Event-Anchored Semantic Augmentation for Video Captioning

1st Langping Wang
College of Computer Science
Chongqing University
Chongqing, China

2nd Bin Fang*
College of Computer Science
Chongqing University
Chongqing, China
Email: fb@cqu.edu.cn

3rd Mengdi Li
College of Computer Science
Chongqing University
Chongqing, China

4th Ningkai Zhong
College of Computer Science
Chongqing University
Chongqing, China

***Abstract*—To address insufficient fine-grained perception and noise interference in video captioning, this paper proposes EAS-Cap (Event-Anchored Semantic Captioning), an event-aware semantic-enhanced model. EAS-Cap decomposes videos into atomic events and aligns visual action features with textual event representations via cross-modal knowledge distillation, enhancing semantic perception. A Visual-Text Alignment (VTA) module suppresses noise and improves coherence by mapping visual features into a textual semantic space with contrastive learning. A Visual-to-Event (V2E) constraint further ensures comprehensive event capture while filtering irrelevant interference. Experiments on MSVD and MSR-VTT show that EAS-Cap outperforms existing methods, achieving CIDEr scores of 120.7 and 58.9 respectively, and generates accurate, semantically rich captions.**

***Index Terms*—video captioning, feature alignment, atomic events, contrastive learning, event-anchored, cross-modal knowledge distillation**

I. Introduction

Video captioning aims to generate descriptive narratives by summarizing visual content, with significant applications in visual retrieval [12], [34]–[36], visual question answering [2], and assistive accessibility [27]. Current approaches fall into two categories: methods that learn holistic video representations often overlook fine-grained details, resulting in ambiguous captions, while those leveraging fine-grained visual elements (e.g., objects and actions) may introduce irrelevant noise that degrades performance [14], [27], [28]. To address these limitations, we propose EAS-Cap, a model designed to enhance fine-grained perception while effectively suppressing visual interference.

As shown in Fig. 1, EAS-Cap employs atomic events—basic semantic units extracted via part-of-speech analysis—as anchors. Using a pre-trained, frozen Phrase-BERT as the teacher, we perform cross-modal knowledge distillation: the Event Encoder transforms visual action features, and the Event Decoder aligns them with textual features using learnable queries. This process transfers prior semantic knowledge to the visual encoder, enhancing its discriminative power. Furthermore, the Visual-Text Alignment (VTA) module enhances global semantics via contrastive learning, and the Visual-to-Event (V2E) constraint ensures local features remain consistent with this global context. This closed-loop interaction between local extraction and global optimization establishes a robust foundation for caption decoding.

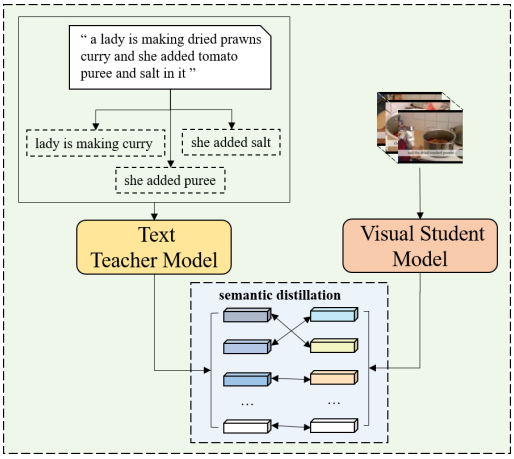

Fig. 1. Diagram of cross-modal distillation mechanism. The process begins by extracting atomic events from the ground-truth caption. Using the text encoder as a teacher model, semantic knowledge is transferred to the visual encoder through feature matching.

