PanViS: Panoptic Video Scene Graph Generation for Zero-Shot Visual Comprehension

Mona Gandhi, Abhinay Putta, Sriram Sai Ganesh

The Ohio State University

{gandhi.255, putta.12, saiganesh.3}@osu.edu

https://github.com/Sriram-Sai-Ganesh/PanViS

Abstract

Scene graphs offer a structured way to represent visual scenes by modelling objects, attributes, and their relationships, greatly benefiting downstream vision-language tasks. Extending this to videos, Panoptic Video Scene Graph Generation (PVSG) aims to capture fine-grained spatial and temporal interactions through pixel-level segmentation and coherent dynamic graphs. However, current methods rely heavily on costly annotations and struggle to generalise to novel concepts. We introduce PanViS, a modular and interpretable framework for zero-shot PVSG that leverages foundation models to recognise unseen objects and relationships without requiring scene graph supervision. PanViS builds temporally consistent scene graphs through four stages: object identification, panoptic segmentation, relationship identification, and relationship verification. By unifying pretrained models via lightweight prompting and refinement, PanViS achieves accurate, scalable, and generalizable scene graph generation. We further propose tailored evaluation metrics to assess scene graph completeness and their utility for downstream tasks like video question answering. Experiments on the STAR dataset would show its effectiveness in addressing challenges of entity disambiguation, segmentation consistency, and relational reasoning across time.

1. Introduction

Scene Graphs (SGs), first introduced in the Visual Genome [4] dataset, offer a structured and semantically rich representation of visual scenes by modeling objects, their attributes, and the relationships between them. This structured abstraction closely aligns with human visual understanding and has been demonstrated to enhance performance in a variety of downstream tasks, including visual question answering and image retrieval. Motivated by these successes, significant research efforts have been devoted to developing automatic Scene Graph Generation (SGG) methods.

Extending this paradigm to the video domain, Video Scene Graph Generation (VSGG) [10] seeks to capture not only spatial relationships within individual frames but also temporal relationships across frames, thereby enabling a coherent modeling of interactions and changes over time. More recently, Panoptic Video Scene Graph Generation (PVSG) [14] has been proposed, combining the benefits of panoptic segmentation with scene graph representations. By leveraging pixel-level segmentation masks, PVSG enables fine-grained, temporally consistent identification of both foreground and background entities and their evolving relationships.

However, despite these advancements, current scene graph generation approaches suffer from several critical limitations. Most notably, they rely heavily on large-scale, manually annotated datasets such as Visual Genome, constraining their applicability to in-domain distributions and limiting their ability to generalize to novel scenes. Furthermore, these models typically operate over a closed vocabulary, restricting their capacity to detect unseen objects, relations, and attributes. Consequently, existing methods offer limited scalability and adaptability to real-world, diverse video data.

To address these challenges, recent work like [15] has explored zero-shot scene graph generation by leveraging foundation models capable of recognizing novel concepts without requiring task-specific annotations. Building on this direction, we propose a novel framework for Panoptic Video Scene Graph Generation.

We introduce PanViS, a modular and interpretable framework designed to enable scalable and generalizable video scene understanding. In particular, PanViS:

- Supports zero-shot scene graph generation, eliminating the reliance on costly human-annotated datasets;

- Provides interpretability by producing explicit intermediate outputs throughout the pipeline, thereby facilitating transparent and traceable scene graph construction;
- Demonstrates high generalizability across diverse video domains and datasets.

By addressing the limitations of prior work, PanViS establishes a robust and scalable foundation for downstream applications such as video question answering, video captioning, and broader video understanding tasks.

2. Related Work

Scene Graph Generation has emerged as a method to efficiently model visual scenes by representing objects, their attributes, and their relationships [4]. Early SGG methods were developed primarily for static images, improving tasks such as visual question answering by providing semantic representations [4]. However, these models often struggle with generalization and typically require supervision, requiring large scale annotations [2].

Video Scene Graph Generation (VSGG) extends this idea by modeling both spatial and temporal relationships in video data [10]. Approaches in this area aim to track objects over time while also capturing evolving interactions between entities across frames [2]. While effective for capturing dynamic relationships, many VSGG methods depend heavily on annotated video data [2].

