

Supplementary Materials: Adaptively Building a Video-language Model for Video Captioning and Retrieval without Massive Video Pretraining

Zihao Liu¹, Xiaoyu Wu¹, Shengjin Wang², and Jiayao Qian¹

¹State Key Laboratory of Media Convergence and Communication, Communication University of China

²Department of Electronic Engineering, Tsinghua University

In this supplementary we first provide detailed experimental settings in Section A. We also provide more comparisons in Section B. Further ablation experiments are described in Section C. Finally, we present more qualitative results and attention visualizations in Section D.

A EXPERIMENTAL SETTINGS

A.1 Datasets

- The **MSVD** dataset [2] consists of 1970 videos and nearly 80K caption annotations. We followed the standard train / validation / test split of 1200/100/670 videos.
- The **MSR-VTT** dataset [27] consists of 10K videos with 20 captions per video. For video captioning, we used the standard train/validation/test split of 6513/497/2990 videos. For video retrieval, we use 9K videos and 18K captions for training and 1K video-caption pairs for testing, following [18].
- The **VATEX** dataset [24] is a bilingual annotated dataset with 35K videos, each having 10 English annotations. Due to unavailability of some videos, we utilized 100% of the training set, 100% of the validation set, and 90% of the public test set.
- The **VALOR-32K** dataset [3] focuses on audio-based video captioning and comprises 32K videos, each having an audio-visual caption. It also has a pretrained version called VALOR-1M, containing 1M videos, which is not yet available. The videos in this dataset are sampled from AudioSet, ensuring that they contain valid audio information. The captions are manually annotated and encompass both visual and audio content descriptions. Some videos, amounting to approximately 4% of the dataset, are no longer available.
- The **DiDeMo** dataset [9] contains 10464 videos, each annotated with multiple sentences describing different moments in the video. We follow other works [3, 18] to use standard split and concatenate multiple sentences into long paragraphs for retrieval.

A.2 Feature Extraction

For vision, we uniformly sample 8/16 frames from the video and follow [12] to extract patch-level features before pooling using CLIP-ViT-L/14¹ [20]. For audio, we divide the audio into 1-second non-overlapping segments and use AST² [8] to extract the features

pooled from each segment, which are then concatenated to form a sequence. We pre-extract these features to accelerate training.

A.3 Model Architecture

Due to the need for inheriting weights, the architecture of our model is set the same as the Q-Former in BLIP-2 ViT-L OPT [12], which is a 12-layer 768-dim 12-head Bert-like Transformer encoder. Note that we use the Q-Former weights from its representation learning stage, not the large language model with billions of parameters. On this basis, the new parameters we add include the audio FFN, temporal embedding, and audio feature projection; for the former, we duplicate the weights of the visual FFN, and for the latter two, we reinitialize them.

A.4 Training

The hyperparameters used during the training phase are listed in Table 1. We conducted all experiments using the PyTorch and LAVIS³ [11] frameworks on two NVIDIA A40 GPUs. Most training experiment can be completed within half a day.

Table 1: Hyperparameter settings during training.

Dataset	MSR-VTT	VATEX	MSVD	VALOR	DiDeMo
optimizer	AdamW				
lr	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	5×10^{-6}
scheduler	linear_warmup_cosine_lr				
warmup lr	1×10^{-8}				
warmup steps	100				
weight decay	0.05				
epoch	5	5	10	10	10
batch size	128 for captioning, 64 for retrieval				
max seq length	50				

A.5 Inference

For the video captioning task, we employ beam search with a beam size of 3, set a minimum length of 5, and a maximum length of 30, without using sampling or other post-processing methods. For the video retrieval task, we follow the inference pipeline in [12–14], which involves selecting the top 50 items based on the cosine similarity of features and then obtaining finer-grained scores using the matching head. Additionally, we use DSL [6] to post-process the scores.

