Cross-Modal Prior Knowledge Guided Dual-Path 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*—Video captioning aims to generate accurate textual descriptions from videos, bridging visual and linguistic understanding. While existing methods have advanced multimodal representation learning during training, they often lack explicit cross-modal semantic guidance during inference, leading to inaccurate or generic descriptions in complex scenes. To address this, we propose CMG-DP, a novel framework that introduces textual prior knowledge for real-time guidance at inference. CMG-DP consists of three key components: a Prior Knowledge Synergy mechanism to retrieve relevant textual descriptions, a Semantic Enhancement module to refine visual semantics, and a Dual-Path Cross-Attention decoder for adaptive fusion. Experiments on MSR-VTT and MSVD benchmarks show that CMG-DP achieves state-of-the-art performance, with CIDEr scores of 58.3 and 124.6, respectively, demonstrating its effectiveness in enhancing caption accuracy and robustness.**

***Index Terms*—video captioning, cross-modal guidance, prior knowledge retrieval, semantic enhancement, dual-path attention**

I. Introduction

Video captioning, the task of generating accurate and coherent textual descriptions from video content, serves as a vital bridge between visual and linguistic understanding. Its applications span a wide range of real-world domains, including video summarization, content-based retrieval, assistive technologies for the visually impaired, and human-computer interaction systems [2], [6], [15], [22], [23].

Recent advances in this field have largely focused on enhancing the model's perception of fine-grained visual information. Predominant approaches include leveraging multi-source features from pre-trained models [10], [14], [26], [29], fine-tuning large-scale vision-language models on extensive corpora [4], [11], [19], and employing contrastive learning (e.g., CLIP [17]) to align visual and textual features in a shared semantic space [13], [24], [27].

While these strategies have significantly improved multimodal representation learning during training, a pivotal yet often overlooked limitation persists at inference: models operate within a visual-only paradigm, lacking a mechanism to incorporate external linguistic knowledge in real time. Consequently, in complex, ambiguous, or out-of-distribution scenes, they often produce inaccurate, generic, or hallucinated descriptions. The root cause lies in the inference mechanism itself: although trained on aligned visual-textual representations, the model must retrieve these alignments solely from visual input at test time—a process that fails when visual evidence is weak or ambiguous. To overcome this, we posit that inference should be augmented with explicit textual priors, retrieved in real time to provide contextualized language patterns that complement visual perception. This serves as a dynamic semantic prompt, akin to a human describer consulting relevant references before writing, guiding generation toward more accurate and specific captions.

To address this critical deficit in inference-stage semantic guidance, we propose a novel framework named CMG-DP (Cross-Modal prior knowledge Guided Dual-Path). The core innovation of CMG-DP lies in its dedicated mechanism to explicitly retrieve and leverage cross-modal (textual) prior knowledge during inference, thereby providing the model with crucial semantic cues that are absent in traditional visual-only decoding. This is achieved through three synergistically designed components that form a coherent pipeline: retrieval, refinement, and fusion.

First, we introduce a Prior Knowledge Synergy (PKS) mechanism. It constructs a textual corpus from training captions and, for an input video, retrieves the top-K semantically similar textual descriptions in real-time. This directly injects relevant linguistic priors into the decoding process, offering a "semantic blueprint" to guide caption generation, especially for challenging or novel scenes. However, naively injecting external text risks introducing semantic noise—descriptions that are globally similar but contain entity or action mismatches with the specific input video.

Therefore, secondly, we design a Semantic Enhancement (SE) Module to sharpen the model's perception of the actual visual content and counteract potential noise from PKS. Inspired by masked language modeling [7], the SE Module is trained to reconstruct masked entity words in ground-truth captions using only the visual input. This forces the encoder within SE to distill and reinforce fine-grained, video-specific semantic representations, enhancing its ability to ground descriptions in observed visual evidence rather than spurious textual associations.

*: Corresponding author.

