# One Identity, Many Roles: Multimodal Entity Coreference for Enhanced Video Situation Recognition

Balaji Darur[1] Amanmeet Garg[2] Makarand Tapaswi[1]
[1]CVIT, IIIT Hyderabad, India [2]Amazon Prime Video, Seattle
https://katha-ai.github.io/projects/cinemec/

## Abstract

*Video Situation Recognition (VidSitu) addresses the challenging problem of "who did what to whom, with what, how, and where" in a video. It tests thorough video understanding by requiring identification of salient actions and associated short descriptions for event roles across multiple events. Grounding with VidSitu requires spatio-temporal localization of key entities across shots and varied appearances.*

*We posit that coherent video understanding requires* consistent identification *of entities that play different roles. We propose* Multimodal Entity Coreference *(MEC) to unite entity descriptions in text with grounding across the video. Towards this, we introduce CineMEC, a multi-stage approach that unites event role mention groups with visual clusters of entities, without explicit grounding supervision during training. Our approach is designed to exploit the synergy between visual grounding and captioning, where improving one influences the other and vice versa. For evaluation, we extend the VidSitu dataset with grounding annotations. While previous work focuses primarily on descriptions, CineMEC improves consistency across both: captioning (+2.5% CIDEr, +7% LEA)* and *visual grounding (+18% HOTA).*

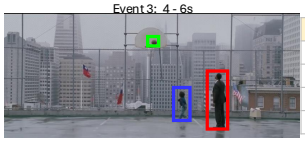
## 1. Introduction

In *The Pursuit of Happyness*, a scene features Chris Gardner watching his son play basketball. Holistic understanding of this video (Fig. 1) requires going beyond a sequence of actions such as *throw* or *walk* and identifying *who is throwing* (the boy), *what* is being thrown (the basketball), *who is watching* (the man), and *where* is the action happening (on the rooftop). Humans are naturally good at this and we also build a narrative around entities (people and objects). Moreover, by watching entities evolve over time, we are able to reason about their interactions and answer questions such as who is doing what and where. However, models struggle with such compositional associations [52].

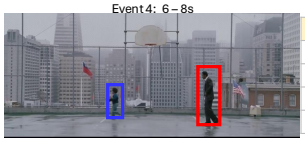
A notable step towards the task of video semantic role labeling (SRL) is Video Situation Recognition (VidSitu) [51].


| | MEC (Ours) | Baseline |
|---|---|---|
| Verb | run | play |
| Arg0 | man in suit | man in blue suit |
| Arg1 | boy in blue shirt | man in blue jacket |
| Arg2 | - | man in blue suit |
| Scn | on roof top | in a park |


| | | |
|---|---|---|
| Verb | throw | shoot |
| Arg0 | boy in blue shirt | man in blue jacket |
| Arg1 | a basketball | a basketball |
| Arg2 | from ground | man in blue shirt |
| Scn | on roof top | in a park |




| | | |
|---|---|---|
| Verb | throw | fall |
| Arg0 | boy in blue shirt | man in blue suit |
| Arg1 | a basketball | - |
| Arg2 | from ground | - |
| Scn | on roof top | in a park |




| | | |
|---|---|---|
| Verb | walk | play |
| Arg0 | man in suit | man in blue suit |
| Arg1 | - | man in blue shirt |
| Arg2 | - | man in blue shirt |
| Scn | on roof top | in front of building |


| | | |
|---|---|---|
| Verb | look | play |
| Arg0 | man in suit | man in grey suit |
| Arg1 | boy in blue shirt | man in blue suit |
| Scn | on roof top | on a sidewalk |

Figure 1. Given a video, we present our model's outputs highlighting Multimodal Entity Coreference (MEC) in comparison to a baseline (GVSR). Each event is tagged with the salient verb and its corresponding semantic roles (*e.g.* Arg0, Arg1). MEC enforces entity-level consistency across events: the boy in blue shirt, man in suit, and basketball are tracked, linked, and described with a *unique caption* across the entire video. In contrast, the baseline produces erroneous and *inconsistent* captions while referring to someone.

Here, a video is described using a structured label space by splitting it into multiple short events, each with a salient action/verb, and tagging corresponding semantic roles (*e.g.* agent, patient) with descriptive arguments. An extension, Grounded Video Situation Recognition (GVSR) [25], argues that there are multiple correct ways to describe the same entity and thus aims to ground them in the video. Specifically, they identify a single most relevant bounding box for each SRL. However, in both, SRLs and their grounding are treated independently across events. They overlook

the entity-centric perspective and do not connect *the entity's identity with multiple roles the entity plays*, neither in SRLs, nor in the visual appearance across the video. This results in inconsistent understanding across video events (see Fig. 1).

To highlight this challenge, we formulate a new task **M**ultimodal **E**ntity **C**oreference (MEC) that aims to unite entity descriptions with corresponding visual tracks across the video. We formalize MEC for VidSitu as 4 sub-tasks: (i) predict salient verbs for each event; (ii) identify entity role groups across events; (iii) cluster visual entity mentions (boxes) throughout the video and link them with role groups; and (iv) generate entity-centric SRLs. Thus, MEC unites multimodal entity mentions (event roles, SRLs, and visual boxes) across the video. For example, Fig. 1 shows that CineMEC recognizes that the *boy in blue shirt* takes on the role of *thrower* (agent) of the *baskbetall* (thing), and later receives the attention (is *looked at*) from the *man in suit*. Simultaneously, all entities are also grounded in the video with boxes that span across all frames.

**Challenges of video MEC.** Recently, MEC has been studied in images [18, 19], where rich image narratives are formed by linking entity mentions to bounding boxes. However, the temporal dimension in videos significantly increases complexity as entities interact dynamically over time and take on different roles. Additional visual and linguistic challenges make the problem even harder in edited videos (movies): (i) Visually, identifying the same entity across shots is hard as their appearance is different from varying viewpoints, may be (partially) occluded, or even absent from the view. Addressing these requires incorporating high-level semantic reasoning about entity roles, context, and their evolution across events. (ii) Linguistically, the same entity may be described in multiple ways due to different shot types (*e.g.* "man in suit" in a long shot or "man with dark hair" in a close-up). However, for coherent understanding, it is crucial to use the same caption for an entity across events. With multiple entities in a video (typical in movies), generating consistent *and* discriminative captions to uniquely identify entities is even more crucial. Note, these challenges are not considered by modern grounding-aware approaches [37, 40].

