USTC-TD: A Test Dataset and Benchmark for Image and Video Coding in 2020s

Zhuoyuan Li, Junqi Liao, Chuanbo Tang, Haotian Zhang, Yuqi Li, Yifan Bian, Xihua Sheng, Xinmin Feng, Yao Li, Changsheng Gao, Li Li, *Member, IEEE,* Dong Liu, *Senior Member, IEEE,* and Feng Wu, *Fellow, IEEE*

***Abstract*—Image/video coding has been a remarkable research area for both academia and industry for many years. Testing datasets, especially high-quality image/video datasets are desirable for the justified evaluation of coding-related research, practical applications, and standardization activities. We put forward a test dataset namely USTC-TD, which has been successfully adopted in the practical end-to-end image/video coding challenge of the *IEEE International Conference on Visual Communications and Image Processing (VCIP)* in 2022 and 2023. USTC-TD contains 40 images at 4K spatial resolution and 10 video sequences at 1080p spatial resolution, featuring various content due to the diverse environmental factors (*e.g.* scene type, texture, motion, view) and the designed imaging factors (*e.g.* illumination, lens, shadow). We quantitatively evaluate USTC-TD on different image/video features (spatial, temporal, color, lightness), and compare it with the previous image/video test datasets, which verifies its excellent compensation for the shortcomings of existing datasets. We also evaluate both classic standardized and recently learned image/video coding schemes on USTC-TD using objective quality metrics (PSNR, MS-SSIM, VMAF) and subjective quality metric (MOS), providing an extensive benchmark for these evaluated schemes. Based on the characteristics and specific design of the proposed test dataset, we analyze the benchmark performance and shed light on the future research and development of image/video coding. All the data are released online: https://esakak.github.io/USTC-TD .**

***Index Terms*—Benchmark, image coding, standardization, test dataset, video coding.**

I. Introduction

Nowadays, with the dramatic growth of data traffic over the internet and the emergent application of versatile image/video formats such as 2K, 4K, high dynamic range, and wide color gamut, there is a pressing demand for storage and transmission. To address this challenge, in recent decades, image/video compression is employed to reduce the amount of data significantly, and several video coding standards have been developed, such as High Efficiency Video Coding (H.265/HEVC) [1], Versatile Video Coding (H.266/VVC) [2], Audio Video Standard (AVS1, AVS2, AVS3) [3], AOMedia Video 1 and 2 (AV1 [4], AV2 [5]).

End-to-end image/video compression has been a research focus on visual data compression for both academia and industry for over six years [6]–[31]. A number of technologies have been developed, such as expressive auto-encoder neural networks, precise probability estimation neural networks, and conditional end-to-end coding frameworks, and so on. Until recently, the performances of both end-to-end image and video compression schemes have surpassed that of the advanced H.266/VVC under certain test conditions [10], [11], [26], [27].

For the evaluation of these image/video compression schemes in the practical application and standardization, they are usually benchmarked with objective and subjective quality metrics to evaluate their rate-distortion (RD) performance and a trade-off between coding efficiency and reconstructed quality. To sufficiently consider the effectiveness of these quality assessments, the test results of representative image/video test datasets are the key to reflecting the practicability and generalization of the researcher's scheme.

In this paper, a new image/video dataset, named USTC-TD, is proposed for testing and evaluating the practical image/video coding algorithms. USTC-TD contains 40 images and 10 video sequences with a wide content coverage. For the image dataset, each image is captured with a high spatial resolution (4K) and converted into RGB, YUV444/420 color space, and PNG/YUV file format. For the video dataset, each video sequence consists of 96/300 frames, and each frame is captured at 30 frames per second (fps) with 1080p spatial resolution and converted into RGB, YUV444/420 color space, and PNG/YUV file format. For the construction of image and video datasets, the data is collected with the specific design of different content factors (environmental/imaging-related factors), which aims to cover as close as possible to the real-world coding transmission scenes. Compared with the common test image/video datasets [32]–[38], we use different quantitative criteria to comprehensively evaluate the diversity of the proposed USTC-TD from the perspective of spatial, temporal, colorfulness, and lightness information, demonstrating its excellent coverage and effectively compensating for the shortcomings of existing datasets.

In addition, we establish baselines and evaluate the classic standardized compression schemes [1]–[5] and recently learned image/video compression schemes [6]–[12], [16], [19]–[23], [25]–[31] under objective quality metrics (*PSNR*, *MS-SSIM* [39], *VMAF*[1] [40]) and subjective quality metric (*MOS*) [41]–[44]), and then benchmark and analyze their performance on the proposed dataset to shed light on future research and development of image/video coding. The benchmark data and test scripts are available online with the proposed dataset and released on the open-sourced website for researchers to reproduce conveniently.

Date of current version March 13, 2025. This work was supported by the Natural Science Foundation of China under Grant 62021001. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC. *(Z. Li and J. Liao contributed equally to this work.) (Corresponding authors: Dong Liu and Feng Wu.)*

The authors are with the MOE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei 230027, China (e-mail: {zhuoyuanli, liaojq, cbtang, zhanghaotian, lyq010303, esakak, xhsheng, xmfeng2000, mrliyao}@mail.ustc.edu.cn; {changshenggao, lil1, dongeliu, fengwu}@ustc.edu.cn).

[1]Available online at https://github.com/Netflix/vmaf.

TABLE I
COMMON TEST DATASETS OF IMAGE COMPRESSION

Dataset	Resolution	Number	Color Space	Bit Depth	Setting	Characteristic
Kodak [32]	768×512	24	RGB	24	-	Rich Texture
Tecnick [33]	1200×1200	100	RGB	24	Sampling	Various Scenery
CLIC [34]	1189×1789 (AVG)	41	RGB	24	Valid-Professional	Appropriate Exposure

TABLE II
COMMON TEST DATASETS OF VIDEO COMPRESSION

Dataset	Resolution	Number	FPS	Length	Characteristic
UVG [35]	3840×2160	16	50/120	5s-12s	Fast/Slow Motion
MCL-JCV [36]	1920×1080	30	-	5s	Diverse Video Scenes
HEVC CTC [37], ***VVC CTC*** [38]	240P-4K	41	24-60	150-600 Frames	Appropriate Exposure

We hope the proposed test datasets allow researchers to make more well-informed decisions under efficient evaluation, and guide the innovation and improvement of future schemes and experiments. In summary, our contributions are as follows:

- We build a new image/video compression test dataset (named USTC-TD), which focuses on the diversity of various content factors.
- We conduct a comprehensive evaluation of the proposed dataset by using different quantitative criteria, demonstrating the excellent compensation of USTC-TD for existing image/video datasets.
- We conduct a comprehensive evaluation of the advanced image/video compression schemes on the proposed dataset, and establish an extended baseline for the evaluative image/video coding schemes benchmarked on USTC-TD.
- Taking a close look at USTC-TD, we analyze the benchmarked performance and shed light on the future research and development of image/video coding.

The remainder of the paper is structured as follows. Section II mentions the background of previous compression-related test datasets. Section III summarises the data collection process of the proposed dataset. Section IV introduces the construction of the image and video dataset of USTC-TD, and discusses the characteristics and utilization of the proposed dataset. Section V presents the experimental configuration and the evaluation of the advanced compression schemes on USTC-TD, and further analyzes their performance, limitation, and inspiration. Section VI concludes the paper and presents some suggestions for future work.

II. BACKGROUND

In the past twenty years, with the rapid development of multimedia data over the advanced exhibition devices, resolutions, frame rates, dynamic range and viewpoints, the transmission/storage quantity of multimedia data is progressively accompanied by dramatic increases in the requirement of users. As the powerful multimedia data transmission/storage tool, lossy/lossless image/video compression has become the primary driver for reducing the internet bandwidth and storage. For the standardization activities and research of compression-related systems, image/video test (evaluation) dataset is a critical component for optimizing the performance and reflecting the practicability and generalization of different compression schemes. Here we review the image/video test dataset commonly used by standards and researchers in the past, and summarize their characteristics.

Image Compression Test Dataset. For the evaluation of previous image compression schemes, *Kodak* [2] [32], *Tecnick* [33] (sampling setting), *CLIC* (professional setting) [34] are commonly used, the setting is mentioned in Table I, and the introduction is summarised below:

- ***Kodak*** [32] is a commonly used true color set of images released for various testing purposes and benchmarks, it contains 24 images with RGB format. The images are all photographic type and continuous tone. Many sites use them as a standard test suite for compression testing.
- ***Tecnick*** [33] is a huge collection of sample images designed for quality assessment of different kinds of displays and image processing techniques. The sampling setting is widely used on testing resampling algorithms.
- ***CLIC*** [34] is a high-quality image set collected from *Unsplash* [3], and contains images of similar quality from potentially different sources. It has been successfully applied in the workshop and challenge on learned image compression (*CLIC*) of *IEEE/CVF Computer Vision and Pattern Recognition Conference* (*CVPR*) and *Data Compression Conference* (*DCC*).

In these image datasets, characteristics mainly focus on the different resolutions with more scene types, most of the test images are captured by high-definition lens in specific scenes. These datasets aim to evaluate the basic ability of image compression algorithms to remove intra-frame redundancy for different scenarios, but the limited diversity of content factors makes it difficult to evaluate the robustness of algorithms.

Video Compression Test Dataset. For the evaluation of previous video compression schemes, *UVG* [35], *MCL-JCV* [36], *HEVC Common Test Conditions* (*CTC*) [37], *VVC CTC* [38] are commonly used, the settings are mentioned in the Table II, and the introduction is summarised below:

- ***UVG*** [35] contains 16 test video sequences. They are captured with Sony F65 video camera in 16-bit F65RAW-HFR format and converted to YUV420 videos by *ffmpeg* tool [4]. It is widely used in the evaluation of advanced learned video compression methods [12], [16], [19]–[23], [25]–[31].
- ***MCL-JCV*** [36] is a compressed video quality assessment dataset based on the just noticeable difference (JND) model. All provided video sequences are available to the public with measured raw JND data for each test subject and allow users to do their own processing.

[2] Available online at https://r0k.us/graphics/kodak/.
[3] Available online at https://unsplash.com/.
[4] Available online at https://ffmpeg.org/.

TABLE III
CAMERA CAPTURED PARAMETERS OF USTC-TD 2022

***Nikon-D3200* Specifications**	
Sensor Type	CMOS (Nikon DX format)
Sensor Size	23.2mm×15.4mm
Effective Pixels	24.7million
Largest Image Size	6016×4000

TABLE IV
CAMERA CAPTURED PARAMETERS OF USTC-TD 2023

***Nikon-Z-fc* Specifications**	
Sensor Type	CMOS (Nikon DX format)
Sensor Size	23.5mm×15.7mm (APS-C)
Effective Pixels	20.9million
Largest Image Size	5568×3712

- ***HEVC CTC*, and *VVC CTC*** [37], [38] define the common test conditions and test sequences for the standardization activities of H.265/HEVC [1] and H.266/VVC [2], and protect the core experiments in a well-defined rule. It promotes the upgrading of many technologies in standardization, and has been widely used in compression-related systems.

In these video test datasets, the characteristics mainly focus on the various video contents, including simple/complex motion and poor/high capture quality. Most of the test videos can only evaluate the basic ability of video compression-related algorithms to remove inter-frame redundancy for different scenarios under different video coding configurations, like motion estimation (ME), motion compensation (MC), and rate allocation/control (RC) technologies in low-delay (LD) and random access (RA) configurations, but these video contents with the limited types of temporal correlation make it difficult to evaluate the robustness of temporal property-related algorithms in video-based compression applications.

III. DATA COLLECTION

In this section, we introduce the hardware, format and collection configuration of dataset collection.

A. Camera and Format Configuration

The images and video sequences are captured by using *Nikon-D3200* and *Nikon-Z-fc* for USTC-TD 2022 and 2023 datasets, and the specific camera parameters are shown in Table III and Table IV. For the format of images and video sequences in the dataset, they are transcoded from Raw camera format (DNG, MOV) and then converted to RGB, YUV444/420 color space/format by using the *ffmpeg* tool and the conversion standard of color space (BT.601 [45]).