Panoptic Video Scene Graph Generation (PVSG) [14] introduces a benchmark and model specifically for the panoptic understanding of videos. PVSG unifies panoptic segmentation, scene graph generation, and object tracking to make coherent scene graphs across frames. However, the method relies onn large-scale manual annotations for supervision and complex architectures, making interpretability and changing domains challenging.

Long-range contextual reasoning has been explored through the STAR benchmark [12], which focuses on evaluating visual question answering based on semantic relations found in videos. STAR emphasizes reasoning over symbolic information while PVSG focuses on generating rich PVSGs. Our work seeks to bridge these two ideas, generating rich and human interpretable scene graphs for downstream reasoning task.

Recent research has explored low supervision and zero-shot approaches to scene graph generation. Zhao et al. [15] demonstrated that zero-shot scene graph generation can be enabled using foundational models like SAM and BLIP without requiring extensive manual labels. TD2-NET [8] addresses denoising and debiasing in dynamic scene graphs, improving robustness to noisy labels in videos. Also, Chen et al. [2] proposed generatng video scene graphs using only single-frame supervision, highlighting the potential to extract rich temporal relationships from sparse data. Inspired by these trends, our framework adopts a zero-shot, modular approach to Panoptic Video Scene Graph Generation, emphasizing interpretability and generalization.

3. Method

3.1. Overview

In this section, we discuss the four-stage pipeline to address the PVSG problem.
For each of the four sections, we consider an option for a pretrained foundational model that may be used as a baseline.
First, we discuss each of the four stages of PanViS:

1. Object Identification
2. Video Panoptic Segmentation
3. Relationship Identification
4. Relationship Verification

We finally describe the end-to-end pipeline using these four modules that enables Panoptic Video Scene Graph Generation on the STAR dataset.

As is standard for the Panoptic Video Scene Graph Generation task, we consider the input to PanViS to be a video clip $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$, a 3-channel $H \cdot W$ pixel dimension video clip with T timesteps [14].

3.2. Object Identification

The goal of the *Object Identification* stage is to output a set of temporally–consistent instances

$$\mathcal{O} = \left\{(b_i,\, \ell_i,\, \tau_i)\right\}_{i=1}^{N},$$

where $b_i \in \mathbb{R}^4$ is the axis–aligned bounding box in the first frame in which the instance appears, $\ell_i \in \mathcal{C}$ its semantic label from the category set $\mathcal{C}$, and $\tau_i = (t_i^{\text{start}}, t_i^{\text{end}})$ the temporal span during which the instance is visible. Unlike single–frame detection, PVSG requires that each physical entity be assigned a *single* global identifier that is stable across the entire clip for temporal consistency; this is later exploited by the panoptic video segmentation and relation stages.

We consider the **Florence 2** vision model [13], a versatile pretrained image foundation model that excels at a variety of downstream tasks. While Florence 2 shows nontrivial improvements over existing models in a variety of image-related tasks, it is built with a DaViT [3] Vision Transfer image encoder backbone which cannot operate on video input.
We circumvent this limitation of its use in video in order to take advantage of its superior object detection capabilities. While Florence 2 does not produce temporally consistent