**Our approach.** We propose CineMEC, a four-stage approach that follows an entity-centric perspective to address MEC in VidSitu. (i) We cluster visual entity representations across video frames on-the-fly and use them to inform entity role representations. (ii) We group event roles across the video to promote consistent entity descriptions even when an entity takes on different roles across events. (iii) A cluster assignment module facilitates linking visual clusters with these entity role groups. (iv) Finally, the relevant visual cluster is fed to the captioner to encourage generation of a consistent SRL for the entity group throughout the video.

Our approach is *not supervised* with ground-truth visual clustering or localization and derives its weak learning signal purely from SRLs. As the captions are derived from visual cues of tracked entities, the approach includes a **synergistic loop**: supervising SRL captioning improves visual clustering, and better visual clustering in turn results in improved entity role grouping and captioning. Empirically, CineMEC delivers substantial improvements on *captioning entities with consistency* and *grounding and tracking them across shots*.

**Contributions summary.** (i) We propose video Multimodal Entity Coreference that unifies the identity (textual description) of each entity that plays multiple semantic roles, with localization and tracking across the entire video. (ii) We propose CineMEC, a multi-stage architecture for addressing MEC in video situation recognition. Our approach features a synergistic improvement loop between visual clustering, entity role grouping, and captioning while being weakly supervised only with SRL captions. (iii) We empirically demonstrate the effectiveness of CineMEC and see performance improvements across both vision and language metrics: captioning (CIDEr 2.5%), captioning with a unique description (LEA 7%), localization (IoU@0.5 13%), and tracking (HOTA 18%). (iv) We present several experiments showing ablations of each module and highlighting the synergy between entity role captioning and visual clustering. To support evaluation, we extend the VidSitu dataset (val, test) with visual box annotations, released for future work.

## 2. Related Work

**Fine-grained video understanding.** Beyond traditional coarse action recognition [6, 16, 17, 56, 66, 69], text-video retrieval [3, 38, 70], or video captioning [9, 10, 49, 54], fine-grained tasks require a structured or localized interpretation of actions, events, and participating entities. They can be broadly classified as: (i) vision-centric challenges such as temporal event localization [15, 21, 55, 83], spatio-temporal action detection [17, 58], and visual tracking [4, 11, 12, 47, 74]; and (ii) video-language tasks such as dense video captioning [22, 72, 84], video question answering [13, 20, 29, 59, 77], referring expression grounding and segmentation [24, 32, 41, 50, 71].

In parallel, similar to image scene graphs [27], video scene graphs have been explored to associate semantic descriptions with visual entities [7, 23, 34, 42, 64, 75]. However, video scene-graph generation (VidSGG) often lacks support to generate free-form captions or consistently maintain entity identity across repeated instances of the same class. Another structured and holistic task is *video situation recognition* or VidSitu [51], where our work lies.

**Video situation recognition.** VidSitu involves predicting salient actions per event and generating SRLs (or captions) of participating entities in the event roles [51]. A consistent description across events is desirable to preserve entity identity and is quantitatively measured with the LEA [39] score.

Several approaches have worked on VidSitu: HostSG [82], driven from the VidSGG lens, frames the task as pairwise mapping of event roles into an entity graph; OME [73] models verbs as changes in visual states of people; ClipSitu [63] uses unpooled frame features to describe event roles; and TypesDev [68] adopts vision-language models and retrieval-augmented generation to obtain demos during inference.

Closest to our work, GVSR (VideoWhisperer) [25] introduced weakly-supervised grounding of SRLs, but restrict grounding to one box per role. Further, role captions are predicted independently and often fail to unite occurrences of the same entity across events. In contrast, our approach explicitly groups all entity mentions across the video and performs SRL conditioned on its corresponding visual cluster, enabling *identity-aware* and *entity-consistent* outputs.

**Multimodal Entity Coreference.** The language community has extensively studied coreference resolution [5, 14, 60] and its variants [36]. Recently, it has been extended to image narratives [18, 19], where textual mentions are linked with bounding boxes to form coherent, entity-centric descriptions. However, extending MEC to videos is not trivial and requires tracking entities across shot changes, resolving changing entity roles, and aligning mentions with visual tracks. Prior efforts in identity-aware audio description (or multi-video captioning) [43, 46, 49] focused on characters and coreference, but only in text. Differently, our approach jointly resolves textual and visual mentions across events, enabling coherent, entity-aware, and structured video understanding.

**Grounding in MLLMs.** Early works (*e.g.* VideoChatGPT [35], Video-LLaMA [80]) primarily focused on conversational abilities and used visual features without grounding mechanisms. More recent advances (*e.g.* Qwen2.5-VL [45], VideoLLaMA3 [78]) demonstrate temporal understanding through dynamic frame-rate sampling or absolute time encoding and enable event localization in extended videos.

There is also notable progress in MLLMs developed for unifying visual segmentation with language [2, 8, 28, 40, 41, 44, 79]. For example, pixel grounding is achieved through off-the-shelf tracker and grounding modules (PG-Video-LLaVA [41]), or extensions with dual vision encoders that emphasize spatio-temporal details (VideoGLaMM [40]). A complementary approach, Open-o3 Video [37] features explicit spatio-temporal evidence in text highlighting timestamps and bounding boxes to ground responses. However, the above methods struggle with structured predictions, grounding multiple entities, and coreference across multiple text mentions and visual appearances across shots. Our work differs in multiple ways: (i) we focus on entity coreference in both text and video as entities take different roles and appearances across events; (ii) during training, we do not need costly visual segmentation/tracking annotations as supervision is derived from SRL captions; and (iii) our model enables structured prediction for VidSitu.

## 3. CineMEC for Video Situation Recognition

MEC for VidSitu is characterized by four interlinked tasks. Given a video $V$ with multiple short events $\mathcal{S} = \{s_i\}$, the tasks are defined as follows. (i) *Verb prediction* recognizes the action label $v_i$ associated with each event $s_i$. Next, a set of roles $\mathcal{R}_i = \{r \mid r \in \mathcal{P}(v_i)\}$ is obtained via a deterministic verb-role map $\mathcal{P}$ defined in VidSitu [51]. (ii) *Entity role grouping* identifies a unique set of entities $\mathcal{G} = \{G_j\}_{j=1}^{J}$ in the video by creating identity labels for each event role $\mathcal{M} = \{M_{ik}\}$. $M_{ik}$ indexes an entity for the $k^{\text{th}}$ role of event $s_i$. (iii) *Entity cluster assignment* maps each entity group $G_j$ to a visual cluster through attention scores. The cluster is represented as a subset of boxes among all proposals $\mathcal{B}$ extracted from sub-sampled video frames to form a track. (iv) Finally, *semantic role labeling* describes the entity group $G_j$ with $C_j$, different from the original formulation that has independent captions for each event role pair $C_{ik}$.

