# 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* 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 (scene type, texture, motion, view) and the designed imaging factors (illumination, shadow, lens). 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 the wider coverage and more diversity of the proposed dataset. We also evaluate both classic standardized and recent learned image/video coding schemes on USTC-TD with PSNR and MS-SSIM, and provide an extensive benchmark for the 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]–[27]. 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 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 forty images and ten video sequences with a wide content coverage. For the image dataset, each image has a high spatial resolution (4K) in RGB, YUV444/420 color space, and PNG/YUV file format. For the video dataset, each video sequence consists of 96 frames, and each frame is captured at 30 frames per second (fps) with 1080p spatial resolution in RGB, YUV444/420 color space, and PNG/YUV file format. For the construction of datasets, the data is collected with the specific design of different content elements (environmental/capture-related elements), which aims to cover as close as possible to the real-world coding transmission scenes. Compared with the previous common test image/video datasets [28]–[34], we use different quantitative criteria to comprehensively evaluate the diversity of the proposed dataset from the perspective of spatial, temporal, colorfulness, and lightness information, and demonstrate the wide coverage for various image/video features.

In addition, we establish baselines and evaluate recent advanced state-of-the-art standardization activities [1]–[5], [35], [36] and traditional/learned image/video compression schemes [6]–[12], [16], [19]–[23], [25]–[27] under different metrics (*PSNR*, *MS-SSIM* [37]), and then benchmark and analyze their performance on the proposed dataset to promote the development of new compression algorithms in the future. The test results and scripts are open-sourced with the dataset and released on the website[1] for researchers to reproduce conveniently.

We hope the proposed test datasets allow researchers to make more well-informed decisions under efficient evaluation,

Date of current version September 16, 2024.

*: Zhuoyuan Li and Junqi Liao contribute equally to this work.

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]https://esakak.github.io/USTC-TD/benchmark

TABLE I
COMMON TEST DATASETS OF IMAGE COMPRESSION

| Dataset | Resolution | Number | Color Space | Bit Depth | Setting | Characteristic |
|---|---|---|---|---|---|---|
| *Kodak* [28] | 768×512 | 24 | RGB | 24 | - | Rich Texture |
| *Tecnick* [29] | 1200×1200 | 100 | RGB | 24 | Sampling | Various Scenery |
| *CLIC* [30] | 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*** [31] | 3840×2160 | 16 | 50/120 | 5s-12s | Fast/Slow Motion |
| ***MCL-JCV*** [32] | 1920×1080 | 30 | - | 5s | Diverse Video Scenes |
| ***HEVC CTC*** [33], ***VVC CTC*** [34] | 240P-4K | 41 | 24-60 | 150-600 Frames | Appropriate Exposure |

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 (termed USTC-TD), which focuses on the diversity of various content elements.
- We conduct a comprehensive evaluation of the proposed dataset by using different quantitative criteria, and demonstrate the wide coverage of various image/video features.
- We conduct a comprehensive evaluation of the advanced image/video compression schemes on the proposed USTC-TD dataset, and establish an extended baseline for the compression-related research benchmarked on USTC-TD.
- Taking a close look at USTC-TD, we analyze the benchmarked performance and point out the potential directions for future compression-related research.

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/video dataset of USTC-TD, and analyzes the characteristics of the proposed dataset. Section V presents the experimental configuration and the evaluation of the advanced compression schemes, and further analyzes their performance, limitation, and inspiration. Section VI concludes the paper and presents some suggestions for future work. Finally, we mention the copyright for the usage of USTC-TD.

## 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* [28], *Tecnick* [29] (sampling setting), *CLIC* (professional setting) [30] are commonly used, the setting are mentioned in the Table I , and the introduction is summarised below:

- ***Kodak*** [28] 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*** [29] 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*** [29] is a high-quality image set collected from *Unsplash* [38], 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 elements makes it difficult to evaluate the robustness of algorithms.

**Video Compression Test Dataset.** For the evaluation of previous video compression schemes, *UVG* [31], *MCL-JCV* [32], *HEVC Common Test Conditions* (*CTC*) [33], VVC CTC [34] are commonly used, the settings are mentioned in the Table II , and the introduction is summarised below:

- ***UVG*** [31] 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 [39]. It is widely used in the evaluation of advanced learned video compression methods [12], [16], [19]–[23], [25]–[27].
- ***MCL-JCV*** [32] 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.

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*** [33], [34] 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, MP4) and then converted to RGB, YUV 4:4:4, YUV 4:2:0 color space/format by using the *ffmpeg* [39] tool and the specific conversion standard of color space (s.t. BT.601, BT.709 [26], [40]).

### B. Collection Configuration

To develop a comprehensive and diverse image/video dataset, we consider the various content elements of collected data, including the environmental elements (scene type, texture, motion, view), capture elements (illumination, quality, lens), which cover as close as possible to the real-world coding transmission scenes. For each element, 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 (such as dormitory, library, river bank, institutes, parks, classroom, street, vehicles, *et al*), 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 better one. For each video, we record five minutes for each scene with the same range of joint sense to select the ninety-six frames.

TABLE V
COLLECTION CONFIGURATION OF USTC-TD 2022 AND 2023

| **Collection Configuration** | | |
|---|---|---|
| **Element** | **Category** | **Example** |
| **Texture** | Structural,<br>Natural,<br>Geometric | Scenery,<br>People,<br>Gridding |
| **Motion** | Complex,<br>Medium,<br>Tiny | Occlusion,<br>Walking,<br>Chatting |
| **View** | Upward Level,<br>Horizontal Level,<br>Overhead Level | Building,<br>People,<br>Close Shot |
| **Illumination** | Appropriate Exposure,<br>Underexposure,<br>Overexposure | Natural Light,<br>Dark Light,<br>High Light |
| **Lens** | Moving,<br>Fix | Camera Motion,<br>Surveillance |
| **Quality** | High,<br>Medium | Captured quality<br>parameter of camera |

## IV. DATASET CONSTRUCTION AND ANALYSIS

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 [28]–[30], our proposed dataset aims to cover various scenarios, and try to collect and simulate the data in the real-world coding transmission scenes, which makes the evaluation of image coding schemes more closer to the actual application.

Considering the various content elements, we combine different environmental conditions and captured conditions in the collection process. For the diversity of environmental conditions, we consider the scene type, texture, and view elements. For the diversity of camera parameters, we consider the resolution and quality, illumination elements. In Fig. 1, we show the all collected image data of USTD-TD 2022, and in Table VI, we show the specific configuration of each image, and make it convenient to the researchers' scheme design for different real-world 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 elements in real-world scenes.

Compared to the previous image dataset [28]–[30], more specific real-world scenes and content elements are considered


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 | Quality | Texture | Illumination | View |
|---|---|---|---|---|---|---|
| ***USTC-2022-01*** | Scenery | 4096×2160 | Medium | Geometric | Appropriate Exposure | Horizontal Level |
| ***USTC-2022-02*** | Scenery | 4096×2160 | High | Structural | Appropriate Exposure | Upward Level |
| ***USTC-2022-03*** | Scenery | 4096×2160 | High | Structural | Underexposure | Horizontal Level |
| ***USTC-2022-04*** | Scenery | 4096×2160 | High | Structural | Appropriate Exposure | Horizontal Level |
| ***USTC-2022-05*** | People | 4096×2160 | Medium | Natural | Overexposure | Horizontal Level |
| ***USTC-2022-06*** | Scenery | 4096×2160 | High | Geometric | Appropriate Exposure | Upward Level |
| ***USTC-2022-07*** | Building | 4096×2160 | High | Geometric | Appropriate Exposure | Upward Level |
| ***USTC-2022-08*** | Scenery | 4096×2160 | High | Geometric | Appropriate Exposure | Upward Level |
| ***USTC-2022-09*** | People, Building | 4096×2160 | High | Natural | Underexposure | Horizontal Level |
| ***USTC-2022-10*** | Building | 4096×2160 | Medium | Geometric | Appropriate Exposure | Upward Level |
| ***USTC-2022-11*** | People, Scenery | 4096×2160 | High | Natural | Appropriate Exposure | Horizontal Level |
| ***USTC-2022-12*** | People | 4096×2160 | High | Natural | Underexposure | Horizontal Level |
| ***USTC-2022-13*** | Scenery | 4096×2160 | High | Structural | Appropriate Exposure | Upward Level |
| ***USTC-2022-14*** | People, Vehicle | 4096×2160 | High | Natural | Appropriate Exposure | Horizontal Level |
| ***USTC-2022-15*** | People | 4096×2160 | High | Natural | Underexposure | Upward Level |
| ***USTC-2022-16*** | People | 4096×2160 | High | Natural | Appropriate Exposure | Horizontal Level |
| ***USTC-2022-17*** | People, Building | 4096×2160 | High | Natural | Overexposure | Horizontal Level |
| ***USTC-2022-18*** | People, Building | 4096×2160 | High | Natural | Appropriate Exposure | Horizontal Level |
| ***USTC-2022-19*** | People, Building | 4096×2160 | Medium | Natural | Overexposure | Horizontal Level |
| ***USTC-2022-20*** | People, River | 4096×2160 | High | Natural | Appropriate Exposure | Overhead Level |

TABLE VII
THE CONFIGURATION OF USTC-TD 2023 IMAGE DATASET

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

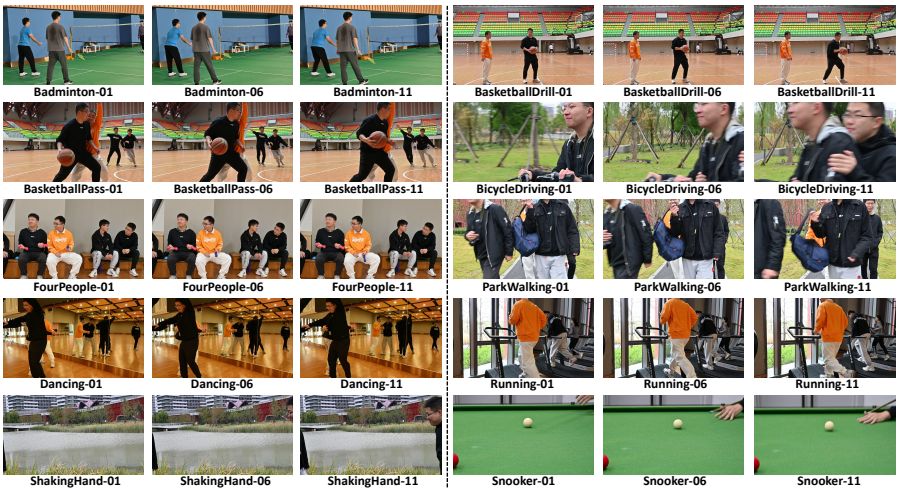



Fig. 2. Illustration of the frame 01, 06, 11 of each video sequence in video test dataset of USTC-TD 2023.

