Pretrained Image-Text Models are Secretly Video Captioners

Chunhui Zhang **Yiren Jian** **Zhongyu Ouyang** **Soroush Vosoughi**[*]
Department of Computer Science, Dartmouth College
{chunhui.zhang.gr, yiren.jian.gr, zhongyu.ouyang.gr, soroush.vosoughi}@dartmouth.edu

Abstract

Developing video captioning models is computationally expensive. The dynamic nature of video also complicates the design of multimodal models that can effectively caption these sequences. However, we find that by using minimal computational resources and without complex modifications to address video dynamics, an image-based model can be repurposed to outperform several specialised video captioning systems. Our adapted model demonstrates top-tier performance on major benchmarks, ranking 2nd on MSR-VTT and MSVD, and 3rd on VATEX. We transform it into a competitive video captioning system by post-training a typical image captioning model BLIP-2 with *only* 6,000 video-text pairs and *simply* concatenating frames—significantly fewer data than other methods, which use 2.5 to 144 million pairs. From a resource optimization perspective, this video captioning study focuses on three fundamental factors: optimizing model scale, maximizing data efficiency, and incorporating human-standard reinforcement supervision. The code, datasets, and models are released: https://anonymous.4open.science/r/ic2vc.

1 Introduction

Vision-language pretraining significantly advances multimodal tasks such as captioning, question answering, and retrieval (Liu et al., 2023b,a; Li et al., 2023b; Dai et al., 2023a; Chen et al., 2023b; Kuo et al., 2023; Xu et al., 2023). Among these, video captioning stands out as it narrates visual concepts and their temporal interactions, reflecting the intricate multimodal processes as humans to perceive and articulate dynamic visual experiences.

Current video-text methods often incorporate intricate designs tailored to video inputs. For instance, some models extend existing frameworks by integrating frame samplers to capture temporal dynamics (Alayrac et al., 2022; Yang et al., 2021; Xu et al., 2021). Other approaches, such as ALPRO (Li et al., 2022a) and VIOLET (Fu et al., 2023), propose end-to-end models that are meticulously trained on large-scale video-text datasets sourced from the Web (Zellers et al., 2021; Bain et al., 2021). Despite their success, video captioning models remain highly resource-intensive, often hitting performance bottlenecks when *(i)* computational resources are constrained, or *(ii)* the task requires specialized priors without clear guidance for model design and training. This raises a critical question: **for simplicity and efficiency, how can we repurpose existing image captioning models for video captioning, without relying on complex, hand-crafted video-specific designs?**

To address this, we revisit fundamental factors in training—**model scale**, **data efficiency**, and **supervision**—that critically influence video captioning while being agnostic to the variants of video-specific designs: First, we find that moderate-sized language models (LMs) when fine-tuned for specific tasks, can meet the demands of video captioning efficiently. This challenges the common belief that larger models are always superior, demonstrating that targeted optimization can outperform sheer model size. Second, using extensive pretraining on image-text pairs, as demonstrated with BLIP-2, is transferable to video tasks. This allows the model to achieve high performance with minimal video usage, offering an efficient alternative to training from scratch. Third, instead of relying on traditional cross-entropy loss, we optimize directly for non-differentiable CIDEr with reinforcement learning, ensuring that the generated captions better align with human-standard video descriptions.

By bypassing complex, specialized video input designs, our experiments demonstrate that BLIP-2 straightforwardly derived from image captioning, can be effectively optimized to deliver competitive video captioning performance. This study

[*] Corresponding author

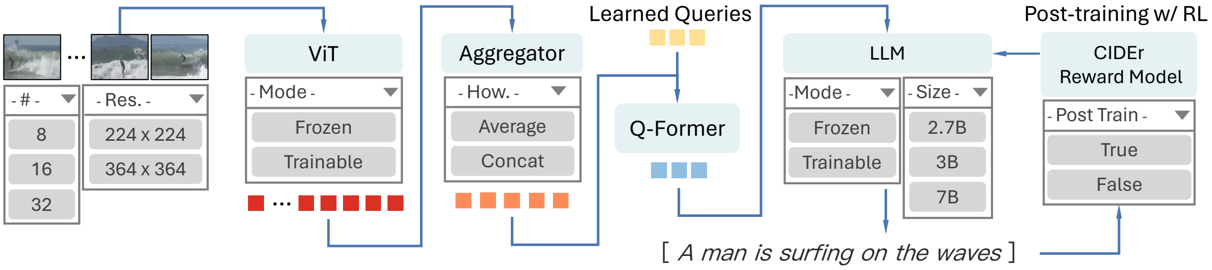

Figure 1: Key factors in recycling BLIP for video captioning: **Model** – assessing the scale and trainability of components like the ViT, LLM, and Q-Former; **Data** – examining the volume, quality, and fusion strategies for image and video-text pairs; **Supervision** – employing reinforcement learning to align generated captions with human language quality standards (CIDEr).

underscores the potential of simplicity and efficiency in advancing multimodal video captioning, providing a streamlined yet stable solution. All code, datasets, and model weights are at: `https://anonymous.4open.science/r/ic2vc`.

2 Recycling BLIP-2 for Video Captioning

As shown in Fig. 1, we adapt BLIP-2, a typical image-text model (details in App. B), for video captioning without any additional parameters. Each video frame is encoded by ViT, which generates visual tokens that are concatenated to form a unified representation (e.g., an 8-frame video produces a token sequence of size 8×256). This unified token sequence is then processed by the Q-former and passed to the LM to generate captions.

3 Training Recipes: Model, Data, and Supervision

According to Tab. 1, our solution has top-level performance on important benchmarks (particularly on the CIDEr metric-the primary ranking measure on Paperswithcode), ranking 2nd on MSR-VTT and MSVD, and 3rd on VATEX, among models with publicly available code. More importantly, it proves to be highly efficient without any video architecture design, using only **6k** video-text pairs—significantly less than the **million-level** datasets required by competing baselines.

