When Video Compression Meets Multimodal Large Language Models: A Unified Paradigm for Cross-Modality Video Compression

Pingping Zhang, Jinlong Li, Kecheng Chen, Meng Wang, Long Xu, Haoliang Li, Nicu Sebe, *Senior Member, IEEE*, Sam Kwong, *Fellow, IEEE*, Shiqi Wang, *Senior Member, IEEE*

.

***Abstract*—Traditional video compression methods perform well at high bitrates but struggle to preserve fine-grained semantic information at low bitrates. Recently, with the blossoming of Multimodal Large Language Models (MLLMs), Cross-modal compression techniques offer prospective solutions for improving video compression under low-bitrate conditions. In this paper, we propose a unified Cross-Modality Video Compression (CMVC) framework that integrates multimodal representations and video generative models. The encoder disentangles video into spatial and temporal components, which are mapped to compact cross-modal representations using MLLMs. During decoding, different encoding-decoding modes are employed to acquire various video reconstruction qualities, including Text-Text-to-Video (TT2V) for semantic preservation and Image-Text-to-Video (IT2V) for perceptual consistency. Additionally, we elaborate on an efficient frame interpolation model using Low-Rank Adaptation (LoRA) to improve the perceptual quality. Experimental results demonstrate that TT2V achieves effective semantic reconstruction, while IT2V ensures competitive perceptual consistency. These findings suggest the potential of leveraging multimodal priors to improve video compression, offering promising future research directions.**

***Index Terms*—Video, multimodal representations, semantic reconstruction, multimodal large language Models.**

I. Introduction

TRADITIONAL video compression [1]–[4] has primarily focused on signal-level reconstruction, optimizing components such as pixel intensities and motion vectors during encoding and decoding. This approach has led to substantial advancements in high-bitrate scenarios, as demonstrated by classical codecs such as H.264/Advanced Video Coding (AVC) [1], High Efficiency Video Coding (HEVC) [2], and Versatile Video Coding (VVC) [3], as well as by most existing learning-based methods like DCVC [5]. These methods predominantly focus on optimizing pixel-level fidelity metrics, leading to excellent results in high-bitrate compression. However, these conventional methods often encounter limitations in low-bitrate scenarios, particularly when it comes to preserving fine-grained semantic information, crucial for applications such as disaster response and remote surveillance. In such contexts, reconstructed video is prone to the loss or blurring of key features.

To address these shortcomings at low bitrates, cross-modal compression (CMC) methods have emerged [6]–[8], which harness the complementary strengths of multiple modalities to enhance image representations under constrained bitrate conditions. Notable approaches, such as VR-CMC [6], SCMC [7], MCM [8], and CMC-Bench [9], exploit the transformation of images into textual formats (I2T) to generate semantically similar content. This transformation not only preserves crucial semantic information but also achieves significant reductions in storage requirements, resulting in higher compression ratios compared to conventional image data representations. However, the potential of cross-modal representations for video compression, particularly in maintaining both semantic and perceptual quality at low bitrates, remains relatively underexplored.

Unlike image compression, which is primarily interested in spatial information, video compression additionally involves the representation of temporal dynamics, significantly increasing its complexity. In this regard, Multimodal Large Language Models (MLLMs) [10], [11] offer unique advantages for video compression due to their ability to understand and interpret temporal relationships in video. Recent studies [12] show that MLLMs are effective at modeling sequential data and capturing dependencies across time. By leveraging these strengths, MLLMs can produce compact and semantically meaningful textual representations of videos, enabling efficient compression while maintaining high reconstruction quality even at low bitrates. Therefore, MLLMs represent a promising direction for achieving both compression efficiency and semantic fidelity.

Motivated by the advantages of MLLMs, we propose a new Cross-Modality Video Compression (CMVC) scheme aimed at optimizing the representation of both spatial and temporal components, thereby enabling high-fidelity semantic and perceptual reconstruction at low bitrates. Capitalizing on the rich potential of multimodal representations, this frame-

Pingping Zhang and Shiqi Wang are with the Department of Computer Science, City University of Hong Kong, Hong Kong, China 999077 (email: ppingyes@gmail.com; shiqwang@cityu.edu.hk).

Jinlong Li and Nicu Sebe are with the Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy. (email: tyronejinlongli@gmail.com; sebe@disi.unitn.it).