Finally, having established two complementary information streams—the retrieved textual guidance from PKS and the enhanced, grounded visual semantics from SE—the challenge becomes their effective integration. Simple concatenation or addition fails to model their complex, dynamic interactions. To this end, we propose a Dual-Path Cross-Attention (DPCA) Decoder. It processes the two streams in parallel via dedicated cross-attention pathways, allowing the model to dynamically query, weigh, and fuse visual details with external semantic context. This structured fusion ensures that the generative process is both anchored in the visual reality and fluently guided by relevant linguistic knowledge.

In summary, CMG-DP tackles the inference-stage guidance problem through a logically sequenced approach: 1) Retrieve relevant cross-modal knowledge (PKS), 2) Refine and ground the model's own visual understanding to resist noise (SE), and 3) Fuse both streams intelligently for balanced and accurate generation (DPCA). Our main contributions are threefold:

- We introduce, for the first time, a paradigm that provides explicit cross-modal (textual) prior knowledge guidance during the inference stage for video captioning, fundamentally overcoming the limitation of visual-only decoding in complex scenes.
- Through the synergistic interplay of the PKS mechanism, the SE Module, and the DPCA Decoder, the model can effectively leverage cross-modal prior knowledge for guidance while filtering out irrelevant semantic noise, thereby enhancing its ability to capture fine-grained semantics.
- CMG-DP achieves outstanding performance on the MSR-VTT [25] and MSVD [3] benchmarks, with CIDEr scores of 58.3 and 124.6 respectively, setting a new state of the art and validating the effectiveness of our framework.

II. Related Work

Video captioning research has evolved along several major trajectories, all aiming to bridge the semantic gap between visual content and linguistic description. These works can be broadly categorized into the following strands, each with its own strengths yet sharing a common limitation regarding inference-stage guidance.

Traditional Encoder–Decoder Frameworks. Early approaches adopted encoder–decoder architectures, using CNNs for visual encoding and RNNs/LSTMs for sequential decoding [8]. Subsequent advancements incorporated attention mechanisms [8] and Transformer-based models [11], [19] to better align spatial-temporal features with text generation. While these methods improved the basic architecture, they primarily operate on frame-level visual features during inference, lacking mechanisms to integrate high-level semantic knowledge beyond what is embedded in the visual encoder.

Object-Centric and Hierarchical Modeling. Another line of work treats objects as fundamental units, modeling their dynamics and interactions through temporal graphs [29] or hierarchical networks (e.g., HMN [26], RL-HMN [10]). These approaches enhance the model's ability to focus on salient visual entities and filter clutter. However, during inference, they remain visually grounded—relying solely on detected objects and their relationships—without access to external linguistic knowledge that could resolve ambiguity or guide description in unseen scenarios.

Semantic Enhancement and Alignment. To reduce the modality gap, many studies focus on semantic alignment during training. This includes global alignment via auxiliary losses [12], [14], fine-grained approaches like SAAT [30] that use syntactic cues, and more recently, contrastive learning with models like CLIP [17] to project visual and textual features into a shared space [13], [24], [27]. These strategies effectively suppress irrelevant visual information during training. Nevertheless, the semantic guidance they provide is implicit and frozen after training; they offer no explicit, real-time textual cues to the decoder during inference, especially when facing novel or complex visual inputs.

Summary and Our Positioning. In summary, prior work has made significant progress in architectural design, object-aware reasoning, and cross-modal alignment. However, a fundamental and common constraint persists: at the critical inference stage, these models operate in a visually-isolated paradigm. They lack a dedicated mechanism to explicitly retrieve and leverage external, language-formatted knowledge to guide the generation process in real-time. This often leads to suboptimal performance in complex, ambiguous, or out-of-distribution scenes where visual cues alone are insufficient.

Our proposed CMG-DP framework directly addresses this inference-stage semantic guidance deficit. Unlike previous methods that align modalities only during training, CMG-DP introduces a dedicated PKS mechanism that actively retrieves and infuses relevant textual knowledge during inference. This core innovation—complemented by SE Module for visual grounding and DPCA Decoder for adaptive fusion—enables explicit cross-modal guidance precisely when it is most needed, bridging a gap that prior architectures have largely overlooked.

