Interaction-Aware Video Narrative Generation for Short-Form Gaming Content

Ari Yu* **Sung-Yun Park*** **Sang-Kwang Lee†**
Electronics and Telecommunications Research Institute
University of Science and Technology
{ari, tjddbs5671, sklee}@etri.re.kr

Abstract

The rapid growth of short-form video consumption underscores the need for a next-generation paradigm in video generation. A key challenge in this paradigm is to design models that can identify dense interaction segments and generate coherent narratives. However, existing video understanding models remain limited in capturing complex interactions, resulting in narratives that often lack coherence. Game videos, in particular, where multiple agents interact in real time to create non-linear storylines, require a deeper understanding of interaction dynamics and narrative coherence. To address this challenge, we introduce an Interaction-Aware Video Narrative Generation (IaVNG) model. Our approach first extracts key interaction segments through kernel density estimation and then produces coherent narratives to generate short-form videos. In experiments, IaVNG shows promise as a generalizable model for next-generation video generation in non-linear domains by selecting key interactions and generating coherent short-form narratives.

1 Introduction

The recent expansion of short-form video consumption has highlighted the necessity for a next-generation paradigm in video generation. This paradigm goes beyond simple scene summaries, requiring technologies that can effectively capture salient interactions and contextual information within a limited timeframe of 15–60 seconds. Therefore, identifying meaningful interaction segments in long videos and reconstructing them into a coherent narrative becomes a key challenge. Motivated by this, research in video understanding has expanded from summarizing entire long videos to detecting semantically meaningful segments and describing them individually [1, 2, 3, 4]. However, complex domains such as sports and games, where multiple agents interact in real time and narratives unfold nonlinearly, remain a limitation for existing models in capturing them effectively. In an attempt to solve this problem, Yu et al. [5] proposed a model that detects event intervals using Detecting Action Proposals (DAPs) [6] for sports videos and constructs narratives by capturing multiple interactions through object segmentation. However, since DAPs rely on visual motion to propose intervals, the extracted event scenes often fail to include meaningful interactions. Additionally, Tanaka et al. [7] segmented full-match game videos into non-semantic clips using automatic scene transition detection and ResNext-50 [8] filtering, resulting in fragmented narratives. As a result, existing approaches fail to preserve meaningful sequential interaction segments and their context. This disrupts the flow of the narrative and makes the composition of short-form content feel unnatural. To address these challenges, we introduce Interaction-Aware Video Narrative Generation (IaVNG), a model tailored for short-form content in game videos. The model follows a two-stage approach including a Key Interaction Segment (KIS) module and a Narrative Generation (NG) module. The KIS module

*Equal contribution.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NextVid Workshop.

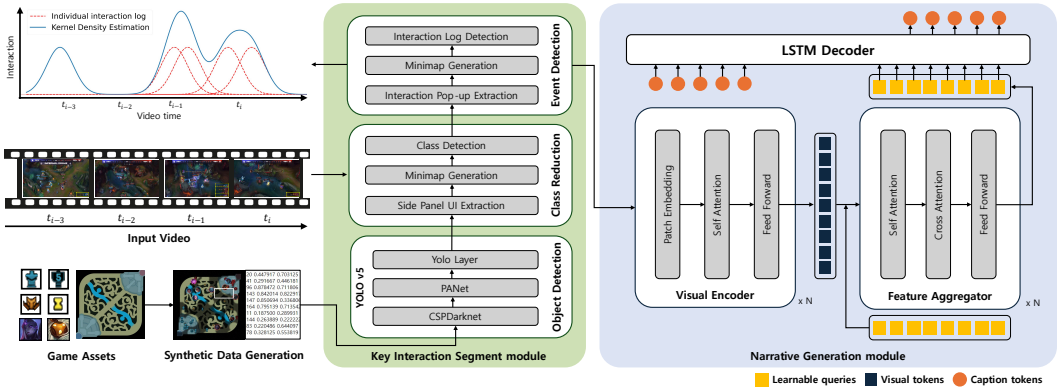


Figure 1: Overview of the proposed IaVNG model

automatically selects interaction segments based on interaction density. The selection must account for the uneven and clustered distribution of interactions in videos. For this reason, we apply Kernel Density Estimation (KDE) [9, 10], which does not assume a fixed distribution, and utilize the resulting segments as KISs. The NG module then produces coherent narratives grounded in these segments. Experimental results demonstrate that the proposed model effectively selects KISs from gaming videos and produces coherent narratives for the generation of short-form content.

2 Methodology

2.1 Dataset

We constructed a large-scale dataset based on League of Legends (LoL), one of the most popular games characterized by a non-linear narrative structure. In LoL, two teams of five players compete to destroy the opposing team's core structures [11]. During a match, in-game events such as neutral object kills (OBJECT_KILL; e.g., Dragon, Baron Nashor) and champion kills (CHAMPION_KILL) frequently occur. We define these as key interaction logs, which both determine match outcomes and shape a complex non-linear narrative through multi-agent interactions. The dataset comprises 952 regular-season match videos from the 2023–2024 Spring and Summer splits of the League of Legends Champions Korea (LCK) and 25,120 timestamped commentary sentences. All videos were collected in 1080p resolution and at 59.94 frames per second, totaling approximately 570 hours. Subtitles were automatically transcribed using the Whisper model [12] and subsequently refined with the OpenAI API. The API removed noise and anonymized entities by converting them into standardized [PLAYER], [TEAM], and [CHAMP] tokens to ensure consistency and accuracy. We present additional comparisons with other datasets in Appendix A.