**Background: GVSR [25].** In contrast to prior works that focus mainly on verb prediction and SRL, GVSR extends SRL with weakly supervised grounding, albeit to a single box for each event role independently. Its architecture, VideoWhisperer, is organized around three key modules. (i) The Video-Object encoder (VO) is a Transformer encoder that aligns event-level video features with object features detected across sampled frames, producing contextualized representations of both events and entities. VO's event embeddings are used directly for verb prediction ($\hat{v}_i$). (ii) The Role-Object decoder (RO) is a Transformer decoder where each event role (query) applies cross-attention to contextualized object embeddings. Grounding is inferred by selecting the object with maximum attention, effectively linking each role to the most relevant entity in one frame. (iii) A Captioner, also a Transformer decoder, generates SRL captions conditioned on outputs of the RO decoder.

While GVSR enables role-wise captioning and single-frame grounding, it does not group multiple roles of a single entity and its grounding is restricted to one frame. This results in fragmented identities across roles and disjoint boxes for the same entity. To overcome these limitations, we extend GVSR's architecture with three new modules (Fig. 2): (i) Visual Clustering groups box proposals into clusters (Sec. 3.1); (ii) Role Grouping links entity mentions across events into groups (Sec. 3.2); and (iii) Entity Cluster Assignment assigns entity groups to visual clusters using cross-attention (Sec. 3.3). These additions enable multi-role reasoning, visual grounding, and captioning, and present an entity-centric framework for holistic video understanding.

### 3.1. Entity Visual Clustering (EVC)

Tracking an entity across an entire video is challenging, particularly due to frequent shot boundaries where conventional trackers fail. The VidSitu dataset features scenes with 2.89

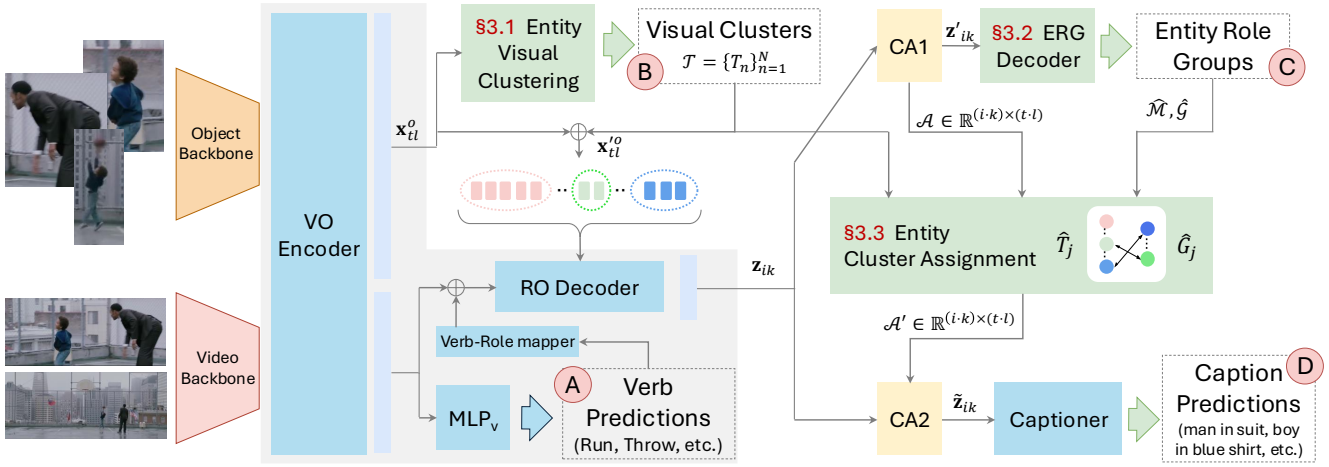



Figure 2. CineMEC extends GVSR's VO encoder, RO decoder, and Captioner modules with entity-specific modules for Visual Clustering (EVC), Role Grouping (ERG), and Cluster Assignment (ECA). We produce four outputs: (A) verb predictions and mapping to specific event-role queries, (B) visual clusters derived from object box proposals, (C) event role mention groups, and (D) entity-consistent captions after assigning entity role groups with the corresponding visual cluster. CineMEC is trained with verb, entity role group, and caption supervision, and the latter provide weak supervision for visual clustering and cluster assignment. The architecture enables synergy between vision-language to produce coherent visual clusters and captions across multiple roles played by an entity.

shot changes on average, high for a 10 second video. We address this challenge with a local-to-global strategy: a tracker first links entities across frames within a shot and subsequent modules cluster tracks on-the-fly.

Similar to GVSR [25], we subsample $\mathcal{F} = \{f_t\}_{t=1}^{F}$ frames from the video $V$. Each frame $f_t$ has up to $L$ proposal boxes that are linked into within-shot tracks using a prompt-free object tracker. Together with event features, the representations of all box proposals are contextualized via the VO encoder to obtain $\mathbf{x}_{tl}^{o}$ (for box $l$ of frame $f_t$).

Next, we link these within-shot tracks across the video by clustering $\mathbf{x}_{tl}^{o}$ using the unsupervised FINCH algorithm [53] (see Appendix E for details). This produces a set of visual clusters $\mathcal{T} = \{T_n\}_{n=1}^{N}$, where each cluster corresponds to a candidate entity spanning the video. To incorporate the clustering knowledge for downstream modules, we add new *cluster id embeddings* to $\mathbf{x}_{tl}^{o}$, yielding cluster-aware box representations $\mathbf{x}_{tl}^{\prime o}$. These encourage event role queries in the RO decoder to attend to visual clusters rather than fragmented per-frame detections.

The RO decoder produces role-specific representations $\mathbf{z}_{ik}$ for event $s_i$ and role $r_k$. An additional cross-attention layer (CA1) is used to compute multimodal entity role features $\mathbf{z}'_{ik}$ and the cross-attention map $\mathcal{A} \in \mathbb{R}^{(i*k)\times(t*l)}$ captures the affinity between event roles and box proposals.

### 3.2. Entity Role Grouping (ERG)

The multimodal entity role features $\mathbf{z}'_{ik}$ from CA1 contain information about roles and visual clusters, unlocking reasoning about entity identity in the video. A key challenge is that the same entity may appear in different roles across events, setting up the *text coreference* problem. Here, ERG's objective is to link all event role mentions that refer to the same entity. Unlike EVC, where ground-truth visual clusters are unavailable, we use the SRL annotations to provide supervision for ERG during training[1].

To group event roles, we adopt an auto-regressive formulation framed as a sequence prediction problem [46]. Our intuition is based on how humans understand a streaming video. As we encounter new events and roles, we associate the actors with previously identified entities or create new ones. Concretely, we train a Transformer decoder that takes the sequence $\mathbf{z}'_{ik}$ as input and predicts role mention identifiers $\hat{M}_{ik}$ (a unique id for each predicted entity). The decoder is trained sequentially with a causal mask so that future predictions do not influence the current step. During training, we stabilize learning using teacher forcing.