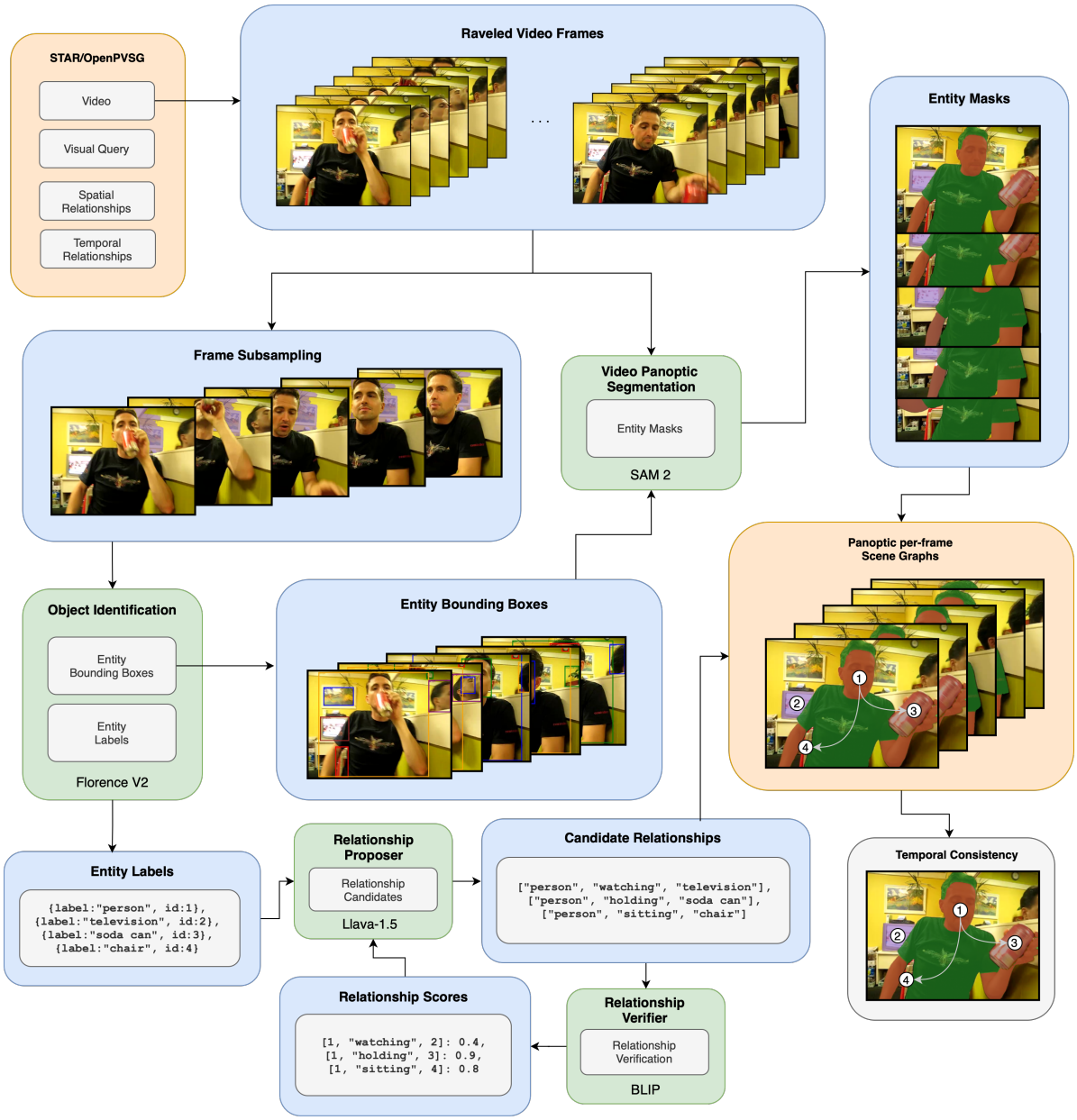


Figure 1. The PanViS framework. Nodes in green are instantiations of four candidate models for each of the stages in PanViS.

bounding boxes for video inputs, we mitigate the effect of this drawback by obtaining object detection labels for every single image frame in T timesteps. We rely on an external temporal consistency module to resolve entities between frames.

3.3. Video Panoptic Segmentation

The *Video Panoptic Segmentation* (VPS) stage assigns every pixel in every frame a unified *thing* instance identifier or *stuff* semantic label while maintaining identity consistency throughout the clip. Formally, for each timestep $t \in \{1, \ldots, T\}$ we predict a panoptic map

$$\mathcal{S}_t \;=\; \big\{(p,\ \mathrm{id}(p)) \mid p \in \Omega\big\},$$

where $\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\}$ is the image lattice and $\mathrm{id}(p) \in \mathbb{N}$ provides a clip–level unique identity for *thing* pixels (set to 0 for *stuff*).

To build a strong yet modular baseline we employ **Segment Anything Model 2 (SAM 2)**. SAM 2 extends the original SAM to video via a transformer with a streaming memory cache that supports real-time processing and zero-shot generalisation. Crucially for PanViS, SAM 2 is *prompt-*

able: segmentation masks are produced in response to sparse spatial prompts such as bounding boxes, key-points, or free-form masks.