EAS-Cap decomposes the task into three objectives: 1) extracting atomic events, 2) distilling textual priors via cross-modal alignment, and 3) enhancing global consistency to suppress noise. Our main contributions are:

- **Cross-modal Knowledge Distillation for Atomic Event Modeling:** We design an Event Module that distills knowledge from text to vision. Learnable queries align fine-grained features with semantic anchors, while the V2E constraint filters out globally inconsistent noise.
- **Global Semantic Enhancement and Denoising:** The VTA Module projects global visual features into a text-aligned space through dimensionality reduction and contrastive learning, enhancing semantics and suppressing noise.
- **Superior Benchmark Performance:** Experiments on MSVD [5] and MSR-VTT [26] demonstrate that EAS-Cap achieves state-of-the-art performance, notably evidenced by leading CIDEr scores.

*: Corresponding author.

II. Related Work

A. Syntax-Based Methods

Syntax-aware methods use syntactic information from ground-truth captions—such as Dependency Parsing (DEP) and Part-of-Speech (POS) tags—to guide attention and improve caption quality. For example, J. Pérez-Martín et al. [7], [14], [17] integrate syntactic tree embeddings with visual features to align grammar with content. B. Wang et al. [24] employ POS sequences as control signals to regulate sentence structure. These approaches enhance grammatical correctness and coherence but often operate at a coarse syntactic level.

B. Spatio-Temporal Feature Modeling

These methods focus on capturing both spatial and temporal information for accurate video descriptions. J. Zhang et al. [1], [19], [28] model hierarchical object–action–scene relationships via semantic clustering. K. Lin et al. [31], [32] adapt contrastive learning to align 3D spatio-temporal features efficiently. Y. Shen et al. [20] propose a lightweight temporal encoder that uses motion vectors and residuals, avoiding heavy optical flow or 3D convolutions. Such works improve temporal awareness but may still miss fine-grained event semantics.

C. Grammar and Knowledge Enhanced Methods

Another direction integrates grammatical rules or external knowledge. Q. Zheng et al. [27], [29] apply syntactic constraints and sentence decomposition, while X. Luo et al. [6], [13] use semantic memory banks or ontologies to enhance representation. These methods show that structured semantic information helps produce coherent captions. In contrast to these approaches, our method employs a cross-modal distillation mechanism that enhances discriminative visual feature extraction through atomic event visual-text alignment and V2E constraint, jointly achieving fine-grained event awareness and noise suppression.

III. Method

As shown in Fig. 2, the workflow of EAS-Cap comprises five key steps. 1) Extract global visual features and encode them to preserve the holistic information of the video. 2) The VTA Module projects these features into a text-aligned semantic space via contrastive learning, thereby enhancing global semantic understanding. 3) The Event Module generates fine-grained embeddings for atomic events based on action features, capturing local details through cross-modal distillation. 4) The V2E constraint guides the local extraction using the aligned global features, ensuring semantic consistency and suppressing noise. 5) The preliminary encodings, VTA-enhanced features, and event features are fused and decoded to produce the final caption. This approach effectively balances fine-grained event perception with global semantic coherence.

A. Visual-Text Alignment Module

As illustrated in Fig. 2, we uniformly sample T keyframes from video v_i and use the CLIP (ViT-B/32) Image Encoder [18] to extract image features from these keyframes, composing a sequence of visual features $\mathbf{B}_i = \{\mathbf{b}_i^n\}_{n=1}^T$ for video v_i, where $\mathbf{b}_i^n \in \mathbb{R}^{d_m}$, and d_m denotes the model dimension. The visual sequence is fed into the VTA Encoder to obtain $\mathbf{G}_i = \{\mathbf{g}_i^n\}_{n=1}^T$, where $\mathbf{g}_i^n \in \mathbb{R}^{d_m}$. Additionally, a mean pooling operation is performed on $\mathbf{G}_i$ to reduce its dimensionality. Therefore, we have $\overline{\mathbf{G}}_i = \mathrm{MeanPooling}(\mathbf{g}_i^1, \mathbf{g}_i^2, \ldots, \mathbf{g}_i^T)$. For the ground-truth caption t_i, we utilize the text encoder of CLIP (ViT-B/32) to obtain its text feature $\mathbf{T}_i \in \mathbb{R}^{d_m}$. Assuming that each batch during training contains N pairs of visual-text features $\{(\overline{\mathbf{G}}_1, \mathbf{T}_1), (\overline{\mathbf{G}}_2, \mathbf{T}_2), \ldots, (\overline{\mathbf{G}}_N, \mathbf{T}_N)\}$, we employ contrastive learning with a symmetric InfoNCE loss function for both video and text features.