During inference, the predicted mapping $\hat{\mathcal{M}}$ for each event role is converted to $\hat{\mathcal{G}}$, a set of entity role groups each mapping to the same entity: $\hat{\mathcal{G}} = \{\hat{G}_j\}_{j=1}^{J}$, where $J$ is the number of predicted entity role groups and the group $\hat{G}_j = \{(i,k) \mid \hat{M}_{ik} = j\}$ is the set of event $s_i$ role $r_k$ indices mapped to the same $j^{\text{th}}$ entity. Next, we use these role groups to compute entity-to-visual cluster assignment.

### 3.3. Entity Cluster Assignment (ECA)

While ERG consolidates all entity role mentions that refer to the same entity, this grouping alone does not establish a consistent link to the visual modality. To bridge this gap, we introduce Entity Cluster Assignment (ECA) that explicitly associates each entity role group with a single visual cluster.

[1]In VidSitu, each entity is annotated with a unique caption across roles. Thus, we derive gold mention clusters with string matching. However, text coreference approaches may also be used.

ECA grounds each entity to the most relevant visual cluster using the attention scores $\mathcal{A} \in \mathbb{R}^{(i*k)\times(t*l)}$ from CA1. We compute this assignment by aggregating attention scores within role groups and visual clusters. First, for each visual cluster $T_n$ (from EVC), we sum the attention scores of its constituent boxes. Then, for each predicted entity group $\hat{G}_j$, we accumulate attention scores across its role mentions. The combined attention map is $\hat{\mathcal{A}} \in \mathbb{R}^{J\times N}$ and each element $\hat{\mathcal{A}}_{jn} = \sum_{g\in\hat{G}_j}\sum_{b\in T_n}\mathcal{A}[g,b]$. Next, for each entity group, $\hat{T}_j = \arg\max_n \hat{\mathcal{A}}_{jn}$ is identified as the maximally attended visual cluster. Thus, $\hat{G}_j \Leftrightarrow \hat{T}_j$ complete the MEC objective of associating each entity role group with a visual cluster.

**Updating role embeddings for captioning.** Given the entity-cluster assignments, we revisit the role-specific embeddings ($\mathbf{z}_{ik}$) to ensure consistency during captioning. Specifically, all role mentions linked to the same entity group $\hat{G}_j$ are constrained to attend *only* to the visual boxes belonging to their assigned cluster $\hat{T}_j$. To enforce this, we reuse parameters of the CA1 layer but replace its attention map by a fixed attention map $\mathcal{A}'$ with uniform weights for indices of mapped role groups and visual clusters and 0 elsewhere. For simplicity, we refer to this layer as CA2. The outputs of CA2 are *visual cluster restricted entity-aware role embeddings* $\tilde{\mathbf{z}}_{ik} = \text{CA2}(\mathbf{z}_{ik}, \mathbf{x}'^{o}_{tl}; \mathcal{A}')$. As all event roles in group $\hat{G}_j$ attend only to boxes in $\hat{T}_j$ (with equal weight), they are encouraged to learn entity consistent visual representations.

**Captioning.** The entity-aware embeddings $\tilde{\mathbf{z}}_{ik}$ are passed to the Captioner (a Transformer decoder) to generate the SRLs. In fact, we can generate them in two ways: (i) independently for each event role using $\tilde{\mathbf{z}}_{ik}$, or (ii) for each entity group $\hat{G}_j$ by mean pooling entity-aware embeddings within the group. During training, we observe that the model benefits from obtaining supervision from both paths, while during inference, we generate captions directly at the entity level.

Our approach yields multiple advantages: (i) during inference, an entity-level caption remains consistent across events and roles played by that entity; (ii) during training, the model is encouraged to learn similar representations for individual roles $\tilde{\mathbf{z}}_{ik}$ within the group $\hat{G}_j$ and the mean pooled representation $\tilde{\mathbf{z}}_j$ as they have the same target caption; and (iii) by controlling $\mathcal{A}'$, the captioner is forced to ground its predictions exclusively in the visual evidence of the assigned track. By avoiding interference from other entities, we improve the quality of generated captions, and address MEC.

### 3.4. Training and Inference

**Training.** CineMEC is trained for verb prediction, entity role grouping, and captioning in a multi-task setup. For verb prediction, we use the standard cross-entropy loss: $L^{\text{v}}_i = \text{CE}(\hat{v}_i, v_i)$. For entity role grouping, we leverage the annotation structure in VidSitu, where an entity is assigned the exact same caption across events. This allows us to derive the ground-truth entity role mapping $\mathcal{M}$. Since ERG typically features a long-tail distribution, we adopt the focal loss: $L^{\text{e}}_{ik} = \text{FL}(\hat{M}_{ik}, M_{ik})$. Finally, we supervise captioning in two ways: First, a single entity *group-level* caption is trained using the mean pooled embedding $\tilde{\mathbf{z}}_j$ with $L^{\text{gc}}_j = \sum_w \text{CE}(\hat{C}^w_j, C^w_j)$, re-weighted by the number of roles in the group. Second, individual *role-level* captions are trained using $\tilde{\mathbf{z}}_{ik}$ with $L^{\text{rc}}_{ik} = \sum_w \text{CE}(\hat{C}^w_{ik}, C^w_{ik})$. Both captioning losses are applied autoregressively over each word $w$. Finally, the overall training objective is the sum of all losses with equal weights: $L = L^{\text{v}} + L^{\text{e}} + L^{\text{gc}} + L^{\text{rc}}$.

Notably, we do not supervise the entity visual clustering (EVC) or entity cluster assignment (ECA) modules, as ground-truth visual clusters with corresponding captions are unavailable. Instead, the captioning and entity role grouping losses act as weak supervision and the model is encouraged to select clusters that align well with entities automatically as only those clusters can support generation of the appropriate caption. In this way, captioning acts as weak supervision for both clustering and assignment.

A further advantage of our formulation is the *synergy* between the linguistic and visual modalities. Accurate role grouping in language helps the model attend to the correct visual cluster, making grounding and captioning more reliable. Conversely, selecting the correct visual cluster reinforces role grouping, since mentions grounded to the same entity in RO are more likely to be assigned consistent entity ids. Thus, improvements in one modality influence the other and vice versa, tightly coupling multimodal entity coreference.