¹<https://github.com/openai/CLIP>

²<https://huggingface.co/MIT/ast-finetuned-audioset-10-10-0.4593>

³<https://github.com/salesforce/LAVIS>

A.6 Efficiency Analysis

In the main text of our paper, we present Table 7 for efficiency analysis, and we will now describe its detailed settings. Our analysis focuses on the model's efficiency during training, with processing time statistics gathered under batch size = 64 and without beam search. To control variables, we calculate the average processing time per batch on the GPU during a single training epoch, excluding the time for data loading and feature extraction.

B MORE COMPARISONS

B.1 VQA Comparisons

For a more comprehensive evaluation of our method, we report our additional open-ended VQA experiments results on two datasets, MSR-VTT-QA and MSVD-QA [25], following [3, 5] in Table 2. During training, the questions and answers are concatenated with learnable queries and are fed to the encoder. Note that we use simplest prompts like *Question: <ques> Answer: <ans>*. Then we calculate cross-entropy loss of the predicted answer part. The hyperparameters are same as other tasks as in Table 1. During inference, only questions (and prompts) are provided, and the model needs to freely generate tokens until it meets the end token. The results shows that our method outperforms others without additional pre-training when $PT \leq 10M$. Compared to models with larger training scales, ours obtains comparable results with higher training efficiency and fewer parameters. For example, our method outperforms GIT2, which has 5.1B parameters.

Table 2: Comparisons with SOTA methods on VQA. PT denotes pre-training video-text pairs. Methods are grouped by PT.

Methods	PT	Params	MSR-VTT-QA Accuracy	MSVD-QA Accuracy
VideoCoCa [28]	8.7M	297M	42.6	53.6
VideoOFA [5]	2.2M	400M	45.4	55.5
VALOR _L [3]	13.5M	593M	49.2	60.0
VAST [4]	297M	1300M	50.1	60.2
JuskAsk [29]	0M	157M	39.6	41.2
GIT _B [23]	0M	129M	41.0	51.2
GIT _L	0M	347M	42.7	55.1
GIT2	0M	5100M	45.6	58.2
Ours	0M	240M	45.9	-
Ours(no audio)	0M	183M	-	55.9

B.2 Comparisons with PEFT methods

We present discussions and comparisons with parameter efficient finetune (PEFT) methods as follows. From a methodological perspective, both our approach and the PEFT methods aim to obtain a model that performs well on a new dataset at a lower cost. However, most PEFT methods (e.g., LoRA [10], Prefix-Tuning [15]) are performed on the same tasks (e.g., QA from general domain to medical domain). Thus, they cannot adapt pre-trained image models to the video domain. Some PEFT methods can achieve that, but are

Table 3: Comparisons with PEFT methods. † denotes method that also inherit BLIP-2 weights. Gray item is estimated by us. Here we report results without any post-processing operations (e.g., DSL [6]) during inference.

Methods	PT	Params	MSR-VTT		DiDeMo
			QA Acc	Ret R@1	Ret R@1
UniAdapter [17]	0M	19M	44.7	50.6	53.7
Aurora [22]	0M	0.1M	44.8	52.4	53.1
Side4Video [30]	0M	22M	-	51.4	-
CrossTVR [†] [7]	0M	200M	-	54.0	55.0
Ours	0M	240M	45.9	59.4	62.1

Table 4: Cross-dataset evaluation. We use CIDEr as the metric. V denotes VATEX. Blue and Green denote multi/single-modal model.

Training ⇒ Testing ⇒	PT	Params	MSR-VTT			MSVD		
			MSVD V-val	V-test	MSR-VTT V-val	V-test	V-test	V-test
SwinBERT [16]	0M	230M	84.2	24.9	20.3	34.7	31.1	25.1
VALOR _B [3]	3.5M	342M	120.7	29.1	23.6	-	-	-
VALOR _L	13.5M	593M	134.4	57.3	50.5	-	-	-
Ours	0M	240M	148.9	50.7	45.8	-	-	-
Ours	0M	183M	-	-	-	69.5	45.8	35.1

Table 5: Cross-dataset novel sentences analysis (MSR-VTT ⇒ VATEX-test).

