MOP-VC: BRIDGING SYNTAX AND SEMANTICS VIA SYNTACTIC KNOWLEDGE TRANSFER WITH MIXTURE-OF-POS EXPERTS FOR VIDEO CAPTIONING

*Haoying Sun, Boyu Qiu, Shuyi Li, Zeyu Xi, Lifang Wu**

School of Information Science and Technology
Beijing University of Technology
100 Pingleyuan, Beijing, 100124, China

ABSTRACT

Video captioning aims to convert video content into natural language descriptions. Despite substantial progress, syntactic-based method such as part-of-speech (PoS) tags or parse trees may overlook the gap between syntax and semantics. In addition, the predicted syntactic tags are frequently noisy, and directly incorporating them into caption generation may hinder model performance. To address these issues, we propose MoP-VC, a video captioning model that transfers syntactic knowledge via a Syntactic Knowledge Transfer (SKT) module and mixture-of-experts. Specifically, SKT captures long-range syntactic representations via a syntactic loss, providing a compact feature space for effective syntactic knowledge transfer. To bridge syntax and semantics while enhancing robustness, we introduce the Mixture-of-PoS (MoPoS) module, which comprises reference, zero, copy, and constant experts. The reference experts rely on an auxiliary Concept PoS Pool, enriched with semantic information and constructed through parameter-free PoS spotting. Finally, syntax features from the SKT space serve as queries for MoPoS, which retrieves and refines relevant knowledge from the pool and subsequently transfers it to the model for caption generation. Experiments on MSVD and MSR-VTT show that MoP-VC surpasses most existing methods.

***Index Terms*—** Mixture-of-Expert, Part-of-Speech, Syntactic Analysis, Video Captioning

1. INTRODUCTION

Video Captioning (VC) generates coherent natural language descriptions from video content, supporting applications such as retrieval and commentary [1]. Encoder–decoder-based captioning methods have improved caption quality, yet aligning complex visual information with the discrete nature of language remains difficult. Prior knowledge (e.g., commonsense, sentiment, and topical cues [2, 3, 4]) has been incorporated to address this issue, but direct injection without syntactic control often produces captions that are semantically rich yet incoherent. Syntax, modeled via part-of-speech (PoS) tags or parse trees [5, 6], can enforce grammatical structure. However, existing approaches often neglect the potential noise in predicted syntax and overlook the gap between syntax and semantics [7], resulting in rigid or incoherent captions (Fig. 1(a)). The Mixture-of-Experts (MoE) framework [8, 9], with its gated expert routing, provides a flexible structure to filter noise and act as a bridge for information transfer.

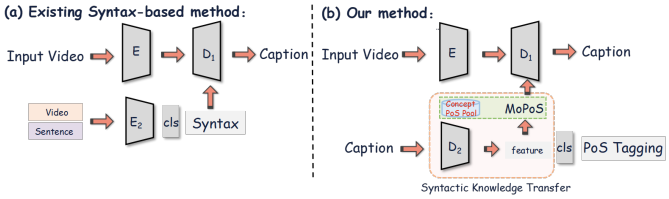

Fig. 1: Comparison between syntax-based methods (a) and our approach (b).

In this paper, we propose MoP-VC, a syntax-transferred video captioning model that efficiently exploits syntactic knowledge and bridges the syntax–semantics gap (Fig. 1(b)). Specifically, we design a dedicated Syntactic Knowledge Transfer (SKT) module that uses a syntactic loss to learn a compact syntax space. To bridge the syntax–semantics gap, the SKT module incorporates a MoPoS module, which consists of reference, zero, copy, and constant experts. Within the syntax space, MoPoS retrieves relevant knowledge from the constructed Concept PoS Pool, ensuring both syntactic correctness and semantic coherence. The contributions of this paper are summarized below:(1) We propose MoP-VC, which employs a meticulously designed Syntactic Knowledge Transfer module to learn a syntax representation space and selectively and dynamically leverage syntactic–semantic knowledge; (2) We propose MoPoS to learn purified syntactic–semantic information. The module selectively transfers knowledge from the constructed Concept PoS Pool into the decoder to facilitate the generation of semantically coherent captions; (3) Experimental results show that MoP-VC performs competitively with most existing approaches.