**Inference.** At inference time, given a video $V$, we predict the salient verb $\hat{v}_i$ for each event $s_i$. Next, we consider two options to obtain roles: (i) based on the ground-truth verb $v_i$, or (ii) based on the predicted verb via the mapper $\mathcal{P}(\hat{v}_i)$. Following previous work, we report these as separate experiments. In the next stage, entity visual clustering and role grouping are performed. We rely on the predicted mention mappings $\hat{\mathcal{M}}$ to compute entity role groups $\hat{\mathcal{G}}$. Finally, we predict a single caption per entity using the mean-pooled embeddings. In summary, inference produces: (i) verb predictions $\hat{v}_i$, (ii) entity coreference across SRLs $\hat{G}_j$, (iii) association between entity role group $\hat{G}_j$ and visual cluster $\hat{T}_j$, and (iv) entity-consistent captions $\hat{C}_j$.

## 4. Experiments

**Implementation details.** For a fair comparison to previous work, we adopt SlowFast [16] as the video backbone for $|\mathcal{S}|$=5 events in the videos. We obtain object tracks within a shot using the YOLOE-11 [65] prompt-free tracker. Detections in subsampled frames are encoded using SigLIP2 [61] (with ROIAlign) as the object backbone. Similar to GVSR [25], we sample frames at 1 fps from a 10 second video, resulting in $F$=11 frames. We cap box propos-

Table 1. Comparing CineMEC against previous work on VidSitu, extended to include localization. Metrics are grouped by subtask. For a fair comparison, we also highlight the backbones used in each baseline: I3D [6], SlowFast (SF) [16], Motifs-TDE (M-TDE) [57], Faster-RCNN (FR) [48], YOLOE-11 tracker [65] + SigLIP2 [61] (YS) features, and BLIP2 [30]. Trends on the test set are similar to the validation set.

| | Method | Backbone | Verb Acc. | | SRL | | | Localization | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | @1 | @5 | CIDEr | LEA | LEA-Soft | IoU@0.3 | IoU@0.5 | HOTA |
| Validation | VidSitu-SlowFast [51] CVPR'21 | SF | 32.64 | 69.20 | 45.52 | 50.48 | 31.99 | - | - | - |
| | OME+OIE [73] AAAI'23 | I3D | 53.36 | 83.94 | 47.16 | - | - | - | - | - |
| | HostSG [82] ACMMM'23 | M-TDE | 56.15 | 86.33 | 55.09 | 55.70 | 35.01 | - | - | - |
| | ClipSitu [63] IJCV'25 | X-CLIP | - | - | 61.93 | 37.77 | - | - | - | - |
| | TypesDev-ucofia [68] ICMR'25 | BLIP2 | 47.23 | - | **90.12** | 38.36 | 44.02 | - | - | - |
| | VideoWhisperer [25] NeurIPS'22 | FR + SF | 45.06 | 75.59 | 68.23 | 48.22 | 43.93 | 41.29 | 17.43 | 11.28 |
| | CineMEC (Ours, old backbone) | FR + SF | 49.86 | 78.36 | 68.38 | 54.93 | 48.91 | 46.74 | 23.02 | 15.90 |
| | VideoWhisperer [25] NeurIPS'22 | YS + SF | 46.02 | 76.49 | 73.73 | 48.37 | 48.87 | 51.49 | 42.82 | 16.23 |
| | CineMEC (Ours) | YS + SF | 49.32 | 79.83 | **76.34** | **55.78** | **52.45** | **60.15** | **55.93** | **34.22** |
| | Human | - | - | - | 84.85 | 72.10 | 70.33 | - | - | - |
| Test | VidSitu-SlowFast [51] CVPR'21 | SF | 33.94 | 70.54 | 47.25 | 50.88 | 33.50 | - | - | - |
| | VideoWhisperer [25] NeurIPS'22 | FR + SF | - | - | 68.04 | 48.77 | 44.57 | 40.36 | 16.89 | 10.81 |
| | VideoWhisperer [25] NeurIPS'22 | YS + SF | - | - | 73.21 | 48.61 | 48.36 | 51.26 | 42.13 | 16.02 |
| | CineMEC (Ours) | YS + SF | - | - | **75.73** | **55.06** | **51.16** | **59.64** | **55.33** | **33.81** |
| | Human | - | - | - | 83.68 | 71.77 | 70.60 | - | - | - |

als to $L$=15 for each frame, resulting in up to $|\mathcal{B}|$=165 box tokens per video. We support up to 6 roles per event, resulting in 30 event role queries, embeddings, entity maps ($\mathcal{M}$) and captions. We also annotate entity boxes for the val and test sets, details and resulting statistics are in Appendix F.

All Transformer modules[2] have the same configurations with 3 layers, 8 attention heads, and hidden $d$=2048. The ERG decoder is set to 2 layers. Visual clusters are obtained through two levels of FINCH [53] clustering. The verb classifier is a single linear layer from $d$=2048 to 1560 verbs. The entity id classifier in ERG is also a single linear layer from $d$=2048 to a maximum of 30 entity IDs. We only consider coreference for Arg0, Arg1, Arg2, and Location/Scene roles as suggested in VidSitu [51]. We use the Adam optimizer [26] with a learning rate of $10^{-4}$ and train the whole model end-to-end on a single L40 GPU with batch size 32.

**Metrics.** For verb prediction, we report Acc@K, *i.e.* event-level action accuracy against all ground-truth (GT) verb annotations. For SRL we report CIDEr [62], and test consistent captioning of the same entity across roles with LEA [39] and LEA-Soft [51] (a combination of LEA and CIDEr per entity). A high LEA score is obtained when the same caption is given to an entity across all its roles. For localization within a frame we report IoU@$\theta$ as defined by GVSR [25]. While this metric compares box predictions in a frame, we assess models' ability to cluster the entity through the entire video with a tracking metric, HOTA [33]. A detailed discussion of the metrics is in Appendix A. An ideal approach: (i) identifies multiple roles of the same entity across the video (LEA), (ii) performs visual clustering and entity cluster assignment correctly (HOTA), and (iii) generates good captions (CIDEr).

[2]The VO encoder contains self-attention SA+MLP layers while the RO decoder has CA+SA+MLP layers and operates in a non-autoregressive manner. The ERG decoder and the Captioner are autoregressive Transformers.

### 4.1. Comparison to State-of-the-Art

In Tab. 1 we compare CineMEC against previous methods for VidSitu. Closest to our work is the GVSR task and VideoWhisperer (VW) model [25]. Here, we reproduce VW's results with the original Faster R-CNN (FR) proposal box features, and also report results with our new YOLOE-11 tracker box proposals and SigLIP2 features (YS). For other works, we report the scores as provided in their paper while highlighting diverse feature backbones.