Methods	Novel CIDEr ↑	Vocab Size ↑	Novel Sent. (#/pct.) ↑
SwinBERT [16]	5.75	848	4/0.07%
VALOR _B [3]	28.28	956	19/0.36%
VALOR _L	68.51	1569	156/2.9%
Ours	46.74	1514	112/2.1%

limited to non-generative tasks or visual channel [7, 17, 19, 22, 30]. Our method differs from them by adapting and expanding a pre-trained image-text model to a *multi-channel multi-task* video-text model. Moreover, our method improves efficiency through optimizing encoding operations, while PEFT methods focus on number of learnable parameters. We believe PEFT methods are complementary to ours, and it would be our further research. As shown in 3, our method significantly outperforms others. Even for the method using the same pre-training weights (CrossTVR [7]), we get 10.0% and 12.9% improvement on the two datasets, respectively.

B.3 Cross-dataset evaluation

To further validate the generalizability of our method, we conducted cross-dataset experiments as in Table 4 and Table 5. The results in Table 4 are based on open source codes and weights. Our model outperforms the traditional model (SwinBERT) and is comparable to the model with a larger size and more pre-training (VALOR), showing our strong generalization. Moreover, following [21], we count the vocabulary sizes of the predictions, find the sentences containing novel words, and present CIDEr scores of these novel

sentences to measure the correctness (Table 5). Correct novel sentences indicate that the model leverages the generalization ability from pre-training. For instance, with our model, trained on the MSR-VTT training set and tested on the VATEX test set, we found a vocabulary size of 1514. Among the results, 112 sentences (2.1%) contain new words not present in the MSR-VTT training set. The accuracy of these new sentences, measured by the CIDEr metric, reached 46.74. These results indicate that our model obtains strong generalizability comparable to much larger models.

C MORE ABLATIONS

C.1 Ablations of GTA

We explore the impact of different window sizes on the proposed model through ablation experiments on DeDiMo [9] dataset. As depicted in Table 6, we gradually increase the number of queries, and find that the number of queries used in BLIP-2 (32) was not enough to represent the rich information in the video, and the optimal number we found is 64. Row 2-5 explore the case where a set of queries applies attention globally over space and locally over time. We achieve the best result by restricting a group of queries to apply attention to two adjacent frames. According to our sparse frame sampling strategy, a single frame represents about 1.8s. In such a short time range, it is easier for queries to capture motion information (e.g., a person running from the left side to the right side). While there may be shot cut in a longer time range, which can lead to noise from other shots interfering with the query to capture key information. We also try to spatially limit the attention area of the queries, and the results indicate that this would degrade the performance. We analyze that this is because the query is responsible for extracting the key information from a complete image during pretraining, and limiting it spatially would lead to the domain shift problem.

C.2 Ablations of ITA

The hyperparameters in ITA include the weight for the loss (λ_{ita}) and the number of selected token-wise similarities for each pair of modalities. We conduct ablation experiments on both captioning and retrieval tasks. As shown in the upper part of Figure 1, since the value range of this loss is similar to the others, there is no need to scale the weight. Moreover, for ease of comparison, we kept the proportion of K token-wise similarities selected from different modal pairs the same, and used this proportion as a variable for comparison. As depicted in the lower part of Figure 1, the optimal performance is achieved when this proportion is 5%, corresponding to $K = 32, 160, 25$ for visual-audio, visual-text, and audio-text, respectively.

C.3 Other baseline with VPA

To substantiate the effectiveness of VPA, we integrated this module into another baseline. We apply VPA to mPLUG-2's two universal layers [26] and train it for 1 epoch on MSVD [2]. As shown in Table 7, despite applying it only on limited layers, VPA enhances the performance of the new baseline while reducing training costs. It should be noted that, in theory, our method can be applied to other VLMs that incorporate multi-query attention pooling (e.g., CoCa [31]).