We interface SAM 2 with the bounding boxes predicted by the Object Identification stage (3.2). For each object instance $(b_i, \ell_i, \tau_i) \in \mathcal{O}$ we create a single 4-point box prompt in the first frame where the instance appears ($t = t_i^{\text{start}}$). SAM 2 then generates an initial mask proposal $\mathbf{m}_{i,t}$ that is refined by the model's internal streaming memory as frames advance; no additional prompts are required. This temporally propagated mask assists the temporal consistency module in resolving inter-frame object dependencies.

Although SAM 2 internally propagates masks, object identities must be reconciled with those from 3.2. We simply reuse the global identifiers id_i assigned during Object Identification, attaching them to SAM 2's masks at each timestep.

The VPS module outputs the sequence $\{\mathcal{S}_t\}_{t=1}^{T}$ together with a dictionary mapping instance IDs to pixel-level masks in every frame.

3.4. Relationship Identification

Given the temporally consistent object set

$$\mathcal{O} = \{(b_i, \ell_i, \tau_i)\}_{i=1}^{N},$$

the *Relationship Identification* (RI) stage proposes *entity–level* candidate relations

$$\mathcal{R}_{\text{id}} = \Big\{ (s,\ p,\ o,\ t) \ \Big|\ s, o \in \mathcal{O},\ p \in \mathcal{P},\ t \in \tau_s \cap \tau_o \Big\}, \quad (1)$$

where s and o denote subject- and object-entity indices, p is a predicate from the fixed set $\mathcal{P}$ supplied by STAR and t is the frame index in which the relation holds. The RI module is deliberately permitted to *over-generate* candidate relationships, a high recall is prioritised, with precision delegated to the downstream Relationship Verification stage (section 3.5).

To obtain semantically rich relational cues, we employ the recent Image-Text-to-Text multimodal model LLaVA [6] with `Qwen-1.5-{0.5,3,7}B` backbones [1]. LLaVA couples a pretrained LLM with an image-encoder via visual query tokens, enabling open-vocabulary grounding and reasoning.

For every ordered pair (s, o) of distinct entities whose visible spans overlap temporally, we sample from LLaVA a set of *candidate* predicates:

$$\mathcal{P}_{s,o} = \Big\{ p \ \Big|\ p \in \mathcal{P};\ \tau_s \cap \tau_o \neq \varnothing \Big\},$$

uniformly thinning to at most K timesteps.

We prefix an instruction header

```
<image> Given objects A and B,
answer {yes/no} as to whether the
following relations hold:
```

to guide the binary decision.

For each (s, p, o, t) LLaVA returns logits $\big(\ell_{\text{yes}}, \ell_{\text{no}}\big)$. We convert them to probabilities by a softmax and retain the "yes" score $\pi^t_{s,o,p}$. The final confidence in (1) is the maximum

$$\sigma = \max_{t \in \mathcal{T}_{s,o}} \pi^t_{s,o,p}.$$

A relation is admitted into $\mathcal{R}_{\text{id}}$ if $\sigma > \tau_{\text{rel}}$; we set $\tau_{\text{rel}} = 0.3$ to favour recall. At most the top M predicates per pair are kept, mitigating explosion in dense scenes with many possible pairwise relationships.

The output $\mathcal{R}_{\text{id}}$ is forwarded to the Relationship Verification stage 3.5, which filters implausible triples and enforces temporal and semantic consistency across the clip.

3.5. Relationship Verification

The goal of the *Relationship Verification* (RV) stage is to refine the high-recall yet noisy set of candidate relations $\mathcal{R}_{\text{id}}$ produced into a compact, high-precision set

$$\mathcal{R}_{\text{ver}} = \big\{ (s, p, o, t, \hat{\pi}^t_{s,o,p}) \big\},$$

where $\hat{\pi}^t_{s,o,p} \in [0, 1]$ is a confidence score that the predicate p truly holds between subject s and object o at frame t.

We adopt **BLIP** [5], a vision–language pre-training framework that establishes a new state-of-the-art in the CIDEr image captioning benchmark [11], as our verification engine.