Kecheng Chen and Haoliang Li are with the Department of Electrical Engineering and the Center for Intelligent Multidimensional Data Analysis, City University of Hong Kong, Hong Kong, China 999077. (e-mail: cs.ckc96@gmail.com; haoliang.li@cityu.edu.hk.)

Long Xu is with Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China 315211 (lxu@nao.cas.cn)

Meng Wang and Sam Kwong are with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong, China 999077 (email: mengwang7@ln.edu.hk; samkwong@ln.edu.hk)

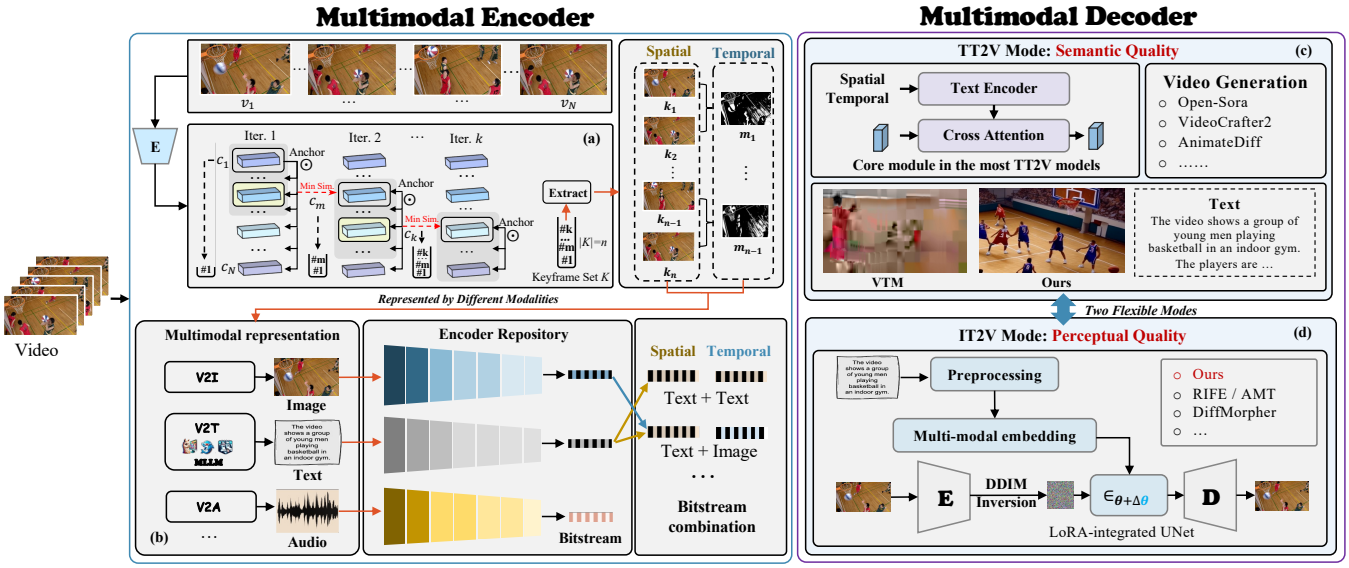

Fig. 1. The framework of the proposed **CMVC** scheme. This framework operates by first segmenting the video into distinct clips using a keyframe selection strategy (a), allowing for the extraction of both spatial and temporal components from each video segment. Subsequently, MLLMs are employed to generate multimodal representations of these components. For instance, spatial information can be represented through text or images, while temporal dynamics may be encoded using text or audio modalities. These multimodal representations are then encoded via their respective encoders, resulting in compressed bitstreams for each component. The bitstreams corresponding to different components are then combined and transmitted to the decoder. In the decoder, we provide two exemplary modes, including TT2V (c) and IT2V (d) modes, for video reconstruction. In this framework, both semantic and perceptual quality at relatively high compression ratios can be preserved effectively for the video compression task.