Table 6: Ablation experiments of local window size and number of queries on DiDeMo dataset.

#	Number of queries	depth	height	width	DiDeMo R@1
1	32	1	16	16	57.8
2	64	1	16	16	58.9
3	64	2	16	16	60.0
4	64	4	16	16	58.8
5	64	8	16	16	58.4
6	64	2	8	8	58.6
7	64	8	4	4	57.2
8	64	4	4	4	54.3
9	128	1	16	16	59.4
10	128	2	16	16	58.6

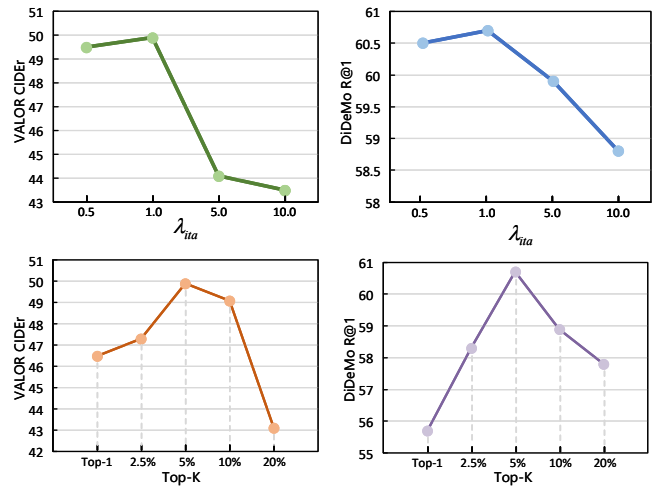


Figure 1: Ablation experiments on loss weight and the proportion of the selected topK similarities in ITA on MSR-VTT and DiDeMo datasets. Top-1 indicates that only the largest similarity is used.

Table 7: Ablations with mPLUG2 on MSVD.

Methods	Training Time	GPU Mem.	METEOR	CIDEr
mPLUG2 [26]	4.4h	23.2G	39.2	128.9
mPLUG2+VPA	3.5h(-20%)	21.8G(-6%)	40.2(+1.0)	131.3(+2.4)

D QUALITATIVE RESULTS

D.1 Video Captioning Results

Figure 2 and Figure 3 are the qualitative examples of our model. In Figure 2, each example is generated from the model trained on the corresponding dataset, and the examples are randomly sampled from the validation set or test set. For datasets with multiple ground truths, we sample three. We also show the predictions of two other models [16, 23] for comparison, but some models trained on specific sets have not been published. In the first example, our

model successfully infers from the audio modality what the woman was talking about in the video, while others provide either a vague or incorrect description, confirming that our model is able to use additional audio modality. In the last example, where other models predict similarly simple descriptions, our model predicts more fine-grained descriptions, which we attribute to our use of fine-grained features and the fine-grained objective function. In the remaining examples, our model also successfully recognizes the subjects, actions, locations, and relationships, and provides syntactically correct generated results.

In Figure 3, we randomly download some videos in the wild. All results are generated by the model trained on the VATEX dataset [24]. The first two examples show that our model is able to recognize landmarks in movies and music videos, which we attribute to the knowledge learned from large-scale image-text pairs. The GIT [23] also has been pretrained on a large number of image-text pairs, which does not provide a satisfactory description, showing that it may be forgetful when fine-tuned with videos. In the third example, our model detects an accident based on a surveillance video, demonstrating the potential of the our model to be applied in real-world tasks. In the last example, our model recognizes fine-grained targets (e.g., the poster on the wall in Andy's cell in the movie). Overall, our proposed model generalizes well and has excellent performance in real application scenarios.