$$s_{ij} = \frac{\overline{\mathbf{G}}_i \mathbf{T}_j^\top}{\|\overline{\mathbf{G}}_i\| \, \|\mathbf{T}_j\|}, \tag{1}$$

$$\mathcal{L}_{\overline{\mathbf{G}} \to \mathbf{T}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ij})}, \tag{2}$$

$$\mathcal{L}_{\mathbf{T} \to \overline{\mathbf{G}}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ij})}, \tag{3}$$

$$\mathcal{L}_C = \frac{1}{2} \left(\mathcal{L}_{\overline{\mathbf{G}} \to \mathbf{T}} + \mathcal{L}_{\mathbf{T} \to \overline{\mathbf{G}}} \right), \tag{4}$$

where s_{ij} denotes the cosine similarity between the ith visual feature and the jth text feature, and s_{ii} corresponds to the correctly paired video-text feature. We optimize the total loss $\mathcal{L}_C$ to align visual and text features for global semantic enhancement.

B. Event Module

The Event Module follows an encoder-decoder architecture akin to the Transformer and is designed to perform cross-modal knowledge distillation from text to vision. It incorporates learnable queries to generate atomic event embeddings that align with textual descriptions, thereby capturing fine-grained semantics in videos. Specifically, we first extract verb-based atomic events from the ground-truth caption through syntactic dependency parsing using RoBERTa [11] in spaCy [9]. The textual embeddings of these events, denoted as $\mathbf{H}_i = \{\mathbf{h}_i^n\}_{n=1}^L$ ($\mathbf{h}_i^n \in \mathbb{R}^{d_a}$), are obtained via Phrase-BERT [25], where L is the number of events and d_a denotes the dimension of the phrase embeddings. On the visual side, we employ XCLIP (ViT-B/16) [15] to extract action-aware features from T keyframes of video v_i. These features are fed into the Event Encoder to produce fine-grained visual event features $\mathbf{A}_i = \{\mathbf{a}_i^n\}_{n=1}^T$ ($\mathbf{a}_i^n \in \mathbb{R}^{d_m}$). To ensure semantic consistency between local events and the global video context, we introduce a V2E constraint that leverages the enhanced global features $\mathbf{G}_i$ from the VTA Module. By fusing $\mathbf{G}_i$ with $\mathbf{A}_i$, we form a composite representation $\mathbf{A}_i' = [\mathbf{G}_i; \mathbf{A}_i]$, which allows the model to simultaneously attend to local event