work supports the development of diverse encoding-decoding modes tailored to specific reconstruction requirements. We propose two exemplary modes alternatively: TT2V (Text-Text-to-Video) mode for semantic reconstruction at ultra-low bitrate (ULB) and IT2V (Image-Text-to-Video) mode for perceptual reconstruction at extremely low bitrate (ELB). In the TT2V mode, inspired by the workflow of Cross-Modality Image Compression (CMIC) [6], [7], [9] and the stunning generation capability of existing TT2V models for video reconstruction, we first extract the representative text constructed with our keyframes selection strategy, effectively encoding video content as the spatial component and motion as the temporal component. Then, the video generation model is utilized to regenerate the video conditioned on textual representations. The rationale behind this strategy lies in that compact yet effective text representations from the encoder encapsulate semantic details to enable high-quality semantic reconstruction for the decoder. Different from TT2V, the IT2V mode is designed to enhance perceptual reconstruction, since images provide richer visual contexts compared to text, benefiting perceptual consistency. This is achieved by inputting similar text representations with the TT2V mode and extra selected keyframes from the encoder to the decoder for better perceptual video reconstruction. To further improve perceptual smoothness across consecutive frames, an efficient adaptation tuning in a frame-interpolation manner via Low-Rank Adaptation (LoRA) tuning is tailored to fully exploit the semantic cues and visual contexts from both input texts and keyframes to facilitate high-quality perceptual consistency for video reconstruction. This comprehensive paradigm adeptly accommodates diverse modality representations within video compression, by tapping into foundational MLLMs and video generation models, which sheds light on future video compression works. Our contributions are threefold:

- To the best of our knowledge, the proposed unified paradigm for CMVC is the first to leverage MLLMs and video generation models for video compression.
- We elaborate multiple encoding-decoding modes to achieve optimal trade-off video reconstruction quality for specified decoding requirements, including TT2V mode to ensure high-quality semantic information and IT2V mode to achieve superb perceptual consistency.
- Extensive experiments demonstrate that the proposed CMVC pipeline obtains competitive video reconstructions on HEVC Class B, C, D, E, UVG, and MCL-JCV benchmarks while maintaining high compression ratios.

II. CMVC Scheme

A. Overview

We propose a CMVC paradigm for efficient video compression with high semantic and perceptual quality, especially at low bitrates, as illustrated in Fig. 1. Given a video $V \in v_i$, where v_i denotes video frames and $i \in \{1, \ldots, N\}$, which consists of spatial (keyframe) and temporal (motion) components, the goal is to compress these components into compact multimodal representations. We leverage MLLMs, specifically Video-to-Text (V2T) models, to map both keyframes and motion into textual representations, which are then encoded using specialized encoders. These multimodal representations are then compressed using dedicated encoders, yielding compressed representations of keyframe and motion. The video is reconstructed by a decoder operating in one of two modes: TT2V, which prioritizes semantic consistency, and IT2V, which focuses on perceptual quality. This approach enables high compression ratios while preserving both semantic information and perceptual quality.

B. CMVC Encoder

As shown in Fig. 1, we focus on efficiently representing spatial and temporal information of videos through keyframes and motion. We first extract the keyframes through the keyframe selection strategy, which can be found in the supplementary material. After that, the keyframes are denoted as $K = \{k_1, k_2, ..., k_n\}$, where each k_i is a keyframe. The motion information m_j between two consecutive keyframes is represented by $M = \{m_1, m_2, ..., m_{n-1}\}$. Keyframes and motion are transformed into multimodality representations as follows: $T_{k,i} = f(k_i)$ and $T_{m,j} = g(m_j)$. Here, $f(*)$ and $g(*)$ denote the process of cross-modality representation for keyframes and motion, respectively. As illustrated in Fig. 1, keyframes and motion can be transformed into textual and visual representations. Thus, the total bitrate is given by:

$$R_{total} = \sum_{i=1}^{n} R_k(T_{k,i}) + \sum_{j=1}^{n-1} R_m(T_{m,j}), \tag{1}$$

where R_k and R_m are the entropy coding modules for keyframes and motion, respectively. The bitrates can be adjusted by n and the compression ratio of keyframes and motion.

C. CMVC Decoder

In the decoder, we utilize the decoded keyframe $\hat{K}$ and motion $\hat{M}$ to achieve video reconstruction, as follows:

$$\hat{V} = \mathcal{G}(\hat{K}, \hat{M}), \tag{2}$$

where $\mathcal{G}(*)$ is a video generation model and $\hat{V}$ is the reconstructed video. Based on different modality representations for keyframe and motion, we designed two modes, including the TT2V mode and the IT2V mode for ULB and ELB coding, respectively.