*Corresponding author. Email: lfwu@bjut.edu.cn

This work is supported by the National Natural Science Foundation of China (62236010, 62306021)

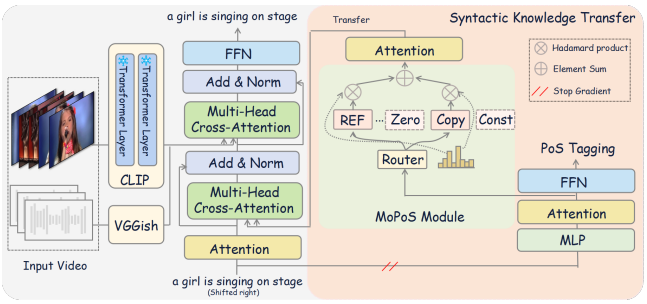

Fig. 2: The proposed MoP-VC framework. The model learns and transfers syntactic knowledge while bridging the syntax–semantics gap.

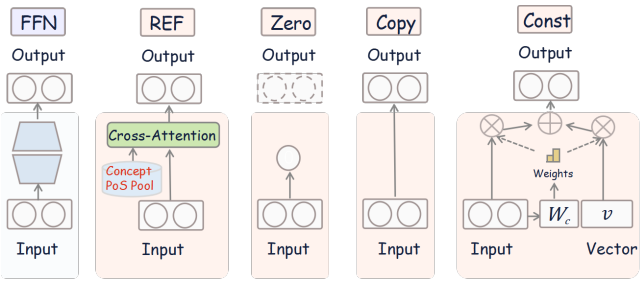

Fig. 3: Vanilla FFN expert and our MoPoS experts.

2. METHODOLOGY

As shown in Fig. 2, the syntax-aware Mixture-of-Experts video captioning model (MoP-VC) extends the conventional encoder–decoder framework by adding a Syntactic Knowledge Transfer module that refines and transfers grammatical knowledge. This branch uses intermediate compact syntax features to retrieve and refine semantic knowledge from the Concept PoS Pool, ultimately transferring it to the decoder. The chapter is organized as follows: Section 2.1 presents the baseline, Section 2.2 the Syntactic Knowledge Transfer module, Section 2.3 MoPoS, and Section 2.4 training/inference.

2.1. Baseline

Without loss of generality, we introduce a Transformer baseline here. Sampled video frames are denoted as $\varsigma = \{\varsigma_1, \varsigma_2, ...\varsigma_{N_f}\}$, encoded by a pre-trained backbone (e.g. CLIP) into $N_f \times d_b$ features $\mathbf{F}^{(\mathrm{B})} = \mathrm{Backbone}(\varsigma)$, where d_b is backbone dimension. The Transformer encoder $\mathbb{E}$ produces contextual features $\mathbf{F}_\varsigma = \mathbb{E}(\mathbf{F}^{(\mathrm{B})}) \in \mathbb{R}^{N_f \times d_h}$, where d_h denotes the encoder dimension. Audio features $\mathbf{F}_a \in \mathbb{R}^{N_f \times d_h}$ are extracted by VGGish. Concatenating both gives $\mathbf{F}$, which is fed into decoder $\mathbb{D}$ to predict caption $s = \mathbb{D}(\mathbf{F})$. Training minimizes cross-entropy:

$$\mathcal{L}_{\mathrm{CE}}(s, \varsigma) = -\sum_{t=1}^{T_{\max}} \mathbb{I}(s_t) \log p_\theta(s_t | s_{:t-1}; \varsigma). \quad (1)$$

where $T_{\max}$ is the max length of caption sentence, p is the probability of predicted word, θ is the trainable parameters of the model, s_t is the sentence which has generated at the t^{th} time step, and $\mathbb{I}(s_t) \in \mathbb{R}^{|\mathcal{V}|}$ is the one-hot encoding of s_t, where the value equals to 1 only at the position t, $|\mathcal{V}|$ is the length of vocabulary.