2.2 Interaction-Aware Video Narrative Generation Model

Our objective is to automatically identify dense interaction segments from non-linear game video content and to generate coherent narratives that effectively capture their context. To this end, we propose a two-stage IaVNG model, consisting of a KIS module and an NG module, as illustrated in Figure 1. Let $\mathcal{V} = \{V_1, V_2, \ldots, V_M\}$ denote a set of League of Legends match videos, where each video V_i comprises a sequence of k frames, represented as $V_i = [v_i^1, v_i^2, \ldots, v_i^k]$. The KIS module processes this input by performing graphical user interface-based log detection $\text{LogDet}(\cdot)$ and dense segment selection $\text{Seg}(\cdot)$, yielding n_i KISs.

$$\{KIS_1^i, KIS_2^i, \ldots, KIS_{n_i}^i\} = \text{Seg}(\text{LogDet}(V_i)), \quad i = 1, 2, \ldots, M \tag{1}$$

Specifically, the $\text{LogDet}(\cdot)$ step extracts interaction logs by classifying icons displayed in the log area of a frame v_i^t in video V_i. These icons, representing champions or neutral objects, appear whenever key interactions such as champion kills or neutral object kills occur. However, in broadcast videos, these icons are often low-resolution and noisy, which reduces detection reliability [13]. To address this, we adopt YOLOv5 extra-large [14] to detect minimap icons, a secondary user interface (UI)

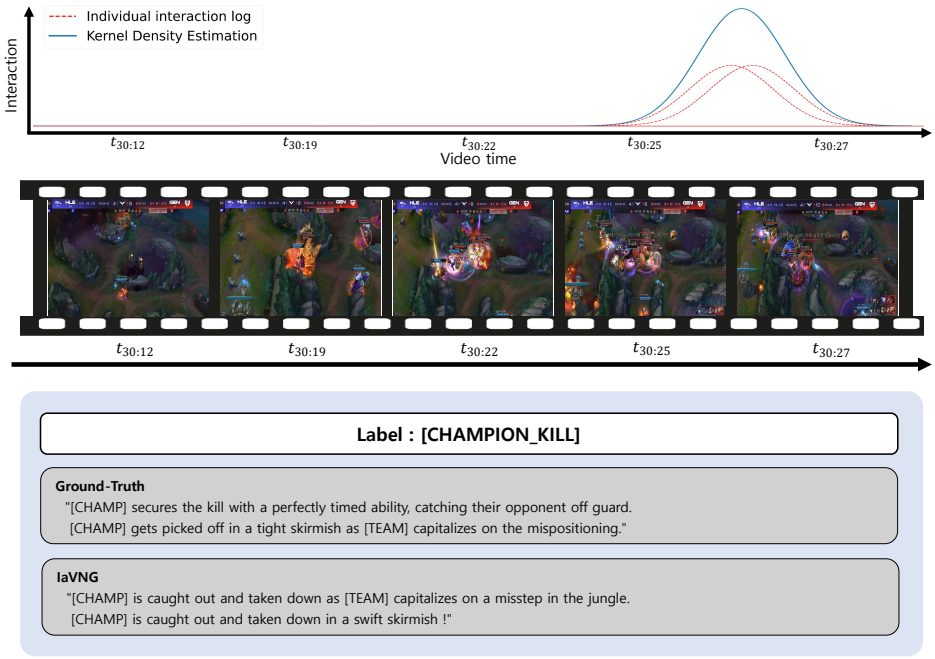


Figure 2: Qualitative results of the IaVNG model

area. The model is trained on synthetic data generated from minimaps and game assets, annotated with icon locations and classes. After training, we map log area icons onto the minimap and classify them, enabling reliable interaction log detection. Implementation details appear in Appendix B.1.1.

In the Seg$(\cdot)$ step, KDE is applied to the interaction logs extracted by LogDet$(\cdot)$, retaining only high-density regions as interaction segments. Formally, for the interaction log set $\{\tau_1^i, \ldots, \tau_{N_i}^i\}$ of V_i, KDE is defined as follows.

$$\hat{f}^{\,i}(x) = \frac{1}{N_i h} \sum_{k=1}^{N_i} K\left(\frac{x - \tau_k^{\,i}}{h}\right), \quad K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \tag{2}$$

To determine the bandwidth h, we first compute the first quartile Q_1^i of intervals between interaction logs for each video V_i. A global representative value Q_1^* is then obtained as the median across all $\{Q_1^i\}_{i=1}^M$, and the bandwidth h is determined by scaling this value. Subsequently, a mean-based threshold is introduced, defined as a constant multiple of the average value of the estimated density $\hat{f}^{\,i}(x)$. This threshold serves as the primary criterion for segment determination. Furthermore, each segment is required to contain a minimum number of interaction logs. Consequently, only the segments satisfying both criteria are retained as key interaction segments $\{KIS_j^i\}_{j=1}^{n_i}$. The detailed parameter settings for Seg$(\cdot)$ are provided in Appendix B.1.2.

