Critical Analysis: Video Event Extraction via Tracking Visual States of Arguments

Kai-Yu Lu

1 Methodological Strengths

1.1 Event-as-state-change formulation and argument-centric design

A central strength of the paper lies in the explicit linkage between linguistic theories of events and the model design. Events are treated as changes of state of arguments, and the entire architecture is constructed to operationalize this view by tracking visual state trajectories of objects and their interactions across time. The event representation

$$e = \{v,\ \langle r_0, a_0\rangle,\ \langle r_1, a_1\rangle,\ \dots\} \tag{1}$$

with a verb v and a set of role-argument pairs aligns directly with the VidSitu ontology and provides a clear target structure for both verb classification and semantic role prediction tasks.

The representation is explicitly argument-centric rather than purely clip-centric. Instead of relying only on global clip features, the framework decomposes visual state change into three components:

- in-bounding-box pixel changes;
- bounding-box displacements;
- multi-argument interaction regions.

This decomposition grounds the design of Object State Embedding (OSE), Object Motion-aware Embedding (OME), and Object Interaction Embedding (OIE) and provides a principled route to encode three major sources of visual evidence for events.

1.2 Systematic decomposition of object motion and interaction

The methodological treatment of object motion is detailed and physically motivated. From a kinematic perspective, object dynamics are decomposed into:

- coordinate displacement of bounding boxes, representing translational motion at the object level;
- pixel-level changes inside bounding boxes, representing local appearance change, including pose changes and non-rigid motion.

OSE combines pooled grid features inside the projected bounding box with a linear coordinate embedding

$$s_{ji} = [\, p_{ji},\ c_{ji}\,], \tag{2}$$

yielding a single-frame state vector for object o_j in frame f_i. OME then aggregates the sequence $\{s_{ji}\}_{i=0}^{F}$ into a trajectory-level embedding m_j using either an LSTM or average pooling. This explicit separation

between per-frame state and temporal aggregation clarifies the role of each component and permits ablation on the choice of temporal aggregator.

Interaction modeling is constructed in a similarly systematic manner. A union bounding box over all objects in a frame is defined, and the same pixel-plus-coordinate encoding is applied to obtain per-frame interaction states, which are then aggregated into OIE. Although the union region is coarse, the design is conceptually clean and reuses the same mechanisms as individual object modeling.

1.3 Argument Interaction Encoder and multi-scale integration

The Argument Interaction Encoder aggregates global, object-level, and interaction-level information with a Transformer layer:

$$e = \text{Trans}\big([\, g;\ m_0; \ldots; m_O;\ i\,]\big), \tag{3}$$

where g is the globally pooled SlowFast feature, $\{m_j\}$ are OME embeddings, and i is OIE. This design exhibits several methodological strengths:

- The Transformer provides a permutation-friendly mechanism for multi-argument aggregation and allows each element to attend to others, rather than enforcing a fixed hand-designed weighting.
- Both slow and fast pathways of SlowFast are preserved and concatenated before projection, ensuring that the event embedding e is multi-scale in time and combines low-frame-rate semantics with high-frame-rate motion cues.
- The maximum number of objects O is treated as a hyperparameter, and empirical results indicate that the self-attention layer can downweight uninformative objects, which supports the robustness of the design.

1.4 Experimental design and evaluation protocol

The experimental design on VidSitu is comprehensive relative to the benchmark.

Dataset and task coverage. The model is evaluated on two related but distinct tasks:

- verb classification over 2 154 verb senses, with multi-label annotations (ten verbs per clip) in validation and test splits;
- semantic role prediction as free-form argument generation for roles such as ARG0, ARG1, ARG2, ARGLOC, and ARGSCENE.

Dataset statistics are clearly reported, with separate counts for clips, verb instances, and role instances across train, validation, and two test splits, which clarifies the size and difficulty of the benchmark.

Metrics and ranking-based evaluation. For verb classification, the study reports Accuracy@1, Accuracy@5, Recall@5, and F1@5, which together characterize both precision and coverage of multi-label predictions. For semantic role prediction, CIDEr, CIDEr-Verb, CIDEr-Arg and ROUGE-L are used, and the distinction between micro and macro CIDEr provides insight into performance across frequent versus rare verbs and roles.

Baselines and fairness. The paper compares against:

- the original VidSitu baselines using I3D and SlowFast (with Non-Local blocks) for both tasks;
- a transformer-based video model, TimeSformer, for verb classification;
- a GPT-2 text-only decoder for semantic role prediction.