For each candidate tuple $(s, p, o, t) \in \mathcal{R}_{\text{id}}$ we synthesise a templated caption

$$\text{caption}(s, p, o) = \texttt{<}s\texttt{>}\ \texttt{is}\ \texttt{<}p\texttt{>}\ \texttt{<}o\texttt{>}$$

Here `s` and `o` are replaced by the category labels ℓ_s, ℓ_o of the subject & object entities.

To focus BLIP on visual evidence, we provide an **entity-pair crop** $x^t_{s,o} = \text{crop}\big(I_t,\ b^t_s \cup b^t_o\big)$, where I_t is the RGB frame at time t and b^t_s, b^t_o are the temporally propagated boxes from SAM 2 (3.3).

BLIP predicts an Image-Text Match (ITM) logit ℓ_{ITM} given $\big(x^t_{s,o}, \text{caption}(s, p, o)\big)$.

We convert all logits for a pair of entities (s, o) to a probability

$$\hat{\pi}^t_{s,o,p} = \sigma\big(\ell_{\text{ITM}}\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}$$

A relation is accepted if $\hat{\pi}^t_{s,o,p} > \tau_{\text{ver}}$, with $\tau_{\text{ver}} = 0.5$.

RI (see 3.4) and RV operate under complementary objectives, we perform a round of *iterative refinement* (see fig. 2):

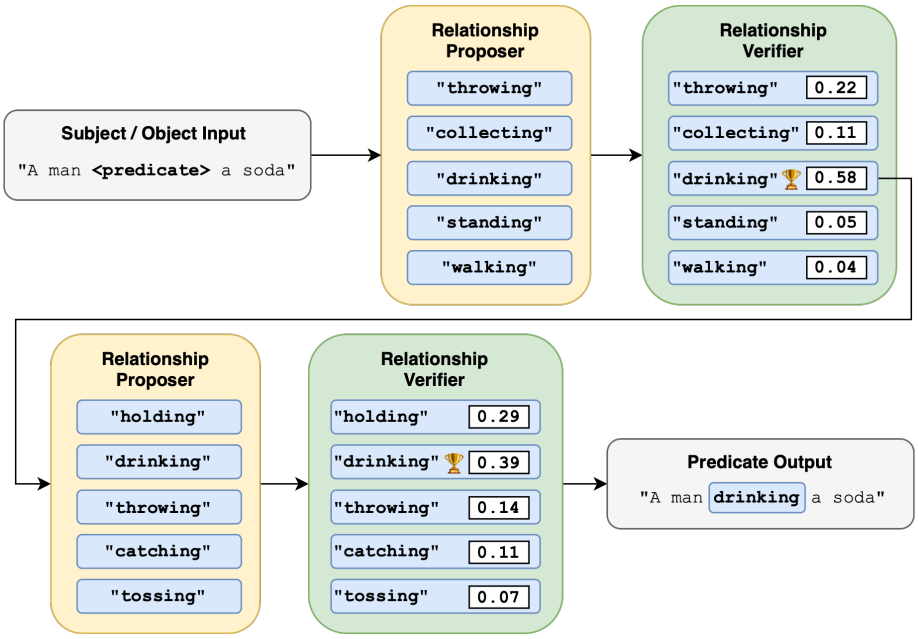


Figure 2. One round of *Iterative refinement* viewed sequentially. We pass high-confidence relationships back to the Relationship Proposer to produce additional semantically grounded predicates, which we re-evaluate with the Relationship Verifier.

1. **Seed.** The top-K verified relations for each entity pair (ranked by $\hat{\pi}^t_{s,o,p}$) are fed back as *prompt seeds* to RI, biasing LLaVA toward plausible predicates in the second pass.
2. **Re-identify.** RI re-samples candidate relations under these seeds, producing an updated set $\mathcal{R}^{(2)}_{\text{id}}$ with markedly higher quality.
3. **Re-verify.** RV re-scores $\mathcal{R}^{(2)}_{\text{id}}$ using the same BLIP procedure, yielding the final set $\mathcal{R}_{\text{ver}}$.

RV returns both the pruned relation set $\mathcal{R}_{\text{ver}}$ and the calibrated confidences $\hat{\pi}$, which are processed by the Scene Graph generation module (3.6) to construct the final Panoptic Video Scene Graph.