Additional background is in App. A. The settings are detailed in App. C, and further experiments (**ablations, other datasets, and other video tasks**) supporting the following analysis are in App. D.

3.1 Model Scale

Trainability: modal connector $>$ LLM $>$ ViT To evaluate the adaptability of various components within the video captioning model, we conducted ablation studies using three setups: training all components, freezing the ViT only, and training the Q-Former only. The results, illustrated in Fig. 2(a) and supported by training curves in Fig. 4 (see App. D.1.1 for detailed discussions), reveal a clear performance hierarchy: freezing the ViT (configurations ii and iii) yields higher performance than training all components (configuration i).

Configurations with a frozen ViT allow the Q-Former and LLM to effectively leverage the pre-trained visual features, leading to better alignment in video captioning tasks. Conversely, training the ViT alongside other components introduces potential overfitting and alignment issues, resulting in suboptimal performance. The analysis establishes a hierarchy of trainability: Q-Former $>$ LLM $>$ ViT. The Q-Former shows the highest adaptability during training, followed by the LLM, which benefits from fine-tuning language data. In contrast, the ViT demonstrates the least trainability, as updating its parameters often disrupts the alignment between visual features and language output.

Supporting figures indicate that the Q-Former configuration achieves the most stable performance, reaching peak validation CIDEr scores without significant overfitting (Fig. 4). This pattern aligns with additional observations in App. D.1.1, confirming that focusing on training the modal connector and LLM while freezing the ViT optimizes the model's performance on video captioning tasks.

Mid-sized LLMs offer trainability for video captioning We analyzed the impact of LM size on video captioning by comparing three models: OPT-2.7B, Flan-T5-XL-3B, and Vicuna-7B (see Fig. 2(b) and Fig. 5). The results demonstrate that **Flan-T5-XL-3B, a mid-sized model, achieves superior performance in generating video captions**, outperforming both the smaller OPT-2.7B and the larger Vicuna-7B on key metric CIDEr.

Model	MSR-VTT (Xu et al., 2016)				MSVD (Chen and Dolan, 2011)				VATEX (Wang et al., 2019)				Code	# video -text
	C.	M.	R.	B4.	C.	M.	R.	B4.	C.	M.	R.	B4.		
IcoCap	60.2	31.1	64.9	47.0	110.3	39.5	76.5	59.1	67.8	25.7	53.1	37.4	No	-
MaMMUT	73.6	-	-	-	195.6	-	-	-	-	-	-	-	No	-
VideoCoCa	73.2	-	68.0	53.8	-	-	-	-	77.8	-	54.5	39.7	No	144.7M
VALOR	74.0	32.9	68.0	54.4	178.5	51.0	87.9	80.7	95.8	29.4	57.4	45.6	Yes	1.18M
VLAB	74.9	33.4	68.3	54.6	179.8	51.2	87.9	79.3	-	-	-	-	No	10.7M
GIT2	75.9	33.1	68.2	54.8	-	-	-	-	-	-	-	-	Yes	-
VAST	78.0	-	-	56.7	-	-	-	-	99.5	-	-	45.0	Yes	27M
mPLUG-2	80.0	34.9	70.1	57.8	165.8	48.4	85.3	70.5	-	-	-	-	Yes	2.5M
Ours	79.5	34.2	68.3	52.4	168.0	48.3	85.8	73.5	87.1	29.1	56.7	43.3	Yes	6K

Table 1: Overall comparison. The results for MSR-VTT, MSVD, and VATEX are from the `PaperswithCode` open leaderboard. The abbreviations C., M., R., and B4. stand for CIDEr, METEOR, ROUGE-L, and BLEU-4, respectively. We choose CIDEr as the most referential metric, following the `PaperswithCode`. Tab. 2 has details about configs and references.

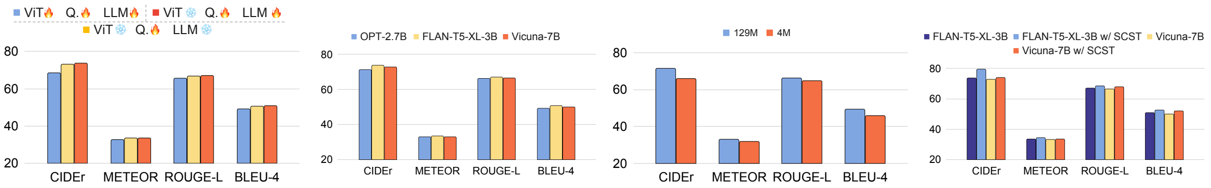

Figure 2: Comparisons of different setups for models on the *MSR-VTT* dataset: (a) freezing modules, (b) scales of LLMs, (c) usage of image-text pairs in pretrained BLIP-2, and (d) supervision with and without SCST. We also replicate the comparisons and ablations on other datasets (e.g., *MSVD* and **VATEX**) in App. D.4.

This challenges the notion that larger LMs always yield better results in multimodal tasks.

Training dynamics further support the advantages of mid-sized LLMs. As shown in Fig. 5, the smaller OPT-2.7B model requires 20 epochs to reach peak performance and fails to overfit, indicating limited expressiveness. On the other hand, Vicuna-7B converges rapidly within 5 epochs but quickly shows signs of overfitting, suggesting that its added complexity may not translate into meaningful improvements for video captioning. Flan-T5-XL-3B strikes a balance, reaching peak validation within 14 epochs and maintaining a better trade-off between generalization and overfitting.

These findings and training procedure analysis in App. D.1.2 indicate video captioning tasks benefit more from models capable of descriptive processing rather than advanced conversational or reasoning abilities. Thus, mid-sized LMs like Flan-T5-XL-3B effectively balance trainability, efficiency, and performance in video captioning tasks.