III. Method

The CMG-DP framework is designed to enhance video captioning by integrating retrieved textual prior knowledge with fine-grained visual semantic enhancement within a dual-path decoding architecture. As illustrated in Fig. 1, the pipeline consists of three stages:

(1) **Feature Encoding and Prior Knowledge Retrieval**: The input video is encoded by a visual backbone to obtain global visual features. Simultaneously, a textual corpus built from training captions provides a source of prior knowledge. For each video, the Prior Knowledge Synergy (PKS) mechanism retrieves the top-k semantically relevant textual features from the corpus based on cosine similarity.

(2) **Fine-Grained Semantic Enhancement**: The global visual features are fed into SE Module, which is trained via a masked language modeling (MLM) objective to reconstruct masked entity words using only visual input. This encourages

the model to capture detailed video-specific semantics and resist potential noise from retrieved text.

(3) **Dual-Path Fusion and Decoding**: The enhanced visual features and the retrieved textual priors are processed in parallel by DPCA Decoder, which dynamically integrates visual details with external semantic guidance to generate accurate and fluent captions.

A. Semantic Enhancement Module

Given a video uniformly sampled into T frames, we extract global visual features $\mathcal{G} = \{g_i\}_{i=1}^{T}$ with $g_i \in \mathbb{R}^{d_m}$. To enhance fine-grained semantic perception, we employ a visual-conditioned MLM task. First, a dedicated SE Encoder extracts fine-grained features $\mathcal{S} = \{s_i\}_{i=1}^{T}$ from $\mathcal{G}$. These are combined with $\mathcal{G}$ to form the enhanced visual features $\mathcal{G}'$ via feature addition (corresponding to "V2S" in the Fig. 1):

$$\mathcal{G}' = \mathcal{G} + \mathcal{S}. \tag{1}$$

During training, we mask entity words in a ground-truth caption sequence $X = \{x_1, \ldots, x_L\}$ to obtain X', which is encoded into queries $\mathcal{Q} = \{q_i\}_{i=1}^{T}$. The SE Decoder then reconstructs the masked tokens using $\mathcal{G}'$ as key and value:

$$\mathcal{H} = \text{SE\_Decoder}(\mathcal{Q}, \mathcal{G}', \mathcal{G}'), \tag{2}$$

where $\mathcal{H} = \{h_i\}_{i=1}^{L}$ and $h_i \in \mathbb{R}^{d_m}$. A causal mask is disabled to allow full-context access. Each h_i is projected to vocabulary space via a linear layer followed by softmax, yielding token probabilities p_i. The MLM loss is computed only over masked positions:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|L|} \sum_{i \in L} \log p_i(x_i). \tag{3}$$

During inference, only the SE Encoder is used to produce $\mathcal{G}'$.

B. Prior Knowledge Synergy Mechanism

To provide explicit semantic guidance during inference, we construct a textual prior corpus from training captions. For N video-text pairs $\{(v_i, \{t_j^{(i)}\}_{j=1}^{M_i})\}_{i=1}^{N}$, each caption $t_j^{(i)}$ is encoded as:

$$\mathcal{T}_{i,j} = \text{Text\_Encoder}(t_j^{(i)}) \in \mathbb{R}^{d_m}. \tag{4}$$

The retrieval corpus $\mathcal{C}$ excludes the current sample's own captions to prevent leakage:

$$\mathcal{C} = \{\mathcal{T}_{i',j'} \mid (i', j') \neq (i, j),\ \forall i', j'\}. \tag{5}$$

For a video feature g_i, we retrieve the top-k most similar textual features via cosine similarity:

$$\mathcal{R} = \text{TopK}\left(\{(\mathcal{T}_{i',j'}, \text{sim}(g_i, \mathcal{T}_{i',j'}))\}_{\mathcal{T}_{i',j'} \in \mathcal{C}}\right), \tag{6}$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$. The retrieved set is $\mathcal{R} = \{r^1, \ldots, r^k\}$ with $r^m \in \mathbb{R}^{d_m}$.