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 | Medium | Natural | Horizontal Level | Moving |
| ***USTC-BasketballDrill*** | YUV420, 444, RGB | Medium | People, Sport | 1920×1080 | High | Natural | Horizontal Level | Moving |
| ***USTC-BasketballPass*** | YUV420, 444, RGB | Medium | People, Sport | 1920×1080 | High | Natural | Horizontal Level | Moving |
| ***USTC-BicycleDriving*** | YUV420, 444, RGB | Complex | People, Daily Life | 1920×1080 | High | Natural | 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 | Horizontal Level | Moving |
| ***USTC-Running*** | YUV420, 444, RGB | Complex | People, Sport | 1920×1080 | High | Natural | Horizontal Level | Moving |
| ***USTC-ShankingHand*** | YUV420, 444, RGB | Complex | People, Daliy life | 1920×1080 | High | Natural | 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 | Medium | Natural | Horizontal Level | Fix |

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 [41]. In *USTC-2023-16*, we capture the scenes with the object occlusion and the pixel/spatial-wise correlation becomes low, which is a challenge for traditional intra-prediction schemes [42] 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.

### *B. Construction of USTC-TD 2023 Video Dataset*

Based on the characteristics of previous video datasets [31]–[34], 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 conditions 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 [43]–[46]. 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 researchers' scheme design 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.

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 [47] and many optical flow-based video compression schemes [12], [16], [19]–[23], [25]–[27], [48] (the performance 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 [49], [50], 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.

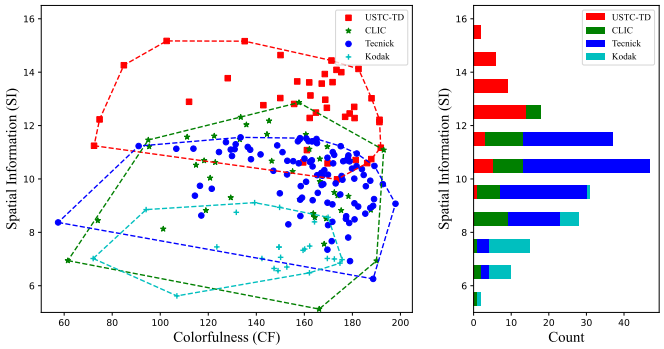



Fig. 3. The visualization of the evaluation of spatial information (SI) and colorfulness (CF) features on different image test datasets. Scatter diagram represents the SI versus CF, and corresponding convex hulls indicates the coverage of different datasets. The histogram represents the number of images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

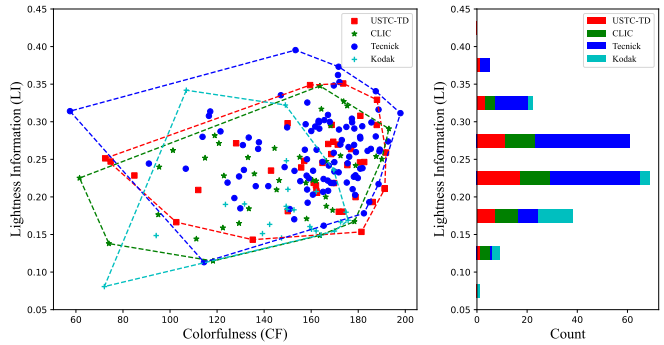



Fig. 4. The visualization of the evaluation of lightness information (LI) and CF features on different image test datasets. Scatter diagram represents the LI versus CF, and corresponding convex hulls indicates the coverage of different datasets. The histogram represents the number of images under different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

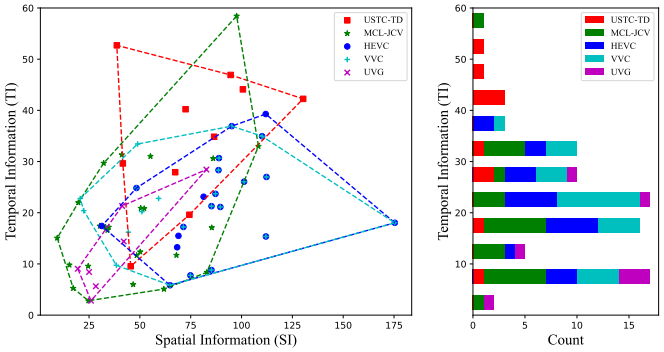



Fig. 5. The visualization of the evaluation of temporal information (TI) and SI features on different video test datasets. Scatter diagram represents the TI versus SI, and corresponding convex hulls indicates the coverage of different datasets. The histogram represents the number of videos under different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

### C. Analysis of USTC-TD Dataset

In the above subsections, the construction of each dataset of USTC-TD is introduced, and we also point out some specific characteristics of the proposed dataset. To comprehensively verify the wide coverage of our proposed dataset for various content elements 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* [28], *CLIC* [30], *Tecnick* [29], video datasets: *HEVC CTC* [33], *VVC CTC* [34], *MCL-JCV* [32], *UVG* [31]). For analysis of image/video features, we select the spatial information (SI) [51], colorfulness (CF) [51], lightness information (LI) [52], and temporal information (TI) [31] to characterize each dataset along the dimensions of space, color, lightness, and temporal correlation, which are commonly used to evaluate the quality of dataset [31], [53], [54]. The definitions of the evaluation of these features can be found in [31], [51], [52], [55], [56], and the detailed calculation schemes are as follows:

- ***Spatial information (SI)***: SI is used as a representation of edge energy [57]. Followed by [51], 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 is the total number of pixels in the filtered image, and L is 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 [31], 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 [31], [58], 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 [56], 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 [31], [55], 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.

TABLE IX
QUANTITATIVE RESULTS OF THE USTC-TD 2022 AND 2023 IMAGE DATASETS. NOTE THE HIGHEST VALUES ARE REPRESENTED IN RED, AND THE LOWEST VALUES 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 HIGHEST VALUES ARE REPRESENTED IN RED, AND THE LOWEST VALUES ARE REPRESENTED IN BLUE

| *USTC-TD 2023* | | |
|---|---|---|
| ***Sequence*** | ***SI*** | ***TI*** |
| ***USTC-Badminton*** | 67.37 | 27.91 |
| ***USTC-BasketballDrill*** | 130.31 | 42.25 |
| ***USTC-BasketballPass*** | 94.63 | 46.92 |
| ***USTC-BicycleDriving*** | 38.66 | 52.70 |
| ***USTC-Dancing*** | 74.34 | 19.61 |
| ***USTC-FourPeople*** | 45.49 | 9.54 |
| ***USTC-ParkWalking*** | 72.48 | 40.22 |
| ***USTC-Running*** | 86.52 | 34.84 |
| ***USTC-ShakingHands*** | 100.67 | 44.11 |
| ***USTC-Snooker*** | 41.62 | 29.63 |

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 [28]–[30], 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 the USTC-TD video dataset, 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 [31]–[34], the USTC-TD exhibits a wide range of temporal variation with generalized spatial variation, ranging from 5 to 55. It further compensates for the absence of the 40 to 55 (higher) range in the temporal variation of other video datasets, which enables a wide coverage of temporal diversity for the comprehensive assessment of video compression-related algorithms.

In addition, we present the detailed variety distribution of these different features of USTC-TD image/video datasets, as shown in Table IX and Table X. The higher/lower scores of different evaluative features are represented in different colors. Benefiting from the specific design of various content elements (environmental elements, capture-related elements) of each image/video, we can find that the typical features of different content elements are evenly distributed in the captured image/video to ensure the diversity. For example, first, as mentioned in Section IV.A and Table VI, VII, IX, the setting of illumination (overexposure/underexposure) of *USTC-2022-17*, and *USTC-2023-05*, *USTC-2023-09* enables the highest/lowest score of LI evaluative features, which demonstrates its specific design of capture-related elements. Second, as mentioned in Section IV.B and Table VIII, the setting of complex motion (high-speed/tiny-speed moving objects) and lens (moving/fix) of *USTC-BicycleDriving* and *USTC-FourPeople* enables the higher/lower score of TI evaluative features, which demonstrates its specific design of environmental elements. We hope these specific design can help the researchers analyze and discover the bottlenecks of their research. The above analysis results and related codes are open-sourced with the dataset and released on the website[2].

## V. EXPERIMENTS

In this section, first, we present the experimental configurations employed for the evaluation of compression schemes. Second, we evaluate the recent advanced image/video compression schemes and standardization activities 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 schemes.

### 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 methods. Second, we introduce the evaluation metrics for these methods on our proposed dataset.

*1) Selection of Evaluative Methods:* For advanced image compression schemes, we select several common traditional and advanced learned schemes. For traditional codecs, we select *BPG* [35] and *Versatile Video Coding (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

[2]https://esakak.github.io/USTC-TD/analysis

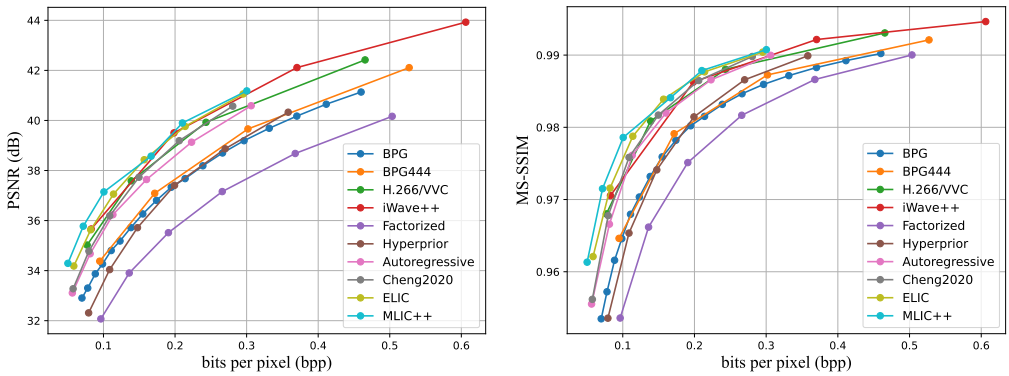



Fig. 6. Overall rate-distortion (RD) curves of advanced image compression schemes on *PSNR* and *MS-SSIM* metrics, the results are evaluated on USTC-TD image dataset 2022 and 2023.