2.2. Syntactic Knowledge Transfer Module

Previous works directly feed syntax into the decoder [5, 10, 6], often introducing noise and undermines robustness. To address this, we propose SKT. Given token embeddings $\mathbf{H} \in \mathbb{R}^{T_{max} \times d_t}$, we first project via an MLP into $\mathbf{H}^{(0)} \in \mathbb{R}^{T_{max} \times d}$, then apply attention:

$$\mathbf{Q} = \mathbf{H}^{(0)}\mathbf{W}_Q,\ \mathbf{K} = \mathbf{H}^{(0)}\mathbf{W}_K,\ \mathbf{V} = \mathbf{H}^{(0)}\mathbf{W}_V, \quad (2)$$

$$\mathbf{H}^{(1)} = \mathrm{Attention}\big(\mathbf{H}^{(0)}\big) = \mathrm{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\Big)\mathbf{V} \in \mathbb{R}^{T_{max} \times d}, \quad (3)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable linear projection matrices.

The attention output is bifurcated: one is passed through a feed-forward network (FFN) and a classification head to predict the next token's PoS tag; the other is routed into the MoPoS module for expert selection and feature enhancement before being transferred into the decoder.

We train the classification head using a standard cross-entropy loss:

$$\mathcal{L}_{\mathrm{syn}} = -\sum_{t=1}^{T_{max}} \sum_{c=1}^{C} y_{t,c} \log \hat{y}_{t,c}\,, \quad (4)$$

$$\hat{\mathbf{y}}_t = \mathrm{softmax}\big(W_{\mathrm{cls}}\,\mathbf{h}_t^{(1)} + b_{\mathrm{cls}}\big) \quad , \quad t = 1, \ldots, T_{max}, \quad (5)$$

$\mathcal{L}_{\mathrm{syn}}$ computes the cross-entropy between the predicted probability $\hat{y}_{t,c}$ and the ground-truth one-hot label $y_{t,c}$ over all tokens and PoS categories, where $c \in 1, 2, \ldots, C$ and C is the total number of PoS tags. This objective guides the model to learn contextually relevant and compact syntactic features for VC.

To avoid interfering with baseline parameters, we apply a stop-gradient operation before the linear layer:

$$\widetilde{\mathbf{H}}^{(1)} = \mathrm{sg}\big(\mathbf{H}^{(1)}\big) \quad \Longrightarrow \quad \widetilde{\mathbf{H}}^{(1)}; \xrightarrow{\text{MoPoS}}; \text{Expert Fusion} \quad (6)$$

Here, $\mathrm{sg}(\cdot)$ blocks gradients from SKT from propagating back to the baseline. MoPoS then routes $\widetilde{\mathbf{H}}^{(1)}$ for feature refinement and knowledge selection, thereby isolating the optimization paths and preserving baseline representations.

Algorithm 1 Concept PoS Pool Construction

Require: Training corpus $\mathcal{D}$, number of keywords per PoS category N_k, CLIP text encoder $\text{CLIP}_{\text{text}}(\cdot)$, visual mean feature vector $\mathbf{v}$ extracted by CLIP visual encoder for the current video.

