MISC: Ultra-low Bitrate Image Semantic Compression Driven by Large Multimodal Model

Chunyi Li, Guo Lu, Donghui Feng, Haoning Wu, Zicheng Zhang, Xiaohong Liu, Guangtao Zhai, *Senior Member, IEEE*, Weisi Lin, *Fellow, IEEE*, Wenjun Zhang, *Fellow, IEEE*

(a) Ground Truth. (24 bpp) (b) CHENG-2020. (0.057 bpp) (c) PICS-2023. (0.041 bpp) (d) Proposed Method. (0.036 bpp)

Fig. 1: An image compressed by existing algorithm and MISC. In ultra-low bitrate, (b) has no detail which leads to a poor perceptual quality; (c) generates details but inconsistent with ground truth; (d) achieves consistency and perception altogether.

***Abstract*—With the evolution of storage and communication protocols, ultra-low bitrate image compression has become a highly demanding topic. However, all existing compression algorithms must sacrifice either consistency with the ground truth or perceptual quality at ultra-low bitrate. During recent years, the rapid development of the Large Multimodal Model (LMM) has made it possible to balance these two goals. To solve this problem, this paper proposes a method called Multimodal Image Semantic Compression (MISC), which consists of an LMM encoder for extracting the semantic information of the image, a map encoder to locate the region corresponding to the semantic, an image encoder generates an extremely compressed bitstream, and a decoder reconstructs the image based on the above information. Experimental results show that our proposed MISC is suitable for compressing both traditional Natural Sense Images (NSIs) and emerging AI-Generated Images (AIGIs) content. It can achieve optimal consistency and perception results while saving 50% bitrate, which has strong potential applications in the next generation of storage and communication. The code will be released on https://github.com/lcysyzxdxc/MISC.**

The work was supported by the National Natural Science Foundation of China under Grant 62301310, and by the Shanghai Pujiang Program under Grant 22PJ1406800. Corresponding author: Xiaohong Liu, Guangtao Zhai.

Chunyi Li, Guo Lu, Donghui Feng, Zicheng Zhang, Guangtao Zhai, and Wenjun Zhang are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (email: lcysyzxdxc,luguo2014,faymek,zzc1998,zhaiguangtao,zhangwenjun@sjtu.edu.cn)

Xiaohong Liu is with the John Hopcroft Center, Shanghai Jiao Tong University, Shanghai 200240, China (email: xiaohongliu@sjtu.edu.cn)

Haoning Wu and Weisi Lin are with the S-Lab, Nanyang Technological University, Singapore 639798, Singapore (email: haoning001@e.ntu.edu.sg, wslin@ntu.edu.sg)

Manuscript received Mar 11, 2024.

***Index Terms*—AI-Generated Content, Image Compression, Perceptual Quality, Ultra-Low Bitrate.**

I. Introduction

Image compression serves as the foundation for visualizing volumetric signals [1]–[3]. This technology effectively reduces the storage space and transmission bandwidth required for visual signals without significantly compromising their quality. With the recent advancements in 5G [4] and 6G [5], the integration of numerous embedded devices and Internet-of-Things (IoT) devices into communication protocols has posed challenges due to their limited storage resources and extreme channel conditions. This scenario has made ultra-low bitrate image compression a challenging and demanding research topic, which compresses images to one-thousandth of their original size or even more. To achieve such extreme compression ratios, the focus shifts from low-level fidelity to semantic consistency with the reference image.

However, a trade-off exists between perception and consistency in image compression [6]. For low bitrate image compression (< 0.1 bpp), the compression algorithm provides a rough encoding of the original image, necessitating the decoder to add details. Inadequate detail leads to poor perceptual quality, while excessive detail results in inconsistency with the original image, as illustrated in Fig. 1. As bitrates decrease further to ultra-low levels (< 0.024 bpp, one-thousandth of the original), the conflict between these two objectives becomes even more intensified [7].

TABLE I: The Structure of existing image compression metric, including en/decoder, format of compressed content, minimum bitrate (Normal/Low/Ultra-low: $> 0.1/0.024 \sim 0.1/ < 0.024$ bpp), and optimization goal.