The $\{KIS_j^i\}_{j=1}^{n_i}$ selected by the KIS module are passed to the NS module for narrative generation. The NS module consists of three main components: a visual encoder $F(\cdot)$, a feature aggregator $G(\cdot)$, and a text decoder $D(\cdot)$. First, $F(\cdot)$ extracts visual features from frames within each KIS_j^i. To this end, we crop each frame to the central region to exclude peripheral UI elements (e.g., health bars and item icons) that are often irrelevant to key interactions. The cropped frames are transformed into fixed-dimensional embeddings using a pre-trained timm-based model [15, 16, 17], and the embeddings are subsequently passed to $G(\cdot)$. $G(\cdot)$ maps the frame embeddings into the same representation space as the text input, allowing the text decoder to jointly process visual and linguistic information. To achieve this, Transformer decoder layers with learnable query embeddings aggregate visual information through cross-attention with the frame embeddings. $D(\cdot)$ takes the visual features aggregated by $G(\cdot)$ along with text tokens as input and generates word sequences in an autoregressive manner. To capture both long-term and short-term context, a vanilla LSTM-based decoder [18] is

Table 1: Evaluation results of interaction log detection

Interaction Log	TP	FP	FN	Precision	Recall	F1-score
Overall	24,382	828	732	96.71	97.08	96.90
CHAMPION_KILL	19,767	701	607	96.57	97.02	96.79
OBJECT_KILL	4,615	127	125	97.32	97.36	97.34

Table 2: Evaluation results compared with baseline model

Model	Method	BLEU@4 ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑
[7]	VTransformer	3.14	12.03	16.57	-
	MART	**3.56**	12.98	15.39	-
Ours	KIS + LSTM	2.87	**14.31**	**17.45**	13.37

used, thereby producing contextually coherent narratives. More details are specified in Appendix B.2 and C.

3 Experiments

Evaluation method. We evaluate the proposed model in terms of its two main modules. For the KIS module, evaluation is based on interaction log detection metrics including Precision, Recall, and F1-score, to measure how accurately the model identifies interaction logs against ground-truth logs. For the NS module, which generates narratives for the KIS, evaluation is conducted using standard captioning metrics including BLEU-4 [19], METEOR [20], ROUGE-L [21], and CIDEr [22], in comparison with the baseline model. Additionally, qualitative analysis compares generated captions with ground-truth commentary to assess narrative coherence and contextual relevance.

Quantitative results. The performance of interaction log detection is reported in Table 1. Metrics were computed using TP, FP, and FN, while TN was excluded due to the disproportionately large proportion of non-interactions across the videos. The proposed model achieves overall results of 96.71% precision, 97.08% recall, and 96.90% F1-score, demonstrating performance comparable to the ground-truth interaction logs. Detailed results for per-class detection and additional KIS extraction are provided in Appendix D.1 and Appendix D.2, respectively. These results demonstrate that the model can reliably select KISs in non-linear video narratives, providing a solid foundation for short-form content reconstruction. The performance of the NS module is presented in Table 2. Compared with the game captioning baseline of Tanaka et al. [7], the proposed model achieves substantial improvements, reaching 14.31% on METEOR and 17.45% on ROUGE-L. These results demonstrate that an interaction segment-based approach significantly enhances narrative coherence and contextual relevance compared to the scene transition-based baseline.

Qualitative results. The results of a qualitative comparison between the captions generated by the model and the ground-truth captions are shown in Figure 2. Ground-truth describes the situation by emphasizing the mispositioning and ensuing skirmish that the team utilized. The proposed model generates a comparable narrative structure, moving beyond kill detection to produce context-aware descriptions such as “capitalizes on a misstep in the jungle” and “swift skirmish”. These results show that the proposed model effectively captures the meaning and contextual dynamics of interactions within the selected segments. Additional qualitative results are provided in Appendix D.3.

4 Conclusion

In response to the need for a next-generation paradigm in video generation, we presented the IaVNG. Our model reconstructs game videos with non-linear narrative structures into short-form content. IaVNG selects interaction segments through the KIS module and generates contextually coherent narratives with the NG module. Experimental results demonstrated that our model effectively extracts key interactions even from complex non-linear game videos and generates coherent short-form narratives. These findings highlight the potential of IaVNG as a generalizable model for short-form

generation across various non-linear domains. Future work will focus on validating the model's adaptability by further categorizing interaction types and expanding its application to diverse game genres.

Acknowledgments and Disclosure of Funding

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of generative AI-based esports service automation platform technology to improve esports operation efficiency, Project Number: RS-2024-00441523, Contribution Rate: 100%)

References

[1] Aadit Barua, Karim Benharrak, Meng Chen, Mina Huh, and Amy Pavel. Lotus: Creating short videos from long videos with abstractive and extractive summarization. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 967–981, 2025.

[2] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208, 2024.

[3] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024.

[4] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13904, 2024.

[5] Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6006–6015, 2018.

[6] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pages 768–784. Springer, 2016.

[7] Tsunehiko Tanaka and Edgar Simo-Serra. Lol-v2t: Large-scale esports video description dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4557–4566, 2021.

[8] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[9] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[10] Richard A. Davis, Keh-Shin Lii, and Dimitris N. Politis. *Remarks on Some Nonparametric Estimates of a Density Function*, pages 95–100. Springer New York, New York, NY, 2011.

[11] Adrián Mateo-Orcajada, Raquel Vaquero-Cristóbal, and Lucía Abenza-Cano. Performance and heart rate in elite league of legends players. *Multimedia Tools and Applications*, 82(19):30151–30176, 2023.

[12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[13] Seung-Jin Hong and Sang-Kwang Lee. Detecting in-game play event in live esports stream. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1929–1931, 2022.

[14] Glenn Jocher. YOLOv5 by Ultralytics. `https://github.com/ultralytics/yolov5`, 2020. Version 7.0, AGPL-3.0 License.

[15] Ross Wightman. Pytorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[17] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[20] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[22] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[23] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017.

[24] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[25] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641, 2013.

[26] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition*, pages 184–195. Springer, 2014.

