CREMA: Multimodal Compositional Video Reasoning via Efficient Modular Adaptation and Fusion

Shoubin Yu* **Jaehong Yoon*** **Mohit Bansal**

UNC Chapel Hill

Abstract

Despite impressive advancements in multimodal compositional reasoning approaches, they are still limited in their flexibility and efficiency by processing fixed modality inputs while updating a lot of model parameters. This paper tackles these critical challenges and proposes CREMA, an efficient and modular modality-fusion framework for injecting any new modality into video reasoning. We first augment multiple informative modalities (such as, *optical flow*, *3D point cloud*, *audio*) from given videos without extra human annotation by leveraging existing pre-trained models. Next, we introduce a query transformer with multiple parameter-efficient modules associated with each accessible modality. It projects diverse modality features to the LLM token embedding space, allowing the model to integrate different data types for response generation. Furthermore, we propose a fusion module designed to compress multimodal queries, maintaining computational efficiency in the LLM while combining additional modalities. We validate our method on video-3D, video-audio, and video-language reasoning tasks and achieve better/equivalent performance against strong multimodal LLMs, including BLIP-2, 3D-LLM, and SeViLA while using 96% fewer trainable parameters. We provide extensive analyses of our CREMA, including the impact of each modality on reasoning domains, the design of the fusion module, and example visualizations.[1]

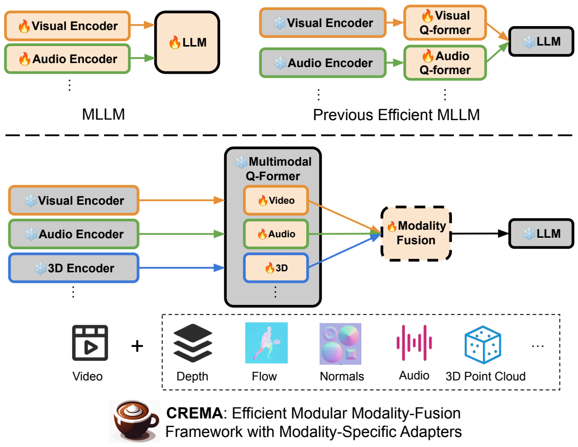

Figure 1: We present CREMA, an efficient and modular modality-fusion framework. We utilize a single multimodal Q-Former with a set of lightweight modality-specific adapters, hence allowing video frames, optical flow, 3D, etc.

1 Introduction

We humans understand the world through various senses, such as sight, sound, touch, and heat, allowing us to understand our environment and act accordingly. This concept has inspired the field of multimodal learning that connects various perceptions, including vision-language (Alayrac et al., 2022; Li et al., 2023; Zang et al., 2023; Radford et al., 2021), audio-video (Han et al., 2020; Tang et al., 2022), and 2D-3D joint vision (Li et al., 2020; Hou et al., 2021, 2023). In particular, recent Multimodal Large Language Models (MLLMs) (Yu et al., 2023b; Li et al., 2023; Liu et al., 2023a; Tang et al., 2023) have shown promising versatility in handling multiple forms of input data, such as vision, audio, and text. These models are crucial in real-world applications that require a comprehensive understanding of multiple modalities to make decisions in various contexts. For example, autonomous vehicles rely on road signs, sirens, and LIDAR for navigation and safe driving. Similarly, educational AI enhances the learning experience by integrating diverse information, such as videos, speech, and textbooks.

Despite their recent advancements, deploying a generic MLLM that handles multiple diverse modalities is still very challenging in terms of *cost* and *flexibility*. For different types of inputs, MLLMs have required extremely large computational budgets to update the LLM with individual

*Equal Contribution

[1] Code available at: https://CREMA-VideoLLM.github.io/.

encoders for modalities (Figure 1 top left). Alternatively, recent efficient MLLMs using separate projection modules (Zhang et al., 2023; Sun et al., 2023; Li et al., 2023) (Figure 1 top right) provide a more efficient and flexible way for multimodal reasoning. However, as each modality module contains hundreds of millions of parameters for training, this approach is still computationally intensive, and balancing as well as fusing various types of inputs becomes even more complex and costly when more modalities are introduced. Such challenges also exist in very recent pioneering works (Liu et al., 2023b; Panagopoulou et al., 2023; Lu et al., 2023); these models aim to integrate more diverse sensory data for compositional understanding via partial updates to the models, yet still require notable training resources to adapt to different modalities (7B for Unified-IO 2 (Lu et al., 2023)). Moreover, they focus primarily on fixed modality pairs (like 3D-text and visual-text), limiting their adaptability to new data forms and broader applications.