3.6. PanViS Framework

Figure 1 presents a bird's-eye view of our proposed ***Pan**optic **Vi**deo **S**cene Graph* (PanViS) framework, which assembles the four modules in Sections 3.2–3.5 into a single end-to-end pipeline.

Given a video clip $\mathcal{V}$, the Object Identification (OI) module first extracts the temporally-consistent entity set of objects $\mathcal{O}$, together with axis-aligned bounding boxes $\{b^t_i\}$ corresponding to boxes across time for each object.

These boxes act as spatial *prompts* to the Video Panoptic Segmentation (VPS) module, which returns dense pixel-accurate masks $\{\mathcal{S}_t\}^T_{t=1}$ containing both *thing* and *stuff* categories.

Subsequently, Relationship Identification (RI) samples high-recall candidate triples $\mathcal{R}_{\text{id}}$ whose entity indices align with $\mathcal{O}$, and Relationship Verification (RV) filters them to the compact, high-precision set $\mathcal{R}_{\text{ver}}$.

The final *Panoptic Video Scene Graph* (PVSG) for a video $\mathcal{V}$ is a dynamic graph $\mathcal{G}_{\mathcal{V}} = \big(V,\ E\big)$, where

1. The vertex set $V = \mathcal{O} \cup \big(\bigcup^T_{t=1} \Omega_{\text{stuff},t}\big)$ contains all *thing* instances and per-frame *stuff* regions. Vertices inherit semantic labels ℓ and (for *thing* nodes) a global identity. Vertices also store a temporal window of validity $(\tau_s,\ \tau_e)$.
2. The edge set $E = \mathcal{R}_{\text{ver}}$ encodes verified subject–predicate–object relations. Edges also inherit a temporal window $(\tau_s,\ \tau_e) \subset \big(\,(\tau^s_s,\ \tau^s_e) \cap (\tau^o_s,\ \tau^o_e)\,\big)$ which is a subset of the validity windows of their incident subject (s) and object (o) nodes.

All objects in $\mathcal{G}_{\mathcal{V}}$ store a set of confidence scores $\{c_i\}^t_{i=1}$ across window.

PanViS executes OI and VPS only once, but iterates between RI and RV for one additional pass (see section 3.5). In practice, a single refinement suffices: the second-round RI is seeded by top-K relations from RV, and the subsequent RV produces high-confidence relationships for graph generation.

4. Experimental Setup

4.1. Datasets

The STAR Benchmark paper [12] introduces the **STAR dataset** for Situated Reasoning, which presents challenging question-answering tasks grounded in symbolic situation descriptions and logic-based diagnosis of real-world video

scenarios. It includes four types of questions: Interaction, Sequence, Prediction, and Feasibility. The dataset features 111 action classes, 37 entity classes, and 24 relation classes. It also provides frame-level annotations highlighting relevant parts of the video. As a baseline, we evaluate the quality of our generated scene graphs using STAR's validation set and corresponding situated descriptions, which we convert to video scene graphs compatible with our framework. Furthermore, we aim to leverage our detailed scene graphs to address the situated questions as a downstream task.

For the Panoptic Video Scene Graph Generation task, we use the **PVSG dataset** [14], which contains 400 videos of varying lengths (averaging 76.5 seconds), captured from diverse perspectives and exhibiting substantial camera and object motion. The dataset provides rich annotations, including Video Panoptic Segmentation and Temporal Scene Graphs over 150K frames, as well as video-level dense captions and QA pairs. We conduct a similar evaluation on PVSG as with STAR, analysing both the quality of scene graph generation and performance on the downstream QA task.

4.2. Evaluation

4.2.1. Video Panoptic Segmentation

In the context of video object segmentation, the overall performance is often evaluated using the combined $\mathcal{J}\&\mathcal{F}$ metric, which averages two complementary measures: the region similarity $\mathcal{J}$ and the contour accuracy $\mathcal{F}$ [7].

The region similarity $\mathcal{J}$ measures the intersection-over-union (IoU) between the predicted mask M and the ground truth mask G:

$$\mathcal{J}(M, G) = \frac{|M \cap G|}{|M \cup G|}$$

where $|\cdot|$ denotes the cardinality of the set.