[27] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[28] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5803–5812, 2017.

[29] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022.

[30] Chengxi Li, Sagar Gandhi, and Brent Harrison. End-to-end let's play commentary generation using multi-modal video representations. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–7, 2019.

[31] Dae-Wook Kim, Sung-Yun Park, Seong-Il Yang, and Sang-Kwang Lee. Real-time player tracking framework on moba game video through object detection. *IEEE Transactions on Games*, 17(2):498–509, 2025.

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PmLR, 2020.

[34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Interaction-Aware Video Narrative Generation for Short-Form Gaming Content

Supplementary Material

A Dataset Comparison

We present a comparison of the proposed IaVNG dataset with existing benchmarks in Table 3. IaVNG comprises approximately 570 hours of video, representing the largest scale within the video game domain. In terms of overall scale, it is also comparable to large datasets from other domains. Furthermore, IaVNG leverages in-game narration to construct sentence-level captions, providing a valuable resource for language-based research on game videos.

Table 3: **Comparison of existing video-language datasets.** Our IaVNG dataset significantly surpasses existing video game datasets in both the number of videos and total duration. *Narration* indicates whether sentences are derived from spoken narration within the video.

Name	Domain	# Video	Duration (hr)	# Sentences	Narration
ActivityNet-Caption [23]	Open	20k	849	100k	-
Youcook2 [24]	Cooking	2k	176	15.4k	-
TACOS [25] / TACOS-ML [26]	Cooking	127/185	15.9/27.1	18.2k/52.5k	-
Ego4D [27]	Ego	9.6k	3,670	3.85M	-
DiDeMo [28]	Open-huma	10k	88.7	40.5k	-
MAD [29]	Movie	650	1207.3	384.6k	✓
Getting Over It [30]	Video game	8	1.8	2.27k	✓
LoL-V2T [7]	Video game	157	76	63k	✓
IaVNG (Ours)	Video game	952	570	25.12k	✓

B Model and Implementation Details

B.1 KIS Module

B.1.1 LogDet Step

Object Detection. LogDet$(\cdot)$ extends the minimap-based detection model proposed by Kim et al. [31]. The original model was restricted to detecting champion icons on the minimap. This limited its applicability to interaction log analysis, where both champion interactions and neutral objects play decisive roles. To address this limitation, we extend the detection targets to include not only champion icons but also neutral object icons. To support this broader coverage, we doubled the synthetic dataset to 200,000 training and 20,000 test samples and trained YOLOv5. The synthetic data incorporates resolution degradation and icon occlusions, essential for robustness under video noise.

Class Reduction. In practice, the YOLOv5 is trained on all icon classes (169 champions and 8 neutral objects, 177 in total), but each match video V_i contains only 10 champion classes alongside the neutral objects. To address this mismatch and improve detection efficiency, we introduce a Class Reduction step that removes icon classes absent from the match, as shown in Figure 3. Specifically, to determine the actual champions in a given video, we crop side panel UI icons during the first 100 frames $[v_i^1, \ldots, v_i^{100}]$ and map them onto the minimap using Algorithm 1, resulting in a synthetic minimap. The outputs are then fed into the trained YOLOv5 for per-frame predictions. Aggregating predictions across 100 frames yields a stable set of 10 champion classes.

Event Detection. In the video V_i, interaction pop-up logs appear at the bottom-right corner as $(\text{killer}, \text{victim})$ icon pair in Figure 3, with up to four rows simultaneously displayed at fixed locations. After Class Reduction, we crop the icons displayed in the pop-up logs at 2 fps starting from frame 100 and project them onto the minimap using Algorithm 1, generating synthetic inputs. These inputs are then fed into the reduced-class model for classification, producing class predictions. Since pop-up