3.2 Data Efficiency

Image-Text pretraining offers transferability to video tasks We examine the effect of image-text pretraining on video captioning by comparing the performance of two BLIP-2 models pre-trained on different dataset sizes: one on 129 million pairs **(officially released)** and the other on 4 million pairs **(reproduced in-house)**. As depicted in Fig. 2(c), the model pre-trained with 129M pairs achieves a significantly higher CIDEr score (71.3) compared to the model trained with only 4M pairs (65.7), underscoring the advantages of using a larger dataset.

Fig. 6 (in App. D.2.1) further reveals that the model trained on 129M pairs converges faster and achieves higher performance than the model trained on fewer pairs. This suggests that video captioning tasks require robust grounding, with larger datasets significantly enhancing the model's ability to map visual concepts to language.

These results further underscore the *efficiency* of reusing extensively pre-trained image-text models for video tasks. Large-scale data exposure improves the model's comprehension of visual content, making it more suitable for generating accurate video captions. For a detailed analysis of the training process, refer to App. D.2.1.

Lower resolution efficiently supports video captioning We examined the impact of video resolution on training video captioning models by comparing two settings: 224×224 and 364×364. As

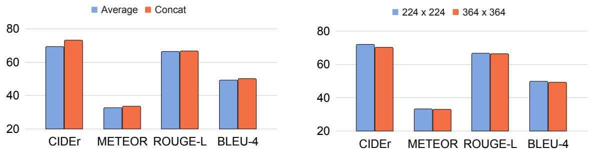

Figure 3: (a) temporal fusion by average v.s. concatenation; (b) different resolutions.

shown in Fig 3(b) and 7, models trained with lower-resolution videos (224×224) achieve competitive performance compared to those trained with higher resolution (364×364), despite exhibiting slightly more fluctuating training curves.

The results reveal that when basic frame aggregation techniques such as averaging or concatenation are used, lower resolution proves to be not only sufficient but also more efficient for generating accurate captions. The competitive CIDEr obtained with 224×224 resolution indicates that coarse visual information is adequate for the model to perceive and generate descriptive captions effectively.

Moreover, Fig. 7 demonstrates that while higher resolution (364×364) can lead to more stable training dynamics, the benefits are minimal when sophisticated frame aggregation is not applied. These findings suggest that adopting lower resolution offers practical advantages, including reduced computational requirements, without compromising captioning performance. For further insights, see the detailed analysis in App. D.2.2.

Frame concatenation effectively captures temporality We evaluate two approaches for temporal fusion in video captioning: frame averaging and frame concatenation. Frame averaging computes the average of visual tokens across sampled frames, maintaining a fixed dimension. In contrast, frame concatenation extends the token sequence by concatenating visual tokens from each sampled frame, preserving more granular temporal information. These fused tokens are subsequently processed by the Q-Former for caption generation.

The training dynamics, illustrated in Fig. 8 and Fig. 3 (a), show that models using frame concatenation consistently outperform those using frame averaging on CIDEr. The model with frame concatenation reaches peak validation performance around epoch 8 (Fig. 8), indicating that this method effectively retains temporality. In contrast, frame averaging shows significant performance oscillations after epoch 5, suggesting that it fails to capture sufficient temporal details for stable training.

These findings indicate that frame concatenation is more effective for capturing temporal information in video captioning, as it retains detailed visual context across frames. This approach allows the LM to access a richer set of visual concepts, resulting in more accurate and coherent captions. For additional analysis, see App. D.2.3.

3.3 Training Supervision

Reinforcement learning aligns captioning with human preference Traditional video captioning methods often rely on cross-entropy loss, which fails to fully align with human preferences for natural sentence generation. To address this, we use SCST (Rennie et al., 2017), which directly optimizes toward the human-like CIDEr metric. SCST leverages policy gradients from the non-differentiable CIDEr objective to guide updates to the Q-Former, LLM, and LoRA layers, enhancing alignment with human evaluation standards.

Fig. 2(d) and 9 show that SCST improves CIDEr scores by approximately 6.5% for Flan-T5-XL-3B and 3.4% for Vicuna-7B, while also boosting other metrics such as METEOR and ROUGE-L. Additionally, Fig. 9 illustrates a decoupling effect between training loss and validation CIDEr; models trained with SCST achieve higher CIDEr scores despite fluctuations in training loss. This shift reflects a prioritization of metrics aligned with human judgment over mere loss minimization.

The smaller improvement for Vicuna-7B likely results from its prior alignment training, which already incorporates reinforcement-based methods. Overall, SCST effectively aligns the training process with human-centered metrics, demonstrating its value for improving video captioning models. See App. D.3 for further details.

4 Discussion and Conclusion

This study stands out from existing video captioning research by identifying three factors—**model scale, data efficiency, and training supervision**—that are critical for effectively adapting image captioning models to video tasks. By using these insights to reuse the image-based BLIP-2 model for video tasks, our solution with minimal resource usage ranks *2nd, 2nd, and 3rd* on MSR-VTT, MSVD, and VATEX. This **open-source guide** provides a foundation for future research aimed at optimizing resource allocation in video captioning and refining post-training techniques.

Limitations

Our open-source solution is currently tailored specifically for video captioning tasks due to the page constraints of this short track. While this focus allows for a detailed and resource-efficient guide, it has not shown immediate applicability to other tasks. However, the methods presented can still be extended to broader applications, in particular to facilitate large-scale pseudolabeling for videotext datasets.

This approach is particularly valuable in specialized domains where annotated data is scarce, providing an efficient way to significantly expand video-text data resources. Similar to how the LAION dataset has advanced the image-text field by leveraging BLIP-1 for large-scale pseudolabeling (Li et al., 2022b; Schuhmann et al., 2022), our work aims to bring comparable improvements to video-text integration, enabling further research and development in this area.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

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 *IEEE International Conference on Computer Vision*.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Annual meeting of the association for computational linguistics: human language technologies*.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023a. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023b. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An opensource chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023a. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023b. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2023. An Empirical Study of End-to-End Video-Language Transformers with Masked Visual Modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xingjian He, Sihan Chen, Fan Ma, Zhicheng Huang, Xiaojie Jin, Zikang Liu, Dongmei Fu, Yi Yang, Jing Liu, and Jiashi Feng. 2023. Vlab: Enhancing video language pre-training by feature adapting and blending. *arXiv preprint arXiv:2305.13167*.

