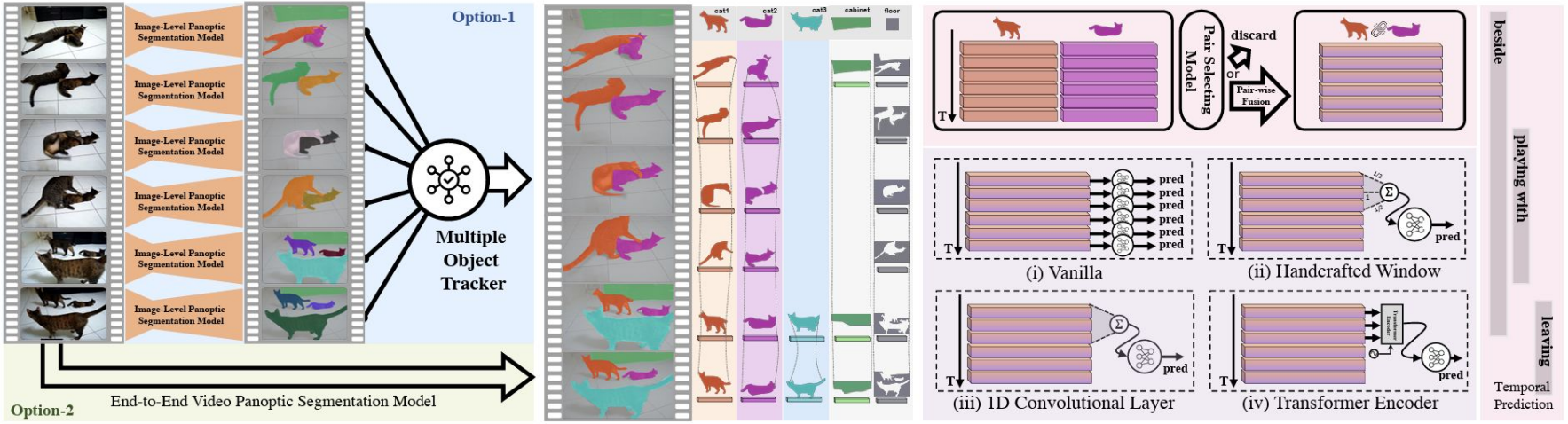
Predict relations of each feature pair:

- *Vanila*: Straightforward fully connected layers on the fused features, multi-label classification with binary cross entropy loss.
- *Handcrafted Filter*: Design simple kernel to gather information from context: [¼, ½, 1, ½, ¼] and window size - 5.
- *1D-Convolutional Layer*: Improve handcrafted filter → learnable convolutional layer, kernel size - 5
- *Transformer Encoder*: Naturally suited for temporal data, positional encoding in entire fused query feature.



(b) Stage 2: Relation Prediction

Methodology



(a) Stage-1: For Feature Tube and Mask Tube Output **(b) Stage 2: Relation Prediction**

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.

Experiments

- 90/10 train/test video split.
- Stage 1:
  - VPS underperforms IPS+T
  - Long & dynamic videos in PVSG
  - Object tracking performance hurts VPS R/mR performance
- Stage 2:
  - Transformer encoder achieves best performance.

Method		PVSG Metrics		
Stage-1	Stage-2	R/mR@20	R/mR@50	R/mR@100
IPS+T [3, 44]	Vanilla	2.35 / 1.22	2.71 / 1.31	2.94 / 1.45
	Handcrafted Window	2.56 / 1.24	2.78 / 1.35	3.05 / 1.54
	1D Convolution	2.79 / 1.24	2.80 / 1.47	3.10 / 1.59
	Transformer Encoder	**4.02 / 1.75**	**4.41 / 1.86**	**4.88 / 2.03**
VPS [3, 25]	Vanilla	0.52 / 0.24	0.60 / 0.24	0.63 / 0.24
	Handcrafted Window	0.54 / 0.27	0.61 / 0.29	0.62 / 0.29
	1D Convolution	0.60 / 0.27	0.73 / 0.28	0.76 / 0.29
	Transformer Encoder	**0.75 / 0.36**	**0.91 / 0.39**	**0.94 / 0.40**

Strengths

- It combines the strengths of video scene graphs and panoptic segmentation
- Pixel-Accurate tracking improves spatial accuracy and supports background objects like the ground and sky better than with bounding boxes
- Contribute a new dataset with annotations and benchmarks to further the field
- Predicts when relationships start and end by analyzing pairs of object for their relationship and its start and end time vs generating a scene graph every 10 or 20 frames