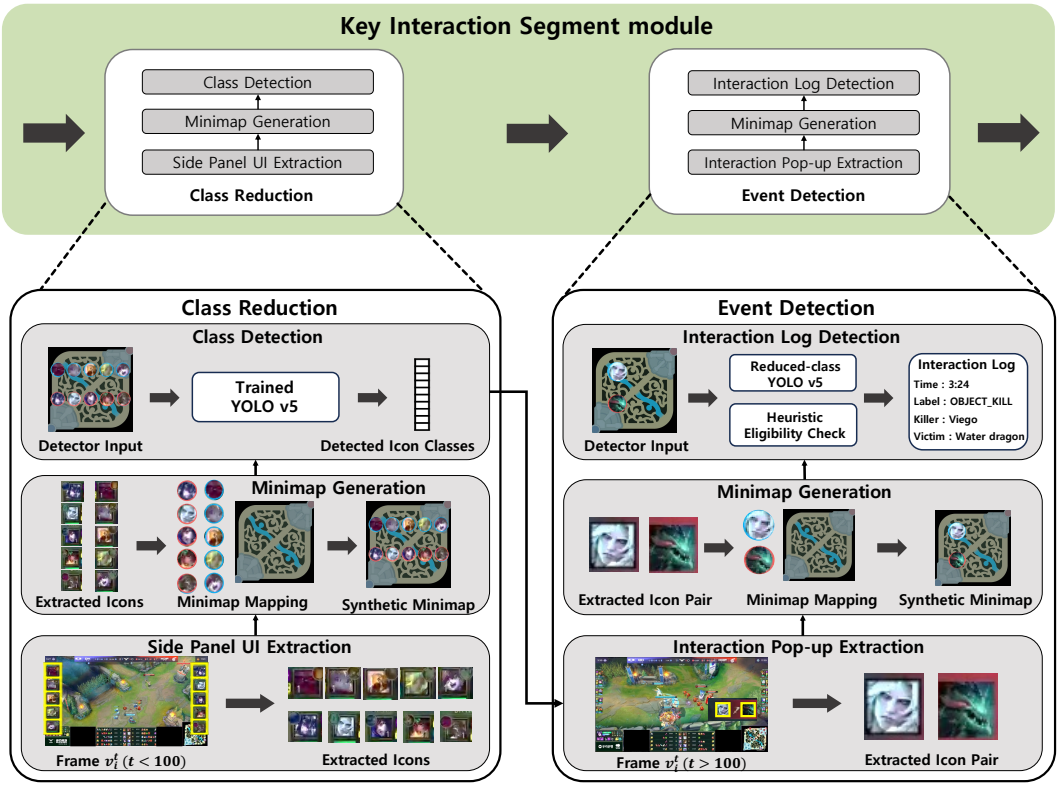


Figure 3: Detailed implementation of Class Reduction and Event Detection

logs may span multiple rows, for efficiency, the third-row pair is examined only if the second-row pair is valid. This restriction is necessary because invalid pairs often arise when cropped regions capture background instead of icons in the absence of interactions. In addition, since pop-up logs persist for several seconds, the same interaction can be redundantly detected. To address this, raw outputs are first filtered into valid pairs V, and only those that satisfy both health and temporal constraints are promoted to eligible pairs I. This design leverages the temporal persistence of pop-up logs to suppress redundant detections while ensuring stable interaction log capture. The complete procedure is summarized in Algorithm 2.

Algorithm 1: Minimap mapping

Input: Icon image r_i
Output: Synthetic minimap M'
$M \leftarrow$ load base minimap from game assets
$r_i \leftarrow$ resize r_i to fixed size and apply circular mask
$(x, y) \leftarrow$ compute placement coordinates from team/slot
Overlay r_i onto M at (x, y) and draw team-colored outline
return M'

B.1.2 Seg step

Parameter Settings. The Seg$(\cdot)$ step was implemented with the following parameter settings:

- **Bandwidth** h: computed as $h = 0.8 \times Q_1^*$, where Q_1^* is the median of first quartiles Q_1^i across all videos. In our experiments, Q_1^* was calculated as 4.0 seconds, yielding $h = 3.2$ seconds. We scaled Q_1^* by 0.8 to obtain a narrower bandwidth, improving sensitivity to dense interaction segments.

- **Density threshold factor** T^i: the threshold was set to $\alpha \cdot \text{mean}(\hat{f}^{\,i}(x))$, with $\alpha = 5.0$. The threshold was set to 5 times the mean, since non-interaction durations vastly outnumber interactions.
- **Minimum interaction count**: at least two interaction logs were required for a segment to be retained. A single interaction within a segment was considered anomalous, as it occurred independently of multi-agent interactions.

Algorithm 2: Interaction pop-up detection

Input: Frame v_i^t, detection model M_{YOLO}, timestamp memory T initialized as $T[c] = -\infty$ for all classes c, threshold $\theta = 0.9$
Output: Detected interaction log I_t at frame t

$I_t \leftarrow \emptyset$ `// Eligible pair`
if $t \bmod 30 \neq 0$ **then**
  return I_t
end
$P \leftarrow$ crop popup icon pairs from v_i^t
$V \leftarrow \emptyset$ `// Valid pair`
for $r = 0$ **to** 3 **do**
  if $r \geq 2$ ***and*** $P[r-1] \notin V$ **then**
    break `// Stop detecting upper rows if lower row is invalid`
  end
  $(k, v) \leftarrow P[r]$ `// (k: killer icon, v: victim icon)`
  $M' \leftarrow$ Minimap mapping(k, v) `// See Algorithm 1`
  $D \leftarrow M_{YOLO}.detect(M')$
  $d_k, d_v \leftarrow$ extract detections for (k, v) from D
  if $d_k.conf > \theta$ ***and*** $d_v.conf > \theta$ **then**
    $V \leftarrow V \cup \{(k, v)\}$ `// conf: confidence`
    if $d_v.class \neq c_{object}$ **then**
      $h_v \leftarrow$ health_point of d_v
      $\Delta t \leftarrow t - T[d_v.class]$
      if $h_v < 0.05$ ***and*** $\Delta t > 10s$ **then**
        $I_t \leftarrow I_t \cup \{(d_k.class, d_v.class)\}$ `// Eligible Pair: CHAMPION_KILL`
        $T[d_v.class] \leftarrow t$
      end
    else
      $\Delta t \leftarrow t - T[d_v.class]$
      if $\Delta t > 30s$ **then**
        $I_t \leftarrow I_t \cup \{(d_k.class, d_v.class)\}$ `// Eligible Pair: OBJECT_KILL`
        $T[d_v.class] \leftarrow t$
      end
    end
  end
end
return I_t

B.2 NG Module

Visual Encoder. We crop each frame to a resolution of 760×760 centered on the original image. The cropped frames are sampled at 1 fps and used as input to the encoder.

Feature Aggregator. We employ a 4-layer transformer decoder with 4 attention heads and 8 learnable queries. To capture temporal dynamics, positional embeddings are added to the video-frame embeddings before cross-attention.

Text Decoder. We adopt a 4-layer LSTM decoder with a hidden size of 512 and 768-dimensional word embeddings. Projected video features from the aggregator initialize the decoder states. The

model then generates word sequences autoregressively, with outputs mapped to the vocabulary space through a linear layer.