C. Dual-Path Cross-Attention Decoder

The DPCA decoder fuses the enhanced visual features $\mathcal{G}'$ and retrieved textual features $\mathcal{R}$ via dual cross-attention pathways. The input token sequence X is embedded and processed by self-attention to obtain queries $\mathcal{Q}' \in \mathbb{R}^{T \times d_m}$. Two parallel attention blocks compute:

$$\mathcal{F}_1 = \text{softmax}\left(\frac{Q_l \mathcal{R}^\top}{\sqrt{d_k}}\right)\mathcal{R}, \tag{7}$$

$$\mathcal{F}_2 = \text{softmax}\left(\frac{Q_l \mathcal{G}'^\top}{\sqrt{d_k}}\right)\mathcal{G}', \tag{8}$$

where Q_l is a linear projection of $\mathcal{Q}'$. The outputs are summed:

$$\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2. \tag{9}$$

Finally, $\mathcal{F}$ passes through a feed-forward network and a classification head to produce token probabilities $P' = \{p_i'\}_{i=1}^{L}$. The cross-entropy captioning loss is:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{L} \log p_i'(y_i), \tag{10}$$

and the total training loss combines both objectives:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{MLM}}. \tag{11}$$

IV. Experimental Results

A. Experimental Details

We implement CMG-DP using PyTorch and conduct all experiments on a single NVIDIA RTX-4090 GPU. During training, we set the batch size to 256, the learning rate to 1×10^{-4}, and the dropout rate to 0.2. Both the SE Encoder, SE Decoder, and DPCA Decoder consist of 2 layers, with a model dimension $d_m = 512$ and 8 attention heads. Each textual description is truncated to at most 40 words. To ensure statistical stability, we run each experiment under five different random seeds and report the average performance. For the MSR-VTT dataset, we uniformly sample 30 keyframes per video; for MSVD, we sample 20 keyframes. All visual and textual features are extracted using CLIP (ViT-B/32) as the backbone encoder.

B. Comparison with State-of-the-Art Methods

To evaluate the effectiveness of CMG-DP, we compare it with recent state-of-the-art video captioning models on two widely used benchmarks: MSR-VTT [25] and MSVD [3]. We adopt four standard automatic metrics: BLEU-4, METEOR, ROUGE-L, and CIDEr. Among these, CIDEr correlates most closely with human judgment and thus serves as our primary performance indicator. The results are summarized in Table I. CMG-DP achieves the best performance across almost all metrics on both datasets. Notably, on MSR-VTT, CMG-DP attains a CIDEr score of 58.3, which surpasses the previous best result (SEM-POS) by a substantial margin of +5.2 points. On MSVD, our model reaches a CIDEr score of 124.6, outperforming the strongest baseline (CoCap) by +11.6 points. These consistent improvements demonstrate that explicitly injecting