The contour accuracy $\mathcal{F}$ evaluates the quality of object boundaries by computing the F-measure between the contours of M and G. It is defined as:

$$\mathcal{F}(M, G) = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision and Recall are computed by comparing the contour points of the predicted and ground truth masks.

Finally, the overall *J&F* score is computed as the average of $\mathcal{J}$ and $\mathcal{F}$ over all frames and all objects:

$$\mathcal{J}\&\mathcal{F} = \frac{1}{2}\left(\overline{\mathcal{J}} + \overline{\mathcal{F}}\right)$$

where $\mathcal{J}$ and $\mathcal{F}$ represent the average region similarity and contour accuracy, respectively [7].

4.2.2. Complete Video Scene Graph

To evaluate the quality of the generated scene graphs, we introduce a set of metrics comparing situated scene graphs from STAR (SG_{STAR}) with those produced by our framework (SG_{PanViS}).

Specifically, we define **Missing-X** metrics to assess recall, verifying whether entities, relations, and actions present in SG_{STAR} are also captured in SG_{PanViS}. For each video frame, we compute the number of **Missing Entities** and **Missing Relations**.

Additionally, to evaluate temporal consistency across the entire video, we introduce the **Missing Actions** metric, which measures the absence of temporal relations. It is important to note that SG_{PanViS} may contain additional objects and relationships beyond those annotated in SG_{STAR}, as our system operates in a more open-world setting.

Therefore, our evaluation emphasizes recall rather than strict precision, focusing on ensuring coverage of the annotated entities and relations.

Beyond scene graph quality, we further assess the utility of SG_{PanViS} for downstream tasks through **Question Answering (QA) Evaluation**. Given a question from the STAR dataset, we will first identify relevant frames based on the objects mentioned in the question. Then, extract a subset of the scene graph corresponding to these objects and their interactions. We will prompt a large language model (LLM) using this extracted subset to generate an answer. This evaluation provides an additional measure of the scene graphs' completeness and utility for video understanding tasks.

5. Results

We report findings for each of the four constituent modules of the PanViS framework.

5.1. Object Identification

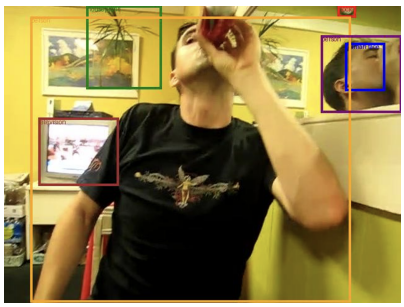
- Florence 2 correctly identifies all items in frame for high enough resolution image frames. Running on the STAR dataset with some videos as low as 360×480px, there are instances when objects are detected in some frames sporadically, as in the the painting in the backgrounds of fig. 3a - fig. 3c.
- Missed detections arise primarily from heavy occlusion and extreme motion blur; false positives stem from short-lived background distractors (e.g. transient reflections, see fig. 3b).

Towards mitigating these errors, we union entity sets from $k = 3$ consecutive frames to form our overall entity set $\mathcal{O}$.

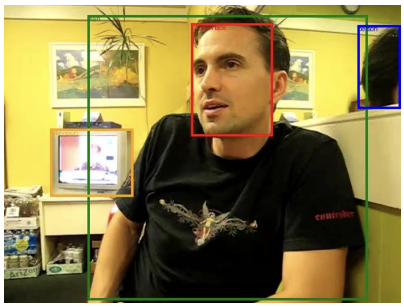
5.2. Video Panoptic Segmentation

- SAM 2 is evaluated with the $\mathcal{J}\&\mathcal{F}$ metric, and is able to consistently generate generate high-quality video segmentation masks for the semi-supervised Video Object Segmentation task, where a prompt is supplied only for the first frame of the input. SAM 2 achieves a single-bounding-box $\mathcal{J}\&\mathcal{F}$ score of **72.9** [9].

(a) Florence 2 fails to capture the man's face.



(b) Florence 2 is confused by reflections.



(c) Florence2 fails to capture background items.

Figure 3. Florence 2 captures slightly different entity sets O in individual frames.

(a) SAM 2 fails to separate nearby objects when faced with similar bounding box prompts.

(b) SAM 2 displays robust object tracking between frames.

