MoP-VC: Mixture-of-PoS experts for video captioning via syntactic constraints (Supplymentary)

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

School of Information Science and Technology, Beijing University of Technology, 100 Pingleyuan, Beijing, 100124, China
{sunhaoying97, qiuboyu@}@163.com,syli2022@bjut.edu.cn,
xi961226@163.com,lfwu@bjut.edu.cn

A Related Work

A.1 Video Captioning

Deep learning–based video captioning methods predominantly employ an encoder–decoder architecture [19, 26], which typically extracts information such as motion [17, 7], objects [12], and optical flow [16] from videos to facilitate caption generation. However, relying solely on intra-video information is insufficient to provide the model with adequate cues; prior-based methods aim to leverage various prior distributions or commonsense knowledge extracted from training corpora to furnish the model with additional priors for video understanding. Commonly used priors include concepts [25], sentiment [17], commonsense [7], topics [3], and syntax [23]. Among these, syntax-prior methods—which emphasize sentence structure—are often employed for syntax-controlled caption generation, such as using the syntactic tree of a given example sentence [18] or syntax representation [27], or constraining caption generation via predicted part-of-speech tags from the video or reference sentences [8, 21, 5]. However, these syntax-controlled approaches seldom account for potential noise in the syntax priors and predominantly emphasize sentence structure while neglecting semantic coherence.

Unlike the aforementioned approaches, to avoid potential noise in the syntax knowledge, we innovatively employ a Mixture-of-Experts (MoE) mechanism that routes and selects highly relevant syntactic priors, thereby minimizing the impact of noise on the model. Simultaneously, by leveraging the constructed Concept PoS Pool, we supply semantic information to the Mixture-of-PoS module (MoPoS), enhancing both structural accuracy and semantic coherence.

A.2 Mixture-of-Experts Models

Mixture-of-Experts was originally proposed to improve model expressiveness and efficiency by selectively activating a subset of expert parameters through a gating mechanism. Shazeer et al. [15] introduced MoE layers between LSTM units,

* Corresponding author.

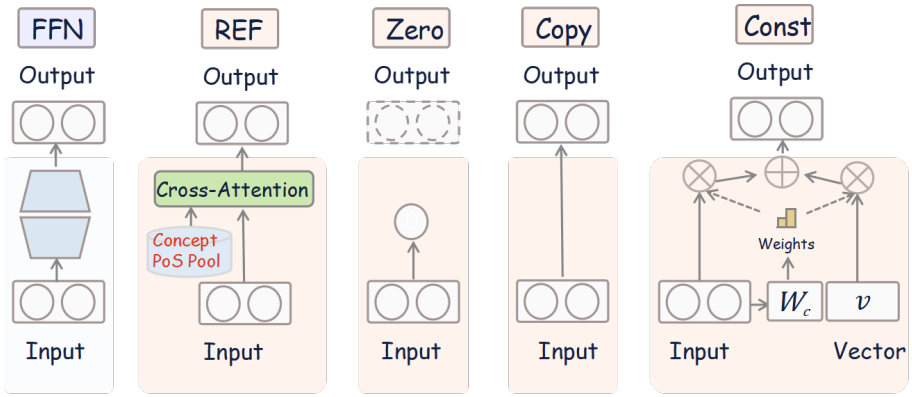

Fig. 1. Vanilla FFN expert and our MoPoS experts.

achieving significant gains in language modeling and machine translation. This mechanism was later adapted to the Transformer architecture by replacing the position-wise feed-forward sublayers. More recently, MoE has been instrumental in reducing computation in large language models [4, 28]. In video captioning, MoE has been explored in text-only settings for modeling distributions over scale-and-shift experts [9], and in zero-shot scenarios to enable knowledge transfer from seen to unseen actions [22]. Building on this, we incorporate a Concept PoS Pool into the MoE to dynamically refine relevant syntactic knowledge, effectively reducing noise and semantic sparsity, and ultimately enhancing the structural and semantic quality of generated captions.