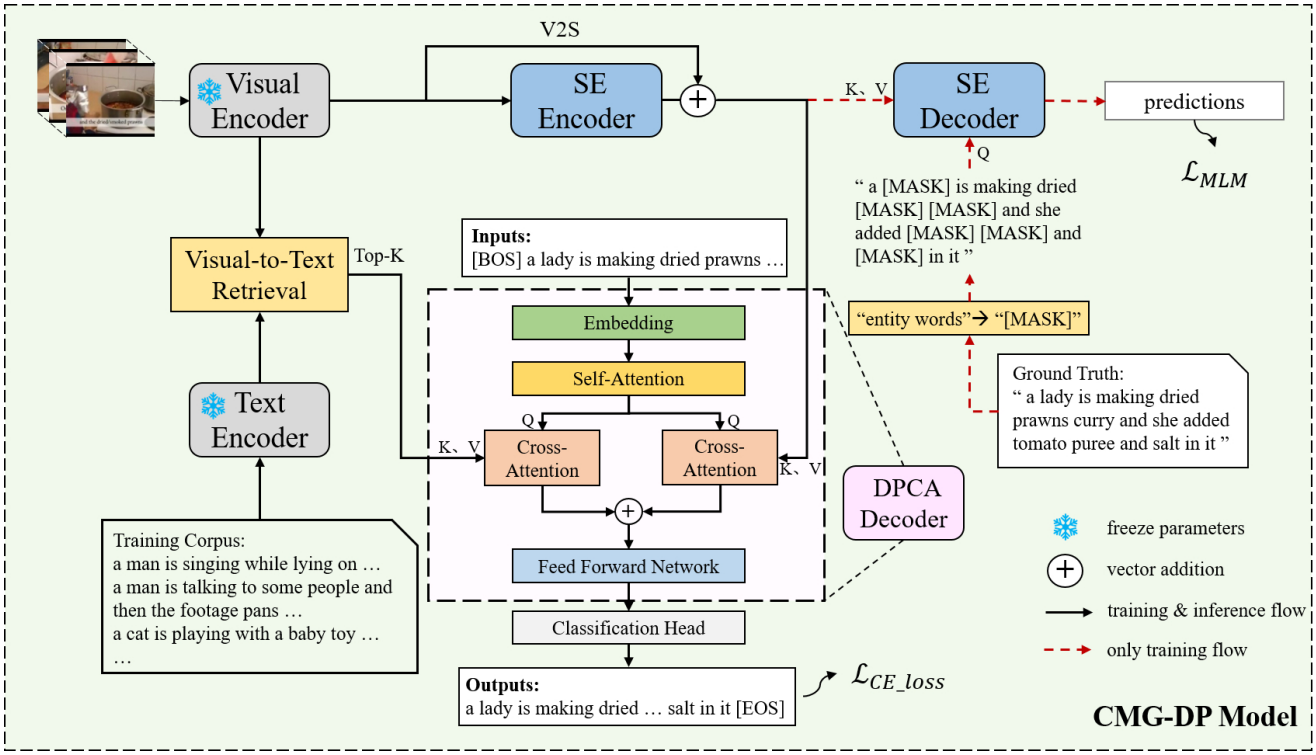

Fig. 1. Overall framework of CMG-DP. It consists of three core components: PKS mechanism, SE Module, and DPCA Decoder. Both visual and textual encoders are based on CLIP.

cross-modal prior knowledge during inference, as realized by our PKS mechanism, effectively enhances the model's ability to describe complex and unseen scenes.

C. Ablation Studies

To validate the contribution of each proposed component, we conduct a series of ablation experiments on both datasets by first establishing a baseline model (which feeds global visual features into a standard Transformer to generate captions) and then ablating each module from the full CMG-DP framework. The results are reported in Table II.

Effect of individual modules. Removing any of the three modules leads to a noticeable drop in performance across all metrics. Specifically, discarding the PKS mechanism ("w/o PKS") causes the most significant decline in CIDEr, highlighting the importance of explicit textual guidance during inference. Removing the SE Module ("w/o SE") also degrades performance, especially on entity-related metrics (as further analyzed in Table III), indicating that SE plays a crucial role in grounding descriptions to actual visual content and filtering out noisy textual priors. Ablating the DPCA Decoder ("w/o DPCA") and simply concatenating visual and textual features results in inferior fusion, confirming that our dual-path attention design is essential for dynamically integrating the two information streams.

Influence of the number of retrieved textual priors (k). The PKS mechanism retrieves the top-k semantically closest textual descriptions from the training corpus. To study the impact of k, we vary it from 4 to 20 and evaluate the CIDEr score on the validation set. As illustrated in Fig. 2, the best performance is achieved at $k = 12$. Although increasing the value of k can retrieve more textual descriptions that are semantically relevant, providing richer context for the generation process, it also introduces more candidates with lower similarity, which may be irrelevant or noisy. A moderate k value (e.g., 12) strikes the best balance between obtaining sufficient semantic guidance and suppressing noise, thereby yielding the most accurate video captions.

Entity grounding capability of the SE Module. To verify whether the SE module indeed helps the model focus on visually present entities and resist noise from retrieved text, we measure the entity-word accuracy in generated captions.