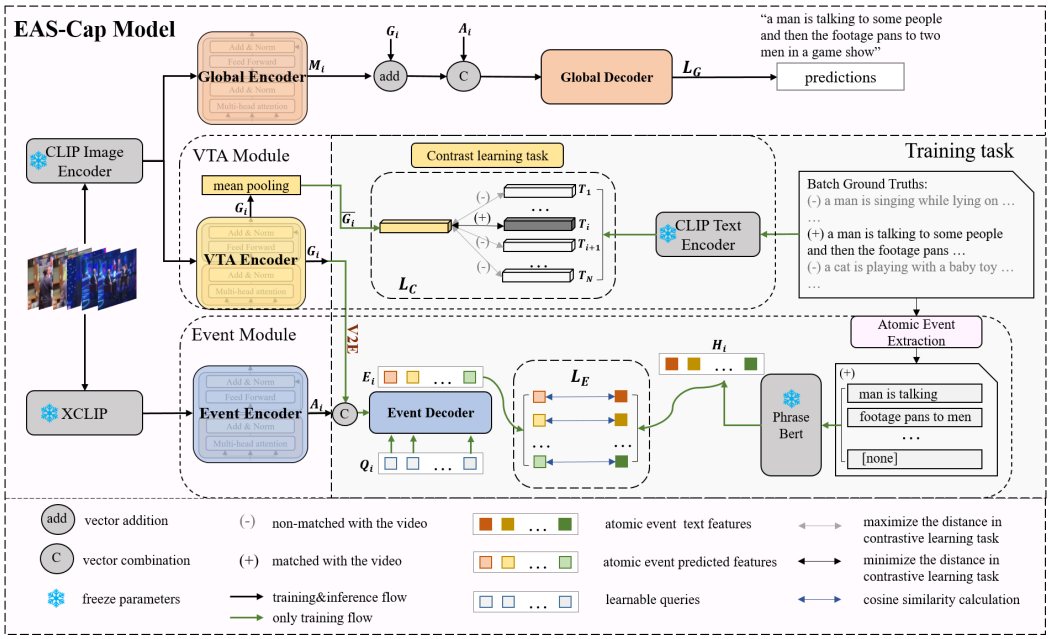

Fig. 2. The overall framework diagram of EAS-Cap. Training task only involves the training process. CLIP, XCLIP, and Phrase-BERT are used with frozen parameters.

details and globally refined semantics. Following the design of DETR [4], a set of learnable queries $\mathbf{Q}_i = \{\mathbf{q}_i^1, \mathbf{q}_i^2, \ldots, \mathbf{q}_i^L\}$ is introduced in the Event Decoder. Through cross-attention with $\mathbf{A}_i'$, these queries generate structured atomic event representations $\mathbf{E}_i = \{\mathbf{e}_i^n\}_{n=1}^L$ ($\mathbf{q}_i^n, \mathbf{e}_i^n \in \mathbb{R}^{d_m}$). Formally, the decoding process is expressed as:

$$\mathbf{E}_i = \text{EventDecoder}\left(\mathbf{Q}_i, \mathbf{A}_i', \mathbf{A}_i'\right), \tag{5}$$

where $\mathbf{Q}_i$ serves as the query and $\mathbf{A}_i'$ acts as the key and value. This design enables the module to distill text-aligned semantic knowledge into the visual encoder, enhancing discriminative feature extraction while filtering out irrelevant noise.

Finally, we compute the cosine similarity loss between $\mathbf{E}_i$ and $\mathbf{H}_i$. However, since $\mathbf{h}_i^n \in \mathbb{R}^{d_a}$ and $\mathbf{e}_i^n \in \mathbb{R}^{d_m}$ have different dimensionalities, we first reduce the dimensionality of $\mathbf{e}_i^n$ to $\mathbb{R}^{d_a}$ via a fully connected layer before calculating the ultimate loss $\mathcal{L}_E$:

$$\mathbf{E}_i' = \text{FC}\left(\mathbf{E}_i\right), \tag{6}$$

$$\mathcal{L}_{\text{E}} = \frac{1}{L}\sum_{j=1}^{L}\left(1 - \frac{\mathbf{e}_{ij}' \cdot \mathbf{h}_{ij}}{\|\mathbf{e}_{ij}'\| \, \|\mathbf{h}_{ij}\|}\right). \tag{7}$$

Guided by the globally enhanced visual features (via the V2E constraint), the Event Module performs semantic distillation: it learns to distill text-aligned atomic event visual representations from the raw action features. The gradients from this distillation task propagate backward to refine the Event Encoder, enabling it to extract more discriminative fine-grained features—effectively filtering out noise while preserving caption-critical semantics. Thus, the module not only aligns local events with text but also progressively purifies the visual representation through this closed-loop, semantics-guided distillation.