C Training Details

We train our model for 30 epochs with a batch size of 32, using the AdamW optimizer [32] with $(\beta_1, \beta_2) = (0.9, 0.999)$. The initial learning rate is 1×10^{-4}, and a cosine decay scheduler [33, 34] is employed to progressively reduce it throughout training. The teacher forcing ratio is set to 0.7. During inference, we adopt beam search with a beam size of 4 and a brevity penalty of 0.6. All experiments are conducted on an NVIDIA RTX A6000 GPU with 48GB memory.

D Additional Experimental Reults

D.1 Per-class Detection Results

In the IaVNG dataset, 149 distinct champion and object icon classes were identified across all match videos. Their per-class detection performance using the YOLOv5-based model is summarized in Table 4. Most classes achieved precision and recall above 95%, resulting in consistently high F1-scores. Such performance was vital for the accurate selection of KIS.

D.2 KIS Extraction Results

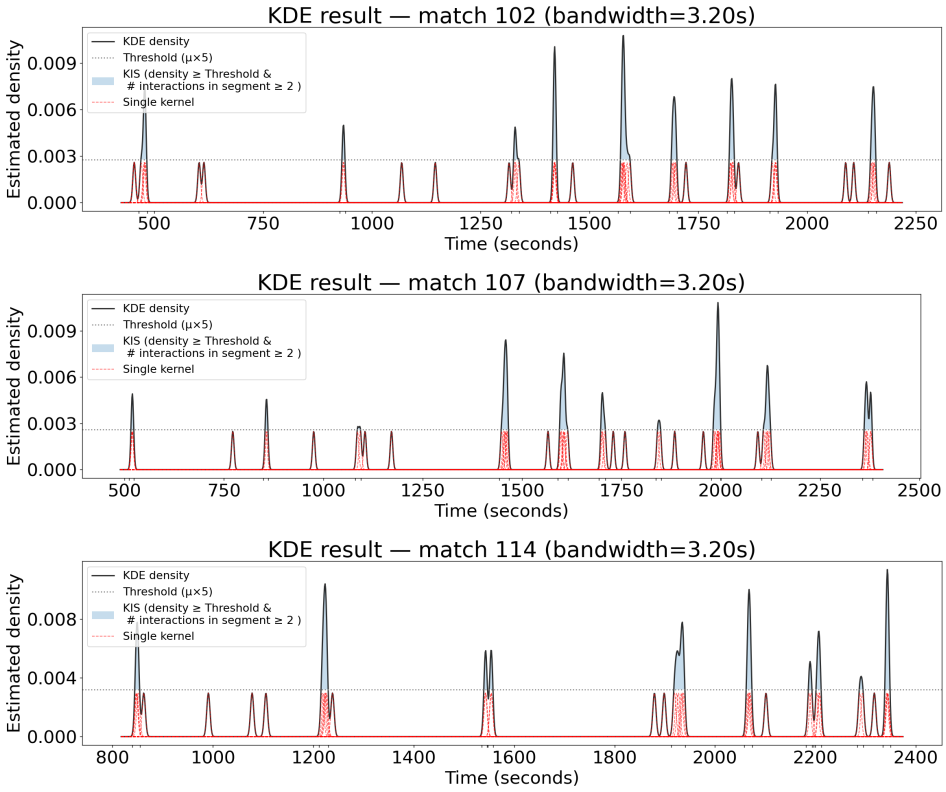
Figure 4 illustrates the extraction of KIS from nonlinearly distributed interactions using KDE with a bandwidth of $h = 3.2$ (Section B.1.2). The density functions are normalized to integrate to one, with the x-axis denoting video time (seconds) and the y-axis showing the estimated interaction density. A mean-based threshold is indicated by a dashed line, and regions exceeding this criterion are highlighted as shaded areas. Panel (a) presents three examples of dense-interaction matches, while panel (b) presents three sparse-interaction cases, demonstrating the variability of interaction distributions across videos. To address this variability, we combine thresholding with an interaction-counting rule: the threshold sharpens dense regions in (a), while the count constraint prevents isolated interactions from being selected in sparse scenarios (b). This integration enables reliable KIS selection across diverse interaction distributions.