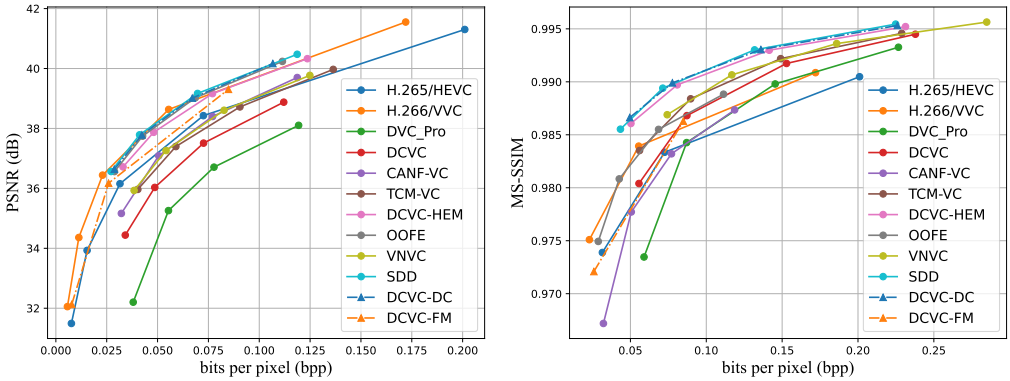



Fig. 7. Overall rate-distortion (RD) curves of advanced video compression schemes on *PSNR* and *MS-SSIM* metrics, the results are evaluated on USTC-TD video dataset 2023.

*MLIC++* [11]. For advanced learned compression standardization activities, we select the high-profile model of *IEEE 1857.11* [36] (*iWave++*).

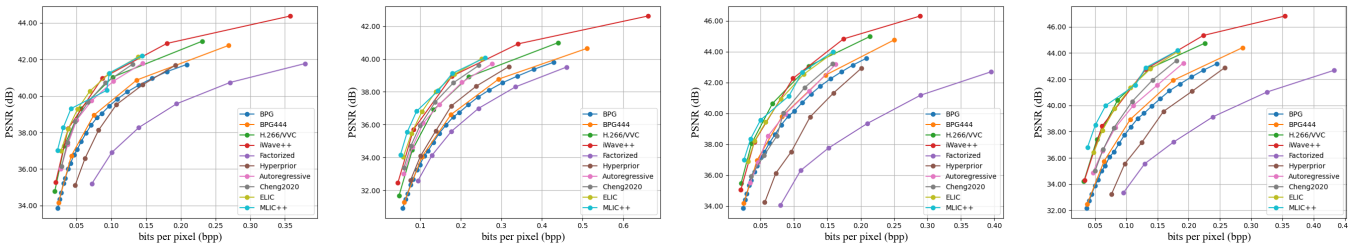
For video compression schemes, we also select common traditional and advanced learned schemes. For traditional codecs, we select the *High Efficiency Video Coding (H.265/HEVC)* [1], and *Versatile Video Coding (H.266/VVC)* [2]. For learned video compression schemes, we select the *DVC_Pro* [12], [48], *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 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 (*VVC* reference software) are chosen. For *BPG*, the default configuration is used, and the internal color space is set to YUV444/YUV420 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 [28]–[30]), we convert them to the YUV444 color space for the input of VTM-17.0 by using the *ffmpeg* [39] tool.

*3) Testing Configurations of Evaluative Traditional Video Compression Schemes:* For testing, the officially released HM-16.20 (*HEVC* reference software) and VTM-13.2 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. The internal color space is set to YUV444. For the different source formats of these testing datasets (USTC-TD and [31]–[34]), we also convert them to the YUV444 color space as the input of the above traditional codecs by using the *ffmpeg* [39] tool.

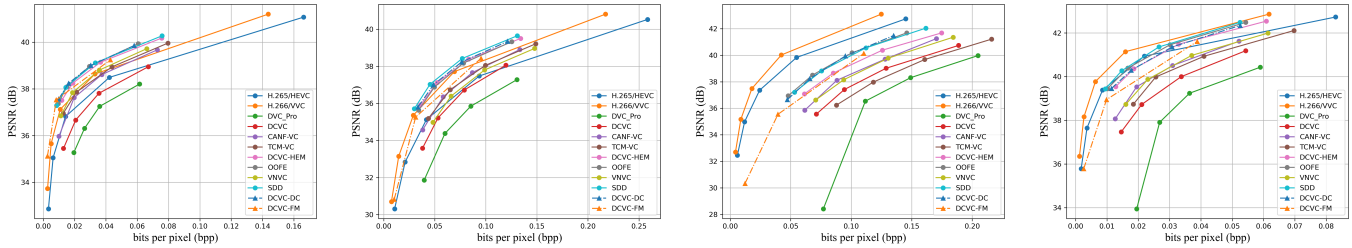
*4) Training and Testing Configurations of Evaluative Learned Image Compression Schemes:* For training, these learned image compression models are optimized by mean squared error (*MSE*) or multi-scale structural similarity index measure (*MS-SSIM*), and the *Flicker2W* [59] dataset is used as the training dataset. These models are optimized by using the Adam Optimizer [60], with a batch size of 8 and a patch size of $256 \times 256$. 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 $\lambda$ 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++* and *iWave++*, and the reproduced model of *ELIC* are used. The model of other methods is provided by *CompressAI* [61].

*5) Training and Testing Configurations of Evaluative Learned Video Compression Schemes:* For training, these learned video compression models are mainly optimized by mean squared error (*MSE*) or multi-scale structural similarity index measure (*MS-SSIM*), and the *Viemo-90k* [62] is used as the training set. 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 source formats of these testing datasets, we convert them from




(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 videos under *PSNR* metric.

YUV444 to the RGB / YUV420 color space as the input of different learned video codecs by using the BT.709 conversion standard (adopted in JPEG AI [40]) / *ffmpeg* [39] tool.

*6) Evaluation Metrics:* We use *PSNR* and *MS-SSIM* [37] to measure the quality of the coded frames in the comparison to the original frames. Bits per pixel (bpp) is used to measure the number of bits for encoding each pixel in each image or video frame. The Bjontegaard Delta bitrate (BD-rate) [63] 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 and video compression schemes, the *PSNR* and *MS-SSIM* are both calculated and compared in RGB color space. For image, the conversion process of different color spaces is performed by using the *ffmpeg* tool [39]. For video, the conversion process is aligned to the setting mentioned in *DCVC-DC* [26] by using the BT.709 conversion standard.

### *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) Evaluation and Analysis of Advanced Image Compression Schemes on USTC-TD Image Datasets:* Taking bpp as the horizontal axis and the reconstructed *PSNR/MS-SSIM* 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 partial 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 further analyze their performance from the perspective of content elements 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 the Section II of supplementary material. Based on the performance comparison of these schemes and feature analysis of proposed datasets (Section IV.C), the conclusions mainly include the following four aspects:

- (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 elements (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 [28]–[30] 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/capture-related elements 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 real-world 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 datasets [28]–[30], the proposed datasets can collaborate with other datasets to handle a wide coverage

TABLE XI
BD-RATE (%) COMPARISON FOR PSNR. THE ANCHOR IS VTM.

| Dataset | *VTM* | *HM* | *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** | **0.0** | **46.9** | **286.7** | **133.0** | **69.3** | **69.8** | **63.9** | **17.0** | **9.5** | **4.5** | **7.1** | **23.1** |
| **Average** | **0.0** | **41.1** | **229.6** | **139.4** | **71.3** | **44.5** | **36.3** | **1.5** | **-9.6** | **-11.1** | **-16.3** | **-10.2** |

TABLE XII
BD-RATE (%) COMPARISON FOR MS-SSIM. THE ANCHOR IS VTM.

| Dataset | *VTM* | *HM* | *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** | **0.0** | **47.9** | **38.6** | **-18.1** | **77.5** | **-41.1** | **-62.1** | **-62.3** | **8.4** | **-66.3** | **-68.6** | **37.1** |
| **Average** | **0.0** | **39.4** | **71.9** | **28.8** | **41.9** | **-23.2** | **-30.3** | **-47.9** | **-11.3** | **-50.3** | **-55.1** | **-10.7** |

of performance evaluation, which also demonstrates the efficiency of specific design of the proposed datasets' different elements/features.

*2) Evaluation and Analysis of Advanced Video Compression Schemes on USTC-TD Video Dataset:* Taking bpp as the horizontal axis and the reconstructed *PSNR/MS-SSIM* 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 XI and Table XII. From the overall results of *PSNR* metric, we can find that the performance of the advanced traditional schemes is better than that of the all advanced learned schemes on the proposed dataset, which is opposed to the performance on other datasets [31]–[34]. From the overall results of *MS-SSIM* metric, the performance of the traditional schemes is lower than that of advanced learned schemes. Here we further analyze their performance from the perspective of typical characteristics of different test video contents 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 the Section II of 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 three aspects:

- (1) The traditional schemes show the robust generalization ability on various real-world scenarios. As shown in Table XI, different from the performance of advanced compression schemes on other datasets [31]–[34], the traditional schemes still achieve the state-of-the-art performance of these schemes on USTC-TD. 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 specific design of lens motion (mentioned in 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, such as the *USTC-FourPeople* with lower scores of TI, and the *USTC-BasketballPass* with higher scores of TI. These scenarios commonly appear in the previous test datasets [31]–[34], 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 analysis of video-related features (Section IV.C), our proposed dataset can make up more typical motion/temporal correlation-related real-world elements 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 elements/features. Furthermore, compared to the performance of other datasets [31]–[34] of different schemes mentioned in [7]–[11], as shown in Table XI and Table XII, the different phenomenon between the performance of our proposed dataset and other datasets also verifies the efficiency of the proposed dataset.