C. Caption Generation

The Caption Generator adopts a Transformer-based encoder-decoder architecture, denoted as Global Encoder and Decoder in Fig. 2. Specifically, the Global Encoder encodes $\mathbf{B}_i$ from the CLIP Image Encoder into sequential features $\mathbf{M}_i = \{\mathbf{m}_i^n\}_{n=1}^T$. Then, enhanced global visual features $\mathbf{G}_i$ (from VTA) are added to $\mathbf{M}_i$. Finally, we concatenate the sum of $\mathbf{G}_i$ and $\mathbf{M}_i$ with the fine-grained visual features $\mathbf{E}_i$ from the Event Module to form the final multimodal features $\mathbf{K}_i = \{\mathbf{k}_i^n\}_{n=1}^T$ as:

$$\mathbf{K}_i = \left[\mathbf{M}_i + \mathbf{G}_i \,;\, \mathbf{E}_i\right], \tag{8}$$

where $\mathbf{m}_i^n, \mathbf{k}_i^n \in \mathbb{R}^{d_m}$. In the Global Decoder, $\mathbf{K}_i$ guides the decoding process to produce distributions over the vocabulary for the final captions, resulting in a standard sequence-to-sequence loss $\mathcal{L}_G$ during training (i.e., the cross-entropy loss).

D. Training

The final loss function of our model integrates these three losses to jointly control the iterative optimization process.

TABLE I

PERFORMANCE COMPARISON ON MSR-VTT AND MSVD DATASETS. "GLOBAL": GLOBAL VISUAL FEATURE EXTRACTED BY BACKBONE; "M": MOTION FEATURE; "O": OBJECT FEATURE. R200, R151, AND R101 STAND FOR RESNET-200, RESNET-151, AND RESNET-101 [8], AND IRV2 REPRESENTS INCEPTION-RESNET-V2 [21]. "†" INDICATES FROZEN PARAMETERS WITHOUT FINE-TUNING. COCAP USES CLIP (VIT-B/16), WHILE SNCL AND EAS-CAP USE CLIP (VIT-B/32). BOLD VALUES REPRESENT THE BEST PERFORMANCE, UNDERLINED VALUES INDICATE THE SECOND-BEST.

Model	Year	Features			MSR-VTT				MSVD			
		Global	M	O	B@4	M	R	C	B@4	M	R	C
OABTG [28]	2019	R200	–	✓	41.4	28.2	–	46.9	56.9	36.2	–	90.6
GRU-EVE [1]	2019	IRv2	✓	✓	38.3	28.4	60.7	48.1	47.9	35.0	71.5	78.1
SAAT [29]	2020	IRv2	✓	✓	40.5	28.2	60.9	49.1	46.5	33.5	69.4	81.0
SemSynAN [17]	2021	R152	✓	–	<u>46.4</u>	30.4	<u>64.7</u>	51.9	**64.4**	<u>41.9</u>	<u>79.5</u>	111.5
SGN [19]	2021	R101	✓	–	40.8	28.3	60.8	49.5	52.8	36.5	72.9	94.3
SHAN [7]	2021	IRv2	✓	–	40.3	28.8	61.2	54.1	50.9	35.1	72.4	94.5
HMN [27]	2022	IRv2	✓	✓	43.5	29.0	62.7	51.5	59.2	37.7	75.1	104.0
SEM-POS [14]	2023	IRv2	✓	✓	45.2	<u>30.7</u>	64.1	53.1	60.1	38.5	76.0	108.3
GSEN [13]	2024	IRv2	✓	✓	42.9	28.4	61.7	51.0	58.8	37.6	75.2	102.5
VCRN [30]	2025	R101	✓	–	41.5	28.1	61.2	50.2	59.1	37.4	74.6	100.8
CoCap [20]	2023	$\text{CLIP}_{\text{ViT-B/16}}$	–	–	43.1	29.8	62.7	56.2	55.9	39.9	76.8	113.0
SNCL [6]	2024	$\text{CLIP}_{\text{ViT-B/32}}$	–	–	43.8	29.6	63.6	54.6	59.7	38.5	77.2	106.3
EAS-Cap	2025	$\text{CLIP}^{\dagger}_{\text{ViT-B/32}}$	–	–	45.5	29.6	62.9	<u>57.1</u>	61.0	40.5	77.3	<u>114.1</u>
EAS-Cap	2025	$\text{CLIP}^{\dagger}_{\text{ViT-B/32}}$	✓	–	**46.8**	**30.7**	**64.8**	**58.9**	<u>63.3</u>	**42.0**	**79.9**	**120.7**