Weicheng Kuo, AJ Piergiovanni, Dahun Kim, xiyang luo, Benjamin Caine, Wei Li, Abhijit Ogale, Luowei Zhou, Andrew M. Dai, Zhifeng Chen, Claire Cui, and Anelia Angelova. 2023. MaMMUT: A simple architecture for joint learning for multimodal tasks. *Transactions on Machine Learning Research*.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. 2023a. LAVIS: A one-stop library for language-vision intelligence. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. 2022a. Align and prompt: Video-and-language pre-training with entity prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Yuanzhi Liang, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2023. Icocap: Improving video captioning by compounding images. *IEEE Transactions on Multimedia*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.

Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *IEEE conference on computer vision and pattern recognition*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. In *Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

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 *IEEE/CVF international conference on computer vision*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022b. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. 2023. mplug-2: a modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*.

Hu Xu, Gargi Ghosh, Po-Yao (Bernie) Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Conference on Empirical Methods in Natural Language Processing*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE conference on computer vision and pattern recognition*.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2022. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *Preprint*, arXiv:2212.04979.

J. Yang, Y. Bisk, and J. Gao. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *International Conference on Computer Vision*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. MERLOT: Multimodal neural script knowledge models. In *Advances in Neural Information Processing Systems*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Empirical Methods in Natural Language Processing: System Demonstrations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*.

A Related Work and Background

Image-Text Models Large-scale pretraining has revolutionized the field of image-text models, enabling significant advances. Models such as CoCa (Yu et al., 2022) and SimVLM (Wang et al., 2022b), which are trained from scratch on billions of image-text pairs, have set new benchmarks in generative tasks such as open-ended visual question answering (VQA) and visual captioning. BLIP-2 addresses the computational demands of pretraining from scratch by reusing existing pre-trained parameters from Vision Transformer (ViT) and LLMs and integrating them with a frozen pre-trained state. A key innovation in BLIP-2 is the introduction of the Q-former connector, carefully designed to enhance the interaction between visual and language modalities (Li et al., 2023b). This methodology has inspired subsequent innovations in visual-lingual tuning, with newer models often incorporating the pre-trained Q-former alongside the `eva-vit-g` model from BLIP-2, demonstrating the lasting impact of this methodology (Dai et al., 2023b; Zhu et al., 2024; Li et al., 2023c).

Video-Text Models Video-text models typically extend the capabilities of image-text models by integrating temporal feature aggregation to capture dynamic content, as exemplified by Video-CoCa (Yan et al., 2022). In addition, specialized models such as Video-LLaMA enhance the processing of temporal dynamics by embedding multiple temporal Q-former layers, facilitating nuanced interactions across modalities. Such advances refine the synergy between video Q-formers and LLMs within the model architecture, building on the foundation of BLIP-2 (Zhang et al., 2023). Building on these developments, recent studies, including VideoChat, PandaGPT, Valley, and Video-ChatGPT, investigate the embedding of frozen LLMs into video LMs, pushing the boundaries of the field (Li et al., 2023c; Su et al., 2023; Luo et al., 2023; Muhammad Maaz and Khan, 2023). In our study, we use BLIP-2 as a basic model for captioning, first pre-trained on images and then adapted to video by incorporating a video frame merging mechanism that effectively captures temporal nuances. This simplicity allows us to focus on evaluating the effects of model size, data volume, and training strategies on video captioning performance as we scale.

Difference between Image and Video Captioning The fundamental difference between image and video annotation stems from their source inputs: image annotation processes a single static image, while video annotation requires an understanding of the temporal dynamics over a sequence of frames. When adapted to video, pre-trained image models such as GIT (Wang et al., 2022a), Video-CoCa (Yan et al., 2022), and IcoCap (Liang et al., 2023) show remarkable adaptability to video with only moderate modifications, demonstrating their transferability. Conversely, video-specific models, including Video-LLaMA (Zhang et al., 2023) and VideoChat (Li et al., 2023c), use different sampling techniques to effectively capture temporal dynamics. Furthermore, models such as ALPRO (Li et al., 2022a) and VIOLET (Fu et al., 2023) utilize extensive web-crawled datasets to achieve end-to-end training, enriching their learning process. In our study, instead of emulating the complex adaptations typical of specialized video models, we adopt a streamlined approach that uses averaging or concatenation to merge temporal information from sampled video frames. This method allows us to focus on evaluating the effects of model size, data volume, and training strategies on video captioning performance as we scale.

B Preliminary

To effectively analyze the impact of specialized video adaptations without the confounding effects of architectural design variations, we base our methodology on BLIP-2, a basic image captioning model. We then describe the rationale for selecting BLIP-2 for our study.

Architecture of BLIP-2 BLIP-2 is originally designed to convert images into captions through a simple pipeline consisting of three main components: Vision, Connector, and Language: ***(i)* Vision** ViT serves as the entry into the BLIP-2 architecture, encoding images into a series of visual tokens. For example, a 224×224 image is transformed into 256 different visual tokens, laying the foundation for subsequent processing; ***(ii)* Modal connector** Q-former, positioned between ViT and LLM, bridges the gap between visual and language modalities. Its primary function is to project the sequence of the visual tokens generated by the ViT into a format compatible with language processing. A distinctive feature of the Q-former is its ability to condense the visual token array to a predetermined size, typically 32 tokens, regardless of the original number. This token reduction is not simply a numerical compression, but involves a sophisticated transformation into a language modality, resulting in so-called *soft prompts*. These soft prompts, now in tensor form, are then passed to the LLM for caption generation; ***(iii)* Language** LLM is responsible for generating the textual captions. It interprets the soft prompts from the Q-former and weaves them into a coherent caption that accurately reflects the visual content. This step is the culmination of the BLIP-2 pipeline, which transforms visual input into descriptive language.