B. Collection Configuration

To develop a comprehensive and diverse image/video dataset, we consider the various content factors of collected data, including the environmental factors (*e.g.* scene type, texture, motion, view), imaging factors (*e.g.* illumination, lens, shadow), which cover as close as possible to the real-world coding transmission scenes. For each factor, we categorize it into different types implicit in the construction of our dataset, the categories are shown in Table V. According to these conditions, we choose more than twenty different scene types (*e.g.* dormitory, library, river bank, institutes, parks, classroom, street, vehicles), and adjust different camera parameters to capture. For each image, we take ten shots of the same scene at the same time to select the best one. For each video, we record five minutes for each scene with the same range of joint sense to select the short and long test sequences.

TABLE V
COLLECTION CONFIGURATION OF USTC-TD 2022 AND 2023

Collection Configuration		
Element	**Category**	**Example**
Texture	Structural, Natural, Geometric	Scenery, People, Gridding
Motion	Complex, Medium, Tiny	Occlusion, Walking, Chatting
View	Upward Level, Horizontal Level, Overhead Level	Building, People, Close Shot
Illumination	Appropriate Exposure, Underexposure, Overexposure	Natural Light, Dark Light, High Light
Lens	Moving, Fix	Camera Motion, Surveillance
Shadow	Hard, Soft, Cast	Camera Flash Lamp, Natural Illumination, Building Occlusion

IV. DATASET CONSTRUCTION, ANALYSIS, DISCUSSION

In this section, we introduce the construction of our proposed USTC-TD image and video datasets, and further analyze them based on the comparison with previous common test image/video datasets under different quantitative criteria.

A. Construction of USTC-TD 2022 and 2023 Image Dataset

Based on the characteristics of previous image datasets [32]–[34], our proposed dataset aims to cover various scenarios. Considering the various content factors, we combine different environmental factors and imaging factors in the collection process. For the diversity of environmental factors, we consider the scene type, texture, and view factors. For the diversity of imaging factors, we consider the resolution, illumination, and shadow factors. In Fig. 1, we show all collected image data of USTD-TD 2022, and in Table VI, we show the specific configuration of each image, and make it convenient for the researchers' scheme design for different application scenes. The collected image data and configuration of USTC-TD 2023 are also mentioned in Fig. 1 and Table VII. Based on USTD-TD 2022, USTC-TD 2023 considers more extreme factors in real-world scenes.

Compared to the previous image datasets [32]–[34], more specific content factors are considered in our dataset. For example, in *USTC-2022-09* and *USTC-2022-05*, we capture the low-light image with underexposure and high-light image with overexposure, which is a challenge for the generalization of many researchers' image compression schemes [46], [47]. In *USTC-2023-16*, we capture the scenes with the object occlusion and the spatial-wise correlation becomes low, which is a challenge for traditional intra-prediction schemes [48] in the traditional codec [1]–[5]. We hope these specific testing sets can help the researchers discover the problem related to spatial characteristics in their image compression scheme.

Fig. 1. Illustration of the image dataset in USTC-TD 2022 and 2023.

TABLE VI
THE CONFIGURATION OF USTC-TD 2022 IMAGE DATASET

Images	Scene Type	Resolutions	Texture	Illumination	View	Shadow
USTC-2022-01	Scenery	4096×2160	Geometric	Appropriate Exposure	Horizontal Level	Soft
USTC-2022-02	Scenery	4096×2160	Structural	Appropriate Exposure	Upward Level	Soft
USTC-2022-03	Scenery	4096×2160	Structural	Underexposure	Horizontal Level	Hard
USTC-2022-04	Scenery	4096×2160	Structural	Appropriate Exposure	Horizontal Level	Soft
USTC-2022-05	People	4096×2160	Natural	Overexposure	Horizontal Level	Hard
USTC-2022-06	Scenery	4096×2160	Natural, Geometric	Appropriate Exposure	Upward Level	Cast
USTC-2022-07	Building	4096×2160	Geometric	Appropriate Exposure	Upward Level	Cast
USTC-2022-08	Scenery	4096×2160	Geometric	Appropriate Exposure	Upward Level	Cast
USTC-2022-09	People, Building	4096×2160	Natural	Underexposure	Horizontal Level	Hard
USTC-2022-10	Building	4096×2160	Geometric	Appropriate Exposure	Upward Level	Soft
USTC-2022-11	People, Scenery	4096×2160	Natural	Appropriate Exposure	Horizontal Level	Soft
USTC-2022-12	People	4096×2160	Natural	Underexposure	Horizontal Level	Hard
USTC-2022-13	Scenery	4096×2160	Nature, Structural	Appropriate Exposure	Upward Level	Cast
USTC-2022-14	People, Vehicle	4096×2160	Natural	Underexposure	Horizontal Level	Cast
USTC-2022-15	People	4096×2160	Natural	Underexposure	Upward Level	Hard
USTC-2022-16	People	4096×2160	Natural	Appropriate Exposure	Horizontal Level	Soft
USTC-2022-17	People, Building	4096×2160	Natural	Overexposure	Horizontal Level	Hard
USTC-2022-18	People, Building	4096×2160	Natural, Geometric	Appropriate Exposure	Horizontal Level	Soft
USTC-2022-19	People, Building	4096×2160	Natural, Geometric	Overexposure	Horizontal Level	Soft
USTC-2022-20	People, River	4096×2160	Natural	Appropriate Exposure	Overhead Level	Cast

TABLE VII
THE CONFIGURATION OF USTC-TD 2023 IMAGE DATASET

Images	Scene Type	Resolutions	Texture	Illumination	View	Shadow
USTC-2023-01	People, Room	3840×2160	Natural, Structural	Appropriate Exposure	Horizontal Level	Hard
USTC-2023-02	Scenery	3840×2160	Geometric, Structural	Underexposure	Overhead Level	Cast
USTC-2023-03	Scenery	3840×2160	Natural, Structural	Underexposure	Horizontal Level	Soft
USTC-2023-04	Bicycle, Dense Objects	3840×2160	Geometric	Appropriate Exposure	Horizontal Level	Soft
USTC-2023-05	Plant, Dense Textures	3840×2160	Structural	Overexposure	Horizontal Level	Soft
USTC-2023-06	Water Wave	3840×2160	Geometric	Appropriate Exposure	Overhead Level	Cast
USTC-2023-07	Building	3840×2160	Geometric	Overexposure	Upward Level	Soft
USTC-2023-08	Plant, Dense Textures	3840×2160	Structural	Appropriate Exposure	Overhead Level	Soft
USTC-2023-09	Scenery, People	3840×2160	Natural	Underexposure	Horizontal Level	Hard
USTC-2023-10	People	3840×2160	Natural	Overexposure	Overhead Level	Soft
USTC-2023-11	People	3840×2160	Natural	Overexposure	Horizontal Level	Soft
USTC-2023-12	People	3840×2160	Natural	Underexposure	Upward Level	Soft
USTC-2023-13	People	3840×2160	Natural	Appropriate Exposure	Horizontal Level	Soft
USTC-2023-14	Building	3840×2160	Geometric	Appropriate Exposure	Horizontal Level	Cast
USTC-2023-15	Plant	3840×2160	Structural	Appropriate Exposure	Horizontal Level	Soft
USTC-2023-16	People, Occlusion	3840×2160	Natural	Appropriate Exposure	Horizontal Level	Soft
USTC-2023-17	Close Shot	3840×2160	Natural	Appropriate Exposure	Horizontal Level	Soft
USTC-2023-18	People	3840×2160	Natural	Appropriate Exposure	Horizontal Level	Soft
USTC-2023-19	Close Shot	3840×2160	Natural	Underexposure	Horizontal Level	Hard
USTC-2023-20	Close Shot	3840×2160	Natural	Appropriate Exposure	Overhead Level	Cast

TABLE VIII
THE CONFIGURATION OF USTC-TD 2023 VIDEO DATASET

Video Sequences	Color Space	Motion	Scene Types	Resolutions	Quality	Texture	View	Lens
USTC-Badminton	YUV420, 444, RGB	Medium	People, Sport	1920×1080	High	Natural	Horizontal Level	Moving
USTC-BasketballDrill	YUV420, 444, RGB	Medium	People, Sport	1920×1080	High	Natural, Geometric	Horizontal Level	Moving
USTC-BasketballPass	YUV420, 444, RGB	Medium	People, Sport	1920×1080	High	Natural, Geometric	Horizontal Level	Moving
USTC-BicycleDriving	YUV420, 444, RGB	Complex	People, Daily Life	1920×1080	High	Natural, Structural	Horizontal Level	Moving
USTC-Dancing	YUV420, 444, RGB	Complex	People, Sport	1920×1080	High	Natural	Horizontal Level	Fix
USTC-ParkWalking	YUV420, 444, RGB	Complex	People, Daily Life	1920×1080	High	Natural, Structural	Horizontal Level	Moving
USTC-Running	YUV420, 444, RGB	Complex	People, Sport	1920×1080	High	Natural, Structural	Horizontal Level	Moving
USTC-ShakingHands	YUV420, 444, RGB	Complex	People, Daliy life	1920×1080	High	Natural, Geometric	Horizontal Level	Moving
USTC-Snooker	YUV420, 444, RGB	Tiny	Sport	1920×1080	High	Natural	Horizontal Level	Moving
USTC-FourPeople	YUV420, 444, RGB	Tiny	People	1920×1080	High	Natural	Horizontal Level	Fix

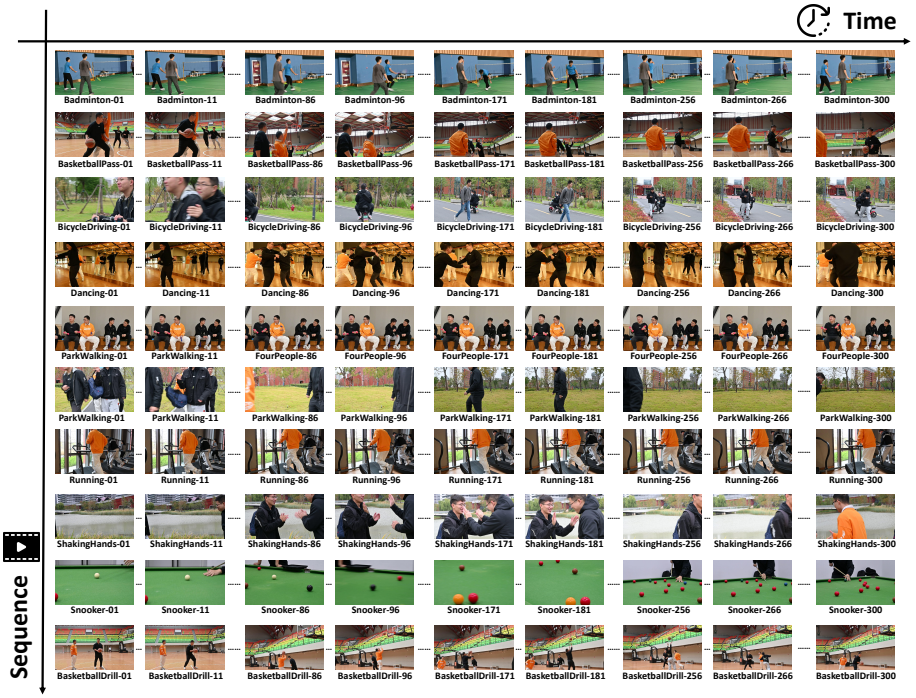


Fig. 2. Illustration of each video sequence in USTC-TD 2023 video dataset. The 0∼96/0∼300 frames correspond to the short and long setting.

B. Construction of USTC-TD 2023 Video Dataset

Based on the characteristics of previous video datasets [35]–[38], our proposed dataset aims to cover more typical characteristics of video content. Compared to the image data, temporal-domain properties are unique to video, especially in the diverse motion types with more environmental and imaging factors in natural videos. There are usually multiple moving objects of arbitrary shapes and various motion types in video frames, leading to complex motion fields, which challenge the video coding schemes [49]–[52]. Therefore, we simulate the video data with various temporal correlation types, including different kinds of motion types and lens motion.