Except VW, other works do not perform localization and their scores are left blank. As VW focuses on grounding to a single box per SRL, they do not report metrics for visual grounding across the entire video (HOTA). We compute HOTA using a post-hoc entity grouping approach: roles with identical predicted captions are grouped into entities, and their detections are merged into visual clusters.

A recent work, TypesDev [68] achieves a high CIDEr score (90.12) on VidSitu (surpassing human performance). However, it follows a different retrieval-augmented generation pipeline to retrieve similar videos and leverage their annotations during inference for prompting MLLMs. Nevertheless, the poor LEA score (38.36) indicates that even VLMs and LLMs struggle to describe an entity that plays multiple roles with a unique caption. Additionally, TypesDev does not perform grounding.

CineMEC establishes new state-of-the-art results across multiple metrics. We improve over VW by +9 points on IoU@0.3 and +12 points on IoU@0.5, and achieve a HOTA score that is +18 points higher, owing to our role grouping and visual clustering. For captioning, CineMEC beats VW

Table 2. Results for the "predicted-verb" setting, where roles are derived using the verb-role mapping with predicted verb $\mathcal{P}(\hat{v}_i)$.

| Method | CIDEr | LEA | LEA-Soft | HOTA |
|---|---|---|---|---|
| VidSitu [51] | 30.33 | 35.92 | - | - |
| TypesDev-ucofia [68] | **73.71** | 30.12 | 34.97 | - |
| TypesDev-GPT4o [68] | 66.56 | 27.77 | 31.89 | - |
| VideoWhisperer [25] | 51.24 | 38.00 | 34.26 | 8.13 |
| CineMEC (Ours) | **60.32** | **46.21** | **42.64** | **24.76** |

Table 3. Comparing instruction-tuned MLLMs on SRL captioning against smaller models like ClipSitu and CineMEC (Ours). Experimental setup and results as reported in ClipSitu [63] IJCV'25.

| Method | VILA | Qwen2-VL | LLaVA-Video | ClipSitu | CineMEC |
|---|---|---|---|---|---|
| CIDEr | 40.44 | 57.28 | 60.10 | 61.93 | **76.34** |

Table 4. Impact of various modules (Mod.) and losses compared to CineMEC (row 10). VideoWhisperer [25] is in row 1 for completeness. The evaluation metrics from left-to-right are Verb Accuracy@1, CIDEr, LEA, LEA-Soft, IoU@0.5, and HOTA.

| # | Mod. | $L^{v}$ | $L^{rc}$ | $L^{gc}$ | $L^{e}$ | A@1 | C | L | L-S | I.5 | H |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | VW | ✓ | ✓ | | | 46.0 | 73.7 | 48.4 | 48.9 | 42.8 | 16.2 |
| 2 | | ✓ | ✓ | | ✓ | 47.8 | 70.1 | 52.1 | 47.8 | 50.3 | - |
| 3 | ERG | ✓ | | ✓ | ✓ | 48.8 | 43.9 | 51.3 | 34.5 | 50.6 | - |
| 4 | | ✓ | ✓ | ✓ | ✓ | 49.0 | 70.9 | 54.1 | 49.4 | 49.4 | - |
| 5 | | ✓ | ✓ | | | 50.0 | 75.9 | 50.4 | 48.5 | 54.2 | - |
| 6 | EVC | ✓ | | ✓ | | 50.4 | 70.6 | 38.7 | 41.5 | 42.8 | - |
| 7 | | ✓ | ✓ | ✓ | | **50.7** | 76.1 | 45.0 | **47.2** | 55.5 | - |
| 8 | ERG | ✓ | ✓ | | ✓ | 48.2 | 75.6 | 55.0 | 52.1 | 53.3 | 29.1 |
| 9 | + | ✓ | | ✓ | ✓ | 49.6 | 74.6 | 55.0 | **54.0** | 51.5 | 28.9 |
| 10 | EVC | ✓ | ✓ | ✓ | ✓ | 49.3 | **76.3** | **55.8** | 52.5 | **55.9** | **34.2** |

Table 5. Ablations for importance of applying the cross-attention after ECA with the hard coded attention map (w/o CA2) and using cluster-aware embeddings as RO's memory (no clst-emb).

| Method | Acc@1 | CIDEr | LEA | LEA-S | IoU-0.5 | HOTA |
|---|---|---|---|---|---|---|
| CineMEC | 49.32 | **76.34** | **55.78** | **52.45** | **55.93** | **34.22** |
| w/o CA2 | **49.68** | 72.16 | 54.83 | 49.29 | 51.40 | 32.68 |
| no clst-emb | 48.37 | 74.23 | 54.17 | 51.62 | 52.36 | 33.61 |

in CIDEr by +2.5 and LEA by a large gap +7.4, reflecting our approach's ability to capture entity coreference across SRLs. Importantly, where prior methods show a trade-off between high CIDEr (*e.g.* VW [25]) or high LEA (*e.g.* HostSG [82]), our approach achieves the best of both worlds. LEA-Soft, a metric introduced by VidSitu, reflects this balance, and CineMEC outperforms the best prior work (VW) by +3.5 points. Thus, *CineMEC demonstrates strength in both tasks, explicit visual grounding of entities and implicit identity preservation through consistent, identity-aware captions.*

**Predicted verb–role evaluation.** Many prior works on VidSitu report results only on the ground-truth role setting (Tab. 1). Here, models predict the verb for each event, but the set of roles are taken from the GT verb. While this setup simplifies the task, it does not capture end-to-end performance. Towards holistic video understanding, we use the predicted verb $\hat{v}_i$ for each event, and determine roles based on the verb-role mapping $\mathcal{P}(\hat{v}_i)$. In this case, the model must predict the correct verb, and then generate entity-level captions and visual clusters for identified roles.

Tab. 2 reports results for this *predicted-verb* setting. Despite the added difficulty and missing oracle roles, our method consistently outperforms VW (and other approaches) across all metrics (+9 CIDEr, +8 LEA, +18 HOTA). This highlights the strength of our approach to generate reliable captions and clusters in a fully end-to-end setting.

**Comparison to captioning with MLLMs.** Tab. 3 reports performance of multiple MLLMs [31, 67, 76, 81] instruction-tuned for structured captioning. We observe poor performance than a simple approach that uses CLIP features (ClipSitu [63]) and a gap of over 15 points CIDEr to CineMEC. A possible reason is that auto-regressive generation is not ideal as event roles are related to each other. Instead, CineMEC benefits from learning entity-specific representations.

## 4.2. Ablations

**Impact of modules.** We study the effects of excluding visual clustering (EVC) or role grouping (ERG) modules and related losses against the VW baseline in Tab. 4.