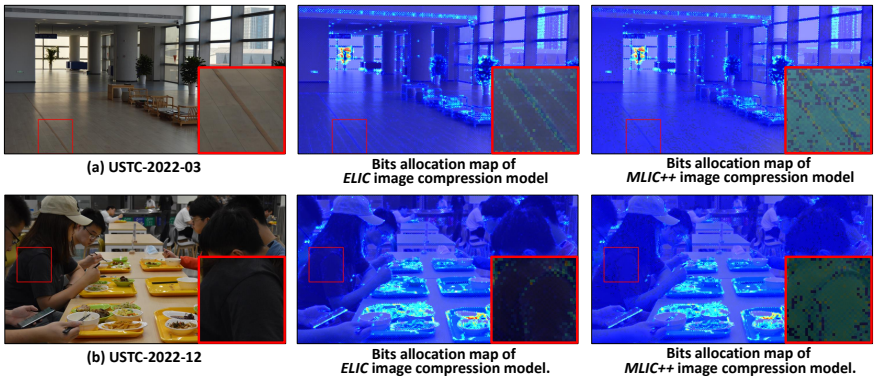



Fig. 10. Visualization of limitation of partial advanced image compression schemes (*ELIC* [10], *MLIC*++ [11]) on *USTC-2022-03* and *USTC-2022-12*.

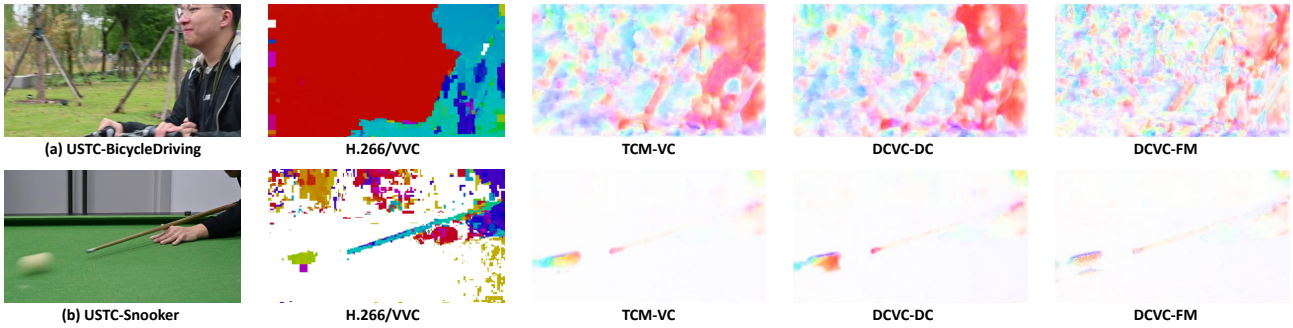



Fig. 11. Visualization of limitation of partial advanced video compression schemes (*TCM-VC* [20], *DCVC-DC* [26], *DCVC-FM* [27]) on *USTC-BicycleDriving* and *USTC-Snooker*.

### C. Limitation and Inspiration

In this subsection, based on the analysis of experimental results, we further 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 optimization direction of deep learning-based manner with the limited evaluation data. Here we further 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 XI, 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

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

| Scheme | BD rate (%) |
|---|---|
| *DCVC-DC* | 7.1% |
| *DCVC-DC + Flow Pre-training* | -1.6% |
| △ | **-8.7%** |

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* [50]) 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 XIII. 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. CONCLUSION

In this paper, we propose a test dataset (USTC-TD) for compression-related research, which covers more diverse content elements. 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 wide coverage. In addition, we evaluate the advanced compression methods 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) [64]–[66], such as object segmentation [67], object detection [68], action recognition [69], *et al*, and the reconstruction dataset of video enhancement, such as image/video super-resolution [70], [71], denoising [72], warping [73], *et al*, for testing of compression-related downstream researches, and provide a comprehensive baseline to promote the development of researchers' compression-related diverse tasks.

## VII. COPYRIGHT

The released images and sequences are captured and processed by the University of Science and Technology of China (USTC). All intellectual property rights remain with USTC.

The following uses are allowed for the contributed dataset: (1) Data (images and videos) may be published in research papers, technical reports, and development events. (2) Data (images and videos) may be utilized for standardization activities (e.g., ITU, MPEG, AVS, VQEG).

The following uses are NOT allowed for the contributed dataset: (1) Do not publish snapshots in product brochures. (2) Do not use video for marketing purposes. (3) Redistribution is not permitted. (4) Do not use it in television shows, commercials, or movies.

## 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 neural 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. 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] Kodak, "Kodak lossless true color image suite (PhotoCD PCD0992)," *https://r0k.us/graphics/kodak/*, 1993.

[29] 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.

[30] "Workshop and Challenge on Learned Image Compression (CLIC) in CVPR," *http://www.compression.cc*, 2020.

[31] 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.

[32] 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.

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

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

[35] F. Bellard, "Bpg Image Format," *https://bellard.org/bpg*, 2015.

[36] "IEEE 1857.11, IEEE C/DC 1857.11 Sub-Working Group," *https://sagroups.ieee.org/fvc/*, 2021.

[37] 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.

[38] "Unsplash: Beautiful Free Images & Pictures," *https://unsplash.com/*.

[39] "FFMPEG," in *https://ffmpeg.org/*.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Z. Li, Z. Yuan, L. Li, D. Liu, X. Tang, and F. Wu, "Object segmentation-assisted inter prediction for versatile video coding," *arXiv preprint arXiv:2403.11694*, 2024.

[45] Z. Li, Y. Li, C. Tang, L. Li, D. Liu, and F. Wu, "Uniformly accelerated motion model for inter prediction," *arXiv preprint arXiv:2407.11541*, 2024.

[46] Z. Li, J. Li, Y. Li, L. Li, D. Liu, and F. Wu, "In-loop filtering via trained look-up tables," *arXiv preprint arXiv:2407.10926*, 2024.

[47] "Practical End-to-End Image/Video Compression Challenge," in *IEEE International Conference on Visual Communications and Image Processing (VCIP), https://vcip2023.iforum.biz/page/goto/*, 2023.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] P. ITU-T RECOMMENDATION, "Subjective Video Quality Assessment Methods for Multimedia Applications," 1999.

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

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

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] J. Li, C. Chen, W. Huang, Z. Lang, F. Song, Y. Yan, and Z. Xiong, "Learning steerable function for efficient image resampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5866–5875.

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



## I. Overview of the Evaluative Image/Video Compression Scheme

Here we mention the characteristics and settings of image/video evaluative compression methods in our experiments (Section V of main text).

### *A. Characteristics of Evaluative Image/Video Compression Schemes*

Here we mention the characteristics of each test image/video compression scheme in detail.

*1) Standardized Schemes:* ***BPG*** [1]: BPG (Better Portable Graphics) is an image format. Its purpose is to replace the JPEG image format when quality or file size is an issue. It is based on a subset of the HEVC and supported by most Web browsers with a small Javascript decoder (gzipped size: 56KB). It supports the same chroma formats as JPEG and has the higher compression ratio than JPEG.

***High Efficiency Video Coding (H.265/HEVC)*** [2]: High Efficiency Video Coding (HEVC), also known as H.265 and MPEG-H Part 2, is a video compression standard designed as part of the MPEG-H project as a successor to the widely used Advanced Video Coding (AVC, H.264, or MPEG-4 Part 10). In comparison to AVC, HEVC offers from 25% to 50% better data compression at the same level of video quality, or substantially improved video quality at the same bit rate. It supports resolutions up to 8192×4320, including 8K UHD, and unlike the primarily 8-bit AVC, HEVC's higher fidelity Main 10 profile has been incorporated into nearly all supporting hardware.

***Versatile Video Coding (H.266/VVC)*** [3]: Versatile Video Coding (VVC), ISO/IEC 23090-3, and MPEG-I Part 3, is a video compression standard finalized on 2020, standardized by the Joint Video Experts Team (JVET), a joint video expert team of the VCEG working group of ITU-T Study Group 16 and the MPEG working group of ISO/IEC JTC 1/SC 29. It is the successor to HEVC. It improves compression performance and supports for a very broad range of applications.

***IEEE 1857.11*** [4]: IEEE 1857.11 provides efficient, neural network-based coding tools for compression, decompression, and packaging of image data, which significantly improve the compression efficiency compared to IEEE Std 1857.4 (intra-picture coding) and IEEE Std 1857.10 (intra-picture coding) under comparable settings, and facilitate the compression and decompression on top of neural network-oriented computing infrastructures like neural network processing units (NPUs). The target applications and services include but are not limited to Internet images, user-generated images, and other image-enabled applications and services such as digital image storage and communications.

*2) Other Learned Image Coding Schemes:* ***Factorized Model*** [5]: Factorized Model introduces the convolutional-based transform network employing generalized divisive normalization (GDN) [6] and the factorized entropy model. It suggests the use of additive uniform noise to address non-differentiable quantization. Notably, this is the first learned image compression method that surpasses JPEG2000 [7] on both RGB BD-rate PSNR and RGB BD-rate MS-SSIM.

***Hyperprior Model*** [5]: The pioneering work proposes a hyperprior entropy model for learned image compression, significantly enhancing compression performance. It has since been widely used and remains influential in the field.

***Autoregressive Model*** [8]: Autoregressive Model proposes combining an autoregressive spatial context model with a hyperprior for more accurate entropy estimation. This method marks the first learning-based image compression approach to outperform BPG [1] on both RGB BD-rate PSNR and RGB BD-rate MS-SSIM.

***Cheng2020*** [9]: Cheng2020 utilizes the discretized Gaussian Mixture Model to estimate the distributions of latent codes, while employing attention modules in transform to enhance performance. Notably, this is the first work to achieve comparable performance with the latest compression standard VVC [3] in terms of both RGB BD-rate PSNR and RGB BD-rate MS-SSIM.

***iWave++*** [10]: iWave++ proposes a versatile learned image compression scheme with a trained wavelet-like transform. It supports both lossy and lossless image compression. This model is accepted as the high-profile model in the IEEE 1857.11 [4] standard.

***ELIC*** [11]: ELIC adopts the space-channel context model in entropy estimation and stacked residual blocks as the non-linear transform. This method not only demonstrates superior performance but also supports fast preview decoding.

***MLIC++*** [12]: MLIC++ utilizes channel-wise, local spatial, and global spatial information to achieve better entropy estimation. Compared to concurrent models like ELIC, this method achieves state-of-the-art compression performance.