Rationale for Choosing BLIP-2 as the Base Model In the field of vision language generative learning, many pre-trained image-based vision LMs are possible candidates besides BLIP-2, such as the LLaVA series, miniGPT-4, OpenCoCa, and OpenFlamingo, each offering different capabilities and features. Given the wide range of options available, our selection of pre-trained BLIP-2 is guided by specific criteria:

First, LLaVA uses a linear projection layer to project visual tokens from ViT and then feeds the projected tokens into LLMs. However, this linear projection layer keeps the visual tokens *consistent*, which means that this connector does not compress the visual token into fewer numbers. Although this redundant representation format does not meet the efficiency bottleneck on a single image as we extend the input modality to a single video containing multiple frames, it may exhaust the maximum token length capacity of an LLM. In contrast, BLIP-2 can reduce the number of tokens for each image/frame to a fixed number (e.g., 32). This efficient design avoids placing additional significant demands on the token length capacity of an LLM. *Second*, mini-GPT4, an instruction-tuned BLIP-2, also uses a linear projection layer to project visual tokens from ViT and then feeds the projected tokens into LLMs. Therefore, it also faces a similar limitation as LLaVA: when processing video frames, mini-GPT4's LLM token capacity also quickly hits a forward-backward bottleneck, limiting the number of frames that can be effectively captioned. *Third*, while Flamingo is easily adapted to video data due to its cross-modal attention design, its open-source reproduction, OpenFlamingo, underperforms BLIP-2 according to Li et al. (2023b)'s experiments. Third, Flamingo's design, which features cross-modal attention, facilitates its straightforward adaptation to video data; however, experiments conducted by Li et al. (2023b) imply that OpenFlamingo, an open-source version of Flamingo, does not perform as well as BLIP-2. Therefore, compared to LLaVA and mini-GPT4, BLIP-2 can be easily applied to video data to process multiple frames by averaging or concatenating the tokens of multiple frames (with a short length for the token of each frame, e.g. 32 tokens). We find that the BLIP-2 is characterized by its generality and simplicity, making it particularly well suited to the task of video captioning. Its design allows for minimal modification, allowing us to focus on the core factors that contribute to the effectiveness of video captioning models. This strategic choice is consistent with our goal of isolating and understanding the key elements that drive effective video captioning.

C Additional Experimental Details

C.1 Setup

Video Dataset Overview Our study uses the **MSR-VTT** dataset (Xu et al., 2016), a comprehensive open-domain video captioning resource. It includes 10,000 video clips across 20 different categories, with each clip annotated with 20 unique English sentences by contributors via Amazon Mechanical Turk. The dataset contains approximately 29,000 different words within the captions. For our experiments, we adhere to the conventional dataset partitioning: 6,513 clips for training, 497 for validation, and 2,990 for testing.

Training Configuration Training is conducted on eight NVIDIA RTX A6000 GPUs, utilizing the MSR-VTT dataset. Optimization is performed using the AdamW algorithm, with a setup that includes a weight decline of 0.05, an initial learning rate of 5×10^{-5}, and a minimum learning rate of 1×10^{-5}. The models are trained with a batch size of 32 over 32 epochs, with learning rate adjustments governed by a cosine annealing scheduler.

C.2 Model Information

Our video captioning model uses the image pre-trained BLIP-2 as its foundation. The BLIP-2 model itself is initially trained from scratch using the MSCOCO (Lin et al., 2014) and CapFilt (Li et al., 2022b) datasets, with additional data from the pseudo-labeled Conceptual Captioning (Sharma et al., 2018), SBU (Ordonez et al., 2011), and LAION (Schuhmann et al., 2022) collections. Our study employs ViT (`eva-vit-g` released from (Fang et al., 2023)) due to its proven effectiveness. In the realm of LM decoders, we investigate the capabilities of `OPT` (Zhang et al., 2022), `Flan-T5` (Chung et al., 2022), and `vicuna-7b` (Chiang et al., 2023). To adapt BLIP-2 for video, we utilize `bert-base-uncased` for the q-former architecture, maintaining parameter consistency with the image-trained version of BLIP-2. Additionally, we implement a frame token concatenation mechanism for aggregating temporal information from videos without increasing the parameter count. We provide the detailed structures, pre-train data, and language backbones in Tab. 2.

D Training Analysis and Results on Other Datasets

D.1 Model Scale

D.1.1 Trainability: modal connector > LLM > ViT

Fig. 4 presents the training curves of the video captioning model on MSR-VTT for different module freezing configurations: (a) ViT frozen, (b) only Q-Former trainable, and (c) all components trainable. The curves highlight the differences in trainability between the modal connector (Q-Former), the LLM, and the vision transformer (ViT).

The training curves indicate that setting (b), **where only the Q-Former is trainable, shows the most stable performance, reaching peak validation CIDEr at epoch 14 without significant overfitting.** In contrast, when additional components are trainable—such as the LLM in setting (c) or the ViT in setting (a)—the models reach peak performance earlier, at 6 and 4 epochs, respectively, but exhibit rapid overfitting afterward. This pattern suggests that increasing the number of trainable components complicates the optimization process, leading to quicker convergence but also accelerated overfitting. Consequently, setting (b) achieves the highest test CIDEr score (73.6), followed by setting (c) (73.0), and setting (a) (68.4).

Training the LLM also proves to be effective for video captioning, as reflected by the higher CIDEr score in setting (c). LLMs benefit from extensive pre-training on structured text, which enhances their ability to reason and assemble concepts. This capability allows them to align seamlessly with other modalities and reorganize visual inputs into coherent captions, making them a crucial component for video captioning tasks.

In contrast, training the ViT module appears suboptimal (or even counterproductive) for video captioning, as shown by the lower performance in setting (a). While large-scale pre-trained vision models like CLIP can capture fine-grained visual details, they often lack the structured representations necessary for composing visual information into coherent descriptions. This limitation affects the ability of the model to generate accurate captions when the ViT is a primary trainable component.