For semantic role prediction, the SlowFast baseline is re-run ten times and mean and standard deviation are reported, rather than using a single run. The proposed model is also evaluated with ten seeds. This addresses the high variance inherent in free-form generation and improves fairness.

Ablation studies. The ablations are carefully structured:

- comparison of OSE-pixel-only versus OSE-pixel-plus-displacement quantifies the contribution of displacement encoding;
- comparison with and without OIE clarifies the incremental effect of interaction modeling on verbs and roles;
- varying the maximum number of objects ($O_{\max} = 2, 4, 8$) assesses sensitivity to object count;
- comparing LSTM versus average pooling in StateAgg isolates the effect of the temporal aggregation operator in OME.

Reproducibility and implementation detail. Training regimes are described with:

- number of epochs (10 for both tasks);
- batch size (8);
- learning rate ranges and optimizer (Adam with explicit hyperparameters);
- object cap (8 objects per clip) and coordinate embedding dimension (128);
- learning-rate shrinkage for pre-trained parameters by 90 percent.

The supplementary section lists the random seeds for semantic role prediction runs. This level of detail substantially improves reproducibility.

1.5 Qualitative analysis

The qualitative examples are tightly aligned with the core hypotheses:

- cases where global motion suggests roll or fall but fine-grained arm pixel changes correspond to fight or hit;
- cases where mouth motion is present but the absence of food interaction shifts prediction from eat to talk;
- cases where minimal positional and pixel change yields look and talk predictions, while baselines predict more motion-oriented verbs.

These analyses illustrate that the argument-level decomposition, especially the combination of pixel changes, displacement, and interaction reasoning, can resolve ambiguities that global 3D features and simple motion cues do not resolve.

2 Key Limitations

2.1 Dataset scope and generalization

The study is conducted exclusively on VidSitu. Although VidSitu is large and challenging, it consists of short clips of approximately two seconds with a fixed ontology and a limited domain. This leads to several limitations:

- No cross-dataset evaluation is presented, so generalization to other video domains such as long-form instructional videos or egocentric recordings is unknown.
- The ontology of 2 154 verb senses and a fixed set of roles restricts evaluation to closed-world event semantics. Open-vocabulary verbs and roles or compositional event structures are not considered.
- Clips are short and typically centered around a single salient event. Scenarios involving multiple overlapping events or long-range dependencies are not evaluated, despite being common in real-world video understanding.

2.2 Dependence on external object detection and tracking

The pipeline relies on VidVRD for object detection and tracklet generation. This introduces several fragility points:

- Missed detections or fragmented tracklets can cause important arguments to be absent or partially observed in OSE and OME. The narrative highlights cases where correct event discrimination depends critically on the detection of items such as bread.
- The approach assumes accurate temporal association of bounding boxes into tracklets. When association fails, OME aggregates inconsistent state embeddings, which corrupts the motion representation.
- The backbone encoder and the tracker are trained separately, and error propagation from detection and tracking into event prediction is not mitigated through end-to-end learning.

No ablation varying tracker quality is provided, so robustness to detection and tracking errors is not quantified.

2.3 Limited exploitation of temporal structure and event boundaries

The model aggregates state embeddings across entire clips via StateAgg and InterAgg and then applies a single Transformer layer to obtain a static event embedding e. This design implicitly assumes:

- a single temporal scale is sufficient to characterize the event, despite the existence of complex temporal patterns such as repeated actions or multi-stage interactions;
- event boundaries are aligned with clip boundaries, since no explicit temporal localization or segment-level modeling is performed;
- one event embedding is adequate for both verb classification and all argument roles, even when different roles are most salient at different time points.

2.4 Interaction modeling limitations

OIE uses a union bounding box over all objects in a frame. This design is simple and computationally efficient but structurally limited:

- All pairwise relations are compressed into a single union region, which cannot distinguish specific object pairs or relation types.
- Union pooling can introduce substantial background content, especially when objects are spatially far apart, leading to potential dilution of interaction signals.
- Empirically, OIE yields only small gains on verb classification and does not consistently improve semantic role prediction, which suggests that the interaction representation is not yet fully expressive or optimally integrated.

2.5 Evaluation and error analysis gaps

Several aspects of evaluation are underdeveloped:

- For verb classification, only set-based metrics such as F1@5 are reported. No average precision or per-verb performance breakdown is provided, so performance on rare or ambiguous verbs remains unclear.
- For semantic role prediction, CIDEr and ROUGE-L measure surface similarity, but role-wise recall or precision metrics are not reported. The degree to which arguments are attached to the correct roles versus merely being plausible phrases is not quantified.
- Error analysis is largely qualitative and limited to a few examples. There is no systematic taxonomy of failure modes such as missed agents, incorrect scenes, or confusion between semantically related verbs.