B Expert Details of MoPoS

The SKT has acquired clear, context-aware syntactic features $\widetilde{\mathbf{H}}^{(1)}$. To endow $\widetilde{\mathbf{H}}^{(1)}$ with richer syntactic–semantic information for injection into the decoder, we propose the Mixture-of-Syntax Experts (MoPoS) module, illustrated in Fig. 1. This module comprises four experts: the reference expert, the zero expert, the copy expert, and the constant expert. Unlike conventional MoE models that employ FFN layers as experts, our reference expert integrates syntactic–semantic priors to enhance the semantic coherence of the captioning model. As described in [10], the zero expert, copy expert, and constant expert correspond to discard, skip, and replace operations, respectively. Each expert is introduced in detail below.

Reference Expert As illustrated in Fig. 1, Reference Expert (REF) contains a Multi-Head Attention layer designed to facilitate interaction between $\widetilde{\mathbf{H}}^{(1)}$ and the Concept PoS Pool. The Concept PoS Pool encodes rich PoS semantic knowledge, providing video-relevant syntactic priors from the dataset to the Syntactic Knowledge Transfer module. The interaction and querying of the syntactic prior pool using $\widetilde{\mathbf{H}}^{(1)}$ is defined by the following equation:

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

where $\mathbf{P}_j$ denotes the syntactic prior 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.

During both training and inference, each Reference Expert is assigned a fixed set of syntactic priors corresponding to a specific PoS category. This fixed assignment enables fine-grained modeling and optimization of PoS-specific features.

Zero Expert The simplest form of a zero expert discards the current input. The output of the zero expert is defined as:

$$E_{\text{zero}}(\widetilde{\mathbf{H}}^{(1)}) = 0 \tag{2}$$

Effectively, the presence of a zero expert can reduce a Top-2 MoPoS module to a Top-1 configuration. Specifically, when the zero expert is selected, its zero-valued output causes the overall output of the Top-2 MoPoS module to be determined solely by the other active expert. Thus, incorporating zero experts into MoPoS provides greater flexibility for processing both simple and complex syntactic tokens.

Copy Expert The copy expert returns the input as output, effectively serving as a shortcut:

$$E_{copy}(\widetilde{\mathbf{H}}^{(1)}) = \widetilde{\mathbf{H}}^{(1)}, \tag{3}$$

Intuitively, the copy expert allows the model to bypass the current MoPoS module. This mechanism is particularly useful when the input syntactic token does not align well with any existing experts, in which case skipping the module may help preserve relevant information and avoid introducing noise.

Constant Expert To further enhance the module's flexibility in handling input tokens, we introduce constant experts, which replace the input token x with a trainable vector $\mathbf{V}$. However, a complete substitution may result in a loss of input information. To mitigate this, we employ a trainable weight matrix W_c to dynamically determine the blending ratio between the original input and the constant vector. Formally, the output of the constant expert is defined as:

$$E_{\text{const}}(\widetilde{\mathbf{H}}^{(1)}) = \alpha_1\widetilde{\mathbf{H}}^{(1)} + \alpha_2\mathbf{V}, \quad [\alpha_1, \alpha_2] = \text{Softmax}(W_c\widetilde{\mathbf{H}}^{(1)}) \tag{4}$$

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

C Experimental setup

C.1 Dataset and Evaluation Metrics

Experiments were carried out on two benchmark datasets: MSR-VTT [24] and MSVD [2]. The MSVD dataset comprises 1,970 videos, each annotated with approximately 40 captions and averaging around 10 seconds in length. Following