D.3 More Qualitative Evaluation

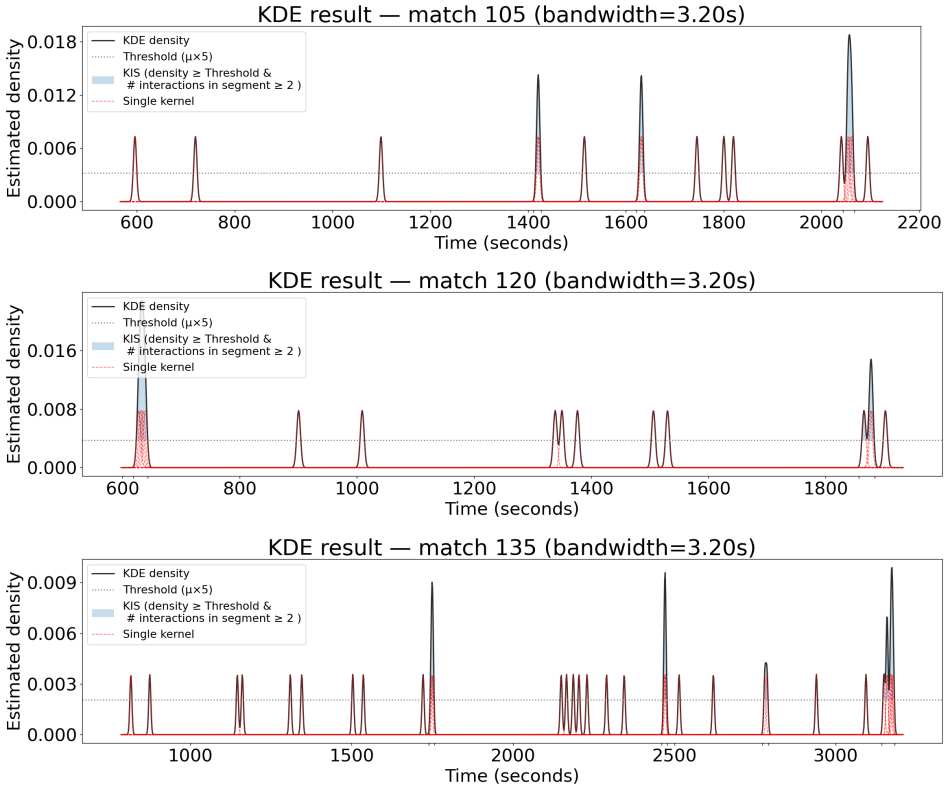
The primary objective of our model is to generate coherent narratives from selected interaction segments in game videos with non-linear narrative structures. Additional qualitative results illustrating this capability are provided in Figure 5. (a) corresponds to a `CHAMPION_KILL` segment. The ground-truth employs tactical terms such as "dive" and "from behind" to convey the complexity of the skirmish, and reinforces the narrative with the phrase "shatters the defense". The proposed model produces a comparable narrative with expressions like "dives into the backline" and "securing a crucial kill", demonstrating its ability to capture decisive actions and structure them into a coherent commentary. (b) illustrates an `OBJECT_KILL` on the dragon followed by a `CHAMPION_KILL`. The ground-truth emphasizes inevitability and sustained dominance through phrases such as "cannot be stopped" and "relentless pressure". The proposed model captures the same sequence, reproducing expressions like "caught out" and "taken down swiftly", while also adding "secures the dragon, gaining a crucial advantage" and "coordinated attack". This demonstrates that IaVNG identifies the key interactions and preserves their sequential flow within a coherent narrative, although it still lacks the expressive detail observed in the ground-truth commentary.