D.2 Video Retrieval Results

Figure 4 presents examples from our video retrieval model tested on the MSR-VTT dataset. We conducted a qualitative comparison with CLIP4Clip [18]. For the first query, the textual description focuses on the purpose of the video rather than the overt actions being performed; that is, the text describes the video as being produced to promote helping others, with specific scenes showing feeding hungry children, etc. This requires the model to possess a solid ability to understand semantics. Our model ranks the correct result at the top, while the other model does not display the correct result within the top 3 results. For the second query, given the presence of multiple similar results in the retrieval library, our model's top-3 retrieved videos largely corresponded to the query, correctly placing the most relevant video at the top (the second-ranked retrieval result featured a classroom and a teacher in an animation, and the third-ranked result also depicted a classroom and a teacher, but without dialogue). In contrast, the other model incorrectly placed the video in the top-2, with its top-ranked video featuring multiple teachers teaching in scenarios similar to the query but did not match in detail. These results demonstrate our model's ability to achieve fine-grained, discriminative representations.

D.3 Attention Visualization

We also visualize the model's attention by the post-processing method in [1]. As shown in Figure 5, in the heat map, each row indicates the attention applied to vision queries and audio tokens when generating a text token, with the first 64 columns being the vision queries and the last 10 columns being the audio tokens in this example. The two sets of images on the left side, from top to bottom, respectively represent the model's attention to various regions of the video when generating *sheep* and the original frames of the

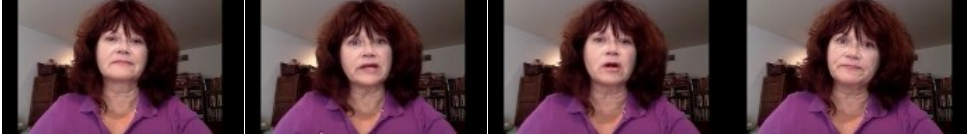
video (three frames are selected for convenient visualization). The attention maps are obtained by averaging the attention weights of the top-5 most attended visual queries. From the results, we can see that the model selectively focuses on key information when generating different concepts. When generating *sheep*, the model mainly focuses on the sheep in the video. When generating *the wind*, the attention to the audio tokens increases. The results demonstrate that our proposed VPA aligns the semantics of different modalities, while VPA is capable of perceiving various visual elements depending on the context.

REFERENCES

- [1] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*. IEEE, Montreal, QC, Canada, 387–396.
- [2] David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, Portland, USA, 190–200.
- [3] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023. VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset. *arXiv preprint arXiv:2304.08345* (2023).
- [4] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*. New Orleans, LA, USA.
- [5] Xilun Chen, L. Yu, Wenhan Xiong, Barlas Ouguz, Yashar Mehdad, and Wen tau Yih. 2023. VideoOFA: Two-Stage Pre-Training for Video-to-Text Generation. *arXiv preprint arXiv:2305.03204* (2023).
- [6] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. 2021. Improving Video-Text Retrieval by Multi-Stream Corpus Alignment and Dual Softmax Loss. *arXiv preprint arXiv:2109.04290* (2021).
- [7] Zuozhuo Dai, Fang Shao, Qingkun Su, Zilong Dong, and Siyu Zhu. 2023. Fine-grained Text-Video Retrieval with Frozen Image Encoders. *ArXiv abs/2307.09972* (2023).
- [8] Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: Audio Spectrogram Transformer. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association 2021*. ISCA, Brno, Czechia, 571–575.
- [9] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing Moments in Video with Natural Language. In *IEEE International Conference on Computer Vision, ICCV 2017*. IEEE Computer Society, Venice, Italy, 5804–5813.
- [10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, Virtual Event.
- [11] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. 2023. LAVIS: A One-stop Library for Language-Vision Intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10–12, 2023*. Association for Computational Linguistics, 31–41.
- [12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023 (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, Honolulu, Hawaii, USA, 19730–19742.
- [13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning (ICML)*. PMLR, Baltimore, Maryland, USA, 12888–12900.
- [14] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 34. Curran Associates, Inc., Virtual Event, 9694–9705.
- [15] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* abs/2101.00190 (2021).
- [16] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. In *2022 IEEE/CVF Conference on Computer*