Ensure: Concept PoS Pool $\mathcal{P} = \{\mathbf{w}_j^*\}_{j=1}^{C}$ for C PoS categories

Initialize $\mathcal{P} \leftarrow \emptyset$
for each PoS category $j \in \{1, \ldots, C\}$ **do**
  Extract word set $\mathcal{W}_j$ from corpus $\mathcal{D}$ based on PoS tag j
  Select top-N_k frequent words $\mathcal{K}_j = \{w_{j,i}\}_{i=1}^{N_k}$ from $\mathcal{W}_j$
  for each keyword $w_{j,i} \in \mathcal{K}_j$ **do**
    Encode to vector: $\mathbf{w}_{j,i} \leftarrow \text{CLIP}_{\text{text}}(w_{j,i}) \in \mathbb{R}^{1\times d}$
    Compute cosine similarity: $s_{j,i} \leftarrow \frac{\langle \mathbf{w}_{j,i}, \mathbf{v} \rangle}{\|\mathbf{w}_{j,i}\| \cdot \|\mathbf{v}\|}$
  end for
  Select $\mathbf{w}_j^* \leftarrow \arg\max_i s_{j,i}$
  Add to Concept PoS Pool: $\mathcal{P} \leftarrow \mathcal{P} \cup \{\mathbf{w}_j^*\}$
end for

2.3. MoPoS Module

MoPoS enriches $\widetilde{\mathbf{H}}^{(1)}$ with syntax-semantic priors via multiple experts (Fig. 3):

Reference Expert. Performs attention between $\widetilde{\mathbf{H}}^{(1)}$ and Concept PoS Pool $\mathbf{P}_j$:

$$E_{ref}\left(\widetilde{\mathbf{H}}^{(1)}, \mathbf{P}_j\right) = \text{softmax}\Big(\frac{\widetilde{\mathbf{H}}^{(1)}\mathbf{W}_Q^p(\mathbf{P}_j\mathbf{W}_K^p)^\top}{\sqrt{d}}\Big)(\mathbf{P}_j\mathbf{W}_V^p). \tag{7}$$

where $\mathbf{P}_j$ denotes the Concept PoS Pool associated with the PoS category j, and $\mathbf{W}_Q^p$, $\mathbf{W}_K^p$, and $\mathbf{W}_V^p \in \mathbb{R}^{d\times d}$ are learnable linear projection matrices. The Concept PoS Pool is constructed via parameter-free PoS spotting: frequent PoS words are extracted from the corpus, encoded with CLIP, and the most video-relevant keywords are selected (Algorithm 1).

Zero Expert [9]. Discards input:

$$E_{\text{zero}}(\widetilde{\mathbf{H}}^{(1)}) = 0. \tag{8}$$

Copy Expert [9]. Passes input unchanged:

$$E_{copy}(\widetilde{\mathbf{H}}^{(1)}) = \widetilde{\mathbf{H}}^{(1)}. \tag{9}$$

Constant Expert [9]. Blends input with trainable vector $\mathbf{V}$:

$$E_{\text{const}}(\widetilde{\mathbf{H}}^{(1)}) = \alpha_1\widetilde{\mathbf{H}}^{(1)} + \alpha_2\mathbf{V}, \quad [\alpha_1, \alpha_2] = \text{Softmax}(W_c\widetilde{\mathbf{H}}^{(1)}). \tag{10}$$

where $W_c \in \mathbb{R}^{2\times d}$ is a trainable weight matrix, and d denotes the hidden dimension of the input token $\widetilde{\mathbf{H}}^{(1)}$. Next, routing is performed:

$$\widetilde{\mathbf{H}}^{(E)} = \sum_{k=1}^{K} G_k \cdot E_k\big(\widetilde{\mathbf{H}}^{(1)}\big) \tag{11}$$

$$G(\mathbf{H}^{(1)}) = \text{Softmax}\big(\text{KeepTopK}(\mathbf{W}_g\widetilde{\mathbf{H}}^{(1)} + \mathbf{b}_g, k)\big) \tag{12}$$

where $\{E_k\} = \{E_{\text{const}}, E_{\text{zero}}, E_{\text{copy}}, E_{\text{ref}}^{(1)}, \ldots, E_{\text{ref}}^{(M)}\}$, M is the number of reference expert. $\text{KeepTopK}(\cdot, k)$ retains the largest k values while setting the others to $-\infty$.