*3) Learned Video Coding Schemes:* ***DVC_Pro*** [13], [14]: The first work proposes the motion-compensated prediction and residual coding framework of end-to-end learned video compression. DVC_pro follows the traditional video compression framework and replaces all the modules with neural networks. When intra period is 10 for HEVC videos and 12 for non-HEVC videos, experimental results show that DVC_pro performs better than H.264 [15] on both RGB BD-rate PSNR

and RGB BD-rate MS-SSIM. Besides, DVC_pro also achieves comparable compression performance with H.265 [2] on both RGB BD-rate PSNR and RGB BD-rate MS-SSIM.

***DCVC*** [16]: The first work proposes the motion-compensated prediction and conditional coding framework of end-to-end learned video compression. DCVC utilizes the learned temporal correlation between the current frame and the predicted frame rather than the subtraction-based residual. When the intra period is 10 for HEVC videos and 12 for non-HEVC videos, its experimental results perform better than DVC_Pro [14] and x265[1] on both RGB BD-rate PSNR and RGB BD-rate MS-SSIM. Specifically, it achieves an average 18.40% and 17.82% RGB BD-rate reduction, respectively, in terms of PSNR compared to x265 and DVC_Pro [14].

***CANF-VC*** [17]: CANF-VC is the first conditional augmented normalizing flows-based end-to-end learned video compression framework. CANF-VC leverages the conditional augmented normalizing flows to learn a video generative model for conditional inter-frame coding, and extends the conditional coding to motion coding, forming a purely conditional coding framework. Under intra period 10/12 configuration, its experimental results perform better than M-LVC [18], DCVC [16] and H.265 [2] on RGB BD-rate PSNR and MS-SSIM. Under intra period 32 configuration, its experimental results perform better than those of M-LVC [18], DCVC [16] on RGB BD-rate PSNR, and H.265 [2] on RGB BD-rate MS-SSIM.

***TCM-VC*** [19]: Based on DCVC [16], TCM-VC further proposes the multi-scale temporal context mining to better utilize the temporal correlation. Under intra period 32 configuration, the experimental results show that the compression performance of neural video codecs (DVC_Pro [13], [14], M-LVC [18], RLVC [20], DCVC [16]) is greatly reduced while their framework still achieves an average 14.4% RGB BD-rate reduction against HM-IPP in terms of PSNR and achieves about 21.1% BD-rate reduction in terms of MS-SSIM.

***DCVC-HEM*** [21]: Based on TCM-VC [19], HEM further designs a parallel-friendly entropy model that explores both temporal and spatial dependencies. Besides, it also supports variable bitrates in a single mode. HEM is the first end-to-end neural video codec to exceed H.266 [22] using the highest compression ratio configuration. Under intra period 32 configuration, the experimental results show that the compression performance of neural video codecs (DVC_Pro [13], [14], M-LVC [18], RLVC [20], DCVC [16], TCM-VC [19]) is greatly reduced while their framework still achieves an average 4.7% RGB BD-rate reduction against VTM-IPP in terms of PSNR and achieves about 46.4% BD-rate reduction in terms of MS-SSIM.

***OOFE*** [23]: OOFE proposes an offline and online optical enhancement strategy for the flow-based end-to-end learned video compression framework, which is integrated into DCVC and DCVC-DC, respectively. Experimental results demonstrate that the proposed offline and online enhancement together achieves on average 13.4% bitrate saving for DCVC [16] and 4.1% bitrate saving for DCVC-DC [24] on RGB BD-rate PSNR when intra period is 12. It's worth noting that, for a fair comparison, we exclusively utilize the offline optical flow enhancement when testing the USTC-TD.

***VNVC*** [25]: The first work proposes a versatile neural video coding framework for both human and machine vision, it uses a single-bitstream, compact coded representation, targeting video reconstruction, video enhancement, and video analysis tasks simultaneously. It reports the experimental results on video reconstruction, video enhancement and video analysis tasks, respectively. For video reconstruction, it regards H.266/VVC official reference software VTM-13.2 with VTM-IPP configuration (one reference frame and flat QP) as the anchor. Under intra period 12 configuration, it achieves an average 7.0% BD-rate reduction against VTM-IPP in terms of PSNR and achieves about an average 49.9% BD-rate reduction in terms of MS-SSIM. Under intra period 32 configuration, the experimental results show that the compression performance of neural video codecs (DVC_Pro [13], [14], DCVC [16], CANF-VC [17], TCM-VC [19]) is greatly reduced while their framework still achieves an average 4.6% BD-rate reduction against VTM-IPP in terms of PSNR and achieves about 50.3% BD-rate reduction in terms of MS-SSIM. For video enhancement and video analysis tasks, their framework achieves excellent performance compared to other learning video coding frameworks.

***SDD*** [26]: Based on the motion-compensated prediction and conditional coding framework, SDD proposes a structure and detail decomposition-based motion model and a long short-term temporal contexts fusion mechanism. Under intra period 32 configuration, the experimental results show that the compression performance of neural video codecs (DVC_Pro [13], [14], M-LVC [18], RLVC [20], DCVC [16], CANF-VC [17], TCM-VC [19], HEM [21]) is greatly reduced while their framework still achieves an average 13.4% RGB BD-rate reduction against VTM-IPP in terms of PSNR and achieves about 48.0% BD-rate reduction in terms of MS-SSIM.

***DCVC-DC*** [24]: Based on HEM [21], DCVC-DC further increases the context diversity in both temporal and spatial dimensions by introducing the group-based offset diversity and quadtree-based partition. Under intra period 32 configuration, the experimental results show that the compression performance of neural video codecs (DVC_Pro [13], [14], DCVC [16], TCM-VC [19], HEM [21]) is greatly reduced while their framework still achieves an average 17.8% RGB BD-rate reduction against VTM-IPP in terms of PSNR and achieves about 47.6% BD-rate reduction in terms of MS-SSIM.

***DCVC-FM*** [27]: Based on DCVC-DC [24], DCVC-FM exploits the training with longer video and proposes a periodically refreshing mechanism. Besides, it modulates the latent feature via the learnable quantization scaler to support a wide quality range in a single model. Under intra period 32 configuration, the experimental results show that the compression performance of neural video codecs (DVC_Pro [13], [14], RLVC [20], DCVC [16], CANF-VC [17], TCM-VC [19], HEM [21], DCVC-DC [24]) while their framework still achieves an average 20.3% RGB BD-rate reduction against VTM-IPP in terms of PSNR. Under intra period 96 configuration, DCVC-FM can also achieve on average 25.3% and 38.3% YUV BD-rate reduction, respectively, against VTM-IPP and

[1]https://www.videolan.org/developers/x265.html

DCVC-DC in terms of PSNR.

### *B. Settings of Evaluative Image/Video Compression Methods*

Here we mention the test instructions of each test image/video compression schemes in detail.

- BPG:
  {*input file name*} --q={*qp*} --f={*chroma format*} --o={*bitstream file name*}
- Factorized Model:
  --inputPath={*data path*} --outputPath={*bin path*} --ckptPath={*checkpoint path*} --model_id={*model ID*}
- Hyperprior Model:
  --inputPath={*data path*} --outputPath={*bin path*} --ckptPath={*checkpoint path*} --model_id={*model ID*}
- Autoregressive Model:
  --inputPath={*data path*} --outputPath={*bin path*} --ckptPath={*checkpoint path*} --model_id={*model ID*}
- Cheng2020:
  --inputPath={*data path*} --outputPath={*bin path*} --ckptPath={*checkpoint path*} --model_id={*model ID*}
- ELIC:
  --inputPath={*data path*} --outputPath={*bin path*} --ckptPath={*checkpoint path*} --model_id={*model ID*}
- MLIC++:
  --inputPath={*data path*} --outputPath={*bin path*} --ckptPath={*checkpoint path*} --model_id={*model ID*}
- iWave++ of IEEE 1857.11:
  --inputPath={*data_dir*} --outputPath={*bin_dir*} --ckptdir={*ckpt_dir*} --model_id={*model ID*} --cfg={*encode_iWave_lossy.cfg*} --outlog={*enc_time_log*}
- VVC Test Model (VTM-13.2, LDB):
  -c {*config file name*} --c={*input sequence cfg*} --InputChromaFormat=444 --FramesToBeEncoded={*encode frame number*} --InputBitDepth=10 --OutputBitDepth=10 --OutputBitDepthC=10 --DecodingRefreshType=2 --IntraPeriod=32 --Level=6.2 --SourceWidth={*width*} --SourceHeight={*height*} --QP={*qp*} --BitstreamFile={*bitstream file name*}
- HEVC Test Model (HM-16.7, RA):
  -c {*config file name*} --c={*input sequence cfg*} --InputChromaFormat=444 --f={*encode frame number*} --InputBitDepth=8 --SourceWidth={*width*} --SourceHeight={*height*} --QP={*qp*} --b={*bitstream file name*}
- DVC_Pro:
  --i_frame_model_name=cheng2020-anchor --i_frame_model_path={*model path*} --model_path={*model path*} --test_config={*test config*} --cuda={*GPU number*} --w={*CPU worker number*} --write_stream=0 --output_path={*output path*}
- CANF-VC:
  --Iframe=BPG --MENet=PWC --motion_coder_conf={*conf_dir*} --cond_motion_coder_conf={*conf_dir*} --residual_coder_conf={*conf_dir*} --dataset={*dataset_dir*} --seq={*seq_dir*} --seq_len={*seq_len_num*} --dataset_path={*dataset_path*} --lmda=2048 --model_dir={*model_dir*} --bitstream_dir={*bitstream_dir*} –action=compress –GOP=32
- DCVC:
  --i_frame_model_name=cheng2020-anchor --i_frame_model_path={*model path*} --model_path={*model path*} --test_config={*test config*} --cuda={*GPU number*} --w={*CPU worker number*} --write_stream=0 --output_path={*output path*} --model_name=DMC_conditional_coding
- TCM-VC:
  --i_frame_model_name=IntraNoAR --i_frame_model_path={*model path*} --model_path={*model path*} --test_config={*test config*} --cuda={*GPU number*} --w={*CPU worker number*} --write_stream=0 --stream_path={*stream path*} --output_path={*output path*} --verbose=0
- DCVC-HEM:
  --i_frame_model_name={*model path*} --model_path={*model path*} --rate_num=4 --test_config={*test config*} --cuda={*GPU number*} --w={*CPU worker number*} --write_stream=0 --output_path={*output path*} --force_intra_period=32 --force_frame_num=96
- OOFE:
  --model_path_i={*model path*} --model_path_p={*model path*} --rate_num=4 --test_config={*dataset json*} --cuda={*GPU number*} --worker={*CPU worker number*} --calc_ssim=1 --write_stream=0 --output_path={*output json*}
- SDD:
  --model_path_i={*model path*} --model_path_p={*model path*} --rate_num=4 --test_config={*dataset json*} --cuda={*GPU number*} --worker={*CPU worker number*} --calc_ssim=1 --write_stream=0 --output_path={*output json*}
- VNVC:
  --i_frame_model_name=IntraNoAR --i_frame_model_path={*model path*} --model_path={*model path*} --test_config={*test config*} --cuda={*GPU number*} --w={*CPU worker number*} --write_stream=0 --stream_path={*stream path*} --output_path={*output path*} --verbose=0
- DCVC-DC:
  --model_path_i={*model path*} --model_path_p={*model path*} --rate_num=4 --test_config={*dataset json*} --cuda={*GPU number*} --worker={*CPU worker number*} --calc_ssim=1 --write_stream=0 --output_path={*output json*}
- DCVC-FM:
  --model_path_i={*model path*} --model_path_p={*model path*} --rate_num=4 --test_config={*dataset json*} --cuda={*GPU number*}