When using only ERG (rows 2-4), visual clusters are unavailable, and we generate captions using mean-pooled embeddings of each event-role group. While this builds coreference across event roles, errors in role-mention grouping cause captions of one entity to be incorrectly assigned to another, resulting in a lower CIDEr and LEA score.

In contrast, when using only EVC (rows 5-7), we treat each role as a separate entity and generate captions from holistic visual clusters. While CIDEr improves as captions are grounded in coherent visual information, LEA remains low as role mentions of the same entity are not linked.

Finally, combining both ERG and EVC (rows 8-10) provides complementary benefits. Referencing the correct visual clusters improves role grouping, while stronger role grouping in turn improves clustering. As a result, both CIDEr and LEA improve over the baseline.

**Dual captioning losses** for role and group complement each other. Comparing row 3-4, 6-7, 9-10, we see a performance drop in CIDEr and LEA when $L^{rc}$ is removed.

**Modified attention map in CA2.** We replace the fixed attention map of CA2 with the original one from CA1. Recall, CA2 constrains the input embedding to the captioner to boxes that are associated with the entity group. Tab. 5 "w/o CA2" results in weaker grounding and lower CIDEr, highlighting that entity-constrained attention is helpful.

**Memory design in RO decoder.** In EVC, we added cluster id embeddings to bbox representations $\mathbf{x}^{o}_{tl}$ from the VO

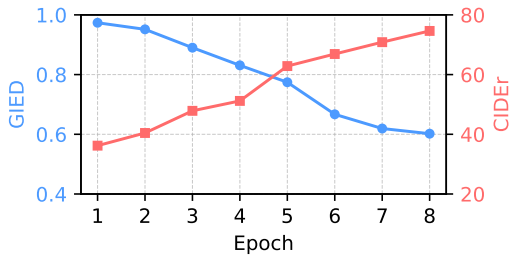



Figure 3. Evolution of Ground-truth Intra-Entity Distance (GIED) for visual clustering and CIDEr during training. Across epochs, GIED decreases and CIDEr increases showing the synergistic behavior of visual clustering with captioning.

encoder. Ignoring these cluster id embeddings results in a small performance drop (Tab. 5 "no clst-emb") in CIDEr and highlights the benefit of incorporating cluster information in the RO decoder's memory representation.

### 4.3. Analysis and Discussion

We refer readers to Appendix B in the supplement for qualitative analysis and example videos. We also include some qualitative outputs from the textual grounding model Open-o3 Video [37] showcasing challenges of MLLMs. Next, we present some diagnostic analyses of CineMEC to understand how individual modules contribute to the approach.

**Visual clustering and captioning improve together.** We study how the captioning metric CIDEr correlates with visual clustering quality by analyzing how proposal boxes that belong to the same entity evolve during training. Specifically, we measure if proposal boxes corresponding to the same entity move closer together in the embedding space as training progresses. We compute ground-truth intra-entity distance (GIED) by associating proposal boxes with the GT entity boxes based on validation annotations with IoU $> 0.3$.

To achieve this, at each frame, we match the GT entity box to the closest proposal box (within a fixed threshold), resulting in a oracle visual cluster based on proposal boxes. Given a cluster with $n$ matched boxes, we compute the average pairwise distance across all $n \cdot (n-1)/2$ box pairs and average this across all entities to obtain GIED. A lower value indicates that boxes belonging to the same entity cluster are coming closer in the learned representation space.

Fig. 3 shows that GIED decreases consistently and correlates with an increase in CIDEr across training epochs. This confirms the *synergistic loop*: improved visual clustering benefits captioning quality and the weak learning signal of captions is able to improve clustering.

**Analyzing entity role grouping.** We evaluate the contribution of the ERG module by comparing inference without grouping or using ground-truth grouping. The experiments with different ERG inference strategies are performed on the same model. As seen in Fig. 4, we compare: (i) no event-role grouping, where each event-role is captioned independently; (ii) predicted grouping, using our ERG decoder

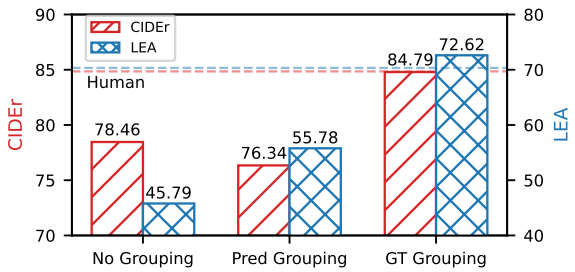



Figure 4. ERG strategies: no grouping, predicted grouping, and ground-truth grouping of event role mentions. Perfect grouping yields high CIDEr and LEA, meeting human performance.

(default CineMEC); and (iii) ground-truth grouping, using the SRL string matched annotations.

Without grouping, LEA drops sharply (-10 points) as captions are generated independently for each role without any consistency across mentions of the same entity. With predicted grouping, LEA improves, as all entity mentions are captioned consistently. Interestingly, CIDEr decreases slightly (-2 points) due to erroneous entity id predictions in ERG resulting in incorrect grouping of some event roles. Finally, with ground-truth grouping, both CIDEr (+8 points) and LEA (+17 points) jump significantly, indicating that our captioning module is capable of producing captions that match human-level performance of CIDEr and LEA.

ERG module's performance is limited by the long-tail nature of the role grouping problem since: (i) the average number of entities per video (5.2) is relatively small but the spread is high, and (ii) few entities (3-5) appear frequently across the video and dominate re-occurrences, also resulting in a skewed distribution. We present additional results and discuss inference speed in Appendix C. Further limitations are discussed in Appendix D in the supplement.

## 5. Conclusion

We proposed Multimodal Entity Coreference (MEC) as a solution for holistic understanding of videos with structured language descriptions and visual grounding. We introduced CineMEC, a multi-stage framework that integrates MEC into video situation recognition (VidSitu) and learns without costly visual grounding supervision. By unifying entity role mention groups with visual clusters, CineMEC enabled consistent identity tracking and entity-aware captions across events, advancing previous role-independent approaches. Extensive experiments on VidSitu and comparison against task-specific works or instruction-tuned MLLMs demonstrated substantial improvements across captioning, localization, and tracking; highlighting the synergy between visual clustering and role grouping.

**Acknowledgments.** We thank funding support from a Google India Faculty Award and SERB SRG/2023/002544 for compute. Amazon did not fund, direct, or influence this research work and all work was done independent of any Amazon involvement.