The final context-aware compact syntactic features transfer to the decoder are denoted as $\widetilde{\mathbf{H}}^{(\text{syn})}$:

$$\widetilde{\mathbf{H}}^{(\text{syn})} = \text{Attention}\big(\widetilde{\mathbf{H}}^{(E)}\big) \tag{13}$$

2.4. Training and Inference Phase

We transfer the rich contextual semantic and syntactic knowledge $\widetilde{\mathbf{H}}^{(\text{syn})}$ obtained from SKT module into the decoder described in Section 2.1; thus, the final loss function in Equation (1) is updated as:

$$\mathcal{L}_{\text{CE}}(s, \varsigma, \widetilde{\mathbf{H}}^{(\text{syn})}) = -\sum_{t=1}^{T_{\max}} \mathbb{I}(s_t) \log p_\theta(s_t | s_{:t-1}; \varsigma; \widetilde{\mathbf{H}}^{(\text{syn})}) \tag{14}$$

The overall loss incorporates the syntactic constraints, resulting in the final loss defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{syn}} \tag{15}$$

$\mathcal{L}_{\text{CE}}$ is During inference, we skip PoS prediction and directly transfer refined compact syntax features, yielding captions that are structurally accurate and semantically coherent.

3. EXPERIMENTS

3.1. Experimental Setup

We evaluate on MSVD [11] (1,970 videos) and MSR-VTT [12] (10,000 clips) using the splits in [13, 12]. Metrics include BLEU@4, METEOR, ROUGE, and CIDEr. Each video is sampled into 20 frames with features from $\text{CLIP}_{\text{ViT-B/32}}$, followed by a two-layer Transformer encoder–decoder ($d = 512$). PoS information ($C = 16$ categories) is extracted via spaCy, with 32 keywords per category. During training, we set the maximum caption length to $T_{\max} = 48$, trained the model for 50 epochs with batch size 512 using the Adam optimizer, and applied a linear learning rate schedule (initial rate 4e-3, warm-up ratio 0.1). We adopt three Reference Experts, and report average results over five runs.

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

To evaluate MoP-VC, we compared it with state-of-the-art VC methods, including ORG-TRL [14], CARE [13], TVRD [7], RSFD [15], VEIN [16], MAN [17], EvCap [18], RLHMN [19], TextKG [2], and C4C [20]. As shown in Table 1, incorporating syntactic priors enables MoP-VC to outperform most methods, demonstrating that bridging syntax and semantics enhances caption quality. These results confirm the effectiveness of transferring syntactic knowledge.

Table 1: Quantitative results on the MSVD and MSR-VTT datasets. † denotes reproduced results.

(a) MSVD dataset

Method	Year	B@4	R	M	C
ORG-TRL	2020	54.3	73.9	36.4	95.2
C4C	2021	-	-	-	-
TVRD	2022	50.5	71.7	34.5	84.3
RSFD	2023	51.2	72.9	35.7	96.7
CARE†	2023	53.8	74.5	37.9	105.3
VEIN	2024	55.7	74.4	37.6	98.9
RLHMN	2024	59.9	74.3	36.2	104.7
MAN	2024	59.7	74.3	36.2	101.5
EvCap	2024	53.6	74.3	36.7	107.2
Ours	-	55.4	**75.4**	**38.0**	**107.6**

(b) MSR-VTT dataset

Method	Year	B@4	R	M	C
ORG-TRL	2020	43.6	62.1	28.8	50.9
C4C	2021	46.1	63.7	30.7	57.7
TVRD	2022	43.0	62.2	28.7	51.8
RSFD	2023	43.4	62.3	29.3	53.1
CARE†	2023	47.7	64.3	30.8	56.9
VEIN	2024	44.1	62.6	30.0	55.3
RLHMN	2024	45.1	63.6	28.8	54.2
MAN	2024	41.3	61.4	28.0	49.8
EvCap	2024	45.5	64.5	30.9	53.8
Ours	-	**48.1**	**64.8**	**31.1**	**58.2**

3.3. Ablation Studies

We thoroughly evaluate the proposed method on MSR-VTT via ablation studies on SKT components and robustness (Table 2), expert types and numbers (Table 3), transfer mechanism, and prior dimensionality (Table 4).