In Fig. 2 , we show the partial frames of all collected video sequences in the USTD-TD 2023 dataset, and the specific configuration of each video in Table VIII , which make it convenient for the scheme design of research in real-world scenes under different temporal correlation types. Note that the different motion types are obviously classified in the Table VIII with different colors. For the construction of USTC-TD video dataset, two kinds of settings are constructed: short setting and long setting. For the short setting (3 seconds), a 96-frame subset is selected from each full-length captured test video sequence (frame rate: 30fps, 300 seconds, 9000 frames in total) to reduce the high complexity associated with long sequences during the practical testing, which can promote the fast evaluation in the research process. For the long setting, we extend from the last frame of the short setting and directly enlarge the short sequence to 300 frames (10 seconds), which aligns with the 10-second setting of test sequence length of HEVC and VVC common test condition (CTC).

Different from the previous video dataset, we add more specific temporal correlation types in our proposed video dataset. For example, in *USTC-BicycleDriving*, we capture the video with a fast scene change (lens motion), high-speed moving objects, and object occlusion, which is a challenge for many inter-alignment schemes of submitted solutions in *VCIP Challenge* 2023 [5] and many optical flow-based video compression schemes [12], [16], [19]–[23], [25]–[27], [53] (the detailed performance analysis is mentioned in Section V.B). For the performance of different schemes of this sequence, the learned video compression schemes are far inferior to the traditional codecs [1]–[5]. In *USTC-Snooker*, we capture the scenes with the tiny motion and fast lens motion, which is also a challenge for optical flow-based schemes. At present, most of the learned video compression schemes use the optical flow-based alignment [54], [55], and the optical flow-based motion estimation is difficult to capture the tiny motion and further influences their performance. Therefore, we put forward our proposed video dataset with the above specific designs, and hope the efficient testing datasets can help the researchers discover the problem related to temporal characteristics on their video compression scheme.

C. Analysis of USTC-TD Image and Video datasets

In the above subsections, the construction of each dataset of USTC-TD is introduced, here we point out the specific characteristics of the proposed dataset and highlight the driving factors behind its benefit, and discuss its practical application.

1) Characteristics of USTC-TD Image and Video datasets: To comprehensively verify the outstanding coverage of our proposed dataset for various content factors and qualitatively analyze the superiority of USTC-TD, we evaluate the USTC-TD on different image/video features and compare it with the previous image/video common test datasets (image datasets: *Kodak* [32], *CLIC* [34], *Tecnick* [33], video datasets: *HEVC CTC* [37], *VVC CTC* [38], *MCL-JCV* [36], *UVG* [35]). For analysis of image/video features, we select the spatial information (SI) [56], colorfulness (CF) [56], lightness information (LI) [57], and temporal information (TI) [35] to characterize

[5]Available online at https://vcip2023.iforum.biz/page/goto/.

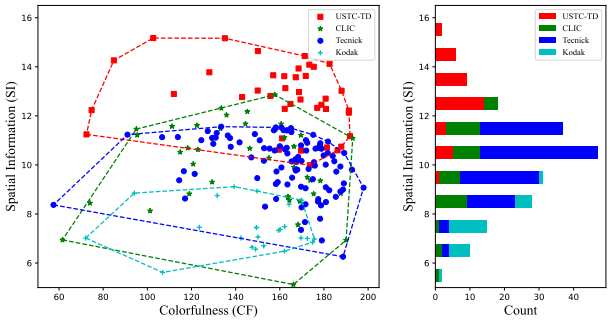


Fig. 3. The visualization of the evaluation of spatial information (SI) and colorfulness (CF) features on different image test datasets. Scatter diagram represents the SI versus CF, and corresponding convex hulls indicate the coverage of different datasets. The histogram represents the number of images under different SI scores.

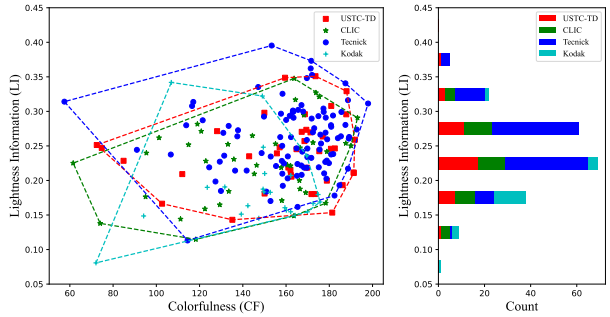


Fig. 4. The visualization of the evaluation of lightness information (LI) and CF features on different image test datasets. Scatter diagram represents LI versus CF, and corresponding convex hulls indicate the coverage of different datasets. The histogram represents the number of images under different LI scores.

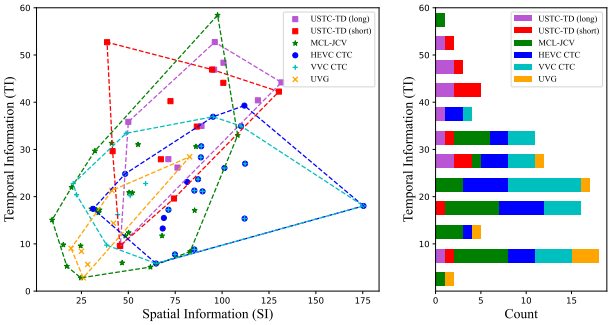


Fig. 5. The visualization of the evaluation of temporal information (TI) and SI features on different video test datasets. Scatter diagram represents the TI versus SI, and corresponding convex hulls indicate the coverage of different datasets. The histogram represents the number of videos under different TI scores.

each dataset along the dimensions of space, color, lightness, and temporal correlation, which are commonly used to evaluate the quality of dataset [35], [58], [59]. The definitions of the evaluation of these features can be found in [35], [56], [57], [60], [61], and the detailed ways are as follows:

- ***Spatial information (SI)***: SI is used as a representation of edge energy [62]. Followed by [56], the SI is defined as the root mean square of edge magnitude over the luma component of an image or a video frame:

$$Score_{SI} = \sqrt{\frac{L}{1080}}\sqrt{\sum \frac{S_r^2}{P}}, \tag{1}$$

where $S_r = \sqrt{S_v^2 + S_h^2}$ indicates the edge magnitude at each pixel. S_v and S_h indicate the images/video frames filtered with vertical and horizontal *Sobel* kernels, respectively. P indicates the total number of pixels in the filtered image, and L indicates the vertical resolution. The normalization factor $\sqrt{\frac{L}{1080}}$ is used to reduce the scale and resolution dependence of SI. For video datasets, followed by [35], SI is taking the maximum of the results of all video frames.

- ***Colorfulness (CF)***: CF is used as a representation of the variety and intensity of colors in the image. Followed by [35], [63], CF is defined as

$$Score_{CF} = \sqrt{\sigma_{rg}^2 + \sigma_{by}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{by}^2} \tag{2}$$

where opponent color spaces (rg, by) are defined in RGB color space. To be special, $rg = R - G$ and $by = 0.5\,(R + G) - B$.

- ***Lightness information (LI)***: LI is used as a representation of lightness variation. To measure the lightness information, we adopt the root mean square (RMS) contrast [61], the LI is defined as the standard deviation of the pixel intensities:

$$Score_{LI} = \sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(I_{i,j} - \hat{I})^2}, \tag{3}$$

where the intensities $(I_{i,j})$ are the i-th and j-th elements of the two-dimensional image of size M by N. $\hat{I}$ is the average intensity of all pixel values in the image. The pixel intensities of the image (I) are normalized in the range $[0, 1]$.

- ***Temporal information (TI)***: TI is used as a representation of temporal variation. Followed by [35], [60], TI is defined as the maximum amount of temporal variation between successive frames F_{n-1} and F_n:

$$Score_{TI} = \max_{1 \le n \le N-1} \left\{ \operatorname*{std}_{\substack{0 \le i \le W-1 \\ 0 \le j \le H-1}} [F_n(i,j) - F_{n-1}(i,j)] \right\} \tag{4}$$

where W, H, N denote the frame width, height and the number of total frames, respectively.

The quantitative evaluation scores of different datasets are shown in Fig. 3 , 4 , 5 . From the comparison with other datasets, we find that USTC-TD can collaborate with other datasets to handle a wide coverage of different image/video features, which verifies the diversity of the proposed dataset.

Specifically, for the evaluation of USTC-TD image dataset, the scores of SI, LI, CF cooperate to evaluate its spatial, colorfulness, and lightness diversity. Compared to the other image datasets [32]–[34], it exhibits SI scores ranging from 9 to 16, and is distinguished from other image datasets, as shown in Fig. 3 . The proposed image dataset incorporates more spatial diversity within the wide range of colorfulness diversity. In Fig. 4 , the proposed image dataset also shows a wide coverage of LI scores, ranging from 0.10 to 0.40, which aligns with the range of other image datasets and demonstrates that the proposed image dataset also exhibits excellent generalization of lightness diversity. For the evaluation of USTC-TD video

TABLE IX
QUANTITATIVE RESULTS OF THE USTC-TD 2022 AND 2023 IMAGE DATASETS. NOTE THE HIGHER SCORES ARE REPRESENTED IN RED, AND THE LOWER SCORES ARE REPRESENTED IN BLUE

USTC-TD 2022				*USTC-TD 2023*			
Image	***SI***	***CF***	***LI***	***Image***	***SI***	***CF***	***LI***
01	13.57	167.24	0.27	***01***	12.81	155.76	0.24
02	10.71	181.37	0.15	***02***	13.13	162.53	0.21
03	11.08	161.07	0.22	***03***	12.22	191.29	0.21
04	12.13	191.26	0.21	***04***	14.44	171.29	0.27
05	12.45	178.87	0.20	***05***	15.17	102.66	0.17
06	10.74	187.86	0.30	***06***	14.12	182.55	0.25
07	13.93	168.51	0.27	***07***	9.99	173.61	0.35
08	11.17	191.82	0.26	***08***	15.16	135.16	0.14
09	12.49	164.82	0.25	***09***	13.02	187.97	0.33
10	10.58	169.67	0.22	***10***	12.97	168.71	0.30
11	13.63	171.74	0.18	***11***	13.65	157.03	0.25
12	12.89	112.00	0.21	***12***	12.28	180.91	0.31
13	10.59	186.17	0.19	***13***	13.77	128.10	0.27
14	14.00	175.38	0.24	***14***	12.67	169.37	0.27
15	12.29	162.23	0.21	***15***	14.64	150.07	0.18
16	11.24	72.44	0.25	***16***	13.03	149.98	0.30
17	10.62	159.53	0.35	***17***	14.26	84.85	0.23
18	14.09	173.38	0.18	***18***	12.23	74.75	0.25
19	12.33	176.98	0.26	***19***	12.76	142.93	0.24
20	12.70	180.76	0.25	***20***	13.62	162.03	0.22

TABLE X
QUANTITATIVE RESULTS OF THE USTC-TD 2023 VIDEO DATASET. NOTE THE HIGHER SCORES ARE REPRESENTED IN RED, AND THE LOWER SCORES ARE REPRESENTED IN BLUE

USTC-TD 2023				
Video Sequence	Short Videos		Long Videos	
	SI	***TI***	***SI***	***TI***
USTC-Badminton	67.37	27.91	71.35	27.94
USTC-BasketballDrill	130.31	42.25	131.66	44.19
USTC-BasketballPass	94.63	46.92	95.74	46.89
USTC-BicycleDriving	38.66	52.70	96.09	52.73
USTC-Dancing	74.34	19.61	7.22	26.16
USTC-FourPeople	45.49	9.54	45.77	9.59
USTC-ParkWalking	72.48	40.22	119.07	40.40
USTC-Running	86.52	34.84	89.02	34.97
USTC-ShakingHands	100.67	44.11	100.74	48.37
USTC-Snooker	41.62	29.63	50.01	35.85

dataset (short/long setting), the scores of TI and SI cooperate to evaluate its spatial and temporal diversity, as shown in Fig. 5. Compared to the other video datasets [35]–[38], the proposed video dataset incorporates more temporal diversity within the excellent generalization of spatial diversity. For the short setting, it exhibits a wide range of temporal variation, ranging from 5 to 55. It compensates for the absence of the 40 to 55 (higher) range in the temporal variation of other video datasets, which enables an excellent coverage of temporal diversity with the collaboration of other datasets. For the long setting, it further extends the coverage of temporal variation, enabling a more robust and comprehensive assessment of video compression schemes. The above analysis results and related codes are open-sourced.