--worker={*CPU worker number*} --write_stream=0 --output_path={*output json*} --force_intra_period=32 --force_frame_num=96

## II. Specific Performance of Evaluative Advanced Image/Video Compression Schemes on USTC-TD

Here we supply the specific rate-distortion (RD) curves of advanced image/video compression schemes on USTC-TD dataset under different metrics (Section V.B of main text). The results of USTC-TD image dataset are mentioned in Fig. 1, Fig. 2, and Fig. 3, Fig. 4. The results of USTC-TD video dataset are mentioned in Fig. 5, and Fig. 6.

## References


[1] F. Bellard, "Bpg Image Format," *https://bellard.org/bpg*, 2015.

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

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

[4] "IEEE 1857.11, IEEE C/DC 1857.11 Sub-Working Group," *https://sagroups.ieee.org/fvc/*, 2021.

[5] 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.

[6] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," *arXiv preprint arXiv:1511.06281*, 2015.

[7] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal processing magazine*, vol. 18, no. 5, pp. 36–58, 2001.

[8] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Transactions on circuits and systems for video technology*, vol. 13, no. 7, pp. 560–576, 2003.

[16] 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.

[17] 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.

[18] J. Lin, D. Liu, H. Li, and F. Wu, "M-lvc: Multiple frames prediction for learned video compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3546–3554.

[19] 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.

[20] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with recurrent auto-encoder and recurrent probability model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 388–401, 2020.

[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] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC)," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021.

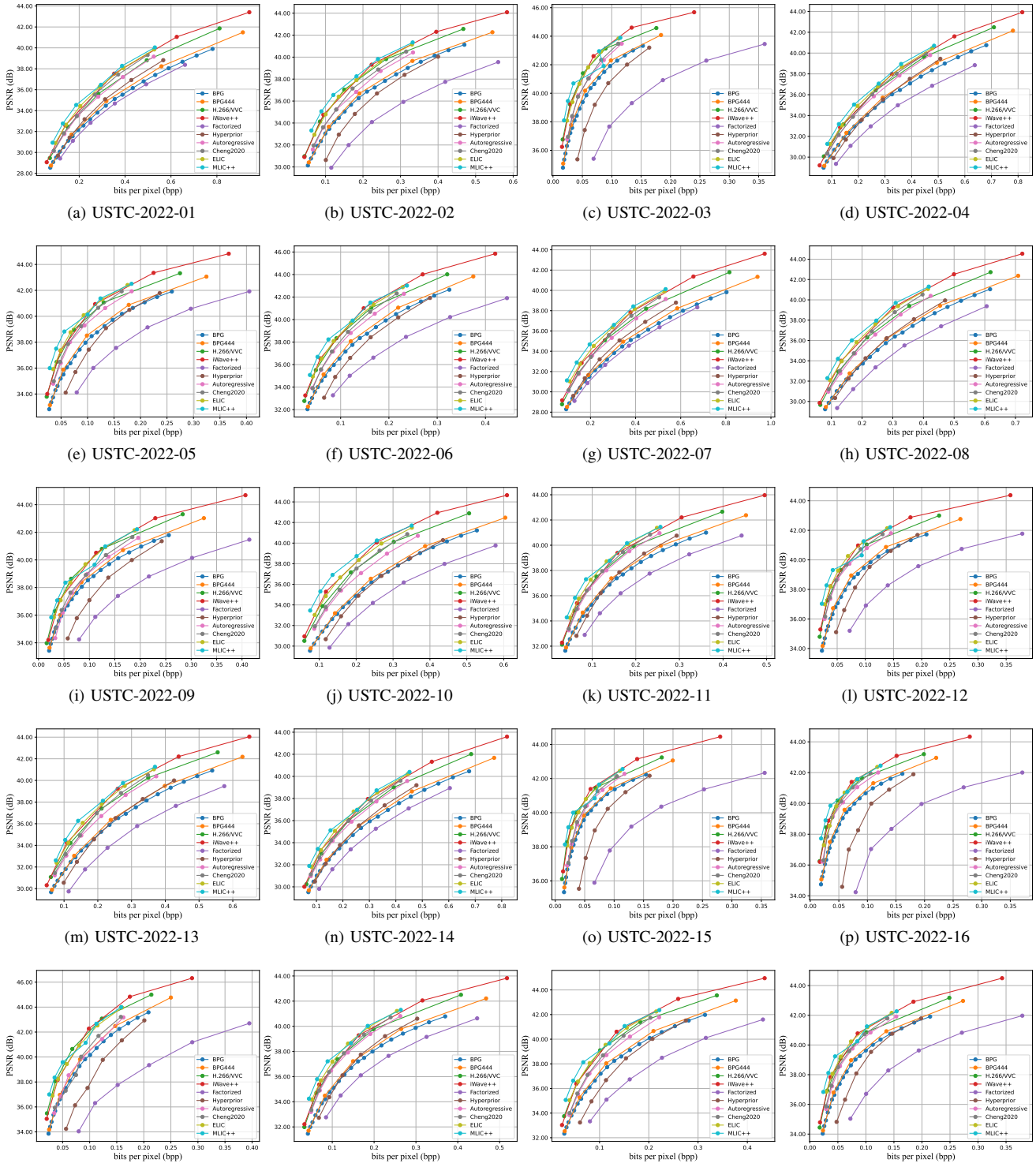
[23] C. Tang, X. Sheng, Z. Li, H. Zhang, L. Li, and D. Liu, "Offline and online optical flow enhancement for deep video compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5118–5126.

[24] 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.

[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] ——, "Spatial decomposition and temporal fusion based inter prediction for learned video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[27] J. Li, B. Li, and Y. Lu, "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.




(a) USTC-2022-01 (b) USTC-2022-02 (c) USTC-2022-03 (d) USTC-2022-04

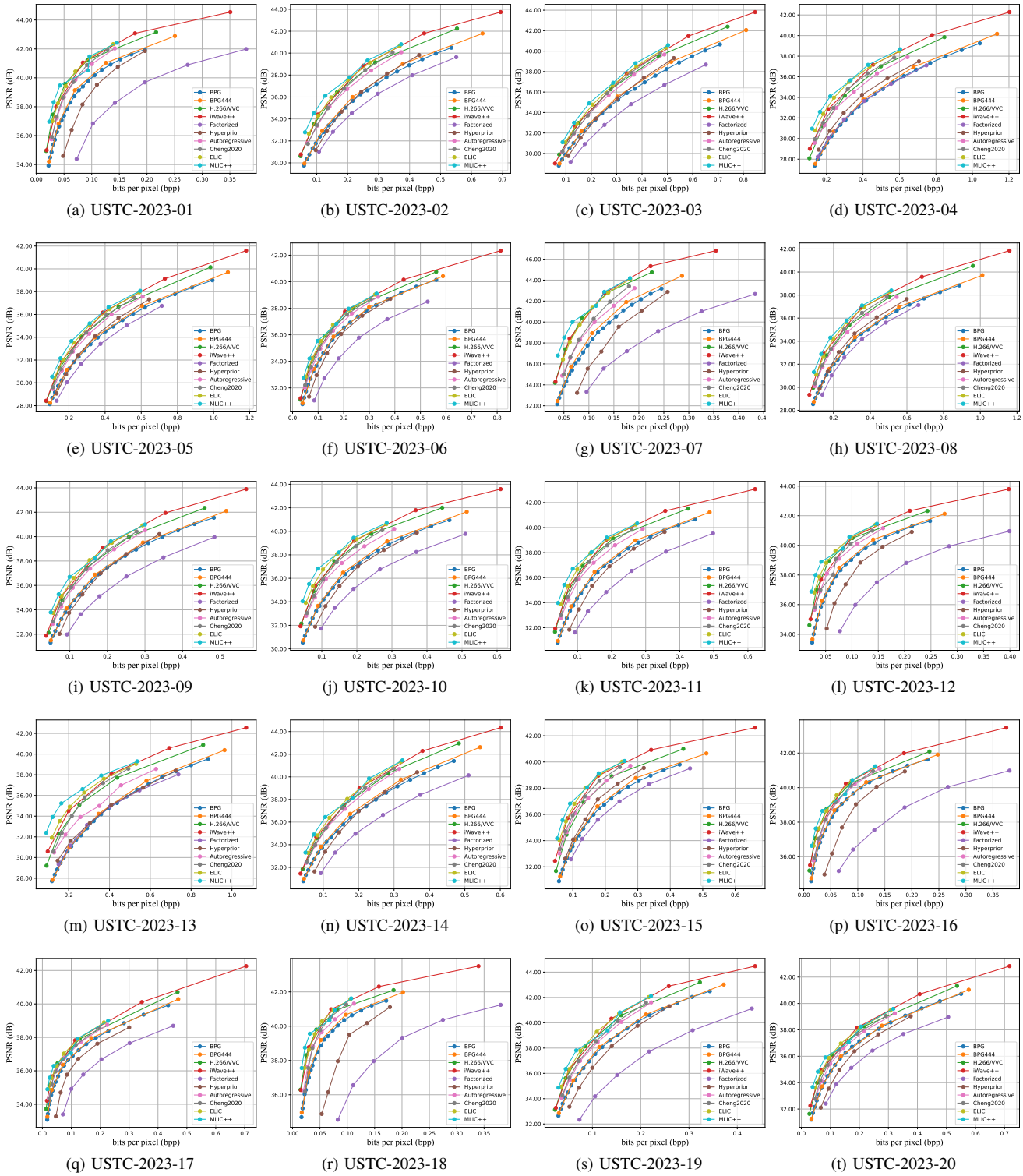
(e) USTC-2022-05 (f) USTC-2022-06 (g) USTC-2022-07 (h) USTC-2022-08

(i) USTC-2022-09 (j) USTC-2022-10 (k) USTC-2022-11 (l) USTC-2022-12

(m) USTC-2022-13 (n) USTC-2022-14 (o) USTC-2022-15 (p) USTC-2022-16

(q) USTC-2022-17 (r) USTC-2022-18 (s) USTC-2022-19 (t) USTC-2022-20

Fig. 1. Specific rate-distortion (RD) curves of advanced image compression schemes on each evaluative image of USTC-TD 2022 under *PSNR* metric.