D.1.2 Mid-sized LLMs offer trainability for video captioning

To validate the advantages of mid-sized LLMs, we present the training dynamics for three different LM sizes in Fig. 5. The training curves indicate that larger models converge more quickly: OPT-2.7B requires 20 epochs to reach peak performance, Flan-T5-XL-3B takes 14 epochs, and Vicuna-7B converges in just 5 epochs. Although OPT-2.7B undergoes the longest training process, it fails to overfit the data, indicating limited model complexity. In contrast, both Flan-T5-XL-3B and Vicuna-7B show signs of overfitting soon after reaching peak performance, reflecting their greater model expressiveness for the video captioning task.

Flan-T5-XL-3B, with fewer parameters than Vicuna-7B, demonstrates sufficient complexity for video captioning tasks while requiring less computational power. Its moderate size avoids the additional burden of excessive parameters, leading to a more balanced and efficient learning process. In conclusion, mid-sized LMs, such as **Flan-T5-XL-3B, provide the optimal balance of trainability and complexity for video captioning, offering more efficient learning and better performance**

Model	# pretrain image-text	#video-text	Vision Backbone	Language Backbone
IcoCap (Liang et al., 2023)	-	-	CLIP-V	Transformer
MaMMUT (Kuo et al., 2023)	1.8B	-	ViT	Transformer
VideoCoCa (Yan et al., 2022)	3B	136M+8.7M	CoCa-V	CoCa-T
VALOR (Chen et al., 2023a)	1.18M	1.18M	CLIP-V/VideoSwin	BERT
VLAB (He et al., 2023)	5M+12M	10.7M	ViT giant	Transformer
GIT2 (Wang et al., 2022a)	12.9B	-	CoSwin	Transformer
VAST (Chen et al., 2023b)	-	27M	ViT	BERT
mPLUG-2 (Xu et al., 2023)	14M	2.5M	ViT-L/14	BERT-L
Ours	129M	6K	EVA-ViT-G	Flan-T5-XL

Table 2: The number of pre-train image-text and video-text pairs, vision backbone, and the language backbone for each video captioning model.

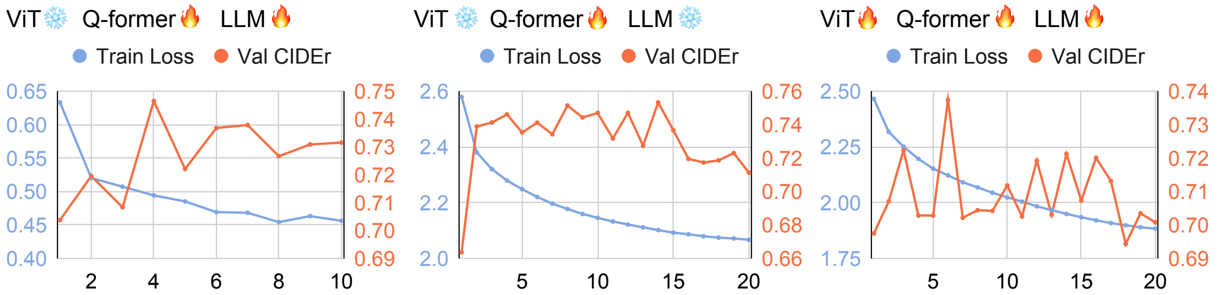

Figure 4: Training curves of the video captioning model on MSR-VTT, with different module freezing configurations. The vision backbone is ViT, and the language backbone is FLAN-T5. The curves represent three settings: (a) ViT frozen, (b) only Q-former trainable, and (c) all components trainable.

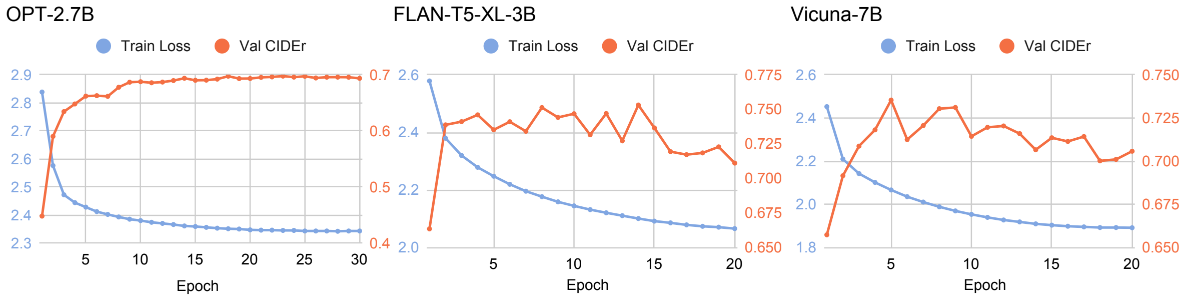

Figure 5: Training curves of a video captioning model with different sizes of LLMs. (a), (b), and (c) show training curves of LLMs with sizes 2.7B, 3B, and 7B respectively.

compared to their larger counterparts.

D.2 Data Efficiency

D.2.1 Image-Text pretraining offers transferability to video tasks

Fig. 6 illustrates that BLIP-2, **when pre-trained on a larger image-text dataset (129M pairs, officially released by the BLIP-2 group), converges faster and achieves a higher performance limit compared to the model trained with 4M image-text pairs.** This difference suggests that video captioning, while not as demanding in reasoning as tasks like VQA, still requires a strong ability to understand and describe visual content accurately. Extensive exposure to large-scale image-text data significantly improves the model's grounding process, enabling it to better understand and articulate visual content in video tasks. Thus, pre-training on extensive image-text datasets enhances the model's ability to map visual concepts from the vision domain to the language domain, making it more effective for video captioning. These results further

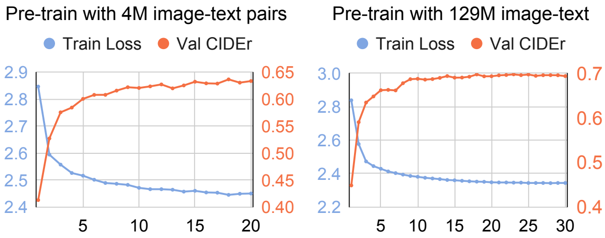

Figure 6: Training curve of a video captioning model with different sizes of pre-trained image-text pairs. (a) and (b) show training curves of models pre-trained with 4M and 129M image-text pairs respectively.