2) Driving Factors Behind the Benefit of USTC-TD: Beyond the above feature analysis, we present the detailed variety distribution of these different features of USTC-TD image/video datasets, as shown in Table IX and X, the detailed scores of different evaluative features are represented with the annotation of different colors. Here we deeply discuss the driving factors behind the benefit of USTC-TD image and video datasets.

For the benefit of USTC-TD image dataset, it compensates for the absence of the higher range of spatial variation within the generalized colorfulness, lightness variation and adds the mixture content diversity, which distinctly differentiates it from existing image datasets [32]–[34]. Compared with the previous image datasets, the driving factors of the benefit mainly come from two aspects: the specific design of various content factors (environmental/imaging-related factors), and the mixture of content diversity (spatial, lightness, colorfulness diversity), which contribute to maximizing the coverage. In detail, first, the samples with higher spatial diversity exhibit the characteristics of complex content factors. For example, the samples with SI scores higher than 13 of Fig. 3 have the above aspect, such as the *USTC-TD-2022-01, 14, 18*, *USTC-TD-2023-04, 05, 08*. According to the Table VI and VII, these samples own the specific design of complex environmental or imaging factors, such as the *USTC-TD-2022-01, 18, USTC-TD-2023-04* with the typical texture (geometric), or *USTC-TD-2022-14/USTC-TD-2023-05* with extreme illumination (underexposure/overexposure), or *USTC-TD-2023-08* with the special captured view (overhead level). Second, the samples with mixture diversity enable the outstanding coverage of different features, such as the *USTC-TD-2022-07, 11, USTC-TD-2023-05, 15*. As shown in Table IX, these samples own the mixture of different kinds of content diversity, such as the *USTC-TD-2022-07* with higher spatial and lightness diversity (higher SI, LI scores), the *USTC-TD-2023-05* with various diversity (higher SI, Lower CF, LI scores). The mutual promotion of different components of content factors has further contributed to the maximization of coverage.

For the benefit of USTC-TD video dataset, it compensates for the absence of the higher range of temporal variation within the generalized spatial variation, which distinctly differentiates it from existing video datasets [35]–[38]. Compared with the previous video datasets, the driving factors of the benefit directly come from the augmentation and extension of motion diversity. For example, according to Table VIII and X, the samples with TI scores higher than 40 of Fig. 5 exhibit the characteristics of diverse motion factors, such as *USTC-BicycleDriving* with high-speed moving objects and scene change, *USTC-ShakingHands* with precise limb motion, *USTC-BaskerballDrill* with abundant athletes' motion and rich lens moving. Compared with the characteristics of existing datasets, the above different kinds of motion factors further compensate for the limitation of temporal correlation, and contribute to the robust evaluation and the maximization of temporal relation-related property coverage.

3) Discussion of Practical Utilization and Application of the Existing Compression Datasets and USTC-TD: Beyond the feature and benefit analysis of USTC-TD, it has been verified that the USTC-TD compensates for the shortcomings of existing image/video datasets, and its collaboration with existing datasets can achieve wider coverage of different image/video features. Although the collaboration of these datasets ensures the robust evaluation of advanced compression schemes, the evaluation of all image/video datasets will cause heavy time/computational complexity in practical use. Here we discuss their desirable collaboration for future compression research, and further put forward the different

TABLE XI
THREE RECOMMENDED PRESETS OF THE EXISTING IMAGE/VIDEO DATASETS [32]–[38] AND THE PROPOSED USTC-TD FOR THE PRACTICAL EVALUATION OF IMAGE/VIDEO COMPRESSION SCHEMES

Preset	Image Dataset	Video Dataset	Characteristic	Purpose
Challenging	*USTC-TD*	*USTC-TD*	Challenging Content	Developing Advanced Coding Algorithms
Desirable	*USTC-TD*, *Tecnick* [33]	*USTC-TD*, *MCL-JCV* [36], *HEVC CTC* [37]	Wider Coverage	Obtaining Reliable Estimation of Coding Performance
Ideal	*USTC-TD*, *CLIC* [34], *Tecnick* [33], *Kodak* [32]	*USTC-TD*, *UVG* [35], *MCL-JCV* [36], *HEVC CTC* [37], *VVC CTC* [38]	Sufficient Coverage	Evaluating Algorithm Robustness and Generalization Ability

recommended presets of USTC-TD and existing image/video datasets with the consideration of practical utilization.

To achieve an efficient evaluation for future research, we discuss the desirable collaboration of existing image/video datasets and USTC-TD from two perspectives: *challenge* and *coverage*, and present three presets of these image and video datasets to support them with the consideration of different uses, including *challenging*, *desirable*, and *ideal* presets, as shown in Table XI. For the challenge, in the development of compression standardization, the coding efficiency on existing test content has gradually reached a bottleneck, such as the performance of hand-crafted intra and inter prediction, making it increasingly difficult to reflect shortcomings and guide future improvements. Thus, incorporating challenging content is essential, as it serves as a continuous driving force for the improvement of next-generation coding standards. For *challenging* preset, typical challenging samples are considered for developing advanced coding algorithms. Correspondingly, USTC-TD samples with the highest spatial/temporal diversity (scores) are chosen for evaluation. For coverage, building upon the *challenging* preset, a broader coverage of diverse content is desirable for the consistent evaluation of algorithm performance. By incorporating a wider range of different kinds of content features, test datasets can effectively evaluate the codec's performance in various scenarios, ensuring its reliability in practical use. For *desirable* preset, the samples with the highest coverage of different kinds of content diversity (spatial, colorfulness, lightness for image samples, temporal, spatial for video samples) are considered for obtaining reliable estimation of coding performance, such as the collaboration of *USTC-TD* image dataset and *Tecnick* [33], *USTC-TD* video dataset, *MCL-JCV* [36], and *HEVC CTC* [37]. These collaborations cover the majority of samples with different kinds of content diversity, with an appropriate number of representative samples supporting comprehensive evaluation in practical use. For the *ideal* preset, given sufficient time/resources, all datasets are ideally tested to ensure a comprehensive evaluation of algorithm robustness and generalization, supporting its use in rigorous benchmarking and evaluation.

V. EXPERIMENTS

In this section, first, we present the experimental configurations employed for the evaluation of compression schemes. Second, we evaluate the classic standardized compression schemes and recent advanced image/video compression schemes on the proposed dataset under different metrics, and benchmark their performance on our proposed dataset. Third, we analyze the benchmarked performance, and further point out some limitations and inspirations among these advanced image/video schemes to shed light on future research.

A. Experimental Settings

In this subsection, first, we present the experimental settings of the evaluative compression schemes, including the selection of advanced image/video compression schemes and the training/testing configurations of these compression schemes. Second, we introduce the evaluative quality metrics for these schemes on our proposed dataset.

1) Selection of Evaluative Compression Schemes: For advanced image compression schemes, we select classic standardized schemes and advanced learned schemes. For traditional codecs, we select *BPG*[6], and *H.266/VVC* [2]. For learned image compression schemes, we select the *Factorized Model* [6], *Hyperprior Model* [6], *Autoregressive Model* [7], *Cheng2020* [8], *iWave++* [9], *ELIC* [10], and *MLIC++* [11]. For learned compression standardized schemes, we select the high-profile model of *IEEE 1857.11*[7] (*iWave++*).

For video compression schemes, we also select classic standardized schemes and advanced learned schemes. For standardized codecs, we select the *H.265/HEVC* [1], *H.266/VVC* [2], *AV1* [4], and *AV2* [5]. For learned video compression schemes, we select the *DVC_Pro* [12], [53], *CANF-VC* [16], *DCVC* [19], *TCM-VC* [20], *DCVC-HEM* [21], *OOFE* [22], *SDD* [23], *VNVC* [25], *DCVC-DC* [26], and *DCVC-FM* [27]. The detailed introduction and test instructions of these methods are mentioned in Section I of supplementary material.

2) Testing Configurations of Evaluative Traditional Image Compression Schemes: For testing, the officially released BPG software, VTM-17.0 (*H.266/VVC* reference software) are chosen. For *BPG*, the default configuration is used, and the internal color space is set to YUV420/444 for the testing of *BPG* and *BPG444*. For VTM-17.0, the *encoder_intra_vtm* configuration is used, and the internal color space is set to YUV444. For the different source formats of these testing datasets (USTC-TD and [32]–[34]), we convert them to the YUV444 color space for the input of VTM-17.0 by using the default *ffmpeg* tool (BT.601 conversion standard).

3) Testing Configurations of Evaluative Traditional Video Compression Schemes: For testing, the officially released HM-16.20 (*H.265/HEVC* reference software), VTM-13.2 (*H.266/VVC* reference software), AV1-3.11.0, and AV2-7.0.0 are chosen. For the setting of HM-16.20, the *encoder_lowdelay_main_rext* configuration is used. For the setting of VTM-13.2, the *encoder_lowdelay_vtm* configuration is used. For the different source formats of these testing datasets (USTC-TD and [35]–[38]), we convert them to the YUV444 color space as the input of the above traditional codecs by

[6]Available online at https://bellard.org/bpg.
[7]Available online at https://sagroups.ieee.org/fvc/.

using the default *ffmpeg* tool, the conversion step is aligned to the setting mentioned in *DCVC-DC* [26].

4) Training and Testing Configurations of Evaluative Learned Image Compression Schemes: For training, these learned image compression models are optimized by mean squared error (*MSE*) or multi-scale structural similarity index measure (*MS-SSIM*), and the *Flicker2W* [64] dataset is used as the training dataset. These models are optimized by using the Adam Optimizer [65], with a batch size of 8 and a patch size of 256×256. They are optimized for around 1.2 million iterations, starting with an initial learning rate of 10^{-4}. The learning rate is reduced to 10^{-5} after 400 epochs and further down to 10^{-6} after 30 epochs. The setting of λ is set to $\{0.001, 0.004, 0.024, 0.080, 0.200\}$ for iWave++, and $\{0.0018, 0.0035, 0.0067, 0.0130, 0.0250, 0.0483\}$ for other schemes. For testing, the officially released model of *MLIC++*/*iWave++*, the reproduced model of *ELIC* are used. The model of other schemes is provided by *CompressAI* [66].

5) Training and Testing Configurations of Evaluative Learned Video Compression Schemes: For training, these learned video compression models are mainly optimized by *MSE* or *MS-SSIM*, and the *Viemo-90k* [67] is used as the training dataset. For the testing of USTC-TD, the officially released models of these schemes are used. For the testing of the variable-rate model, the bitrate points are aligned to that of the traditional codec. For the testing of other models, we directly use the released model. For the different formats of these testing datasets, we convert them to the RGB color space as the input of different learned video codecs by using the default *ffmpeg* tool (BT.601), the conversion step is aligned to the setting mentioned in *DCVC-DC* [26].

6) Testing Configurations of Subjective Quality Evaluation of Image and Video Compression Schemes: For the testing of traditional compression model, we directly use their officially released standardized model, these hand-crafted models are designed and optimized for *PSNR* metric. For the testing of learned compression model, we use the officially released *MS-SSIM* model of the advanced learned image and video compression schemes, these trained models are optimized for *MS-SSIM* metric. Some image compression schemes [9], [10] without the open-sourced *MS-SSIM* model are skipped. For the testing of video compression model, the long USTC-TD video dataset is used. To efficiently conduct the subjective test, we select the compressed images/video results at the intermediate bitrate point among all test bitrate points. For the evaluation of image compression schemes, since most learned approaches do not support the variable bitrate models, we align the target bitrate point of all approaches to approximately 0.12 bits per pixel. For the evaluation of video compression schemes, we align the target bitrate point of all approaches to approximately 0.06 bits per pixel.