2.6 Computational cost and efficiency

The pipeline combines:

- a 3D convolutional backbone (SlowFast with Non-Local blocks);
- VidVRD object detection and tracking;
- per-object and union-region pooling;
- temporal aggregators for objects and interactions;
- a Transformer-based Argument Interaction Encoder and a Transformer-based decoder.

Training time for verb classification (approximately 20 hours on four V100 GPUs) is briefly mentioned, but no detailed analysis of memory footprint, inference latency, or scaling is provided. Comparative resource usage relative to baselines such as plain SlowFast or TimeSformer is not presented, so the trade-off between accuracy gains and computational cost cannot be fully assessed.

3 Technical Bottlenecks

3.1 Information bottlenecks in representation

Several stages of the pipeline compress rich spatio-temporal information into compact vectors:

- Average pooling within bounding boxes discards spatial structure inside objects. Detailed part-level configurations, such as hand–object contact, are reduced to a single vector.
- StateAgg and InterAgg compress per-frame state sequences into single vectors for each object and for the union interaction region, potentially losing ordering and duration information.
- A single Transformer layer aggregates all objects, interactions, and the global feature into one event embedding e, which is used for both verb classification and the entire argument sequence.

These compression steps can be sufficient for short clips with relatively simple events, but they create bottlenecks for more complex temporal patterns or for events that depend on fine-grained spatial relationships.

3.2 Architectural constraints from external modules

The architecture is constrained by the characteristics of SlowFast and VidVRD:

- SlowFast is optimized for dense video classification, not for structured event prediction with roles. Its internal representations may not be ideally aligned with the requirements of argument-level reasoning, and only shallow adaptation is performed through pooling and a single Transformer layer.
- VidVRD is optimized for relation detection and may produce tracklets tailored to object–relation tasks rather than to argument roles in event semantics. Misalignment between VidVRD categories and VidSitu argument ontology may lead to mismatches in coverage of relevant entities.

Since these modules are not integrated end-to-end with the event extraction objective, the system cannot adapt detection or tracking behavior based on event-level supervision.

3.3 Integration challenges between modules

The pipeline involves components trained at different stages:

- verb classification is trained first using the event embedding e;
- semantic role prediction is trained subsequently with the visual encoder frozen and only the Transformer encoder–decoder updated.

This staged training introduces integration challenges:

- The event embeddings e are optimized for verb classification, not specifically for role-sensitive distinctions among arguments. Freezing e in the second stage prevents refinement of the visual representation to better support argument generation.
- No joint loss couples verb and role predictions. As a consequence, object-level features are not explicitly shaped so that argument roles such as ARG0 and ARG1 become more separable in the visual space.

3.4 Trade-offs in temporal aggregation and model capacity

Ablation studies show that average pooling and LSTM aggregators achieve similar performance in OME. This suggests that temporal modeling capacity is not fully exploited. However, it also indicates a trade-off:

- Simple pooling reduces parameters and improves efficiency, but it may limit the ability to capture nuanced temporal patterns such as acceleration or cyclical motion.
- More expressive temporal models, such as deeper recurrent networks or temporal transformers, would increase computational cost and complicate training, especially when stacked on top of SlowFast.

The chosen architecture favors simplicity, which works well on VidSitu but may become a bottleneck on more temporally complex datasets.

4 Research Implications

4.1 Implications for event-centric video understanding

The empirical gains in both verb classification and semantic role prediction indicate that argument-level modeling provides tangible benefits over global clip encodings. Improvements in verb F1@5 and in CIDEr for semantic roles demonstrate that:

- events are more than global motion patterns; they are structured configurations of object states and interactions;
- modeling the dynamics of individual arguments reduces verb and role ambiguities that global 3D backbones do not resolve.

This supports a view of video understanding in which event semantics and argument roles are primary modeling targets, rather than labels attached post hoc to clip-level features.

4.2 Gaps between benchmark settings and real-world deployment

The study highlights several gaps that are relevant for real-world deployment:

- reliance on high-quality object detection and tracking is feasible on curated benchmarks but fragile in uncontrolled environments with occlusion, motion blur, and crowded scenes;
- short clips with known event presence do not capture scenarios where events must be localized in continuous streams, often with overlapping and nested event structures;
- fixed ontologies do not reflect evolving vocabularies and domain-specific event types in applications such as surveillance, sports analytics, or instructional video analysis.

4.3 Connections to structured reasoning in other domains

The representation choices resonate with trends in other areas:

- In natural language processing, event extraction and semantic role labeling emphasize joint modeling of triggers and arguments with structured models. The present results show that an analogous perspective is beneficial in video.