Syntactic Knowledge Transfer Module and masked Concept PoS Pool We validate the SKT module and the model's robustness by randomly masking parts of the Concept PoS Pool (Table 2). Baseline denotes the method in Section 3.1, w/o $\mathcal{L}_{\text{syn}}$ removes the syntactic loss, and w/o attention removes the attention in Eq.(2). '*_mask' indicate the ratio of masked pool. Results show that both syntactic loss and attention improve performance: the loss encourages abstract syntactic feature learning, while attention enhances syntactic representation, enabling the model to capture grammatical structures beyond fluency. Masked pool experiments show no significant performance drop, indicating that MoP-VC possesses inherent robustness and resilience to noise.

Expert types and Number of Reference Experts We vary the expert configurations and the number of reference experts (2–5) ranked by PoS categories (Table 3). Best results occur with three reference experts (nouns, verbs, adverbs), as these are frequent and informative. Additional categories (e.g., adjectives, rare PoS) introduce noise or lack coverage, confirming the central role of nouns, verbs, and adverbs in captioning. Using only standard FFNs or removing other experts leads to performance degradation, demonstrating the effectiveness of the MoPoS design.

Table 2: Ablation study results on SKT and masked pool.

Method	B@4	M	R	C
Baseline	47.68 ± 0.24	30.84 ± 0.08	64.25 ± 0.07	56.85 ± 0.23
w/o $\mathcal{L}_{\text{syn}}$	47.90 ± 0.45	30.88 ± 0.16	64.63 ± 0.28	57.33 ± 0.47
w/o attenion	**48.11** ± 0.26	31.03 ± 0.10	**64.80** ± 0.17	58.05 ± 0.57
ours_0.2_mask	48.22 ± 0.38	31.03 ± 0.14	64.73 ± 0.20	58.03 ± 0.20
ours_0.4_mask	48.07 ± 0.38	30.71 ± 0.14	64.83 ± 0.20	57.53 ± 0.20
ours	48.09 ± 0.38	**31.10** ± 0.14	64.76 ± 0.20	**58.22** ± 0.20

Table 3: Ablation study on experts and reference expert count

No. of REF	B	M	R	C
2	48.12 ± 0.35	30.89 ± 0.14	**64.79** ± 0.17	57.77 ± 0.78
3 (ours)	48.09 ± 0.38	**31.10** ± 0.14	64.76 ± 0.20	**58.22** ± 0.20
4	48.07 ± 0.21	31.02 ± 0.11	64.66 ± 0.11	58.20 ± 0.25
5	48.06 ± 0.37	30.99 ± 0.08	64.78 ± 0.26	57.92 ± 0.44
w/o Zero	48.12 ± 0.31	30.93 ± 0.09	64.78 ± 0.16	57.75 ± 0.51
w/o Const	48.10 ± 0.28	30.95 ± 0.12	64.77 ± 0.16	57.83 ± 0.40
only FFN	**48.33** ± 0.19	31.03 ± 0.69	64.76 ± 0.45	58.13 ± 0.23

Table 4: Ablation study on transfer and dimension.

Method	B@4	M	R	C
Additive Fusion	47.93 ± 0.56	30.93 ± 0.17	64.67 ± 0.15	57.86 ± 0.34
Max Fusion	47.74 ± 0.27	30.96 ± 0.08	64.53 ± 0.40	57.34 ± 0.40
Attention (ours)	**48.09** ± 0.38	**31.10** ± 0.14	**64.76** ± 0.20	**58.22** ± 0.20
512 dimension	47.85 ± 0.22	31.03 ± 0.12	64.52 ± 0.14	57.68 ± 0.44

Transfer Mechanism and Linear Self-Attention Dimensionality We evaluate the transfer mechanism via additive fusion, max pooling, and attention (Table 4), excluding concatenation due to excessive width. Attention-based fusion performs best, likely owing to its adaptability to sequential tasks and superior information preservation, whereas summation and pooling suffer from information loss. In addition, the dimensionality of linear self-attention is critical for SKT: performance is optimal at 256, while increasing to 512 leads to degradation, likely due to overfitting.