7) Evaluative Metrics: For the objective quality evaluation, *PSNR*, *MS-SSIM* [39], and *VMAF* [40] are used to measure the quality of the coded frames in comparison to the original frames. Bits per pixel (bpp) is used to measure the number of bits for encoding each pixel in each image/video frame. The Bjontegaard Delta bitrate (BD-rate) [68] is used to compare the performance of different compression schemes, where negative numbers indicate bitrate saving and positive numbers indicate bitrate increasing. For the evaluation of image/video compression schemes, *PSNR* and *MS-SSIM* are calculated and compared in RGB color space, *VMAF* is calculated and compared in YUV color space. The conversion process of different color spaces is performed by using the default *ffmpeg* tool (BT.601). For *PSNR* and *MS-SSIM* tests, the *MSE* model of these schemes is used for *PSNR*, while the *MS-SSIM* model of these schemes is used for *MS-SSIM*. For *VMAF* test, the *MSE* and *MS-SSIM* models are all used for *VMAF* test (*VMAF MSE/MS-SSIM model* for short). Notably, for some test schemes without the *MS-SSIM* model, their *MSE* model is used for the *MS-SSIM* test, and the *VMAF (MS-SSIM model)* test is skipped.

For the subjective quality evaluation, mean opinion score (*MOS*) [41]–[44] is used to quantify the perceptual quality of the coded frames based on human evaluations, which is calculated by collecting ratings from multiple participants who assess the quality of images/videos on a pre-defined five-point scale (5∼Excellent, 4∼Good, 3∼Fair, 2∼Poor, 1∼Bad). The final *MOS* score is computed as the average of ratings given by all participants across test samples, the formula is as follows:

$$MOS = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N}\sum_{j=1}^{N} S_{i,j}\right) \tag{5}$$

where M is the total number of test samples, N is the number of participants, and $S_{i,j}$ represents the rating given by the j-th participant for the i-th sample. In our setting, 50 participants are selected for the *MOS* test, with testers from USTC and online sources, ensuring diversity in user experience and perception.

B. Experimental Results

In this subsection, we establish the baselines and benchmark the performance of advanced image/video compression schemes on USTC-TD image/video datasets, and further analyze their performance.

1) Objective Quality Evaluation and Analysis of Advanced Image Compression Schemes on USTC-TD: Taking bpp as the horizontal axis and the reconstructed *PSNR, MS-SSIM, VMAF (MSE/MS-SSIM model)* as the vertical axis, we present the rate and distortion curves of different image compression schemes over USTC-TD 2022 and 2023 image datasets in Fig. 6 . From the overall results, we can find that the partially learned schemes (*iWave++* [9], *ELIC* [10], *MLIC++* [11]) can outperform the traditional image compression schemes and achieve better compression performance than *H.266/VVC* on proposed datasets under different metrics, which show its powerful potential. Here we analyze their performance from the perspective of content factors of different test images of USTC-TD. As shown in Fig. 8 , the detailed RD curves of some test images (*USTC-2022-12*, *USTC-2023-15*, *USTC-2022-17*, *USTC-2023-07*) with some special phenomenons are illustrated, and the results of each test image are presented in supplementary material. Based on the performance comparison of these schemes and feature analysis of proposed datasets (Section IV.C), the conclusions mainly include four aspects:

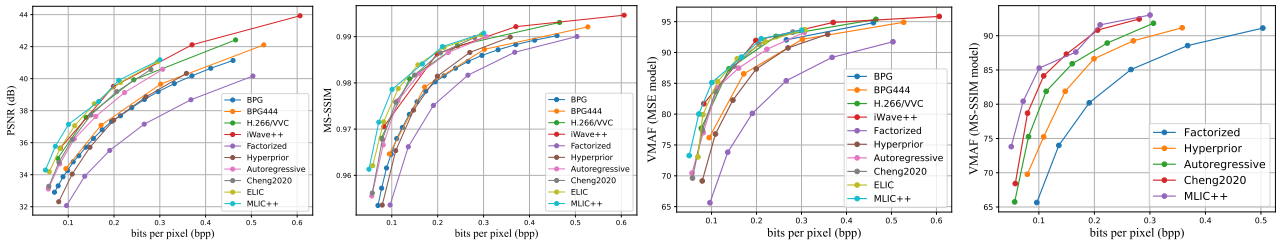
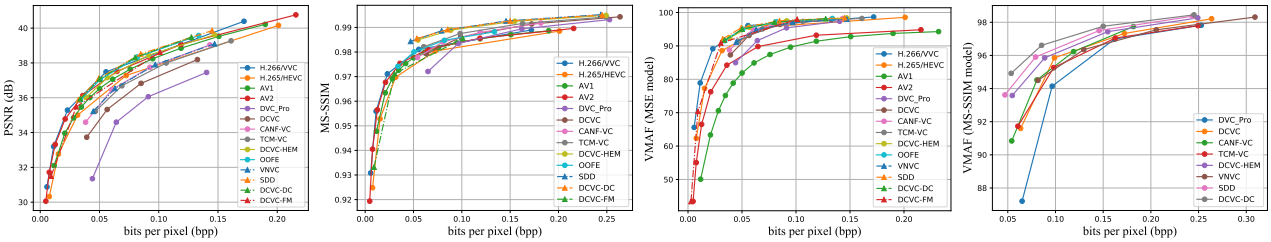
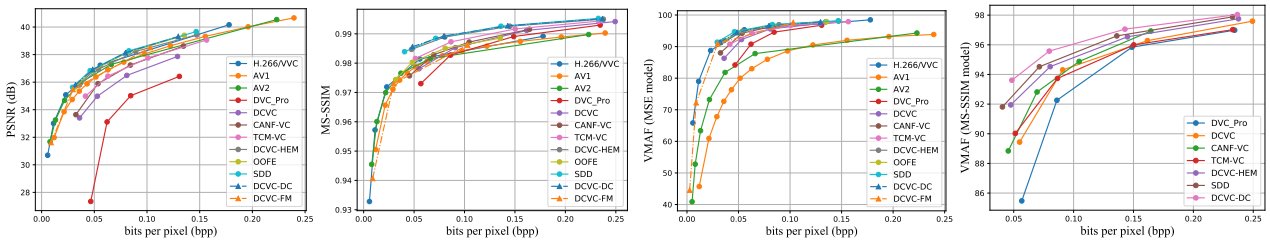


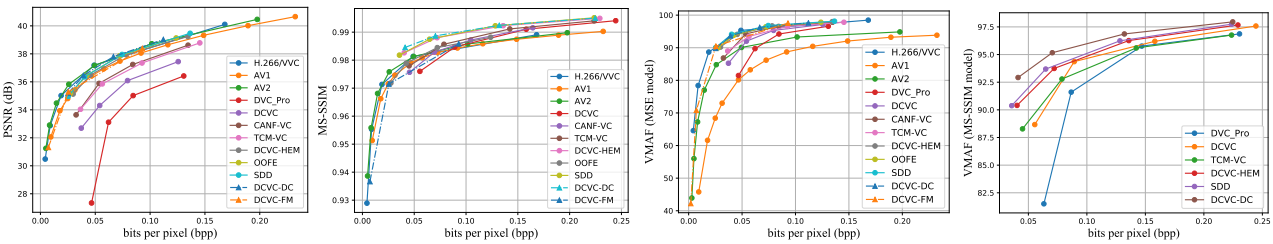
Fig. 6. Overall rate-distortion (RD) curves of advanced image compression schemes on different metrics. From left to right, the results are evaluated by *PSNR*, *MS-SSIM*, *VMAF (MSE model)*, and *VMAF (MS-SSIM model)* metrics on USTC-TD image dataset 2022 and 2023.



(a) Short Setting, Intra Period = 32



(b) Long Setting, Intra Period = 32

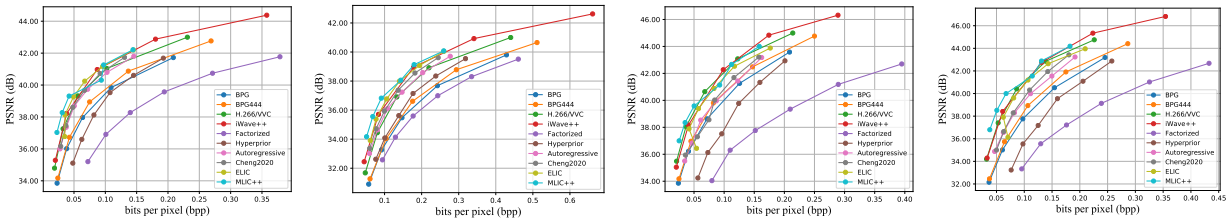


(c) Long Setting, Intra Period = -1

Fig. 7. Overall rate-distortion (RD) curves of advanced video compression schemes on different metrics. From left to right, the results are evaluated by *PSNR*, *MS-SSIM*, *VMAF (MSE model)*, and *VMAF (MS-SSIM model)* metrics on different settings of USTC-TD video dataset.

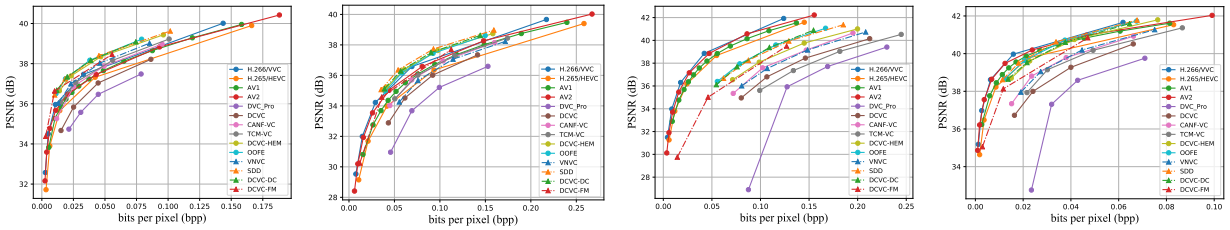
- (1) The learned schemes show the good potential on some complex scenarios. For example, as shown in the Fig. 8 (a) and (b), compared to the overall results (Fig. 6), more learned schemes (*Autoregressive Model* [7], *Cheng2020* [8], *iWave++* [9], *ELIC* [10], *MLIC++* [11]) perform better than the traditional schemes on some test images with specific features of environment-related factors (Table IX), such as the *USTC-2022-12* with the lower scores of CF, the *USTC-2023-15* with the higher scores of SI. Meanwhile, the results also demonstrate the previous image train/test datasets [32]–[34] can guide the researcher to train/evaluate the basic ability of intra-frame redundancy removal of their schemes to some extent.
- (2) The traditional schemes show the powerful generalization ability in some extreme scenarios. For example, as shown in the Fig. 8 (c) and (d), compared to the overall results (Fig. 6), the generalization ability of learned schemes is lacking to handle the evaluative images with extreme mixture features of environment/imaging-related factors well (Table IX), such as the *USTC-2022-17* with the lower scores of SI, CF and the higher scores of LI, the *USTC-2023-07* with the lower scores of SI and the higher scores of LI. Although the performance of learned schemes surpasses the traditional schemes in general, they are still limited to some extreme scenarios.
- (3) Based on the analysis of different features (SI, CF, LI, TI) of proposed datasets (Section IV.C), the detailed characteristics of different test images can efficiently assist the researcher to analyze the detailed bottleneck of their compression scheme.
- (4) Compared to the performance of these schemes on different image datasets [32]–[34], the proposed datasets can collaborate with other datasets to handle a wide coverage of performance evaluation, which also demonstrates the efficiency of the specific design of the proposed datasets' different content factors/features.

2) Subjective Quality Evaluation and Analysis of Advanced Image Compression Schemes on USTC-TD: In Table XII, we present the overall *MOS* results of compressed images of different image compression schemes over USTC-TD 2022



(a) USTC-2022-12 (b) USTC-2023-15 (c) USTC-2022-17 (d) USTC-2023-07

Fig. 8. Specific rate-distortion (RD) curves of advanced image compression schemes on partial evaluative images under *PSNR* metric.