## References


[1] CVAT Annotation Tool. https://github.com/openvinotoolkit/cvat. 3

[2] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One Token to Seg Them All: Language Instructed Reasoning Segmentation in Videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[4] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. XMem++: Production-Level Video Segmentation from Few Annotated Frames. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[5] Bernd Bohnet, Chris Alberti, and Michael Collins. Coreference Resolution through a Seq2Seq Transition-Based System. *Transactions of the Association of Computational Linguistics (TACL)*, 11:212–226, 2023. 3

[6] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6

[7] Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. Joint Multimedia Event Extraction from Video and Article. In *Findings of the Association for Computational Linguistics (ACL)*, 2021. 2

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*, 2023. 3

[9] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[10] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[11] Ho Kei Cheng and Alexander G Schwing. XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[12] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and Track Anything. *arXiv preprint arXiv:2305.06558*, 2023. 2

[13] Rohan Choudhury, Koichiro Niinuma, Kris M Kitani, and László A Jeni. Zero-Shot Video Question Answering with Procedural Programs. In *ECCV*, 2024. 2

[14] Vladimir Dobrovolskii. Word-Level Coreference Resolution. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 3

[15] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. DAPS: Deep Action Proposals for Action Understanding. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 5, 6

[17] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video Action Transformer Network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[18] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Semi-Supervised Multimodal Coreference Resolution in Image Narrations. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 2, 3

[19] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Who Are You Referring To? Coreference Resolution in Image Narrations. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[20] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[21] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[22] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTIMELLM: Empower LLM to Grasp Video MomentsVideo Action Transformer Network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[23] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2

[25] Zeeshan Khan, CV Jawahar, and Makarand Tapaswi. Grounded Video Situation Recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3, 4, 5, 6, 7, 2

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017. 2

[28] Ju-Hee Lee and Je-Won Kang. SRTube: Video-Language Pre-Training with Action-Centric Video Tube Features and Semantic Role Labeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[29] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, Compositional Video Question Answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 2

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*, 2023. 6

[31] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[33] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking. *International Journal of Computer Vision (IJCV)*, 129(2):548–578, 2021. 6, 1

[34] Zelun Luo, Wanze Xie, Siddharth Kapoor, Yiyun Liang, Michael Cooper, Juan Carlos Niebles, Ehsan Adeli, and Fei-Fei Li. MoMA: Multi-Object Multi-Actor Activity Parsing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[35] Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Association of Computational Linguistics (ACL)*, 2024. 3

[36] Kawshik Manikantan, Shubham Toshniwal, Makarand Tapaswi, and Vineet Gandhi. Major Entity Identification: A Generalizable Alternative to Coreference Resolution. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 3

[37] Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. Open-o3 Video: Grounded Video Reasoning with Explicit Spatio-Temporal Evidence. *arXiv preprint arXiv:2510.20579*, 2025. 2, 3, 8, 1

[38] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[39] Nafise Sadat Moosavi and Michael Strube. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-Based Entity Aware Metric. In *Association of Computational Linguistics (ACL)*, 2016. 2, 6, 1

[40] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. VideoGLAMM: A Large Multimodal Model for Pixel-Level Visual Grounding in Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3

[41] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. PG-Video-LLaVA: Pixel Grounding Large Video-Language Models. *ArXiv 2311.13435*, 2023. 2, 3

[42] Trong-Thuan Nguyen, Pha Nguyen, and Khoa Luu. HIG: Hierarchical Interlacement Graph Approach to Scene Graph Generation in Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[43] Jae Sung Park, Trevor Darrell, and Anna Rohrbach. Identity-Aware Multi-Sentence Video Description. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[44] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding Multimodal Large Language Models to the World. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[45] Qwen Team. Qwen2.5-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *Blog post: https://qwenlm.github.io/blog/qwen2.5-vl/*, 2025. 3

[46] Haran Raajesh, Naveen Reddy Desanur, Zeeshan Khan, and Makarand Tapaswi. MICAP: A Unified Model for Identity-Aware Movie Descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 4

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 6

[49] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie Description. *International Journal of Computer Vision (IJCV)*, 123:94–120, 2017. 2, 3

[50] Arka Sadhu, Kan Chen, and Ram Nevatia. Video Object Grounding using Semantic Roles in Language Description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[51] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual Semantic Role Labeling for Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 6, 7

[52] Darshana Saravanan, Darshan Singh, Varun Gupta, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. VELOCITI: Can Video-Language Models Bind Semantic Concepts through Time? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1

[53] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient Parameter-Free Clustering Using First Neighbor Relations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 6, 2

[54] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-End Generative Pretraining for Multimodal Video Captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[55] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal Action Localization in Untrimmed Videos via Multi-Stage CNNs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[56] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-Centric Relation Network. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[57] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and

Hanwang Zhang. Unbiased Scene Graph Generation from Biased Training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[58] Makarand Tapaswi, Vijay Kumar, and Ivan Laptev. Long Term Spatio-Temporal Modeling for Action Detection. *Computer Vision and Image Understanding (CVIU)*, 210, 2021. 2

[59] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, , and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[60] Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. On Generalization in Coreference Resolution. In *Workshop on Computational Models of Reference, Anaphora and Coreference*, 2021. 3

[61] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025. 5, 6

[62] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-Based Image Description Evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[63] Dhruv Verma, Debaditya Roy, and Basura Fernando. Effectively Leveraging CLIP for Generating Situational Summaries of Images and Videos. *IJCV*, 2024. 3, 6, 7

[64] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. MovieGraphs: Towards Understanding Human-Centric Situations from Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[65] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-Time Seeing Anything. *arXiv preprint arXiv:2503.07465*, 2025. 5, 6, 2

[66] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[67] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024. 7

[68] Hanxiao Wei, Bin Wu, Chunjia Wang, Guangyao Su, and Tao Zhou. Demonstration Meets Typed Events: Type Specific Video Semantic Role Labeling via Multimodal Prompting and Retrieval. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 2025. 3, 6, 7, 2

[69] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[70] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[71] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-Temporal Video Grounding with Transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[72] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[73] Guang Yang, Manling Li, Jiajie Zhang, Xudong Lin, Heng Ji, and Shih-Fu Chang. Video Event Extraction via Tracking Visual States of Arguments. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2023. 3, 6, 2

[74] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track Anything: Segment Anything Meets Videos. *arXiv preprint arXiv:2304.11968*, 2023. 2

[75] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, et al. Panoptic Video Scene Graph Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[76] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qi-An Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*, 2024. 7

[77] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019. 2

[78] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, et al. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*, 2025. 3

[79] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. LLaVA-Grounding: Grounded Visual Chat with Large Multimodal Models. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[80] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *EMNLP Demo track*, 2023. 3

[81] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *TMLR*, 2025. 7

[82] Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing Holistic Spatio-Temporal Scene Graph for Video Semantic Role Labeling. In *ACM Multimedia (MM)*, 2023. 3, 6, 7, 2

[83] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal Action Detection with

Structured Segment Networks. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[84] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming Dense Video Captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2