TABLE I

PERFORMANCE COMPARISON ON MSR-VTT AND MSVD DATASETS. BOLD VALUES INDICATE THE BEST RESULTS. "GLOBAL": GLOBAL VISUAL FEATURE EXTRACTED BY THE BACKBONE; "A": ACTION FEATURE; "O": OBJECT FEATURE. RESNET VARIANTS (R101/R151/R200) [9]; IRV2: INCEPTION-RESNET-V2 [21]. CLIP-BASED METHODS USE VIT BACKBONES AND DO NOT USE SEPARATE A/O FEATURES.

Model	Year	Features			MSR-VTT				MSVD			
		Global	A	O	B@4	M	R	C	B@4	M	R	C
OABTG [29]	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 [30]	2020	IRv2	✓	✓	40.5	28.2	60.9	49.1	46.5	33.5	69.4	81.0
SGN [18]	2021	R101	✓	–	40.8	28.3	60.8	49.5	52.8	36.5	72.9	94.3
SHAN [8]	2021	IRv2	✓	–	40.3	28.8	61.2	54.1	50.9	35.1	72.4	94.5
HMN [26]	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	**30.7**	64.1	53.1	60.1	38.5	76.0	108.3
GSEN [12]	2024	IRv2	✓	✓	42.9	28.4	61.7	51.0	58.8	37.6	75.2	102.5
VCRN [28]	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 [5]	2024	$\text{CLIP}_{\text{ViT-B/32}}$	–	–	43.8	29.6	63.6	54.6	59.7	38.5	77.2	106.3
CMG-DP	2025	$\text{CLIP}_{\text{ViT-B/32}}$	–	–	**45.9**	30.3	**64.2**	**58.3**	**62.3**	**41.5**	**79.3**	**124.6**

TABLE II

ABLATION STUDIES ON MSR-VTT AND MSVD DATASETS. "DPCA": DUAL-PATH CROSS-ATTENTION DECODER; "SE": SEMANTIC ENHANCEMENT MODULE; "PKS": PRIOR KNOWLEDGE SYNERGY MECHANISM; "W/O": WITHOUT; "CUT V2S": REMOVE VISUAL-TO-SEMANTIC ADDITION.

Model	MSR-VTT				MSVD			
	B@4	M	R	C	B@4	M	R	C
Baseline	42.1	28.6	61.3	55.7	58.4	38.9	76.1	107.2
w/o DPCA	44.5	29.3	62.2	57.3	61.3	40.4	77.9	117.2
w/o SE	44.8	30.0	63.2	57.1	61.6	41.0	78.2	119.6
w/o PKS	43.2	29.9	62.9	56.5	60.8	40.5	77.3	114.1
Cut V2S	45.5	30.1	63.8	57.9	62.0	41.0	78.9	120.7
Full model	**45.9**	**30.3**	**64.2**	**58.3**	**62.3**	**41.5**	**79.3**	**124.6**

TABLE III

COMPARISON OF ENTITY WORDS GENERATION ACCURACY ON MSR-VTT AND MSVD

SE	PKS	DPCA	MSR-VTT	MSVD
–	–	–	81.9%	82.1%
✓	–	–	84.8%	85.1%
–	✓	–	82.1%	82.9%
✓	✓	–	83.1%	84.1%
–	✓	✓	82.9%	83.5%
✓	✓	✓	83.8%	84.8%

Specifically, we check whether entity words (nouns) in the generated sentence appear in the ground-truth caption's entity set. Table III shows that adding the SE module consistently improves entity accuracy on both datasets. For instance, on MSR-VTT, enabling SE raises the accuracy from 81.9% to 84.8% even without PKS, and further boosts it to 83.8% in the full model. This confirms that the SE Module enhances visual grounding and helps the model rely more on actual video content rather than spurious textual associations.