(a) USTC-FourPeople (b) USTC-BasketballPass (c) USTC-BicycleDriving (d) USTC-Snooker

Fig. 9. Specific rate-distortion (RD) curves of advanced video compression schemes on partial evaluative short videos under *PSNR* metric.

TABLE XII
THE OVERALL *MOS* RESULTS OF COMPRESSED IMAGES OF CLASSIC STANDARDIZED AND ADVANCED LEARNED IMAGE COMPRESSION SCHEMES, WHERE BLUE REPRESENTS THE LOWEST SCORE AND RED REPRESENTS THE HIGHEST SCORE.

Dataset	Scheme Classification	Compression Scheme	MOS Score
USTC-TD 2022 & 2023 Image Dataset	Traditional	*H.266/VVC* [2]	3.67
	Learned	*Factorized Model* [6]	3.44
		Hyperprior Model [6]	3.43
		Autoregressive Model [7]	3.65
		Cheng2020 [8]	3.77
		MLIC++ [11]	3.80

and 2023 image datasets. From the overall results, we can find that recent advanced learned compression schemes [8], [11] outperform the traditional compression schemes, which demonstrates their powerful potential. This can be attributed to the optimization of *MS-SSIM* for perceptual quality and the fact that end-to-end learning methods, compared to hand-crafted approaches, are more flexible and can be easily optimized toward different quality assessment metrics. From the detailed results of compressed images of different schemes mentioned in supplementary material, it is observed that different compression schemes show different *MOS* performance trends on different images. Here we analyze their performance from the perspective of feature analysis of different test images of USTC-TD (Section IV.C), and explore the advantages of different learned compression schemes for varying scenes under the evaluation of subjective metric. The conclusions include the following two aspects:

- (1) The learned schemes show the powerful potential on various scenarios for perception optimization, especially the complex scenarios with mixture features, such as the *USTC-2023-03* with the lower SI and higher CF scores, *USTC-2023-05* with the higher SI, lower CF and LI scores. Combined with the performance of objective quality metrics, these samples highlight the learned schemes' generalization in optimizing across different metrics.
- (2) Despite traditional schemes [2] being hand-crafted and optimized for *PSNR*, their subjective optimization capability remains competitive with certain learned schemes. Moreover, they exhibit a degree of practical applicability for visual perception, achieving comparable performance to some state-of-the-art learned schemes [8], [11] in challenging scenarios, such as *USTC-2022-02* with lower SI and LI scores, and *USTC-2023-15* with higher SI and lower LI scores. Therefore, for the development of learned schemes, merely modifying the optimization metric is not sufficient; it is essential to further integrate perceptual theories into the framework design to achieve more effective improvements.

3) Objective Quality Evaluation and Analysis of Advanced Video Compression Schemes on USTC-TD: Taking bpp as the horizontal axis and the reconstructed *PSNR, MS-SSIM, VMAF (MSE/MS-SSIM model)* as the vertical axis, we present the rate and distortion curves and BD-rate results of different video compression schemes over the USTC-TD 2023 video dataset in Fig. 7 , Table XIII, XIV , XV for short dataset, Table XVI , Table XVII for long dataset. Note that the setting of intra period = -1 is used to enhance comprehensive benchmark for the testing of long sequences, it allows for an accurate evaluation of the video compression scheme's robustness to handle the long video sequences. From the overall results of *PSNR* and *VMAF (MSE model)* metric, we can find that the performance of the advanced traditional schemes is better than that of all advanced learned schemes on the proposed short/long video dataset. Notably, the *PSNR* results of short setting differ from the conclusions drawn from other datasets [35]–[38]. From the overall results of *MS-SSIM* metric, the performance of the traditional schemes is lower than that of advanced learned schemes. Here we analyze their performance from the perspective of typical characteristics of different test video contents and test length of USTC-TD video dataset. As shown in Fig. 9 , the detailed RD curves of some test videos (*USTC-FourPeople*, *USTC-BasketballPass*, *USTC-Snooker*, *USTC-BicycleDriving*) with some special phenomenons are illustrated, the results of each test video are also presented in supplementary material. Based on the performance comparison of advanced schemes and feature analysis of proposed datasets (Section IV.C), the conclusions mainly include the following four aspects:

TABLE XIII
BD-RATE (%) COMPARISON FOR PSNR. SHORT SETTING WITH INTRA PERIOD = 32. THE ANCHOR IS VTM.

Dataset	*VTM*	*HM*	*AV1*	*AV2*	*DVC_Pro*	*DCVC*	*CANF-VC*	*TCM-VC*	*VNVC*	*DCVC-HEM*	*OOFE*	*SDD*	*DCVC-DC*	*DCVC-FM*
HEVC Class B	0.0	39.0	–	–	188.6	115.7	58.2	32.8	24.6	-0.7	-15.3	-13.7	-13.9	-8.8
HEVC Class C	0.0	37.6	–	–	202.8	150.8	73.0	62.1	48.2	16.1	-2.2	-2.3	-8.8	-5.0
HEVC Class D	0.0	34.7	–	–	160.3	106.4	48.8	29.0	22.4	-7.1	-23.0	-24.9	-27.7	-23.3
HEVC Class E	0.0	48.6	–	–	429.5	257.5	116.8	75.8	66.8	20.9	-0.4	-8.4	-19.1	-20.8
HEVC Class RGB	0.0	44.0	–	–	186.8	118.6	87.5	25.4	16.0	-15.6	-17.5	-17.5	-27.9	-18.6
UVG	0.0	36.4	–	–	218.7	129.5	56.3	23.1	18.0	-17.2	-22.3	-19.7	-25.9	-20.5
MCL-JCV	0.0	41.9	–	–	163.6	103.9	60.5	38.2	30.2	-1.6	-5.8	-7.1	-14.4	-7.4
USTC-TD[1]	**0.0**	**46.7**	**34.3**	**11.0**	**284.4**	**124.7**	**64.6**	**67.1**	**60.4**	**16.0**	**8.2**	**3.7**	**6.3**	**24.9**
Average (*Desirable*)	**0.0**	**41.8**	**–**	**–**	**230.9**	**139.7**	**72.8**	**47.2**	**38.4**	**4.0**	**-8.0**	**-10.0**	**-15.1**	**-8.4**
Average (*Ideal*)	**0.0**	**41.1**	**–**	**–**	**229.3**	**138.4**	**70.7**	**44.2**	**35.8**	**1.4**	**-9.8**	**-11.2**	**-16.4**	**-9.9**

[1] The results of USTC-TD indicate the average results of the *Challenging* preset of dataset collaboration mentioned in Table XI .

TABLE XIV
BD-RATE (%) COMPARISON FOR MS-SSIM. SHORT SETTING WITH INTRA PERIOD = 32. THE ANCHOR IS VTM.

Dataset	*VTM*	*HM*	*AV1*	*AV2*	*DVC_Pro*	*DCVC*	*CANF-VC*	*TCM-VC*	*VNVC*	*DCVC-HEM*	*OOFE*	*SDD*	*DCVC-DC*	*DCVC-FM*
HEVC Class B	0.0	36.8	–	–	67.0	35.9	25.5	-20.5	-33.1	-47.4	-15.1	-48.0	-53.0	-12.5
HEVC Class C	0.0	38.7	–	–	61.1	24.9	17.7	-21.7	-29.3	-43.3	-15.9	-49.6	-54.6	-18.0
HEVC Class D	0.0	34.9	–	–	25.3	2.7	1.5	-36.2	-41.1	-55.5	-28.6	-60.0	-63.4	-30.6
HEVC Class E	0.0	38.4	–	–	195.8	90.0	114.9	-20.5	-0.4	-52.4	-9.3	-51.5	-60.7	-32.6
HEVC Class RGB	0.0	37.3	–	–	66.8	43.7	52.9	-21.1	-32.4	-45.8	-16.7	-46.3	-54.4	-16.6
UVG	0.0	37.1	–	–	74.6	11.9	33.1	-6.0	-15.2	-32.7	-10.6	-34.2	-36.7	-7.3
MCL-JCV	0.0	43.7	–	–	46.1	39.1	11.7	-18.6	-29.0	-44.0	-2.5	-46.3	-49.1	-5.0
USTC-TD[1]	**0.0**	**47.7**	**31.8**	**11.0**	**55.8**	**3.9**	**3.4**	**-19.1**	**-23.7**	**-44.3**	**0.4**	**-48.0**	**-48.7**	**39.5**
Average (*Desirable*)	**0.0**	**39.6**	**–**	**–**	**74.0**	**34.3**	**32.5**	**-22.5**	**-27.1**	**-47.5**	**-12.5**	**-50.0**	**-54.8**	**-10.8**
Average (*Ideal*)	**0.0**	**39.3**	**–**	**–**	**74.1**	**31.5**	**32.6**	**-20.5**	**-25.5**	**-45.7**	**-12.3**	**-48.0**	**-52.6**	**-10.4**

[1] The results of USTC-TD indicate the average results of the *Challenging* preset of dataset collaboration mentioned in Table XI .

TABLE XV
BD-RATE (%) COMPARISON FOR VMAF (MSE MODEL), VMAF (MS-SSIM MODEL). SHORT SETTING WITH INTRA PERIOD = 32. THE ANCHOR IS VTM.

Dataset	Evaluative Metric	*VTM*	*HM*	*AV1*	*AV2*	*DVC_Pro*	*DCVC*	*CANF-VC*	*TCM-VC*	*VNVC*	*DCVC-HEM*	*OOFE*	*SDD*	*DCVC-DC*	*DCVC-FM*
USTC-TD	VMAF (MSE model)	**0.0**	**46.5**	**291.9**	**137.1**	**118.5**	**63.0**	**39.6**	**45.8**	**42.2**	**10.6**	**9.5**	**-1.1**	**9.2**	**21.6**
	VMAF (MS-SSIM model)	**–**	**–**	**–**	**–**	**160.6**	**89.4**	**93.7**	**107.3**	**112.8**	**64.8**	**–**	**46.1**	**34.5**	**–**

TABLE XVI
BD-RATE (%) COMPARISON FOR PSNR, MS-SSIM, VMAF (MSE MODEL), VMAF (MS-SSIM MODEL). LONG SETTING WITH INTRA PERIOD = 32. THE ANCHOR IS VTM.

Dataset	Evaluative Metric	*VTM*	*AV1*	*AV2*	*DVC_Pro*	*DCVC*	*CANF-VC*	*TCM-VC*	*DCVC-HEM*	*OOFE*	*SDD*	*DCVC-DC*	*DCVC-FM*
USTC-TD	PSNR	**0.0**	**33.6**	**11.0**	**403.5**	**120.8**	**70.2**	**54.4**	**9.5**	**2.8**	**-1.5**	**0.4**	**13.5**
	MS-SSIM	**0.0**	**28.4**	**7.0**	**43.3**	**5.6**	**5.7**	**-25.3**	**-49.1**	**-3.0**	**-51.5**	**-53.4**	**23.5**
	VMAF (MSE model)	**0.0**	**349.2**	**194.5**	**118.9**	**58.0**	**23.8**	**33.7**	**4.5**	**3.5**	**-6.7**	**2.4**	**10.5**
	VMAF (MS-SSIM model)	**–**	**–**	**–**	**166.8**	**104.6**	**94.8**	**111.0**	**69.9**	**–**	**52.1**	**37.8**	**–**

TABLE XVII
BD-RATE (%) COMPARISON FOR PSNR, MS-SSIM, VMAF (MSE MODEL), VMAF (MS-SSIM MODEL). LONG SETTING WITH INTRA PERIOD = -1. THE ANCHOR IS VTM.