(c) Fine pixel borders (eg. the man's hair) are well-defined.

Figure 4. SAM 2 masks created from an initial frame & then propagated to subsequent ones.

- We observe that while SAM 2 maintains robust tracking across the temporal dimension, smaller model sizes struggle to distinguish adjacent objects when supplied with related bounding box prompts, as is evident in fig. 4a.

5.3. Relationship Identification

- We deliberately set a low threshold for candidate relationships $\tau = 0.3$ to capture multiple candidate relations between entities. We rely on the iterative refinement step – seed, re-identify and re-verify, in order to whittle down the set of relations to the highest confidence predicates.
- Qualitatively, LLaVA-1.5 resolves subtle spatial and entity nuances (e.g. *"drinking soda"* is preferred over *"drinking a soda can"* for fig. 4a), which was the goal with the verifier step. Future work could also see entity relationship predictions refining entity labels in an iterative manner (ie. recognizing the upstream change from *"soda can"* to *"soda"* based on VLM confidence logits)

5.4. Relationship Verification

We see significant utility in the iterative refinement step. Perhaps a side-effect of our choice of the dated Llava family [**?**] as a vision-language model, we observe that even many of the top-5 first-pass relationships that are proposed tend to be quite absurd for a given context (see fig 2). Iterating on these candidate relationships with the RV BLIP module produces more fine-grained and accurate relationships.

6. Challenges & Future Work

While PanViS demonstrates promising capabilities in zero-shot PVSG and all the individual components are working, integrating all components of the framework presented notable challenges. We aim to complete a thorough evaluation using the proposed Missing-X metrics to quantify scene graph completenes and extend this analysis on the PVSG dataset.

Another important challenge unaddressed by existing PVSG methods is the handling of multi-label panoptic segmentation, where a single pixel may correspond to multiple object categories. For instance, in real-world videos, a pixel may simultaneously belong to both a person and their clothing, as illustrated in Figure 4c. Current frameworks, including PanViS and PVSG, assume that each pixel is uniquely assigned to a single object or stuff label, which can lead to a loss of semantic richness. Future work should explore extending segmentation methods to support multi-label pixel assignments, enabling richer and more accurate scene graph constructions in complex, overlapping environments.

Contributions

Here are the contributions of our team:

- **Mona**: Prepared the dataset for the pipeline and evaluation, and contributed to the design and development of the PanViS framework. Wrote abstract and sections 1 & 6.
- **Ram**: Contributed to the ideation & design of the PanViS. Wrote code for the PanViS project, implemented functions & classes for the various steps in the PanViS pipeline. Created figs. 1-4. Wrote sections 3.1-3.6 & 5.1-5.4.
- **Abhinay**: Researched evaluation strategies for scene graph similarity and background information on the state of PVSG. Wrote sections 1, 2 & 6.

Acknowledgements

We would like to thank Prof. Srinivasan Parthasarathy, our TA Kuan Chieh Lo, and our peers from `SP'25 CSE 5245: Network Science` for their valuable feedback and constructive suggestions. Their insights and support were instrumental in helping us concretize and refine our ideas. We are also grateful to the Ohio Supercomputer Center (Project `PAS2030`) and the Department of Computer Science & Engineering at The Ohio State University.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. 4

[2] Siqi Chen, Jun Xiao, and Long Chen. Video scene graph generation from single-frame weak supervision. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[3] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers, 2022. 2

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5295–5303, 2017. 1, 2

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 4

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 4

[7] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation, 2018. 6

[8] Tao Pu, Zhihao Li, Jiahui Geng, Lingxi Xie, Qi Tian, Ya Zhang, and Yanfeng Wang. Td^2-net: Toward denoising and debiasing for dynamic scene graph generation, 2024. 2

[9] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 6

[10] Jing Shang, Tianshui Li, Tong Xu, Liang Zhou, Wai Lam Wong, Xiaowen Chu, and Liang Lin. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7103–7112, 2019. 1, 2

[11] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 4

[12] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos, 2024. 2, 5

[13] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023. 2

[14] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation, 2023. 1, 2, 6

[15] Shu Zhao and Huijuan Xu. Less is more: Toward zero-shot local scene graph generation via foundation models, 2023. 1, 2