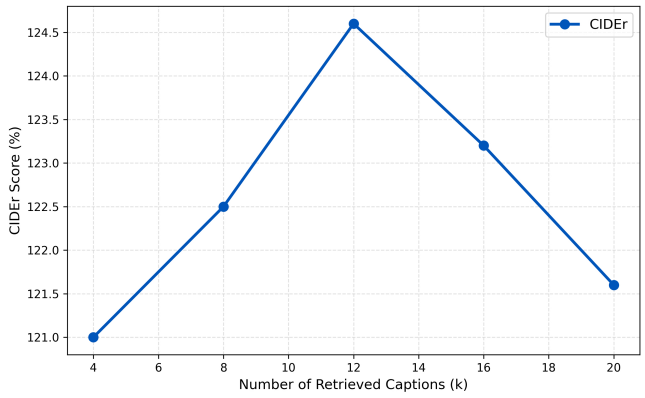

Fig. 2. Performance (CIDEr score) comparison under different number of k

V. CONCLUSION

In this paper, we propose CMG-DP, a novel framework that addresses the lack of explicit cross-modal guidance during inference in video captioning. CMG-DP introduces three core components: a PKS mechanism for retrieving relevant textual guidance, a SE Module for grounding visual semantics and filtering noise, and a DPCA Decoder for adaptive multimodal fusion. Experiments on MSR-VTT and MSVD demonstrate that CMG-DP achieves outstanding performance, with significant improvements in CIDEr scores. Ablation studies confirm the contribution of each component and the model's robustness in complex scenes. Our work establishes a new paradigm for inference-stage semantic guidance, effectively bridging visual perception and linguistic priors. Future work may extend this approach to other vision-language tasks.

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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 12487–12496.

[2] M. Abdar, M. Kollati, S. Kuraparthi, F. Pourpanah, D. McDuff, M. Ghavamzadeh, S. Yan, A. Mohamed, A. Khosravi, E. Cambria, and F. Porikli, "A review of deep learning for video captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2024.

[3] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol. (ACL-HLT)*, 2011, pp. 190–200.

[4] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, B. Lin, Z. Tang, L. Yuan, Y. Qiao, D. Lin, F. Zhao, and J. Wang, "ShareGPT4Video: Improving video understanding and generation with better captions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024, Art. no. 614.

[5] K. Chen, Q. Di, Y. Lu, and H. Wang, "Semantic-guided network with contrastive learning for video caption," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 3830–3884.

[6] Y. Chen, S. Wang, W. Zhang, and Q. Huang, "Less is more: Picking informative frames for video captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 358–373.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186.

[8] 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, 2021.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[10] G. Li, H. Ye, Y. Qi, S. Wang, L. Qing, Q. Huang, and M.-H. Yang, "Learning hierarchical modular networks for video captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1049–1064, 2024.

[11] 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. 17928–17937.

[12] 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, 2024.

[13] L. Momeni, M. Caron, A. Nagrani, A. Zisserman, and C. Schmid, "Verbs in action: Improving verb understanding in video-language models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 15533–15545.

[14] A. Nadeem, A. Hilton, R. Dawes, G. Thomas, and A. Mustafa, "SEM-POS: Grammatically and semantically correct video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2023, pp. 2606–2616.

[15] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1029–1038.

[16] 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.

[17] 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.

[18] 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.

[19] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17938–17947.

[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] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence — video to text," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4534–4542.

[23] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7622–7631.

[24] H. Wazni, K. I. Lo, and M. Sadrzadeh, "VerbCLIP: Improving verb understanding in vision-language models with compositional structures," in *Proc. 3rd Workshop Adv. Lang. Vis. Res. (ALVR)*, 2024, pp. 195–201.

[25] 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.

[26] H. Ye, G. Li, Y. Qi, S. Wang, Q. Huang, and M.-H. Yang, "Hierarchical modular network for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17918–17927.

[27] Y. Zeng, Y. Wang, D. Liao, G. Li, J. Xu, H. Man, B. Liu, and X. Xu, "Contrastive topic-enhanced network for video captioning," *Expert Syst. Appl.*, vol. 237, p. 121601, 2024.

[28] 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, 2025.

[29] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 8319–8328.

[30] 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.