- Vision and Pattern Recognition (CVPR)*. IEEE/CVF, New Orleans, LA, USA, 17928–17937.
- [17] Haoyu Lu, Mingyu Ding, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Masayoshi Tomizuka, and Wei Zhan. 2023. UniAdapter: Unified Parameter-Efficient Transfer Learning for Cross-modal Modeling. *arXiv preprint arXiv:2302.06605* (2023).
 - [18] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval. *Neurocomputing* 508 (2021), 293–304.
 - [19] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding Language-Image Pretrained Models for General Video Recognition. In *European Conference on Computer Vision*.
 - [20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Vol. 139. PMLR, Virtual Event, 8748–8763.
 - [21] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2021. ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 17897–17907.
 - [22] Haixin Wang, Xinlong Yang, Jianlong Chang, Di Jin, Jinan Sun, Shikun Zhang, Xiao Luo, and Qi Tian. 2023. Parameter-efficient Tuning of Large-scale Multi-modal Foundation Model. In *Neural Information Processing Systems*.
 - [23] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *Transactions on Machine Learning Research* (2022).
 - [24] Xin Eric 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE/CVF, 4580–4590.
 - [25] D. Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. *Proceedings of the 25th ACM international conference on Multimedia* (2017).
 - [26] 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, ICML 2023 (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, Honolulu, USA, 38728–38748.
 - [27] 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 (CVPR)*. IEEE, 5288–5296.
 - [28] 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).
 - [29] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Learning to Answer Visual Questions from Web Videos. *IEEE TPAMI* (2022).
 - [30] Huanjin Yao, Wenhao Wu, and Zhiheng Li. 2023. Side4Video: Spatial-Temporal Side Network for Memory-Efficient Image-to-Video Transfer Learning. *ArXiv abs/2311.15769* (2023).
 - [31] 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* (2022).

- *video7072.mp4 MSR-VTT testset*



GT1: a woman in a purple suit begins a course on written english
 GT2: a woman in violet t-shirts explains something on camera
 GT3: a woman is talking about a second part the college writing
SwinBERT: a woman with blonde hair is talking about something
GIT-Large(*vatex): a woman is talking about the benefits of a video
Ours: a woman is talking about **college writing**

- *video9687.mp4 MSR-VTT test set*



GT1: a man chopping lobster and taking off the shell
 GT2: a cook cracks open lobster and rubs butter onto the lobster
 GT3: a man cutting open a crab and taking the meat out to prepare a food dish
SwinBERT: a chef is showing how to marinate a chicken dish
GIT-Large(*vatex): a lobster is shown on a platter and a person is talking about it
Ours: a man in a **black shirt** is cutting a **lobster** on a cutting board

- *FsrD9In_oBM_000002_000012.mp4 VATEX validation set*



GT1: a little girl is performing gymnastic figures in a hall.
 GT2: a coach is giving some instructions to a young gymnast, and then she does a tumbling routine.
 GT3: a man coaches a young girl on how to do gymnastics moves.
SwinBERT: a girl does a series of flips on a mat in a gym
GIT-Large: a man is standing on a mat
Ours: a **young girl** does a **backflip** on a **trampoline** in a **gym**.

- *Y9YhG2CFTz0_20.000_30.000.mp4 VALOR-32k test set*



GT: In the room a child in a blue bib was laughing, and the boy laughed more as the man spoke.
SwinBERT(*vatex): a baby is sitting in a high chair and laughing
GIT-Large(*vatex): a baby is sitting in a high chair and smiling
Ours: **in the room**, a baby in a **blue bib** sat in a **white chair laughing**

Figure 2: Qualitative comparison of the proposed model with other open-source models [16, 23] on three video captioning datasets. The videos are sampled from the MSR-VTT, VATEX, and VALOR datasets. Models not trained on the corresponding datasets are annotated with their training dataset in parentheses. GT is an abbreviation for Ground Truth.