Dataset	Evaluative Metric	*VTM*	*AV1*	*AV2*	*DVC_Pro*	*DCVC*	*CANF-VC*	*TCM-VC*	*DCVC-HEM*	*OOFE*	*SDD*	*DCVC-DC*	*DCVC-FM*
USTC-TD	PSNR	**0.0**	**33.4**	**-1.9**	**520.7**	**221.4**	**95.6**	**99.1**	**23.9**	**16.5**	**15.2**	**9.2**	**22.4**
	MS-SSIM	**0.0**	**24.4**	**-5.0**	**64.9**	**30.7**	**17.5**	**-2.3**	**-41.3**	**10.7**	**-42.5**	**-51.2**	**34.7**
	VMAF (MSE model)	**0.0**	**387.8**	**95.7**	**195.4**	**100.0**	**49.1**	**59.2**	**17.3**	**15.1**	**4.1**	**10.4**	**18.7**
	VMAF (MS-SSIM model)	**–**	**–**	**–**	**277.2**	**132.7**	**146.3**	**158.0**	**99.0**	**–**	**82.9**	**50.8**	**–**

- (1) The traditional schemes show good generalization ability on various real-world video scenarios. As shown in Table XIII , different from the performance of advanced compression schemes on other datasets [35]–[38] , the traditional schemes still achieve the state-of-the-art performance of these schemes on USTC-TD under the PSNR metric. Different from the other datasets, USTC-TD video dataset focuses on various temporal correlation types. Combined the attributes of different test videos of USTC-TD (Fig. 5 , Table VIII), we find that the traditional schemes can handle more scenarios with the various temporal features of motion-related elements (motion type, lens motion), such as the *USTC-BicycleDriving* with higher scores of TI and *USTC-Snooker* with the specific design of lens motion (Section IV.C). The performance of these test videos is shown in Fig. 9 (c) and (d), and the results demonstrate that the traditional schemes can handle these scenes with complex motion types robustly.
- (2) The learned schemes show the optimistic potential on some scenarios with complex motion cases, such as the *USTC-FourPeople* with lower TI scores, and the *USTC-BasketballPass* with higher TI scores. These scenarios commonly appear in the previous test datasets [35]–[38] , such as the *FourPeople* and *BasketballPass* in *HEVC/VVC CTC*, which can guide the design and the optimized target of deep network to handle these motion situations. The results of these specific scenarios demonstrate the basic evaluative ability of the previous datasets and the performance potential of the deep learning-based manner.
- (3) Based on the performance comparison of advanced compression schemes under different test length settings and intra period settings of USTC-TD (Table XIII for

short dataset, Table XVI and Table XVII for long dataset), the performance gap between traditional and learned schemes is amplified in the long setting, it further demonstrates the superiority of traditional schemes. Combined with the attributes of different length settings (Fig. 5), the more powerful temporal diversity of long sequences challenges the robustness of these schemes even further. The extended temporal variation with different intra period tests further exposes the shortcomings of advanced compression schemes, revealing the limitations of learned schemes in the error propagation of prediction chain, motion modeling, and the design of reference mechanism.

- (4) Based on the analysis of video-related features (Section IV.C), our proposed dataset can make up more typical motion/temporal correlation-related real-world factors with other datasets to handle an excellent coverage of performance evaluation, which also demonstrates the efficiency of the specific design of the proposed datasets' different factors/features. Furthermore, compared to the performance of other datasets [35]–[38] of different schemes mentioned in [12], [16], [19]–[23], [25]–[27], as shown in Table XIII and Table XIV , the different phenomenon between the performance of our proposed dataset and other datasets also verifies the efficiency of the proposed dataset.
- (5) Based on the performance comparison of different presets of dataset collaborations under different feature coverage settings, as shown in Table XIII and XIV , the performance trends observed across different presets highlight their distinct testing preferences, and further validate the configuration rationality of different presets. Combining the attributes of each preset, the results from the *Challenging* preset to the other presets gradually validate the performance of these video compression schemes, highlighting the testing effects with the consideration of general, reliable, and comprehensive evaluation, respectively. Specifically, in Table XIII and XIV , the performance of *DVC_Pro* and *CANF-VC* tend to show variational BD-rate changes under different presets, reflecting their limitations in handling specific considerations or coverage settings compared to other methods. On the other hand, the variants of *DCVC* show more consistent performance across different feature coverage settings, demonstrating that these methods are more robust when adapting to various scenarios.

4) Subjective Quality Evaluation and Analysis of Advanced Video Compression Schemes on USTC-TD: In Table XVIII , we present the overall *MOS* results of compressed videos of different video compression schemes over USTC-TD 2023 video dataset. The results of video compression follow the same trend as image compression, where learned schemes outperform traditional schemes, highlighting their advantages in perceptual quality optimization and adaptability to different quality assessment metrics. From the detailed results mentioned in the supplemental material, we present the detailed *MOS* results of compressed videos of different video compression schemes over USTC-TD 2023 video dataset. It is also observed that different schemes show different *MOS* performance trends on different videos. Here we further analyze their performance from the perspective of feature analysis of different test videos of USTC-TD (Section IV.C), and explore the advantages of different learned compression schemes for varying scenes under the evaluation of subjective metric. The conclusions mainly include the following two aspects:

TABLE XVIII
THE OVERALL *MOS* RESULTS OF COMPRESSED VIDEOS OF CLASSIC STANDARDIZED AND ADVANCED LEARNED VIDEO COMPRESSION SCHEMES, WHERE BLUE REPRESENTS THE LOWEST SCORE AND RED REPRESENTS THE HIGHEST SCORE.

Dataset	Scheme Classification	Compression Scheme	MOS Score
USTC-TD 2023 Video Dataset	Traditional	*H.266/VVC* [2]	3.49
	Learned	*DCVC* [19]	3.14
		TCM-VC [20]	3.24
		DCVC-HEM [21]	3.42
		OOFE [22]	3.51
		SDD [23]	3.41
		DCVC-DC [26]	3.51
		DCVC-FM [27]	3.55

- (1) The learned video schemes also demonstrate powerful potential for perceptual optimization, particularly in complex motion scenarios such as *USTC-ShakingHands* and *USTC-BasketballPass*, which exhibit higher TI scores. Combined with the performance of objective quality metrics, these samples further highlight the learned schemes' ability to generalize and optimize across different metrics.
- (2) Despite traditional schemes [2] being hand-crafted and optimized for *PSNR*, they exhibit strong competitiveness compared to learned schemes. Especially in certain sequences with extreme motions, they demonstrate a greater advantage over learned schemes, such as *USTC-BicycleDriving* with the highest TI scores, and *USTC-BasketballDrill* with higher SI and TI scores. Therefore, for the development of learned schemes, merely modifying the optimization metric is also not enough. It is crucial to further incorporate video-related factors into the framework design, such as temporal consistency and motion fidelity, to achieve more effective improvements.

C. Limitation and Inspiration of Advanced Image and Video Compression Schemes

In this subsection, based on the analysis of experimental results, we analyze the limitations of evaluative image/video compression schemes, and point out some limitations and inspirations among these advanced compression schemes.

1) Limitation and Inspiration of Image Compression: Based on the above analysis in Section V.B, we can find that the generalization ability of the learned codec is a major challenge for practical usage. Most existing methods focus on improving the compression performance while neglecting its generalization for various scenarios. As the situations mentioned in the item (1) and (2) of conclusions (Section V.B (1)), the generalization ability is challenged with the effective extension of the evaluation data. These problems mainly arise from the incompletion of training data and the constraint

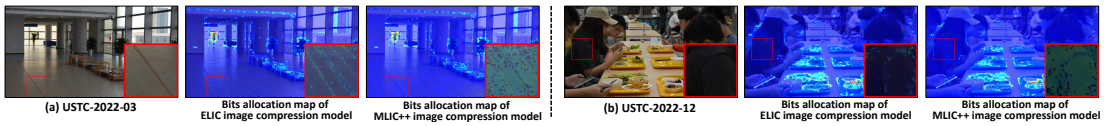


Fig. 10. Visualization of limitation of partial advanced image compression schemes (*ELIC* [10], *MLIC++* [11]) on *USTC-2022-03* and *USTC-2022-12*.

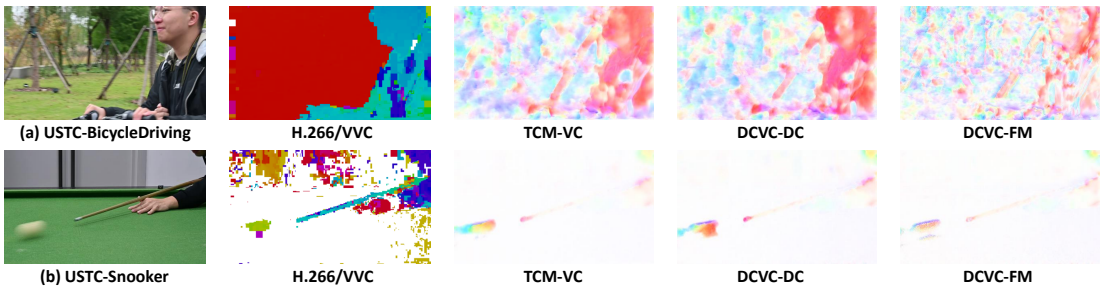


Fig. 11. Visualization of limitation of partial video compression schemes (*TCM-VC* [20], *DCVC-DC* [26], *DCVC-FM* [27]) on *USTC-BicycleDriving/Snooker*.

TABLE XIX
BD-RATE (%) RESULTS OF DIFFERENT TRAINING STRATEGIES OF OPTICAL FLOW-RELATED MODULE OF DCVC-DC [26] ON USTC-TD UNDER PSNR METRIC

Scheme	BD-rate (%)
DCVC-DC	6.3%
DCVC-DC + Flow Pre-training	-1.2%

optimization direction of deep learning-based manner with the limited evaluation data. Here we explore these problems based on the coding process of different compression schemes.

For example, as shown in Fig. 6, the performance of one bpp point of *MLIC++* is lower than that of other schemes, but the performance of other bpp points is better. In detail, we further illustrate the detailed rate-distortion curves of each image, such as the RD curves of *USTC-2022-12*, *USTC-2022-17*, *USTC-2023-07* shown in Fig. 8 (a), (b), and (c), we can find that one bpp model of *MLIC++* all performs poorly on several specific images. To explore it, we visualize the bits allocation map of *ELIC* and *MLIC++* on some test images, such as the cases of *USTC-2022-03* and *USTC-2022-12* shown in Fig. 10. Compared to *ELIC*, *MLIC++* allocates more bits to some flat areas, whereas these areas could be encoded with fewer bits. Inspired by them, the controllable model optimization, domain adaptation, and precise rate allocation of the learned compression models need to be further improved for future practical usage.

2) Limitation and Inspiration of Video Compression: Based on the above analysis in Section V.B, as shown in Fig. 7 and Table XIII, the performance of all learned video codecs is lower than that of traditional video codecs on USTC-TD, which is different from the performance phenomenon on other datasets. The reason mainly comes from that the learned video codecs perform poorly on some sequences with complex motion features. Here we further explore these problems based on the motion-related modules of different schemes.

As mentioned in Fig. 9 (c) and (d), the performance of the state-of-the-art learned video codecs is even lower than that of the *H.265/HEVC* [1]. Based on these observations, we visualize the video reconstructed frames of these video codecs as shown in Fig. 11. From the comparison of different scenarios, we can find that the specific design of motion-related features (high-speed moving objects, object occlusion, and camera motion) bring severe motion blur in the temporal domain, which challenges the optical flow-based motion estimation/compensation module of learned video codecs that is difficult to estimate accurate motion vector prediction. Therefore, we further illustrate the estimated motion vectors of these traditional and learned codecs in Fig. 11, it obviously observes that the motion field of learned video compression schemes performs wrong and disordered, which further demonstrates that the flow-based motion-related modules of learned video codecs are difficult to handle the complex motion situations.

To further verify the performance impact of these problems, we tentatively design the experiment to optimize the optical flow-related module of different learned video codecs. We set the state-of-the-art scheme (*DCVC-DC* [26]) as the anchor, and use the motion vectors of *H.266/VVC* as the optimized target of the optical flow-based motion estimation module (*Spynet* [55]) in the offline pre-training stage, instead of the usage of EPE loss for the training of these optical flow-based modules. The performance is shown in Table XIX. Inspired by the results, it verifies that the motion modeling and training strategy of learned video compression models are necessary to be further improved for practical usage in the future.