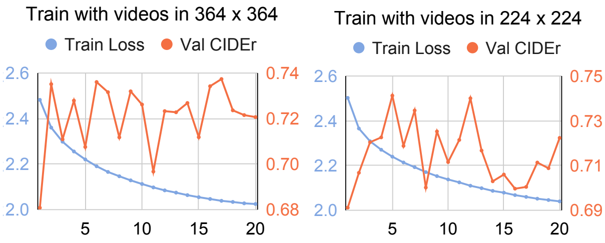

Figure 7: The training dynamics of a video captioning model with videos in different resolutions. (a) and (b) shows training curves of models trained with videos in 364×364 (up-sampling from original resolution 320x240 from MSR-VTT) and 224×224 respectively.

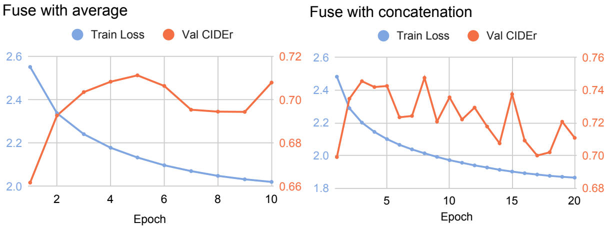

Figure 8: The training dynamics of a video captioning model with different fusion mechanisms for video frames. (a) and (b) show training curves of models that adopt the average and concatenation mechanisms respectively.

highlight the *effectiveness* of reusing extensively pre-trained image-text models for video captioning tasks.

D.2.2 Lower resolution efficiently supports video captioning

Fig. 7 compares the training dynamics of models using different video resolutions, showing that higher resolution videos (364×364) exhibit slightly more stable performance when combined with a stronger frame aggregator. **However, when the video frame aggregator is not highly sophisticated, lower resolution (224×224) proves to be efficient and effective, providing sufficient visual information for the model to perceive and generate accurate captions.** These findings indicate that lower resolution is not only sufficient but also more efficient for video captioning, especially when using basic frame aggregation techniques.

D.2.3 Frame concatenation effectively captures temporality

Fig. 8 illustrates the training dynamics for two fusion mechanisms: frame concatenation and averaging. **The model using concatenation reaches peak validation performance at epoch 8, suggesting that the complex visual tokens retain sufficient temporal information for effective learning.** In contrast, the averaging mechanism demonstrates weaker performance, with significant oscillations after epoch 5, indicating that it fails to provide enough temporal information for stable training. These results indicate that **frame concatenation is essential for effectively preserving temporal information, making it a more suitable approach for capturing visual concepts in video captioning.**

D.3 Training Supervision

D.3.1 Reinforcement learning aligns captioning with human preference

Fig. 9 shows the training dynamics for the Flan-T5-XL-3B and Vicuna-7B models with and without Self-Critical Sequence Training (SCST). The plots illustrate how SCST affects the relationship between training loss and validation CIDEr score. When SCST is applied, the training loss shows more variation, but the validation CIDEr score remains higher compared to models without SCST. For example, Flan-T5-XL-3B with SCST achieves a validation CIDEr score of about 0.82 despite increasing training loss, while Vicuna-7B with SCST maintains a CIDEr score of about 0.77.

Without SCST, both models follow a more conventional pattern where a steady decrease in training loss corresponds to a plateau in validation performance. In contrast, SCST introduces a decoupling effect: **fluctuations in training loss are no longer directly correlated with changes in validation CIDEr, suggesting that SCST promotes learning focused on optimizing human-centered metrics.** These results show that reinforcement learning via SCST effectively aligns the training process with human evaluation standards, prioritizing high-quality label generation that aligns with

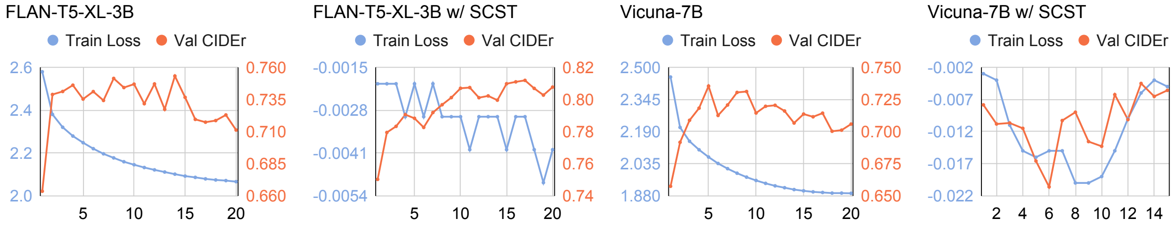

Figure 9: The training dynamics for the model when trained with/without SCST in LLM.

human judgment over simply minimizing training loss.

D.4 Experiments on MSVD and VATEX dataset

The ablation results on the *MSVD* and **VATEX** dataset are provided in Fig. 10 and 11. The experiments on the *MSVD* and **VATEX** dataset are primarily aligned with the analysis based on MSR-VTT presented in Sec. 2, App. D.1, D.2, and D.3.

Fig. 10 and 11 present detailed comparisons of different training setups for video captioning models on the MSVD and VATEX datasets. We use Fig. 10 as the example, and the results provide the following key patterns across four configurations:

- Module freezing (Fig. 10(a)): The results show that freezing various modules has a significant impact on performance. Models with no frozen components achieve the highest CIDEr scores, indicating the benefit of fine-tuning all parts. However, freezing both LLM and ViT results in the lowest performance, suggesting that the trainability of the connector (Q-Former) and LLM is essential for optimal fitting.

