

MoS²: Mixture of Scale and Shift Experts for Text-Only Video Captioning

Anonymous Authors

In the supplementary materials, we analyze the impact of training text domain and the proportion of in-domain text data on video captioning performance, as detailed in 1 and Section 2. Additionally, we provide comprehensive implementation details in 3.

1 TRAINING TEXT DOMAIN

In this section, we conduct a quantitative evaluation of the impact of training across various text domains on captioning performance. Besides our synthesized in-domain data, we also incorporate near-domain and out-of-domain captions from the SMiT [11] and WebVid-10M [2] datasets, respectively. We evaluate on the MSR-VTT [17] dataset and employ the mean of BLEU-4 [12], METEOR [3], ROUGE-L [9], and CIDEr [15] as the primary evaluation metric for robustness. The SMiT captions, manually annotated by human experts, are close to the text style of the downstream MSR-VTT [17] dataset, whereas the WebVid-10M captions, derived from web crawls, are significantly noisier and differ considerably from the MSR-VTT text, representing out-of-domain data. Our experimental results, presented in Fig. 1, demonstrate a significant improvement in video captioning performance when training with our synthesized in-domain text. Training with near-domain SMiT [11] text yields only marginal improvements, while incorporating out-of-domain text from WebVid-10M [2] leads to a noticeable decline in performance. These findings highlight the benefits of employing in-domain text, synthesized via GPT-4 [1], in improving video captioning performance.

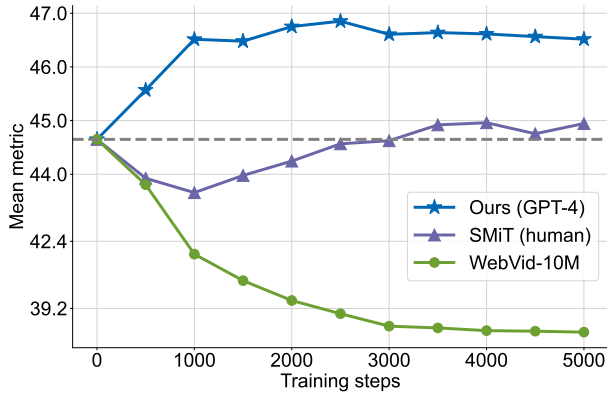


Figure 1: Training across various text domains.

2 IN-DOMAIN TEXT DATA RATIO

In this section, we conduct a quantitative analysis to evaluate the impact of varying in-domain text data ratios on video captioning performance. We maintain a constant data volume to isolate the

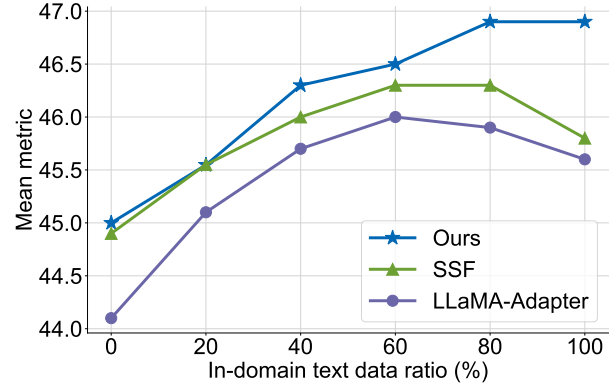


Figure 2: Comparative performance of SSF [8], LLaMA-Adapter [20], and our MoS² across varying in-domain text data ratios on the MSR-VTT [17] dataset.

effect of data composition. We combine synthesized in-domain text with human-annotated, near-domain captions from the SMiT [11] dataset to create training datasets with varying in-domain ratios. We employ the mean of BLEU-4 [12], METEOR [3], ROUGE-L [9], and CIDEr [15] as the primary evaluation metric for robustness. The experiment results are depicted in Fig. 2.

Our results reveal that SSF [8] consistently outperforms LLaMA-Adapter [20] across all in-domain text ratios tested, underscoring the superiority of methods that utilize scale and shift transformations in the text decoder over prompt-based approaches. Our model achieves competitive performance relative to [8] at lower in-domain ratios. Moreover, as the proportion of in-domain text increases, our method begins to outperform both SSF [8] and LLaMA-Adapter, which do not demonstrate consistent performance enhancements at higher ratios.

Our proposed architecture, which integrates a mixture of scale and shift experts, shows significant adaptability across varied data distributions, enabling the assimilation of specialized knowledge from distinct domains. This adaptability results in increased model capacity and improved performance, highlighting the effectiveness of leveraging in-domain text data.

3 IMPLEMENTATION DETAILS

We developed our training pipeline using the DeepSpeed [14] library. We employ fp16 training and ZeRO stage 1 [13] to shard the optimizer state, which significantly reduces memory usage and accelerates the training process. Unless specified, we employ the default configurations provided by DeepSpeed [14]. Our experiments are conducted on four NVIDIA A6000 GPUs.

Table 1: Data size and hyper-parameters for text-only training.

	MSRVTT [17]	MSVD [4]	VATEX [16]
#Downstream text	130.3K	48.8K	259.9K
#In-domain text	65.2K	23.1K	100.1K
Batch size	128		
Learning rate schedule	Linear decay		
Training epochs	10		
Optimizer	AdamW [10], $\beta_1 = 0.9$, $\beta_2 = 0.999$		
Weight decay	0.01		
Gradient clipping norm	1		

Table 2: Hyper-parameters for parameter-efficient learning methods.

Method	Learning Rate	Other Hyper-parameters
SSF [8]	1e-5	-
BitFit [19]	1e-5	-
LLaMA-Adapter [20]	1e-4	#prompts=16
LoRA [6]	1e-6	rank=40
Adapter [5]	1e-6	hidden dimension=160
Full fine-tuning	1e-6	-
MoS ² (ours)	1e-5	-

The hyper-parameters for our text-only training are detailed in Table 1. We synthesize approximately 65.2K, 23.1K, and 100.1K in-domain texts for MSR-VTT [17], MSVD [4] and VATEX [16], respectively. The models are trained with a batch size of 128 and we linearly decay the learning rate to 1e-7 over 10 epochs without incorporating warm-up. Optimization is performed using the AdamW optimizer [10], with betas set at [0.9, 0.999] and a weight decay of 0.01. Gradients are clipped by a norm of 1. We utilize the text prompt "a video segment where". During inference, we perform beam search with a beam width of 5 and apply a length penalty of 1.0, as described in [7]. Additionally, we uniformly sample 8 frames of 224×224 resolution, following [18]. The hyper-parameters for parameter-efficient methods are provided in Table 2.

REFERENCES

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1728–1738.
- [3] Satanjeev Banerjee and Alon Lavie. 2005. 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*. 65–72.
- [4] David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 190–200.
- [5] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. PMLR, 2790–2799.
- [6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- [7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
- [8] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. 2022. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems* 35 (2022), 109–123.
- [9] Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. 605–612.
- [10] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [11] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. 2021. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14871–14881.
- [12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [13] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–16.
- [14] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3505–3506.
- [15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.
- [16] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VateX: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4581–4591.
- [17] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.
- [18] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2022. Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners. *arXiv preprint arXiv:2212.04979* (2022).
- [19] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199* (2021).
- [20] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199* (2023).