We do not assign hyperparameters or weights to each of them, allowing the model to effectively leverage the distinct contributions of each loss function without additional tuning:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_E + \mathcal{L}_G. \tag{9}$$

IV. EXPERIMENTAL RESULTS

A. Dataset and Evaluation Metrics

Datasets. We evaluate on MSR-VTT [26] and MSVD [5]. MSR-VTT contains 10,000 video clips, each with 20 human-written descriptions. Its diverse content and varying lengths challenge model generalization. MSVD has 1,970 short clips with about 40 descriptions each, featuring concise but diverse captions focusing on different details.

Evaluation Metrics. We adopt standard video captioning metrics: BLEU@4 [16], METEOR [3], ROUGE-L [10], and CIDEr [23]. Among these, the CIDEr metric has been shown to align more closely with human judgment, making it a reliable indicator of caption quality.

B. Details

During training, we use batch size 256, dropout rate 0.1, and learning rate 1×10^{-4}. Captions are truncated to 20 words maximum. For reliability, we conduct experiments with five random seeds (1, 2, 3, 4, 5) and report average performance. We sample 20 keyframes for MSR-VTT and 16 for MSVD. Our model uses 2-layer Transformers with 8 attention heads and 512 dimension size. The system is implemented in PyTorch on a single RTX-4090 GPU.

C. Comparison with Existing Methods

To verify the effectiveness of our model, we evaluate the proposed method against state-of-the-art approaches on the MSR-VTT and MSVD datasets. As shown in Table I, our method achieves a comprehensive lead on the MSVD benchmark and obtains the best results on three out of four metrics on the MSR-VTT benchmark. Notably, EAS-Cap does not fine-tune CLIP and still delivers excellent performance even without using motion features, which fully demonstrates the effectiveness of the model itself. Especially on the more challenging MSR-VTT dataset, we achieve a significant lead in CIDEr score (with an improvement of +1.18), a metric known to align most closely with human judgment and thus serves as a strong indicator of caption quality.

TABLE II

ABLATION STUDIES ON MSR-VTT AND MSVD DATASETS. "E": EVENT MODULE; "V": VTA MODULE; "W/O": WITHOUT; "CUT V2E": REMOVE THE VISUAL-TO-EVENT (V2E) PATH.

Model	MSR-VTT				MSVD			
	B@4	M	R	C	B@4	M	R	C
Baseline	44.2	28.9	61.2	55.3	58.9	39.4	76.1	107.2
w/o E	45.5	29.6	62.9	57.1	61.0	40.5	77.3	114.1
w/o V	46.6	30.0	63.6	56.9	61.6	41.3	78.2	117.6
Cut V2E	46.3	30.2	63.8	58.2	63.0	41.8	78.9	118.7
Full model	**46.8**	**30.7**	**64.8**	**58.9**	**63.3**	**42.0**	**79.9**	**120.7**

D. Ablation Study

To further validate the effectiveness of each individual module, we conducted ablation studies on the MSR-VTT and MSVD datasets. Our baseline model comprises only the Global Encoder and Global Decoder. As shown in Table II, compared to the baseline, adding each component leads to significant improvements across all four evaluation metrics, clearly demonstrating their positive contribution.