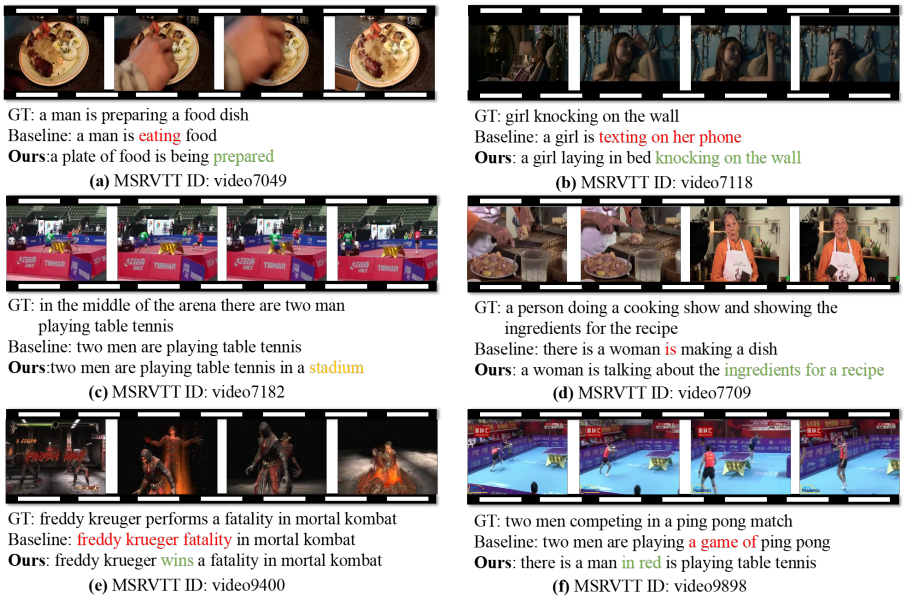

Fig. 2. Qualitative results on MSR-VTT dataset. Accurate keywords, errors and might-be-correct in generated captions are highlighted.

the protocols of [25] and [24], we allocated 1,200 videos for training, 100 for validation, and 670 for testing. The MSR-VTT dataset contains 10,000 video clips, each paired with 20 captions and averaging approximately 15 seconds. Using the official split from [24], we employed 6,153 clips for training, 497 for validation, and 2,090 for testing. To assess our method against prior work, we adopted four standard evaluation metrics: BLEU@4 (B@4) [14], a precision-based measure; METEOR (M) [1], which computes sentence-level alignment; ROUGE (R) [13], quantifying the longest common subsequence between predicted and reference captions; and CIDEr (C) [20], a consensus-based metric.

C.2 Implementation Details

The maximum number of words in the generated caption was set to $T_{\max} = 48$. The model was trained using a batch size of 512 for a total of 50 epochs with the Adam optimizer [11]. The learning rate was initialized to 4e-3 with a warm-up proportion of 0.1, where the learning rate linearly increased during the warm-up period and then linearly decayed to zero. For each video, we uniformly sampled $N_f = 20$ frames. A 12-layer ViT-B/32 with a patch size of 32 [6] (abbreviated as $\text{CLIP}_{\text{ViT}-\text{B}/32}$) was used to extract features. The extracted video features were fed into a two-layer Transformer video encoder with a hidden layer dimension of $d_h = 512$ to obtain enhanced visual features. These were then passed to a two-layer Transformer decoder with the same hidden size to generate captions.

Table 1. A computational cost comparison on the test split of the MSVD dataset.

Model	#Params(M)	Training time(h)	Inference(ms)	Mem(GB)	CIDEr
CLIP4Caption [19]	225.8	0.54	17.0	0.22	104.9
Ours	353.0	0.73	25.6	0.31	107.6

Latent dimension in SKT is $d = 256$. A dropout probability of 0.1 was applied across all attention and other layers. The Spacy toolkit was used to determine word PoS within the captions.

D Qualitative Experiments

As shown in Figure 2, we present six representative cases from the MSR-VTT test set. Two key advantages can be observed: (1) Our method generates more accurate and fine-grained captions aligned with video content. For instance, the baseline misidentifies "preparing food" as "eating" in (a), and "knocking on the wall" as "texting on her phone" in (b), whereas our model produces captions that are more consistent with the ground truth. Moreover, our approach produces more detailed descriptions such as explicitly capturing "stadium" in (c). (2) Our method produces captions with fewer syntactic errors and greater fluency. The baseline produces grammatically incorrect or redundant expressions such as "there is a woman in making a dish" in (d), "freddy krueger fatality" in (e), and the superfluous phrase "a game of" in (f). In contrast, our model consistently generates captions that are both fluent and semantically accurate.