In the TT2V mode, we utilize state-of-the-art (SoTA) video generation models to reconstruct videos from decoded keyframes and motion descriptions. Leveraging advancements in these models, we optimize semantic reconstruction, with our results showing that more detailed descriptions yield higher bitrates and improved semantic quality. In the IT2V mode, keyframe images and motion descriptions are integrated to enhance perceptual quality. In addition to employing existing IT2V models, we propose a generative model utilizing LoRA tuning to ensure superior perceptual consistency at ELB.

The IT2V mode is designed to obtain a reconstructed video conditioning on keyframe images and the text of motion. Thus, we propose a IT2V generative model, which generates a video clip according to two keyframe images (I_0 and I_1) and the description of the motion of this video clip. Specifically, we adopt a stable diffusion model (SD) with LoRA, which fine-tunes the model parameters θ by training a low-rank residual component $\triangle\theta$. This residual can be decomposed into products of low-rank matrices. LoRA demonstrates significant efficiency in generating various samples while maintaining consistent semantic identity across different latent noise traversals. The proposed IT2V model is shown in Fig. 2. We first train two LoRAs ($\Delta\theta_0$ and $\Delta\theta_1$) on the SD UNet ϵ_θ for each of the two images I_0 and I_1. The learning objective of $\Delta\theta_i (i = 0, 1)$ is:

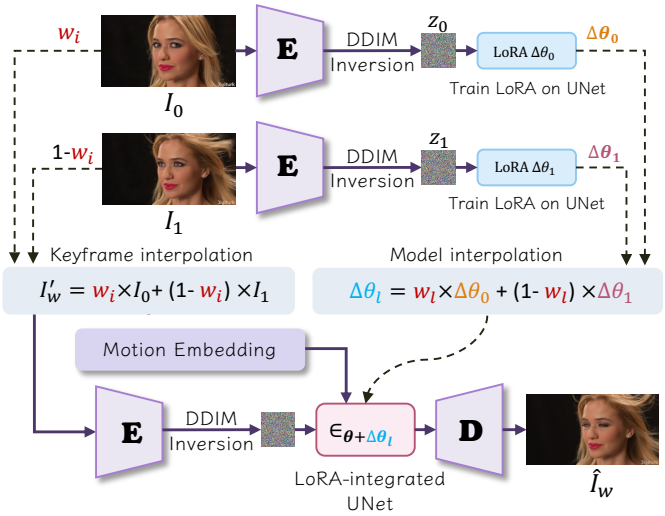

Fig. 2. The workflow of the IT2V generative model. Two LoRAs are trained to fit the two keyframe images (I_0 and I_1), respectively. To generate w-th frame between I_0 and I_1, we interpolate I'_w and the LoRA parameters according to the weights w_i and w_l. Denoising Diffusion Implicit Models (DDIM) inversion is a technique used in diffusion models to reverse the generation process.

$$\mathcal{L}(\Delta\theta_i) = \mathbb{E}_{\epsilon,t}\left[\left\|\epsilon - \epsilon_{\theta+\Delta\theta_i}(\mathbf{z}_{t,i}, t, \mathbf{c}_i)\right\|^2\right], \tag{3}$$

where $\mathbf{z}_{t,i} = \sqrt{\bar{\alpha}_t}\hat{\mathbf{z}}_i + \sqrt{1-\bar{\alpha}_t}\epsilon$ is the noised latent embedding at diffusion step t. $\hat{\mathbf{z}}_i$ is the VAE encoded latent of the I_i image. $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the random sampled Gaussian noise. $\mathbf{c}_i$ is the motion embedding encoded from the motion prompt, which describes the movement between two reconstructed frames. $\epsilon_{\theta+\Delta\theta_i}$ represents the LoRA-integrated UNet. The fine-tuning objective is optimized separately via gradient descent in $\Delta\theta_0$ and $\Delta\theta_1$.

III. Experiments

A. Experimental settings

Datasets. The datasets, including HEVC Class B, C, D, and E, as well as UVG and MCL-JCV, are extensively utilized for evaluating both traditional and neural video codecs. These datasets vary in resolution and content, providing a diverse range of scenarios for comprehensive assessment. To ensure compatibility with various video codecs, we resize videos to dimensions that are multiples of 64 for both width and height.
Comparison methods. We choose two prominent models, namely VideoLLaVA [10] and VideoLLaMA [11], to extract semantic information from videos. This process aligns with the V2T stage depicted in Fig. 1, where the selected models play an important role in extracting semantic descriptions for keyframes and motion. In the TT2V mode, numerous video generation models are available. In this context, we employ advanced video generation models for the purpose of generating videos based on textual input. In the IT2V mode, we conduct a comparative analysis of existing traditional video codecs, such as x264, x265, and VTM [3]. Alongside this, we evaluate our method against deep video codecs such as DCVC [5] and DCVC-DC [13], but these codecs encounter challenges in achieving extremely low bitrate coding.