EAS-Cap extracts atomic events through syntactic analysis and then aligns these events with action features under the constraint of enhanced global visual representations, thereby

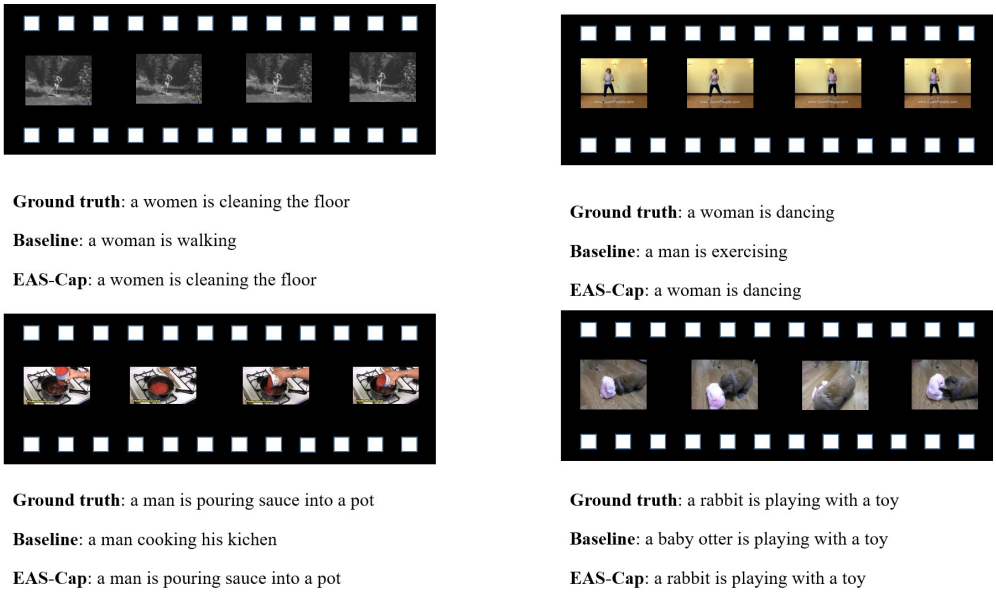

Fig. 3. Quantitative results of model performance.

TABLE III
COMPARISON OF DIFFERENT METHODS FOR APPLYING ACTION FEATURES ON MSR-VTT AND MSVD DATASETS.

	MSR-VTT				MSVD			
Method	B@4	M	R	C	B@4	M	R	C
Predicate	46.3	30.3	63.6	57.2	62.1	41.1	78.0	118.0
Aux+Verb	45.8	30.3	63.4	57.6	61.5	40.7	77.6	116.0
EAS-Cap	**46.8**	**30.7**	**64.8**	**58.9**	**63.3**	**42.0**	**79.9**	**120.7**

capturing fine-grained semantic information in videos. We note that prior work has proposed related yet distinct approaches for utilizing action features. For instance, HMN [27] employs action features to predict predicate components in ground-truth captions, aiming to guide the model toward high-level semantic information in sentences; whereas SEM-POS [14] uses action features to predict part-of-speech tags (e.g., auxiliary verbs, verbs) in ground-truth captions, helping the model focus on key elements within the caption's semantic structure. As shown in Table III, to ensure experimental rigor, we conduct comparative experiments within our Event Module using these two alternative methods, evaluating whether our proposed approach aligns better with the overall framework and task objectives.

Based solely on the evaluation scores from the experimental results, our method is better suited to our work context and can more effectively enhance the model's ability to perceive high-level semantic information.

E. Quantitative Results

As shown in Fig. 3, to provide an intuitive demonstration of our final results, we select several samples from the actual test set to illustrate the generation performance of the EAS-Cap model. The results show that EAS-Cap is able to generate descriptive sentences that closely match the ground-truth captions, regardless of whether the video background is simple or complex, and whether the content depicts real-life scenes or movie clips. By focusing on atomic events, our method not only effectively enhances the model's focus on fine-grained information in the generated sentences, but also helps filter out redundant visual features during the vision-to-text alignment process, thereby improving the overall quality of the generated captions.