Captain America: The First Avenger(2011)



SwinBERT: a man is walking through the streets of a busy city.
GIT-Large: a man walking through a busy city.
Ours: a man is running in the street in **times square**.

Greatest Works of Art MV(2022)



SwinBERT: a man walks around a building and then a light is shown.
GIT-Large: a building that is lit up.
Ours: a city is **lit up at night** with a view of the **eiffel tower**.

Random video of an accident from the internet



SwinBERT: a person is using a machine to extinguish a fire.
GIT-Large: a group of people are playing with a tank and a fire is coming out of the ground.
Ours: a group of people are working in a building that has a lot of **fire coming out of it**.

The Shawshank Redemption(1994)



SwinBERT: a man is talking to a man who is wearing a suit and then a man is talking.
GIT-Large: a man in a uniform is directing a man in a security uniform.
Ours: a group of **police officers** are talking to each other in a **room** with **posters on the walls**.

Figure 3: Qualitative comparison of our proposed model with other open-source models [16, 23] on out-of-domain videos. The videos are sampled from the Internet, including various types (i.e., movies, music videos, surveillance footage). Ground truths are not available. All models are trained on the VATEX dataset.



Figure 4: Qualitative comparison of the proposed model with CLIP4Clip [18] on MSR-VTT testset. GT is an abbreviation for Ground Truth, which is marked by a green box.

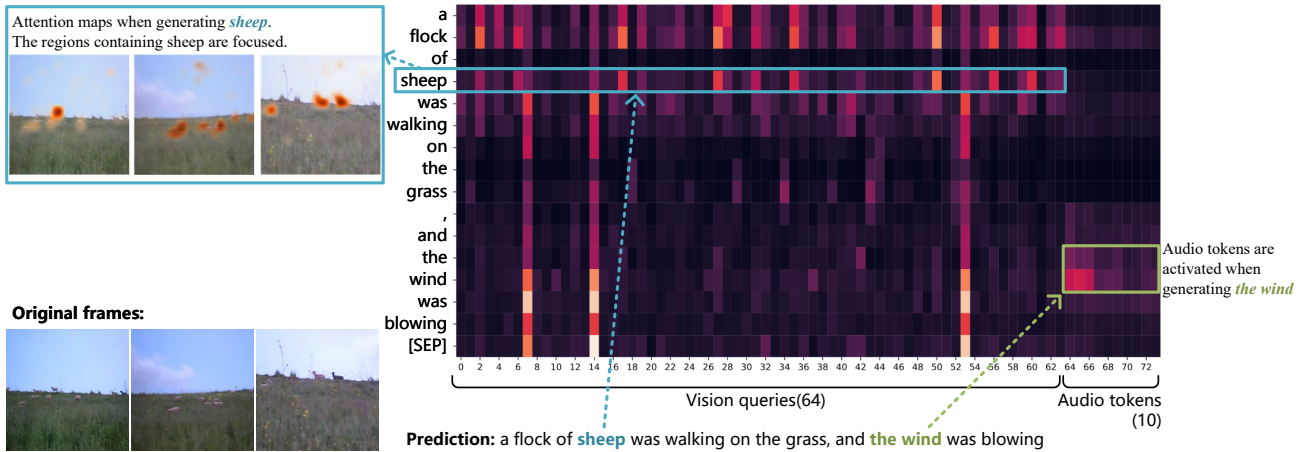


Figure 5: Visualization results of our proposed model. The heatmap matrix shows the attention applied to visual queries and audio tokens when generating each text token. The heatmap overlaid on the images displays the model's attention to various regions of the video when generating *sheep*. The original frames of the video and its prediction result are also presented.