Weaknesses

- An object belongs to only one mask with no overlap. If you segment an object into smaller parts, how effective will it be at recognizing the whole object.
- Time cost to annotate panoptic masks with temporal relations, work to be done for real time analysis
    - Mask2Former
    - Tracking with UniTrack
    - Feature extraction
    - Relationship Predictor (for every pair of objects)
    - Sliding window processing
- If you want to model something niche with less data, hard to learn since it’s rarer (Long-tail heteroscedastic distribution)

Conclusions

Challenges:

- Long-tail heteroscedastic distributions in real-world videos
  - Models should be predictive even of statistically uncommon relations.
- Panoptic Segmentation.
- Data biases.

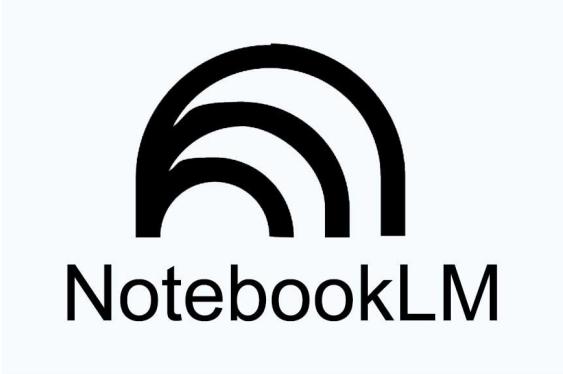
Contributions:

- PVSG Dataset.
- Formalized PVSG task.
- Baseline architectures/metrics.

NotebookLM

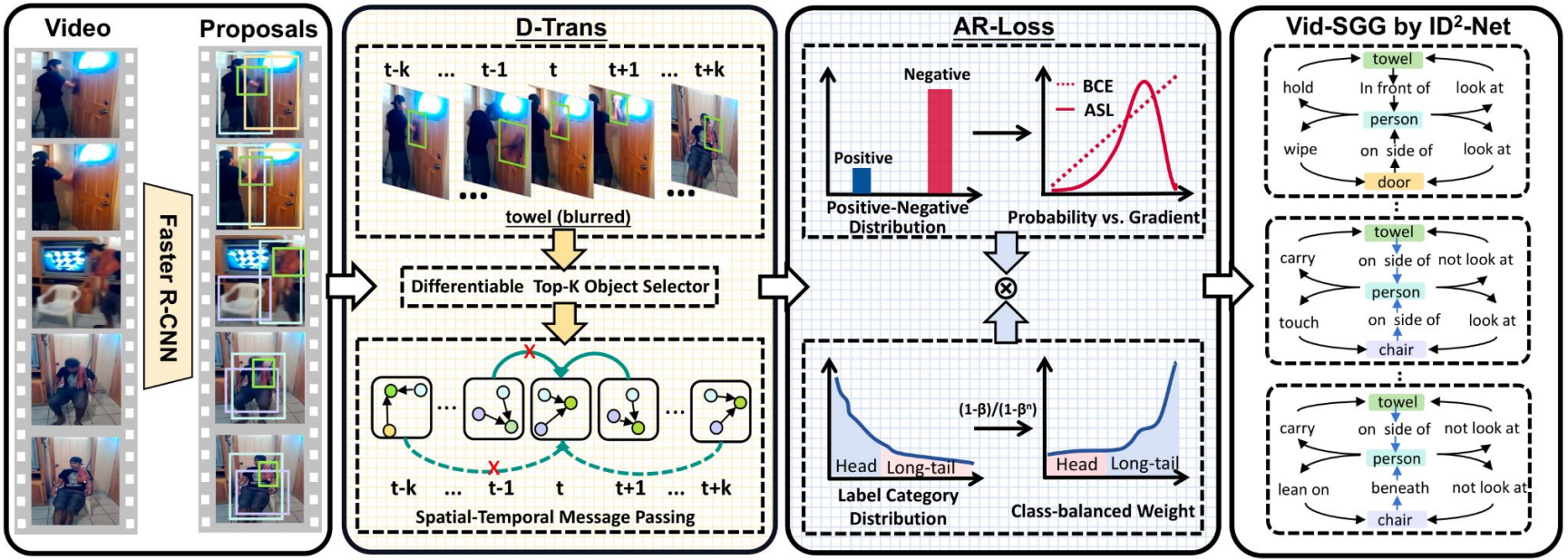
- Prompt: <empty>
  - Explains paper to laymen through a discussion.
  - Points out statistics, models, approach steps.
  - Q/A format for each aspect of the pipeline.