- LLM scales (Fig. 10(b)): Moderate-size LLMs, such as the Flan-T5-XL-3B, provide strong performance across all metrics. Although larger models such as Vicuna-7B offer slight improvements, the gains are modest, likely reflecting MSVD's higher text quality requirements. This finding supports the use of mid-range LLMs as a balanced choice for video captioning tasks.

- Pre-training of image-text pairs (Fig. 10(c)): Models pre-trained on larger datasets (129M image-text pairs) outperform those trained on smaller datasets (4M pairs), especially in terms of CIDEr scores. This result underscores the importance of extensive pre-training for capturing diverse visual-linguistic relationships and improving video captioning performance.

- SCST (Fig. 10(d)): Applying SCST improves the model's ability to generate human-like captions by optimizing directly for the CIDEr metric. Models trained with SCST show noticeable improvements in all evaluation metrics, highlighting its effectiveness in aligning speech generation with human preferences.

Overall, the ablation results confirm that flexible tuning of the connector and LLM components is critical for adapting image-text models like BLIP-2 to video captioning tasks. While moderate-sized LLMs offer a balanced trade-off between performance and computational efficiency, extensive pre-training on large datasets significantly improves model performance. In addition, reinforcement learning via SCST effectively improves the quality of generated captions by aligning the training goal with human-centric evaluation metrics.

D.5 Experiments on MSR-VTT and MSVD Video Question-Answering Datasets

The experiments on video question-answering (VQA) tasks using the MSR-VTT and MSVD datasets are summarized in Table 3. We extend the instruction tuning recipe from LAVIS (Li et al., 2023a) and InstructBLIP (Dai et al., 2023b) by 30K steps to test whether our findings from video captioning are applicable to VQA. The results in Table 3 show that many of the patterns observed in video captioning extend well to video question answering:

- Similar to video captioning, keeping more modules trainable leads to better performance. Specifically, models with all components trainable achieve the highest top-1 accuracy, while freezing only the ViT results in lower performance. This underscores the importance of fine-tuning all components for effective adaptation to VQA tasks.

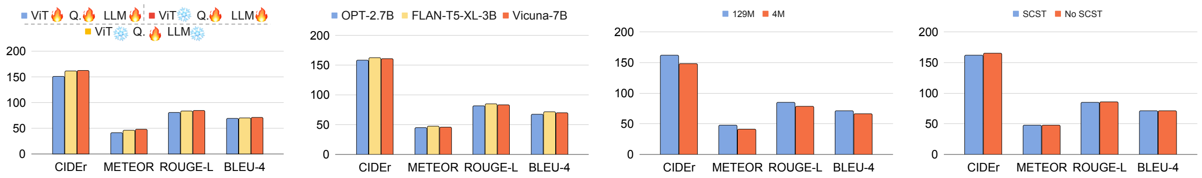

Figure 10: Comparative analysis of different training setups for video captioning models on *MSVD* dataset: (a) freezing modules, (b) scales of LLMs, (c) amount of pre-trained image-text pairs, and (d) models trained with and without SCST.

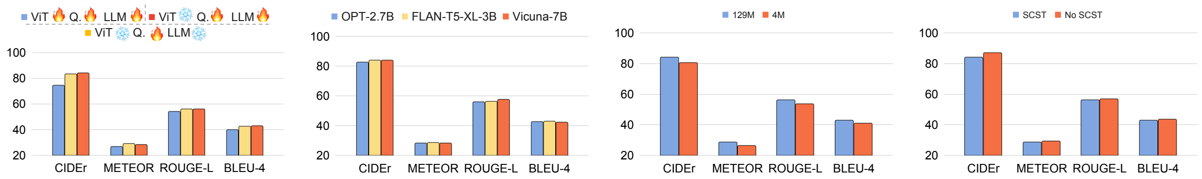

Figure 11: Comparative analysis of different training setups for video captioning models on *VATEX* dataset: (a) freezing modules, (b) scales of LLMs, (c) amount of pre-trained image-text pairs, and (d) models trained with and without SCST.

Category	**MSRVTT-QA**	**MSVD-QA**
Module Trainability		
All modules trainable	18.1	36.2
Unfreeze Q-former only	23.9	38.8
Freeze ViT only	22.5	38.5
RL to Human Standard		
SCST Disabled	23.9	38.8
SCST Enabled	24.1	41.0
Pretrained Image-Text Pairs		
129M	23.9	38.8
4M	18.8	36.2
Language Model Size		
OPT-2.7B	16.5	35.7
FLAN-T5-XL-3B	23.9	38.8
Vicuna-7B	20.2	38.5

Table 3: Top-1 accuracy comparison for different configurations on MSR-VTT and MSVD VQA datasets.

- Applying SCST slightly improves the model's ability to generate human-like responses by directly optimizing the metrics used in scoring. This is consistent with our findings in video captioning, where SCST helped improve CIDEr scores by aligning model outputs with human preferences.

- The use of moderately sized LLMs, such as FLAN-T5-XL, achieves strong performance on both datasets. Although larger models, such as Vicuna-7B, provide slight improvements, the gains are modest, suggesting that mid-range LLMs also provide a good balance between accuracy and computational efficiency for VQA.

- Similar to video captioning, extensive pretraining on large datasets (129M image-text pairs) leads to better performance than on smaller datasets (4M pairs). This reinforces the importance of diverse visual-linguistic pretraining for improving generalization in both video captioning and VQA tasks.

Overall, our experiments show that the key findings from our video captioning experiments are transferable to video question-answering tasks. The tuning of trainable Q-formers and LLMs, the reuse of extensive image-text pre-trained BLIP-2, and the use of reinforcement learning all contribute to improving the performance of video-based models across tasks. This transferability suggests that our summarized guidelines provide a basic but general handbook for building effective multimodal models for video captioning and potentially even other extended tasks.