E Computational Cost

Despite the performance improvements, our approach results in increased cost and model size, as shown in Table 1. Our experiments were conducted on the MSVD dataset using a single RTX 3090 GPU. We report five metrics: the total number of model parameters (#Params(M)), the training time (hours(h)), the average time to generate a caption without mini-batching (Inference(ms)), the memory used during inference without mini-batching (Mem(GB)) and CIDEr. Compared to Clip4Caption, our method uses more parameters, training time, and inference time. However, the performance (CIDEr score) is improved by 2.7%. To enhance the efficiency of our model, we suggest that future work could focus on compressing concepts and syntax. One potential approach is to leverage large language models (LLMs) for encoding and compressing concepts and syntax into latent embeddings, or alternatively, to extract key information and generate concise textual representations.

References

1. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005)
2. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)
3. Chen, S., Jin, Q., Chen, J., Hauptmann, A.G.: Generating video descriptions with latent topic guidance. IEEE Transactions on Multimedia **21**(9), 2407–2418 (2019)
4. Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., et al.: Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066 (2024)
5. Deng, J., Li, L., Zhang, B., Wang, S., Zha, Z., Huang, Q.: Syntax-guided hierarchical attention network for video captioning. IEEE Transactions on Circuits and Systems for Video Technology **32**(2), 880–892 (2021)
6. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Gu, X., Chen, G., Wang, Y., Zhang, L., Luo, T., Wen, L.: Text with knowledge graph augmented transformer for video captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18941–18951 (2023)
8. Hou, J., Wu, X., Zhao, W., Luo, J., Jia, Y.: Joint syntax representation learning and visual cue translation for video captioning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8918–8927 (2019)
9. Jia, H., Xu, Y., Zhu, L., Chen, G., Wang, Y., Yang, Y.: Mos2: Mixture of scale and shift experts for text-only video captioning. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8498–8507 (2024)
10. Jin, P., Zhu, B., Yuan, L., Yan, S.: Moe++: Accelerating mixture-of-experts methods with zero-computation experts. arXiv preprint arXiv:2410.07348 (2024)
11. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Li, L., Gao, X., Deng, J., Tu, Y., Zha, Z.J., Huang, Q.: Long short-term relation transformer with global gating for video captioning. IEEE Transactions on Image Processing **31**, 2726–2738 (2022)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
15. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
16. Shen, Y., Gu, X., Xu, K., Fan, H., Wen, L., Zhang, L.: Accurate and fast compressed video captioning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15558–15567 (2023)
17. Song, P., Guo, D., Yang, X., Tang, S., Wang, M.: Emotional video captioning with vision-based emotion interpretation network. IEEE Transactions on Image Processing **33**, 1122–1135 (2024)
18. Sun, J., Song, P., Zhang, J., Guo, D.: 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. pp. 1–7 (2024)

19. Tang, M., Wang, Z., Liu, Z., Rao, F., Li, D., Li, X.: Clip4caption: Clip for video caption. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4858–4862 (2021)
20. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
21. Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., Liu, W.: Controllable video captioning with pos sequence guidance based on gated fusion network. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2641–2650 (2019)
22. Wang, X., Wu, J., Zhang, D., Su, Y., Wang, W.Y.: Learning to compose topic-aware mixture of experts for zero-shot video captioning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8965–8972 (2019)
23. Wu, B., Niu, G., Yu, J., Xiao, X., Zhang, J., Wu, H.: Towards knowledge-aware video captioning via transitive visual relationship detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(10), 6753–6765 (2022)
24. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
25. Yang, B., Cao, M., Zou, Y.: Concept-aware video captioning: Describing videos with effective prior information. IEEE Transactions on Image Processing (2023)
26. Yang, B., Zhang, T., Zou, Y.: Clip meets video captioning: Concept-aware representation learning does matter. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 368–381. Springer (2022)
27. Yuan, Y., Ma, L., Zhu, W.: Syntax customized video captioning by imitating exemplar sentences. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 10209–10221 (2021)
28. Zhu, T., Qu, X., Dong, D., Ruan, J., Tong, J., He, C., Cheng, Y.: Llama-moe: Building mixture-of-experts from llama with continual pre-training. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 15913–15923 (2024)