- Prompt: “Discuss the key contributions of <paper>… to an audience of experts in <conf>.”
  - Noticeably richer in semantics.
  - Does not spend time on clarifying each term.
  - Reasons about architectural choices & their connection to performance.

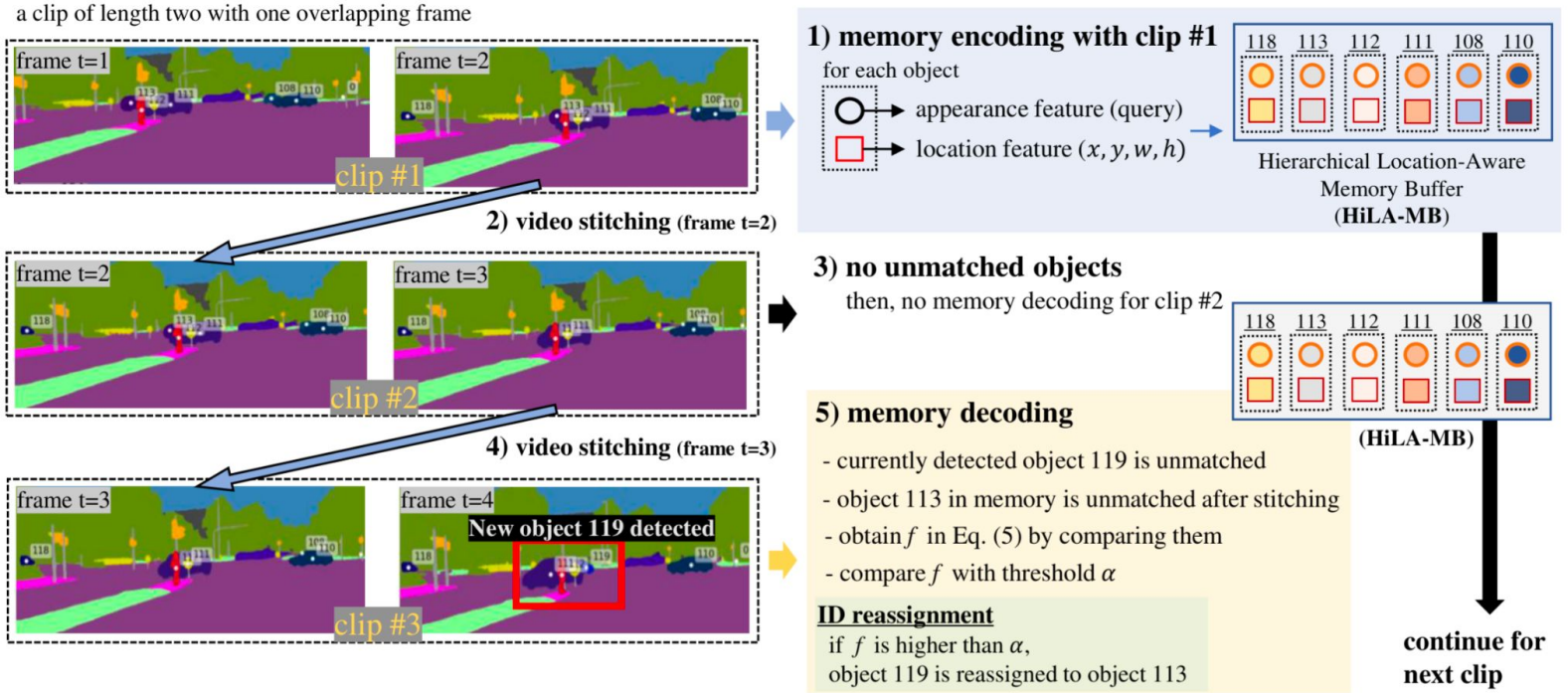


Links: go.osu.edu/{pvsg_notebooklm, pvsg_notebooklm_cvpr, pvsg_notebooklm_3}

Related Work TD^2-Net



Related Work Video-kMaX



Thank you! Questions?

Panoptic Video Scene Graph Generation

Jingkang Yang et al.

CVPR 2023

Paper: arxiv.org/abs/2311.17058

Presented by Abhinay, Mona & Ram