- In vision-language tasks such as visual question answering and image captioning, object-centric representations and relational reasoning have been successful. The results here provide further evidence that explicit modeling of objects and their interactions is a promising direction for multi-modal event reasoning.

4.4 Theoretical insights about state change and event semantics

By demonstrating that explicit state-change modeling yields improvements over strong baselines, the paper provides empirical support for the view that events are fundamentally state transitions of entities. The observation that both displacements and pixel changes within bounding boxes are required for best performance strengthens this view:

- displacement encodes coarse spatial transitions such as moving towards or away from another object;
- pixel changes encode intrinsic state transitions such as changing from standing to lying or modifying posture.

This suggests that event semantics require both relational movement and intrinsic state information, which has implications for future representation design for video reasoning.

5 Potential Research Directions

5.1 End-to-end integration of detection, tracking, and event extraction

One important extension is to integrate detection, tracking, and event extraction into a unified model. Potential directions include:

- replacing VidVRD with transformer-based detectors that directly produce spatio-temporal object queries supervised jointly by detection, tracking, and event-level losses;
- using differentiable association mechanisms that learn tracklets under event supervision rather than relying on precomputed tracks.

Such integration would reduce error propagation and allow the event objective to influence lower-level perception modules.

5.2 Richer interaction modeling with relational architectures

Instead of union bounding boxes, future work can adopt relational architectures:

- graph neural networks where nodes are objects and edges encode spatial and temporal relations;
- multi-head attention layers where query–key pairs correspond to specific object pairs, enabling explicit modeling of who acts on whom;
- relation-specific embeddings that distinguish contact, collision, containment, and other interaction types.

These designs would preserve pairwise and higher-order relations and could improve both verb classification and role assignment, especially for complex multi-agent events.

5.3 Hierarchical and segment-level temporal modeling

To address the limitations of global temporal aggregation, future models could:

- introduce segment-level representations where events are defined over variable-length temporal intervals discovered by attention or change-point detection;
- employ hierarchical temporal transformers that operate at frame, segment, and episode levels, enabling modeling of multi-stage events and repeated actions;
- jointly predict event boundaries and semantics, rather than assuming clip-aligned events.

5.4 Enhanced evaluation and diagnostic metrics

To deepen understanding of model behavior, future evaluations could:

- report per-verb and per-role metrics, including confusion matrices that reveal systematic errors among semantically similar verbs and roles;
- introduce role-wise recall and precision metrics that explicitly assess how well arguments are attached to the correct roles;
- design diagnostic test suites that isolate capabilities such as distinguishing events based primarily on interaction presence versus internal object state change.

5.5 Robustness, uncertainty modeling, and efficiency

The dependence on external modules and heavy backbones suggests opportunities for:

- incorporating uncertainty estimates from the tracker into the event model, for example by weighting object contributions according to detection confidence;
- exploring model compression and distillation strategies that transfer argument-aware capabilities into lighter-weight architectures suitable for real-time processing;
- designing training objectives that encourage robustness to missing or noisy tracklets, for instance through dropout over objects or perturbations of bounding boxes.

5.6 Multimodal and knowledge-guided extensions

Since videos often come with audio or text, future work can:

- integrate audio signals that provide cues for events such as collisions, speaking, or environmental changes;
- incorporate textual descriptions, subtitles, or narrations to disambiguate visually similar events and to refine role assignments;
- use external commonsense knowledge to constrain possible verb–role combinations and improve the plausibility of role prediction.

6 Conclusion

The paper presents a structured, argument-centric approach to video event extraction that operationalizes the idea of events as changes of state of arguments. Through Object State Embedding, Object Motion-aware Embedding, and Object Interaction Embedding, aggregated by an Argument Interaction Encoder, the study demonstrates clear gains on VidSitu for both verb classification and semantic role prediction compared with strong 3D convolutional and transformer-based baselines.

The main methodological strengths include the principled event formulation, the detailed decomposition of object state and motion, the comprehensive ablation studies, and improved reporting of variance in free-form argument generation. At the same time, key limitations remain in dataset coverage, reliance on external detection and tracking, limited temporal modeling, and coarse interaction representations.

The most promising research directions involve end-to-end integration of detection, tracking, and event extraction, richer relational modeling, hierarchical temporal representations, more diagnostic evaluation protocols, and extensions to multimodal and knowledge-enriched settings. Pursuing these directions would move the field from state-change tracking in curated short clips toward robust, interpretable event understanding in complex real-world video environments.