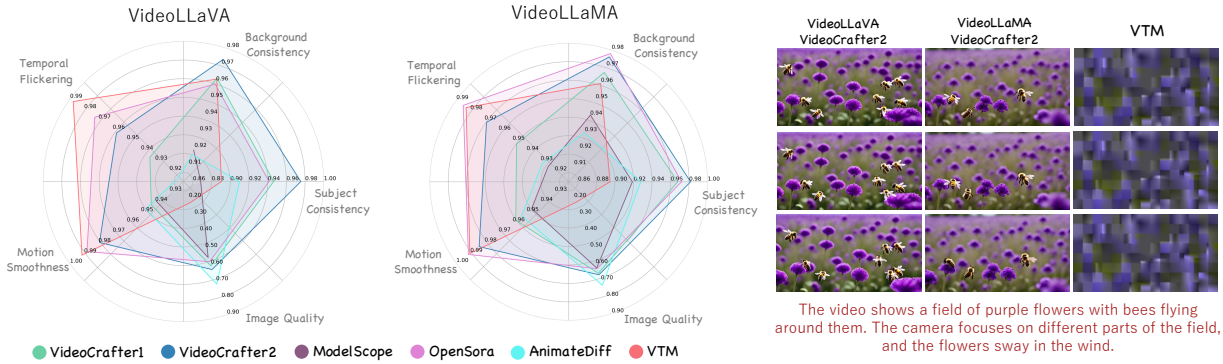

Fig. 3. Left: Comparison results of combination of different V2T models (VideoLLaVA and VideoLLaMA) and TT2V models (VideoCrafter1, VideoCrafter2, ModelScope, OpenSora and AnimateDiff). Right: Visual quality comparison of the TT2V mode and VTM. At ULB, our proposed TT2V mode successfully preserves the semantic quality of the videos. In contrast, VTM brings significant blocking artifacts, which impedes the effective conveyance of semantic information in videos.