4. CONCLUSION AND FUTURE WORK

This paper presents MoP-VC, a Syntactic Knowledge Transfer video captioning model. We construct a Concept PoS Pool and leverages the MoPoS module to dynamically fuse and filter syntax and semantic knowledge, bridging syntax and semantics while enhancing robustness. Future work will focus on more efficient long-tailed concept selection and semantically consistent generation methods for multimodal large models to mitigate hallucinations.

5. REFERENCES

[1] Zeyu Xi, Ge Shi, Haoying Sun, Bowen Zhang, Shuyi Li, and Lifang Wu, "Eika: Explicit & implicit knowledge-augmented network for entity-aware sports video captioning," *Expert Systems with Applications*, vol. 274, pp. 126906, 2025.

[2] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen, "Text with knowledge graph augmented transformer for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18941–18951.

[3] Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang, "Emotional video captioning with vision-based emotion interpretation network," *IEEE Transactions on Image Processing*, vol. 33, pp. 1122–1135, 2024.

[4] Yawen Zeng, Yiru Wang, Dongliang Liao, Gongfu Li, Jin Xu, Hong Man, Bo Liu, and Xiangmin Xu, "Contrastive topic-enhanced network for video captioning," *Expert Systems with Applications*, vol. 237, pp. 121601, 2024.

[5] Jiahui Sun, Peipei Song, Jing Zhang, and Dan Guo, "Syntax-controllable video captioning with tree-structural syntax augmentation," in *Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition*, 2024, pp. 1–7.

[6] Yitian Yuan, Lin Ma, and Wenwu Zhu, "Syntax customized video captioning by imitating exemplar sentences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 10209–10221, 2021.

[7] Bofeng Wu, Guocheng Niu, Jun Yu, Xinyan Xiao, Jian Zhang, and Hua Wu, "Towards knowledge-aware video captioning via transitive visual relationship detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6753–6765, 2022.

[8] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[9] Peng Jin, Bo Zhu, Li Yuan, and Shuicheng Yan, "Moe++: Accelerating mixture-of-experts methods with zero-computation experts," *arXiv preprint arXiv:2410.07348*, 2024.

[10] Qi Zheng, Chaoyue Wang, and Dacheng Tao, "Syntax-aware action targeting for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13096–13105.

[11] David Chen and William B 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.

[12] Jun Xu, Tao Mei, Ting Yao, and Yong Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[13] Bang Yang, Meng Cao, and Yuexian Zou, "Concept-aware video captioning: Describing videos with effective prior information," *IEEE Transactions on Image Processing*, 2023.

[14] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha, "Object relational graph with teacher-recommended learning for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13278–13288.

[15] Xian Zhong, Zipeng Li, Shuqin Chen, Kui Jiang, Chen Chen, and Mang Ye, "Refined semantic enhancement towards frequency diffusion for video captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 3724–3732.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. IEEE, 2006, vol. 2, pp. 1735–1742.

[17] Shuaiqi Jing, Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, and Heng Tao Shen, "Memory-based augmentation network for video captioning," *IEEE Transactions on Multimedia*, vol. 26, pp. 2367–2379, 2023.

[18] Sheng Liu, Annan Li, Yuwei Zhao, Jiahao Wang, and Yunhong Wang, "Evcap: Element-aware video captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[19] Guorong Li, Hanhua Ye, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang, "Learning hierarchical modular networks for video captioning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 2, pp. 1049–1064, 2023.

[20] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li, "Clip4caption: Clip for video caption," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4858–4862.