Table 4: Result of per-class detection performance

Name	precision	recall	f1_score	Name	precision	recall	f1_score	Name	precision	recall	f1_score	Name	precision	recall	f1_score	Name	precision	recall	f1_score
Aatrox	97.79	98.74	98.27	Ezreal	99.44	99.72	99.58	Khazix	93.60	96.69	95.12	Poppy	89.41	88.28	88.84	Thresh	100.00	100.00	100.00
Ahri	97.29	97.51	97.40	FiddleSticks	100.00	100.00	100.00	Kindred	93.88	95.83	94.85	Pyke	97.44	100.00	98.70	Tristana	96.66	97.37	97.01
AirDragon	97.77	98.10	97.93	Fiora	96.92	96.92	96.92	KogMaw	100.00	100.00	100.00	Quinn	93.62	95.65	94.62	Trundle	100.00	100.00	100.00
Akali	92.86	93.32	93.09	FireDragon	97.71	98.03	97.87	Leblanc	94.82	94.82	94.82	Rakan	96.90	97.05	96.97	Tryndamere	100.00	100.00	100.00
Alistar	99.06	98.88	98.97	Galio	86.67	92.86	89.66	LeeSin	98.18	98.18	98.18	Rammus	100.00	100.00	100.00	TwistedFate	97.56	96.15	96.85
Amumu	100.00	100.00	100.00	Gangplank	95.76	94.17	94.96	Leona	98.70	98.89	98.80	RekSai	100.00	100.00	100.00	Twitch	94.12	94.12	94.12
Annie	95.08	95.08	95.08	Garen	100.00	100.00	100.00	Lillia	97.79	98.88	98.33	Rell	97.03	97.81	97.42	Udyr	99.39	100.00	99.70
Aphelios	92.82	94.03	93.42	Gnar	97.58	97.77	97.67	Lissandra	98.53	98.53	98.53	Renata	97.99	99.32	98.65	Urgot	100.00	100.00	100.00
Ashe	98.55	98.91	98.73	Gragas	89.39	89.68	89.53	Lucian	97.21	97.21	97.21	Renekton	91.92	93.34	92.62	Varus	97.07	97.76	97.42
AurelionSol	100.00	100.00	100.00	Graves	100.00	100.00	100.00	Lulu	94.84	92.24	93.52	Rumble	97.22	98.20	97.71	Vayne	98.00	98.00	98.00
Aurora	100.00	100.00	100.00	Gwen	96.92	97.67	97.30	Lux	94.25	98.80	96.47	Ryze	94.12	98.46	96.24	Veigar	96.13	96.13	96.13
Azir	96.25	96.71	96.48	Heimerdinger	100.00	100.00	100.00	Malphite	95.45	98.44	96.92	Samira	87.88	87.88	87.88	Vi	98.05	98.29	98.17
Bard	100.00	100.00	100.00	HextechDragon	99.83	99.01	99.42	Maokai	95.55	96.96	96.25	Sejuani	96.98	97.46	97.22	Viego	96.68	96.39	96.54
Baron	96.06	96.34	96.20	Hwei	100.00	100.00	100.00	Milio	99.32	99.32	99.32	Senna	100.00	100.00	100.00	Viktor	93.40	91.67	92.52
Belveth	100.00	100.00	100.00	Illaoi	100.00	100.00	100.00	MissFortune	99.23	98.85	99.04	Seraphine	98.63	100.00	99.31	Vladimir	100.00	100.00	100.00
Blitzcrank	96.67	95.60	96.13	Irelia	89.29	94.34	91.74	MonkeyKing	96.24	97.65	96.94	Shen	100.00	100.00	100.00	Volibear	97.73	100.00	98.85
Brand	100.00	100.00	100.00	Ivern	91.30	94.38	92.82	Mordekaiser	93.75	100.00	96.77	Shyvana	100.00	100.00	100.00	WaterDragon	97.89	97.73	97.81
Braum	95.00	96.07	95.53	JarvanIV	97.37	97.37	97.37	Morgana	100.00	100.00	100.00	Singed	100.00	100.00	100.00	Xayah	96.21	96.81	96.51
Caitlyn	99.27	99.27	99.27	Jax	98.01	98.01	98.01	Naafiri	100.00	100.00	100.00	Sion	93.10	94.74	93.91	Xerath	85.29	85.29	85.29
Camille	100.00	100.00	100.00	Jayce	96.18	97.42	96.80	Nami	96.89	98.03	97.46	Sivir	100.00	97.92	98.95	XinZhao	99.00	99.00	99.00
Cassiopeia	85.00	94.44	89.47	Jhin	98.59	98.59	98.59	Nasus	100.00	100.00	100.00	Skarner	97.72	100.00	98.85	Yasuo	95.00	95.00	95.00
ChemtechDragon	95.95	95.65	95.80	Jinx	96.49	96.07	96.28	Nautilus	96.84	97.08	96.96	Smolder	99.75	100.00	99.87	Yone	99.13	99.13	99.13
Chogath	85.71	75.00	80.00	KSante	95.45	95.99	95.72	Neeko	97.13	96.67	96.90	Sona	90.00	94.74	92.31	Yuumi	90.32	95.45	92.82
Corki	99.28	99.79	99.54	Kaisa	95.65	95.74	95.70	Nidalee	99.48	100.00	99.74	Soraka	96.00	100.00	97.96	Zac	97.22	97.22	97.22
Darius	97.14	97.14	97.14	Kalista	99.03	99.80	99.42	Nilah	93.33	96.55	94.92	Swain	87.50	87.50	87.50	Zeri	92.89	93.14	93.01
Diana	90.00	90.00	90.00	Karma	98.60	97.78	98.19	Nocturne	95.59	95.59	95.59	Sylas	92.15	91.77	91.96	Ziggs	98.66	98.22	98.44
Draven	94.16	93.84	94.00	Karthus	100.00	100.00	100.00	Olaf	100.00	100.00	100.00	Syndra	95.37	98.10	96.71	Zilean	100.00	100.00	100.00
EarthDragon	96.82	97.34	97.08	Kassadin	95.54	97.27	96.40	Orianna	99.59	100.00	99.79	TahmKench	99.27	98.55	98.91	Zoe	80.00	100.00	88.89
ElderDragon	98.11	98.11	98.11	Kayle	100.00	100.00	100.00	Ornn	96.84	97.45	97.14	Taliyah	96.89	97.38	97.14	Zyra	98.63	100.00	99.31
Elise	96.99	98.47	97.73	Kennen	99.30	98.26	98.78	Pantheon	100.00	100.00	100.00	Taric	100.00	100.00	100.00				



(a) KDE for dense interaction videos



(b) KDE for sparse interaction videos

Figure 4: KDE examples for interaction distributions in videos

Label : [CHAMPION_KILL]

Ground-Truth
"To actually break this first mid turret, they need a dive to connect,
and here comes [PLAYER] from behind—dive in! [CHAMP] charges forward and shatters the defense!"

IaVNG
"[PLAYER] from [TEAM] is looking to secure the kill on [CHAMP], and it looks like they 're going for the win.
[PLAYER] dives into the backline with [CHAMP], securing a crucial kill for [TEAM]."

(a) Example of `CHAMPION_KILL` narrative

Label : [OBJECT_KILL], [CHAMPION_KILL]

Ground-Truth
"If [TEAM] doesn't make a mistake, they can secure the dragon, especially since [CHAMP] can't be stopped.
[CHAMP] is caught out and taken down swiftly, unable to escape the relentless pressure from [TEAM]."

IaVNG
"[TEAM] secures the dragon, gaining a crucial advantage in this match.
[CHAMP] is caught out of position and taken down swiftly by [TEAM]'s coordinated attack."

(b) Example of `OBJECT_KILL` and `CHAMPION_KILL` narrative

Figure 5: Additional qualitative results of the IaVNG model

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims in the abstract and introduction are supported by empirical evidence presented in the Experiments section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations in the Appendix D.3.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explained in detail the model architecture and training details in Appendix B, C.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We described the model architecture and training procedure in Appendix B, C. The source code is available at `https://github.com/code-lab78/IaVNG-interaction-log-detection`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Comprehensive implementation details are included in the main text as well as in the Appendix B, C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experiments are quite stable with multiple runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specified in the Appendix C that the training environment was based on an NVIDIA RTX A6000.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our understanding, the paper conforms to the guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our model can enhance content accessibility and production efficiency by automatically generating coherent short-form narratives from complex gaming videos.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretraionoed language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: We removed all players' faces in appendix figures to prevent unintended exposure.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: To the best of our knowledge, all the assets used are properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We introduced our large-scale dataset, game video, in main paper and Appendix A.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer:[NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: We have not used any LLM.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.