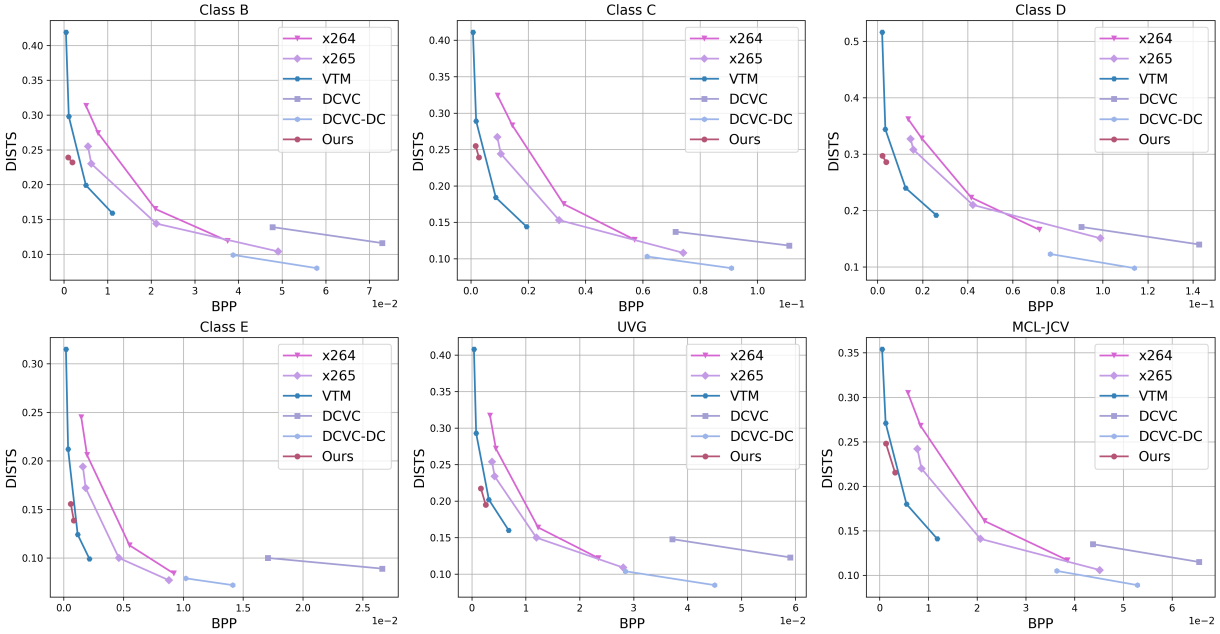

Fig. 4. The R-D performance comparison in the IT2V mode. The comparisons are performed on the Class B, Class C, Class D, Class E, UVG, and MCL-JCV.

B. Experimental results

Comparison in the TT2V mode. We conduct a comparative analysis of two V2T models, VideoLLaVA and VideoLLaMA, both of which are the SoTA MLLMs. Subsequently, we extend our comparison to include five video generation models: VideoCrafter1, VideoCrafter2, ModelScope, OpenSora, and AnimateDiff. Furthermore, we compare these models against VVC Test Model (VTM) at QP=63, which results in a higher bitrate than our proposed scheme. The reported values correspond to the average performance across all testing datasets, and all evaluation values are computed following VBench [14]. The results, illustrated in Fig. 3, show that the TT2V generation models achieve better performance than VTM, exhibiting enhanced frame-wise quality and greater consistency in both background and subject representation.

Comparison in the IT2V mode. We compare our model with traditional codecs (x264, x265, and VTM) as well as deep video codecs (DCVC and DCVC-DC). As presented in Fig. 4, we use DISTS to evaluate perceptual quality. Additional comparisons with other evaluation metrics, such as LPIPS, FID, and PSNR, are provided in the supplementary material. However, the pretrained models provided by deep video codecs have limitations in achieving ELB. We further compare our proposed model against various video generation methods, with detailed results provided in the supplementary material.

IV. Conclusion

We propose a CMVC paradigm that represents a promising advancement in video compression technology. This framework effectively tackles the challenges of preserving semantic integrity and perceptual consistency at ULB and ELB. By leveraging MLLMs and cross-modality representation techniques, the proposed CMVC framework disentangles videos into content and motion components, transforming them into different modalities for efficient compression and reconstruction. Through the TT2V and IT2V modes, CMVC achieves a balance between semantic information and perceptual quality, offering a comprehensive solution at high compression ratios.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h. 264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[4] L. Zhu, Y. Zhang, N. Li, W. Wu, S. Wang, and S. Kwong, "Neural network based multi-level in-loop filtering for versatile video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 12 092–12 096, 2024.

[5] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," in *Conference on Neural Information Processing Systems*, vol. 34, 2021, pp. 0–12.

[6] J. Li, C. Jia, X. Zhang, S. Ma, and W. Gao, "Cross modal compression: Towards human-comprehensible semantic compression," in *ACM International Conference on Multimedia*, 2021, pp. 4230–4238.

[7] P. Zhang, S. Wang, M. Wang, J. Li, X. Wang, and S. Kwong, "Rethinking semantic image compression: Scalable representation with cross-modality transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4441–4445, 2023.

[8] J. Gao, J. Li, C. Jia, S. Wang, S. Ma, and W. Gao, "Cross modal compression with variable rate prompt," *IEEE Transactions on Multimedia*, vol. 26, pp. 3444–3456, 2024.

[9] C. Li, X. Wu, H. Wu, D. Feng, Z. Zhang, G. Lu, X. Min, X. Liu, G. Zhai, and W. Lin, "Cmc-bench: Towards a new paradigm of visual signal compression," *arXiv preprint arXiv:2406.09356*, 2024.

[10] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[11] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[12] B. Li, B. Chen, Z. Wang, S. Wang, and Y. Ye, "Semantic face compression for metaverse: A compact 3d descriptor based approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8978–8982, 2024.

[13] J. Li, B. Li, and Y. Lu, "Neural video compression with diverse contexts," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 616–22 626.

[14] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu, "VBench: Comprehensive benchmark suite for video generative models," in *IEEE/CVF International Conference on Computer Vision*, 2024.