To overcome these limitations, our work presents *Multimodal **C**ompositional Video **R**easoning via **E**fficient **M**odular **A**daptation and Fusion (**CREMA**)*, a highly efficient yet effective multimodal LLM framework for video reasoning that extends existing vision-language models to adapt to any new set of modalities, including *video*, *depth map*, *optical flow*, *surface normals*, *audio*, *3D point cloud*, notably with very few trainable parameters ($<$ 5M for each new modality) as compared to BLIP-2 (Li et al., 2023), 3D-LLM (Hong et al., 2023) ($\sim$188M) and SeViLA (Yu et al., 2023a) ($\sim$376M). Given a frozen pre-trained vision-language backbone, our approach introduces modality-adaptive modules on top of the Q-Former (Li et al., 2023) architecture, including linear projectors, low-rank adapters (Hu et al., 2022), and learnable queries. Our parameter-efficient modular design ensures that the pre-trained backbone remains unchanged and enables updates with new modalities and more advanced LLMs in the future without complex architecture changes. To enrich the input modalities, we utilize public pre-trained models to extract features from raw videos, such as depth map, optical flow, and surface normals.

Furthermore, despite the usefulness of our compositional video reasoning framework for multimodal data, dealing with numerous modalities is not always beneficial, because some modality features may be redundant to each other or unrelated to the target reasoning tasks. Besides, the LLM needs to receive longer input contexts, which include token embeddings from all modality queries, resulting in increased computations to produce responses. Hence, to address these remaining concerns, we introduce a lightweight modality fusion module, dubbed CREMA-*Espresso*, that effectively combines various modality tokens through a novel self-gated attention. As a result, we enable the model to maintain GFLOPs while still achieving competitive performance, even when the LLM processes a larger number of modality inputs.

We validate CREMA on various video reasoning benchmarks, including conventional VideoQA (NExT-QA (Xiao et al., 2021)), as well as compositional VideoQA including 3D-QA (SQA3D (Ma et al., 2023)) and Audio-QA (MUSIC-AVQA (Li et al., 2022)) that require additional modalities beyond video and text, such as *3D point cloud* or *audio*. CREMA surpasses other modality-specific baselines, improving fine-tuning performance by **+3.3%** on SQA3D, **+1.9%** on MUSIC-AVQA, and **+0.9%** on NeXT-QA with just **2$\sim$4% of the trainable parameters and more modalities**. CREMA also outperforms general-purpose baselines in the zero-shot setting. We further provide comprehensive analyses of varying sets of modalities, different modality fusion strategies, benefits of adding more modality, and qualitative analysis with input/response visualizations to highlight the efficiency and effectiveness of our CREMA framework in compositional video reasoning.

We summarize our contributions as 4-fold:

- We propose a highly efficient and generalizable modality-extensible learning framework, coined *CREMA*, which learns multiple modality-adaptive modules to understand given data through augmented senses.
- CREMA's design allows easy embracing of new modalities by adding additional modality-adaptive modules without any need to modify the existing framework.
- We present a modality fusion module that efficiently weights modalities, integrating useful modality features into response generation.
- We demonstrate the efficacy of our CREMA on multiple video reasoning datasets by achieving better/equivalent performance while reducing about 96% of trainable parameters than BLIP-2, 3D-LLM, and SeViLA.

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 (NeurIPS)*.

Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ji Hou, Xiaoliang Dai, Zijian He, Angela Dai, and Matthias Nießner. 2023. Mask3d: Pre-training 2d vision transformers by learning masked 3d priors. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. 2021. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the International Conference on Computer Vision (ICCV)*.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. 2020. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. 2023b. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. Sqa3d: Situated question answering in 3d scenes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2023. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Fine-grained audio-visual joint representations for multimodal large language models. *arXiv preprint arXiv:2310.05863*.

Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. 2022. Tvlt: Textless vision-language transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-any generation via composable diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023a. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023b. Connecting speech encoder and large language model for asr. *arXiv preprint arXiv:2309.13963*.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2023. Contextual object detection with multimodal large language models. *arXiv preprint arXiv:2305.18279*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.