(a) USTC-2023-01 (b) USTC-2023-02 (c) USTC-2023-03 (d) USTC-2023-04

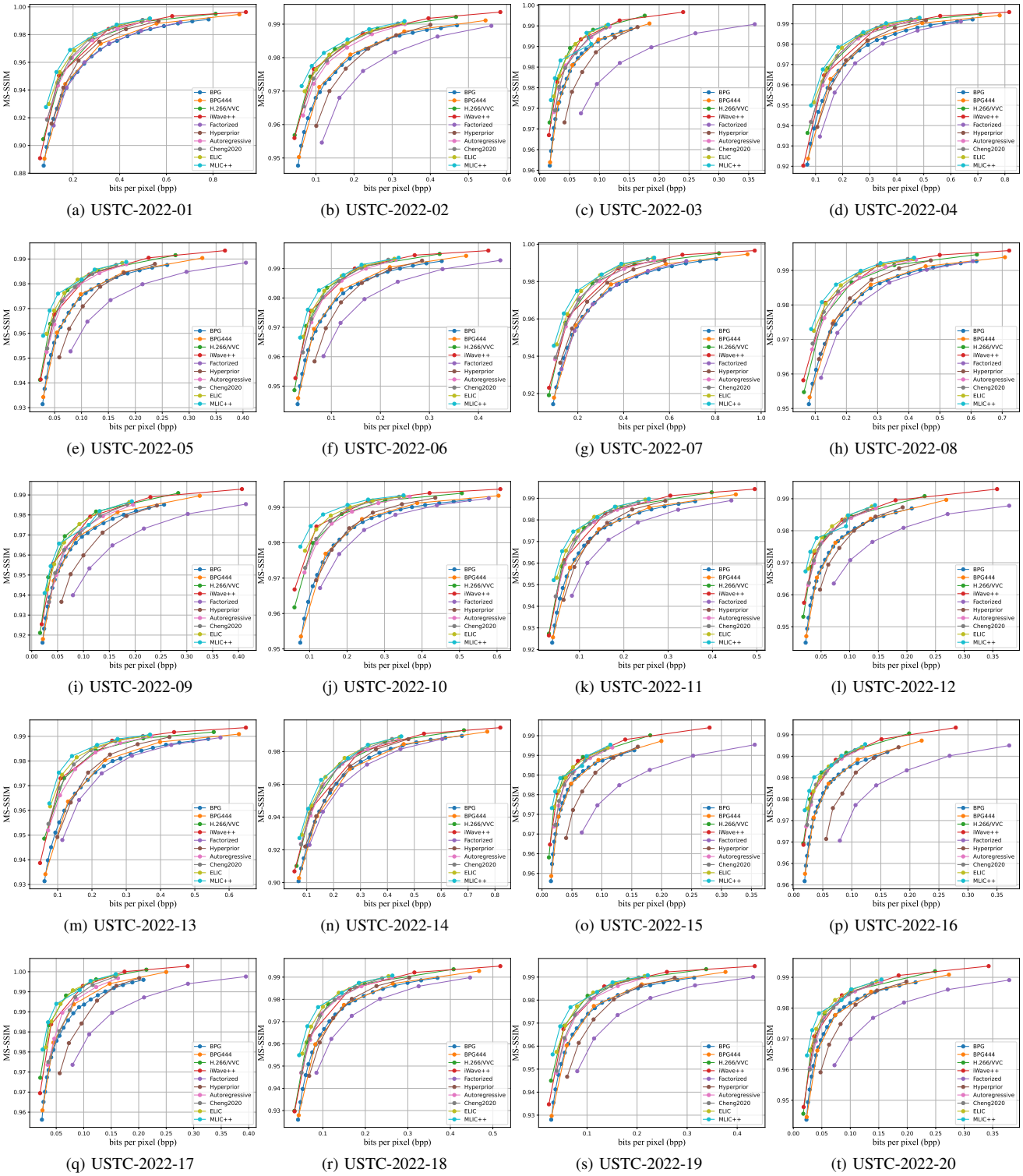
(e) USTC-2023-05 (f) USTC-2023-06 (g) USTC-2023-07 (h) USTC-2023-08

(i) USTC-2023-09 (j) USTC-2023-10 (k) USTC-2023-11 (l) USTC-2023-12

(m) USTC-2023-13 (n) USTC-2023-14 (o) USTC-2023-15 (p) USTC-2023-16

(q) USTC-2023-17 (r) USTC-2023-18 (s) USTC-2023-19 (t) USTC-2023-20

Fig. 2. Specific rate-distortion (RD) curves of advanced image compression schemes on each evaluative image of USTC-TD 2023 under *PSNR* metric.




(a) USTC-2022-01 (b) USTC-2022-02 (c) USTC-2022-03 (d) USTC-2022-04

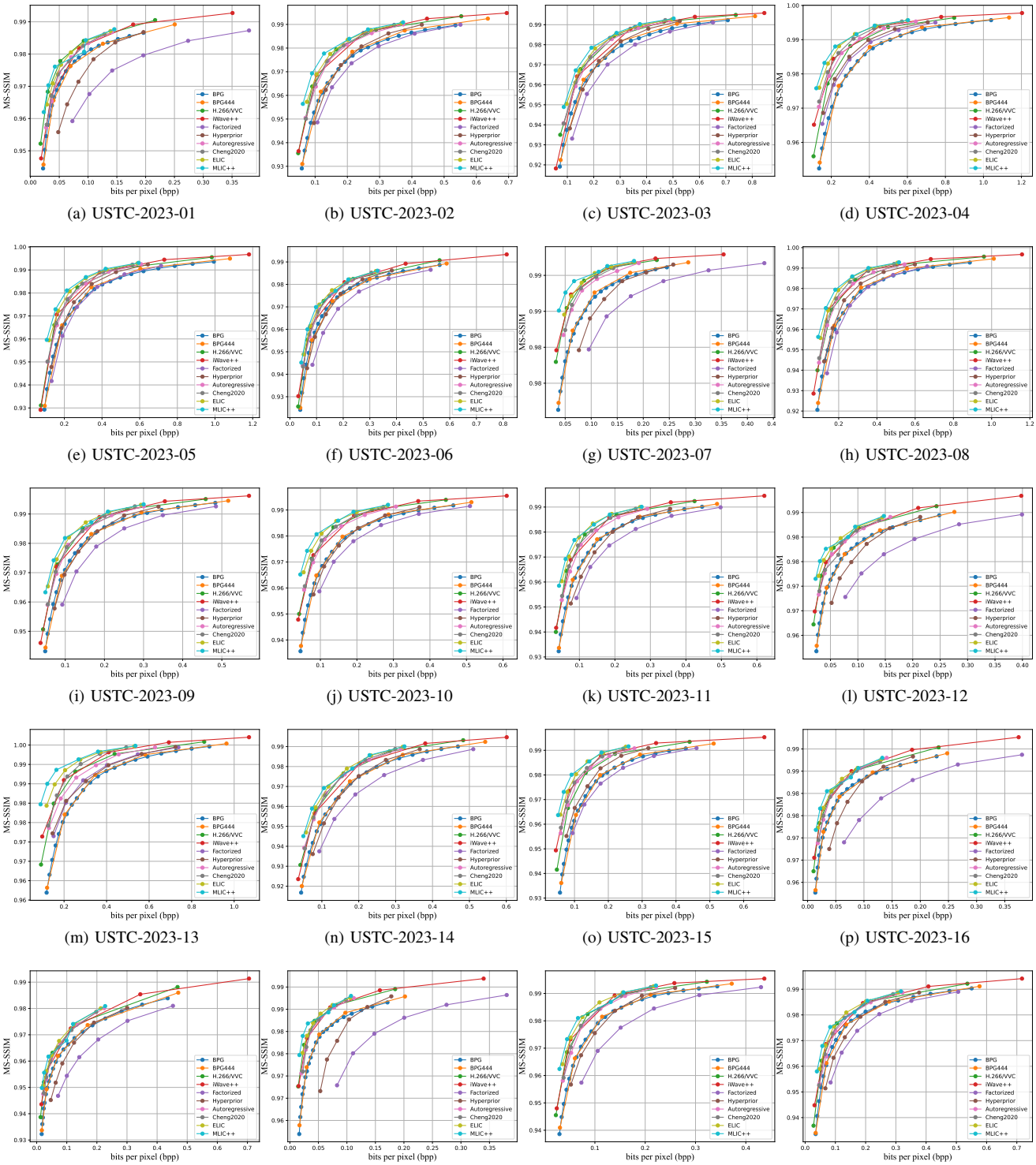
(e) USTC-2022-05 (f) USTC-2022-06 (g) USTC-2022-07 (h) USTC-2022-08

(i) USTC-2022-09 (j) USTC-2022-10 (k) USTC-2022-11 (l) USTC-2022-12

(m) USTC-2022-13 (n) USTC-2022-14 (o) USTC-2022-15 (p) USTC-2022-16

(q) USTC-2022-17 (r) USTC-2022-18 (s) USTC-2022-19 (t) USTC-2022-20

Fig. 3. Specific rate-distortion (RD) curves of advanced image compression schemes on each evaluative image of USTC-TD 2022 under *MS-SSIM* metric.