VI. LIMITATION DISCUSSION

Beyond the existing image/video datasets and USTC-TD, here we further discuss the limitations of these image/video datasets. With the collaboration of these datasets, the overall coverage of different image/video features has become more comprehensive, yet there are still some missing data types.

First, one notable missing type is computer-generated (CG) content, such as animations, cartoon content, and screen content. Unlike the camera-captured content of existing datasets, these special images/sequences differ significantly in visual characteristics and elements, such as vibrant color schemes, exaggerated motion, and mixed textures. For example, as shown in Fig. 5, the particular sample of the existing datasets with the highest temporal diversity is the animated video of *MCL-JCV* [36], which is the only sample from other datasets, aside from USTC-TD, with TI scores higher than 40. Second, AI-generated content [69], [70] is another crucial missing data type, which is created using advanced generative models, such as GANS and diffusion models. These images/videos introduce unique challenges for compression models, as they

often contain novel visual patterns, surreal features, or unusual motion patterns, which significantly differ from the content captured by human creators. The absence of these above non-camera-captured content types reveals a limitation of existing datasets, which mainly focus on natural images/videos captured by cameras. As synthetic media becomes more common in fields like entertainment, gaming, digital art, and AI-generated content, the demand for their compression also increases, creating a need for datasets that include CG and AI-generated media. Therefore, addressing these limitations in the future would enhance compression models' ability to handle a broader range of visual content.

VII. Conclusion

In this paper, we propose a test dataset (named USTC-TD) for compression-related research, which covers more diverse content factors. To evaluate the efficiency of USTC-TD, we qualitatively evaluate the USTC-TD on different image/video features and compare it with the previous image/video common test datasets to verify its excellent compensation. In addition, we evaluate the advanced compression schemes under different metrics benchmarked on USTC-TD, and further analyze their performance to point out the inspirations for future compression-related research.

In the present dataset construction process, we only consider the basic image and video test datasets. In the future, we plan to progressively extend the annotation datasets of USTC-TD for image/video coding for machine (ICM/VCM) [71]–[73], such as object segmentation [74], object detection [75], action recognition [76], *et al*, and the reconstruction dataset of video enhancement, such as image/video super-resolution [77], [78], denoising [79], *et al*, for the testing of compression-related downstream researches, and provide a comprehensive baseline to promote the development of compression-related diverse tasks.

References

[1] 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.

[2] 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.

[3] S. Ma, L. Zhang, S. Wang, C. Jia, S. Wang, T. Huang, F. Wu, and W. Gao, "Evolution of AVS video coding standards: twenty years of innovation and development," *Science China Information Sciences*, vol. 65, no. 9, p. 192101, 2022.

[4] J. Han, B. Li, D. Mukherjee, C.-H. Chiang, A. Grange, C. Chen, H. Su, S. Parker, S. Deng, U. Joshi *et al.*, "A technical overview of AV1," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1435–1462, 2021.

[5] X. Zhao, L. Zhao, M. Krishnan, Y. Du, S. Liu, D. Mukherjee, Y. Xu, and A. Grange, "Study on coding tools beyond AV1," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[6] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[7] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in Nneural Information Processing Systems*, vol. 31, 2018.

[8] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.

[9] H. Ma, D. Liu, N. Yan, H. Li, and F. Wu, "End-to-end optimized versatile image compression with wavelet-like transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1247–1263, 2020.

[10] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5718–5727.

[11] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "Mlic: Multi-reference entropy model for learned image compression," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7618–7627.

[12] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 006–11 015.

[13] Z. Hu, G. Lu, and D. Xu, "FVC: A new framework towards deep video compression in feature space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1502–1511.

[14] G. Lu, C. Cai, X. Zhang, L. Chen, W. Ouyang, D. Xu, and Z. Gao, "Content adaptive and error propagation aware deep video compression," in *European Conference on Computer Vision*. Springer, 2020, pp. 456–472.

[15] G. Lu, T. Zhong, J. Geng, Q. Hu, and D. Xu, "Learning based multi-modality image and video compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6083–6092.

[16] Y.-H. Ho, C.-P. Chang, P.-Y. Chen, A. Gnutti, and W.-H. Peng, "Canf-vc: Conditional augmented normalizing flows for video compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 207–223.

[17] M.-J. Chen, Y.-H. Chen, and W.-H. Peng, "B-canf: Adaptive b-frame coding with conditional augmented normalizing flows," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[18] P.-Y. Chen and W.-H. Peng, "CANF-VC++: Enhancing conditional augmented normalizing flows for video compression with advanced techniques," *arXiv preprint arXiv:2309.05382*, 2023.

[19] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 114–18 125, 2021.

[20] X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, "Temporal context mining for learned video compression," *IEEE Transactions on Multimedia*, vol. 25, pp. 7311–7322, 2022.

[21] J. Li, B. Li, and Y. Lu, "Hybrid spatial-temporal entropy modelling for neural video compression," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1503–1511.

[22] C. Tang, X. Sheng, Z. Li, H. Zhang, L. Li, and D. Liu, "Offline and online optical flow enhancement for deep video compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5118–5126.

[23] X. Sheng, L. Li, D. Liu, and H. Li, "Spatial decomposition and temporal fusion based inter prediction for learned video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[24] Y. Bian, X. Sheng, L. Li, and D. Liu, "LSSVC: A learned spatially scalable video coding scheme," *IEEE Transactions on Image Processing*, 2024.

[25] X. Sheng, L. Li, D. Liu, and H. Li, "Vnvc: A versatile neural video coding framework for efficient human-machine vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

[27] ——, "Neural video compression with feature modulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 099–26 108.

[28] H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava, "Nerv: Neural representations for videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 557–21 568, 2021.

[29] H. Chen, M. Gwilliam, S.-N. Lim, and A. Shrivastava, "Hnerv: A hybrid neural representation for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 270–10 279.

[30] H. M. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "Hinerv: Video compression with hierarchical encoding-based neural representation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[31] ——, "NVRC: Neural video representation compression," *arXiv preprint arXiv:2409.07414*, 2024.

[32] Kodak, "Kodak lossless true color image suite (photocd pcd0992)," vol. 6, p. 2, 1993.

[33] N. Asuni, A. Giachetti *et al.*, "TESTIMAGES: a large-scale archive for testing visual devices and basic image processing algorithms." in *STAG*, 2014, pp. 63–70.

[34] G. Toderici, W. Shi, R. Timofte, L. Theis, J. Balle, E. Agustsson, N. Johnston, and F. Mentzer, "Workshop and challenge on learned image compression (CLIC2020)," in *CVPR*, 2020.

[35] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120fps 4k sequences for video codec analysis and development," in *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020, pp. 297–302.

[36] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1509–1513.

[37] F. Bossen *et al.*, "Common test conditions and software reference configurations," *JCTVC-L1100*, vol. 12, no. 7, p. 1, 2013.

[38] J. Boyce *et al.*, "JVET Common Test Conditions and Software Reference Configurations," *JVET document, JVET-J1010*, 2018.

[39] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[40] T.-J. Liu, Y.-C. Lin, W. Lin, and C.-C. J. Kuo, "Visual quality assessment: recent developments, coding applications and future trends," *APSIPA Transactions on Signal and Information Processing*, vol. 2, p. e4, 2013.

[41] ITU-T, "P. 800.1, mean opinion score (MOS) terminology," *International Telecommunication Union, Geneva*, vol. 2, 2006.

[42] ——, "Methods for subjective determination of transmission quality," 1996.

[43] ——, "Vocabulary for performance, quality of service and quality of experience," 2017.

[44] ——, "Methodology for the subjective assessment of the quality of television pictures," vol. 500, no. 13, 2012.

[45] E. Alshina, J. Ascenso, T. Ebrahimi, F. Pereira, and T. Richter, "AHG 11," *Brief information about JPEG AI CfP status. In JVET-AA0047*, vol. 6, p. 13, 2022.

[46] S. Shen, H. Yue, and J. Yang, "Dec-adapter: Exploring efficient decoder-side adapter for bridging screen content and natural image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 887–12 896.

[47] S. Cai, L. Chen, S. Zhong, L. Yan, J. Zhou, and X. Zou, "Make lossy compression meaningful for low-light images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8236–8245.

[48] J. Pfaff, A. Filippov, S. Liu, X. Zhao, J. Chen, S. De-Luxán-Hernández, T. Wiegand, V. Rufitskiy, A. K. Ramasubramonian, and G. Van der Auwera, "Intra prediction and mode coding in VVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3834–3847, 2021.

[49] L. Li, H. Li, D. Liu, Z. Li, H. Yang, S. Lin, H. Chen, and F. Wu, "An efficient four-parameter affine motion model for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1934–1948, 2017.

[50] Z. Li, Z. Yuan, L. Li, D. Liu, X. Tang, and F. Wu, "Object segmentation-assisted inter prediction for versatile video coding," *IEEE Transactions on Broadcasting*, 2024.

[51] Z. Li, Y. Li, C. Tang, L. Li, D. Liu, and F. Wu, "Uniformly accelerated motion model for inter prediction," in *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2024, pp. 1–5.

[52] Z. Li, J. Li, Y. Li, L. Li, D. Liu, and F. Wu, "In-loop filtering via trained look-up tables," in *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2024, pp. 1–5.

[53] G. Lu, X. Zhang, W. Ouyang, L. Chen, Z. Gao, and D. Xu, "An end-to-end learning framework for video compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3292–3308, 2020.

[54] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[55] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4161–4170.

[56] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

[57] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[58] D. Ma, F. Zhang, and D. R. Bull, "BVI-DVC: A training database for deep video compression," *IEEE Transactions on Multimedia*, vol. 24, pp. 3847–3858, 2021.

[59] A. Mackin, F. Zhang, and D. R. Bull, "A study of high frame rate video formats," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1499–1512, 2018.

[60] ITU-T, "Subjective video quality assessment methods for multimedia applications," 1999.

[61] E. Peli, "Contrast in complex images," *JOSA A*, vol. 7, no. 10, pp. 2032–2040, 1990.

[62] A. Standard, "Digital transport of one-way video signals-parameters for objective performance assessment," *ANSI Standard ATIS-0100801.03*, 2003.

[63] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. SPIE, 2003, pp. 87–95.

[64] J. Liu, G. Lu, Z. Hu, and D. Xu, "A unified end-to-end framework for efficient deep image compression," *arXiv preprint arXiv:2002.03370*, 2020.

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[66] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "Compressai: a pytorch library and evaluation platform for end-to-end compression research," *arXiv preprint arXiv:2011.03029*, 2020.

[67] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.

[68] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU SG16 Doc. VCEG-M33*, 2001.

[69] X. Duan, S. Ma, H. Liu, and C. Jia, "PKU-AIGI-500K: A neural compression benchmark and model for ai-generated images," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2024.

[70] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[71] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Transactions on Image Processing*, vol. 29, pp. 8680–8695, 2020.

[72] C. Gao, D. Liu, L. Li, and F. Wu, "Towards task-generic image compression: A study of semantics-oriented metrics," *IEEE Transactions on Multimedia*, vol. 25, pp. 721–735, 2021.

[73] N. Yan, C. Gao, D. Liu, H. Li, L. Li, and F. Wu, "SSSIC: semantics-to-signal scalable image coding with learned structural representations," *IEEE Transactions on Image Processing*, vol. 30, pp. 8939–8954, 2021.

[74] Z. Yang, J. Miao, Y. Wei, W. Wang, X. Wang, and Y. Yang, "Scalable video object segmentation with identification mechanism," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[75] Y. Xu, Z. Yang, and Y. Yang, "Integrating boxes and masks: A multi-object framework for unified visual tracking and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9738–9751.

[76] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4768–4777.

[77] J. Li, C. Chen, Z. Cheng, and Z. Xiong, "Mulut: Cooperating multiple look-up tables for efficient image super-resolution," in *European Conference on Computer Vision*. Springer, 2022, pp. 238–256.

[78] Z. Xiao, X. Fu, J. Huang, Z. Cheng, and Z. Xiong, "Space-time distillation for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2113–2122.

[79] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.