V. CONCLUSION

This paper introduces EAS-Cap, an event-aware semantic-enhanced model for video captioning. The model decomposes videos into atomic events and performs cross-modal knowledge distillation within the Event Module, aligning visual event features with textual descriptions to capture fine-grained semantics. The VTA Module enhances global representations through contrastive learning, while the V2E constraint ensures consistency between locally distilled events and the global video context. Experiments on MSVD and MSR-VTT show that EAS-Cap outperforms existing methods and achieves state-of-the-art CIDEr scores without fine-tuning CLIP. Ablation studies confirm the contribution of each component.

Future work will extend the framework to dense captioning and multi-modal settings.

VI. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China 62376042, the Natural Science Foundation of Chongqing CSTB2024NSCQ-KJFZZDX0036, the Key Special Project for Technological Innovation and Application Development in Chongqing CSTB2023TIAD-KPX0050 & CSTB2022TIAD-KPX0176.

REFERENCES

[1] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12479–12488.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.

[3] S. Banerjee and A. 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*, 2005, pp. 65–72.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision - ECCV 2020*, 2020, pp. 213-229.

[5] D. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 190–200.

[6] K. Chen, Q. Di, Y. Lu, and H. Wang, "Semantic-guided network with contrastive learning for video caption," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 3830–3884.

[7] J. Deng, L. Li, B. Zhang, S. Wang, Z. Zha, and Q. Huang, "Syntax-guided hierarchical attention network for video captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 880–892, Feb. 2022.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," Zenodo, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3945274

[10] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[11] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12] H. Luo et al., "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval," *Neurocomputing*, vol. 508, pp. 293-304, 2021.

[13] X. Luo, X. Luo, D. Wang, J. Liu, B. Wan, and L. Zhao, "Global semantic enhancement network for video captioning," *Pattern Recognit.*, vol. 145, p. 109906, 2023.

[14] A. Nadeem, A. Hilton, R. Dawes, G. Thomas, and A. Mustafa, "SEM-POS: Grammatically and semantically correct video captioning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 2606–2616.

[15] B. Ni et al., "Expanding language-image pretrained models for general video recognition," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 1–18.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[17] J. Pérez-Martín, B. Bustos, and J. Pérez, "Improving video captioning with temporal composition of a visual-syntactic embedding," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 3039–3049.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[19] H. Ryu, S. Kang, H. Kang, and C. D. Yoo, "Semantic grouping network for video captioning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 2514–2522.

[20] Y. Shen, X. Gu, K. Xu, H. Fan, L. Wen, and L. Zhang, "Accurate and fast compressed video captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 15558–15567.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 31, no. 1, 2017.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.

[23] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575.

[24] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, "Controllable video captioning with POS sequence guidance based on gated fusion network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2641–2650.

[25] S. Wang, L. Thompson, and M. Iyyer, "Phrase-BERT: Improved phrase embeddings from BERT with an application to corpus exploration," *arXiv preprint arXiv:2109.06304*, 2021.

[26] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 5288–5296.

[27] H. Ye, G. Li, Y. Qi, S. Wang, Q. Huang, and M.-H. Yang, "Hierarchical modular network for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17939–17948.

[28] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8327–8336.

[29] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 13096–13105.

[30] P. Zeng, H. Zhang, L. Gao, X. Li, J. Qian, and H. T. Shen, "Visual commonsense-aware representation network for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 1092–1103, Jan. 2025.

[31] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "SwinBERT: End-to-end transformers with sparse attention for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17949–17958.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[33] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, "Memory-attended recurrent network for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8347–8356.

[34] R. Hong, Y. Yang, M. Wang, and X.-S. Hua, "Learning visual semantic relationships for efficient visual retrieval," *IEEE Trans. Big Data*, vol. 1, no. 4, pp. 152–161, Oct.–Dec. 2015.

[35] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3652–3664, Aug. 2017.

[36] J. Song, L. Gao, L. Liu, X. Zhu, and N. Sebe, "Quantization-based hashing: a general framework for scalable image and video retrieval," *Pattern Recognit.*, vol. 75, pp. 175–187, Mar. 2018.