(a) USTC-2023-01 (b) USTC-2023-02 (c) USTC-2023-03 (d) USTC-2023-04

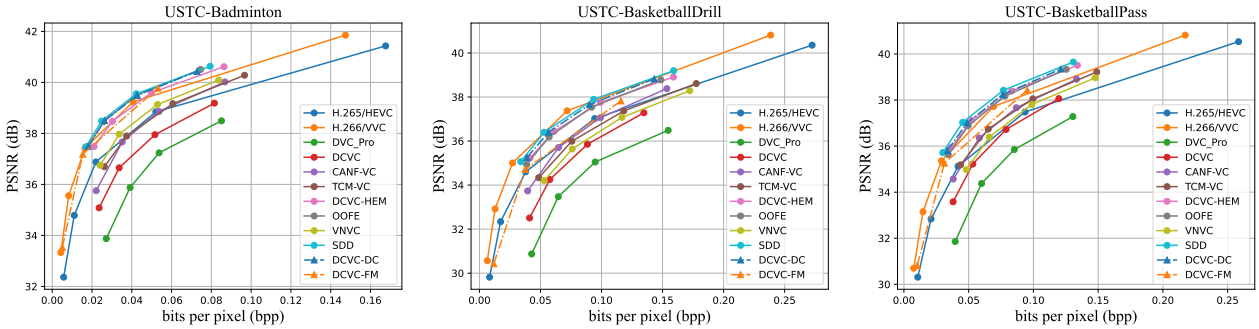
(e) USTC-2023-05 (f) USTC-2023-06 (g) USTC-2023-07 (h) USTC-2023-08

(i) USTC-2023-09 (j) USTC-2023-10 (k) USTC-2023-11 (l) USTC-2023-12

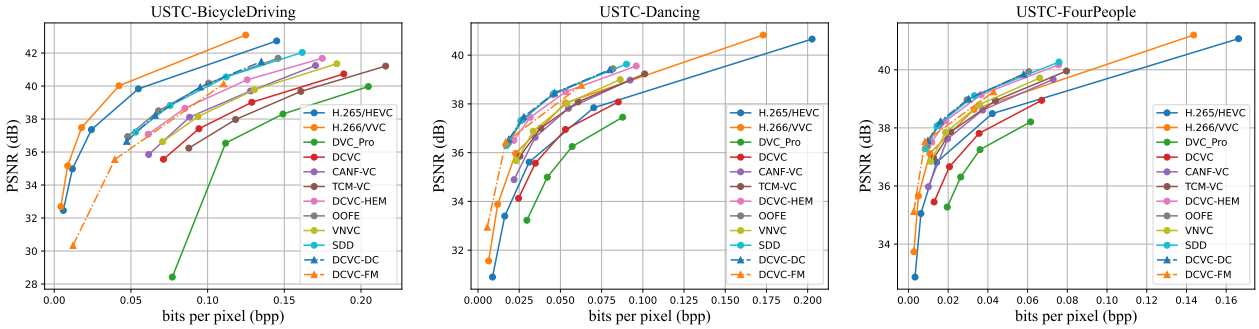
(m) USTC-2023-13 (n) USTC-2023-14 (o) USTC-2023-15 (p) USTC-2023-16

(q) USTC-2023-17 (r) USTC-2023-18 (s) USTC-2023-19 (t) USTC-2023-20

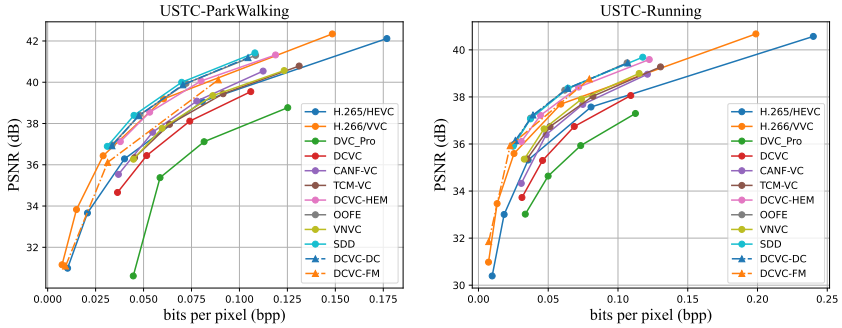
Fig. 4. Specific rate-distortion (RD) curves of advanced image compression schemes on each evaluative image of USTC-TD 2023 under *MS-SSIM* metric.



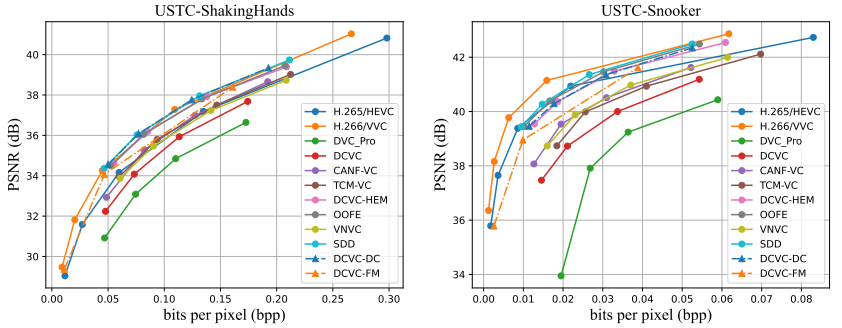

(a) USTC-Badminton (b) USTC-BasketballDrill (c) USTC-BasketballPass




(d) USTC-BicycleDriving (e) USTC-Dancing (f) USTC-FourPeople

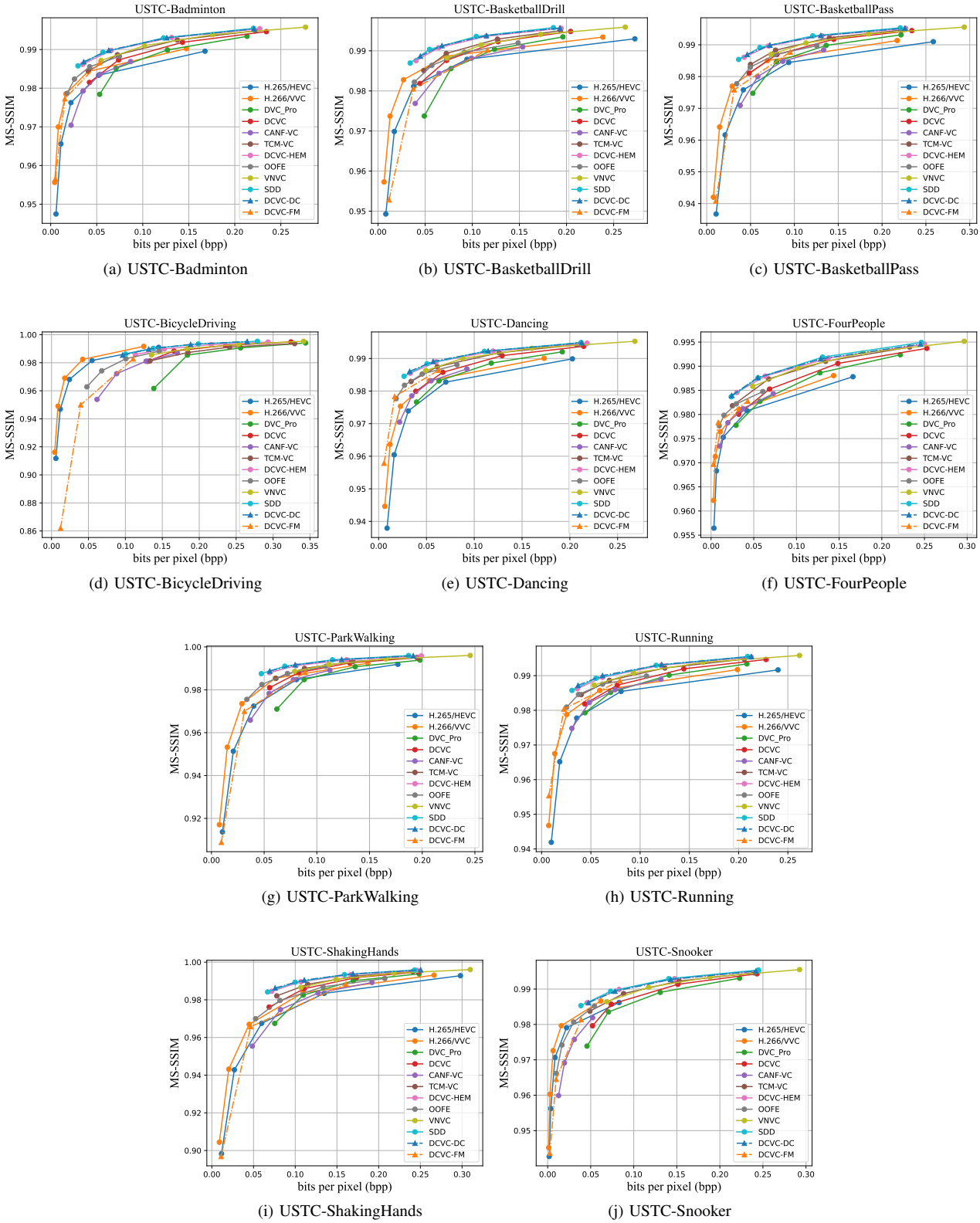



(g) USTC-ParkWalking (h) USTC-Running




(i) USTC-ShakingHands (j) USTC-Snooker

Fig. 5. Specific rate-distortion (RD) curves of advanced video compression schemes on each evaluative video of USTC-TD 2023 under *PSNR* metric.




(a) USTC-Badminton (b) USTC-BasketballDrill (c) USTC-BasketballPass

(d) USTC-BicycleDriving (e) USTC-Dancing (f) USTC-FourPeople

(g) USTC-ParkWalking (h) USTC-Running

(i) USTC-ShakingHands (j) USTC-Snooker

Fig. 6. Specific rate-distortion (RD) curves of advanced video compression schemes on each evaluative video of USTC-TD 2023 under *MS-SSIM* metric.