Type	Mertic	Encoder	Content	Decoder	Min Bitrate	Goal
Traditional	Jpeg [17], Webp [18], etc.	Low-level Processing	Bitstream	Low-level Processing	Normal	Consistency
NIC	Mbt2018 [19], RDO-PTQ [20]	CNN	Feature	CNN	Normal	Consistency
	SReC [21], Gao2020 [22], etc.	CNN	Compressed Image	CNN+SR	Low	Consistency
GIC (GAN)	Generative-comp [23]	CNN	Feature	Conditional GAN	Normal	Consistency
	ULCompress [24]	DWT+GAN	Compressed Image	IDWT+GAN	Ultra-low	Consistency
	HiFiC [25]	CNN	Feature	Conditional GAN	Normal	Perception
	Multi [26]/Vari-realism [27]	HRRGAN	Feature	HRRGAN	Low	Either
GIC (Diffusion)	CDC [28]	Re-DDIM	Feature	DDIM	Normal	Consistency
	Pan2022 [29]	Re-SD1.4	Feature	SD1.4	Ultra-low	Consistency
	CMC [30]/M-CMC [31]	CNN	Canny Edge, Text	DDIM+Controlnet	Ultra-low	Consistency
	SGC [32]	DeepLabV3	Semantic Map	LD	Low	Perception
	HFD [33]	Mean Square Error	Compressed Image	SD1.5+Upscaler	Low	Perception
	Text-Sketch [34]	CLIP	Canny Edge, Text	SD2.1+Controlnet	Ultra-low	Perception
GIC (LMM)	MISC (**proposed**)	GPT-4 Vision	Compressed Image, Text, Semantic Map	SD2.1+Controlnet+Mask	Ultra-low	Both

In recent years, the emergence of AI-Generated Content (AIGC) has revolutionized the field of image compression [8]–[10]. This paradigm shift enables achieving both perception and consistency at ultra-low bitrates. Leveraging the image understanding capabilities of GPT4-Vision [11], Llama [12], and the image generation prowess of Stable Diffusion [13], [14] and DALLE [15] series models, images can be compressed into semantic information, facilitating high-quality reconstruction. Additionally, Large Multi-modal Models (LMMs) have altered the content of images to be compressed. Beyond Natural Sense Images (NSIs), AI-Generated Images (AIGIs) have demonstrated significant commercial value by reshaping the creation and marketing of visual content [16]. Given the low-level differences between AIGIs and NSIs (e.g., AI artifacts, texture distribution), the existing compression methods for NSIs may not be suitable for AIGIs, posing an open question of how to compress this distinct image form. To expand the application of image compression in the AIGC era, we propose Multimodal Image Semantic Compression (MISC) for ultra-low bitrate compression, making the following contributions:

- A new paradigm of image compression driven by LMMs. MISC is the pioneering image compression model that integrates LMMs in both the encoder and decoder. This holistic approach will facilitate the extensive utilization of LMMs in image compression applications.
- A high-quality AIGI database. We collected 500 high-quality AIGIs generated by today's mainstream Text-to-Image models for future evaluation of the performance of AIGI compression algorithms.
- A good balance between consistency and perception. We extensively compared today's mainstream image compression algorithm. Experimental result shows that MISC achieves both satisfactory consistency and perceptual quality for the first time at ultra-low bitrates.

II. Related Works

A. Image Compression Metric

Over the past few decades, the evolution of image compression has progressed through four distinct stages as outlined in Table I: **(i)** The initial stage involved traditional image compression algorithms that utilized low-level visual processing techniques to encode redundancy, such as Macro Blocks (MBs) from H.264 to H.266 [35]–[37]. However, these methods primarily focused on pixel-level information, necessitating relatively high bitrates. **(ii)** Subsequently, with the advancement of deep learning, Neural Image Compression (NIC) emerged as a prominent algorithm. NIC employs end-to-end convolutional neural networks (CNN) in both the encoder and decoder to map the original image to a latent space and restore it. By incorporating Super Resolution (SR) in the decoding process to enhance image details, NIC achieves compression to lower bitrates, but there remains potential for further bitrate reduction. **(iii)** The introduction of Generative Image Compression (GIC) in 2019 marked a significant development. In contrast to NIC, GIC encodes images within specific constraints to guide the decoder in generating images consistent with the ground truth. Early GIC implementations utilized Generative Adversarial Networks (GANs) as decoders, offering the potential for ultra-low bitrate compression. **(iv)** After 2022, GAN is gradually replaced by Diffusion, which can be constrained with multimodal information (such as text, edge), encoded by Contrastive Language-Image Pre-Training (CLIP) [38], and reconstruction models like Denoising Diffusion Implicit Model (DDIM), Latent Diffusion (LD), and Stable Diffusion (SD) have facilitated ultra-low bitrate compression. However, despite the advantages, achieving both consistency and perceptual quality at such low bitrates remains a challenge, necessitating the development of comprehensive metrics that cater to both objectives.

B. Image Compression Databases

Image compression databases play a pivotal role in validating image compression algorithms. For example, Kodak-24 [39] offers a realistic view of compression effects on authentic images, DIV2K [40] focuses on high-resolution scenarios, and CLIC-2020 [41] consists of high-quality images with diverse content. However, all the databases above are NSIs, whose characteristics are significantly different from AIGIs. Considering the quality of the existing AIGI database [42], [43] is already low, no matter what compression algorithm is

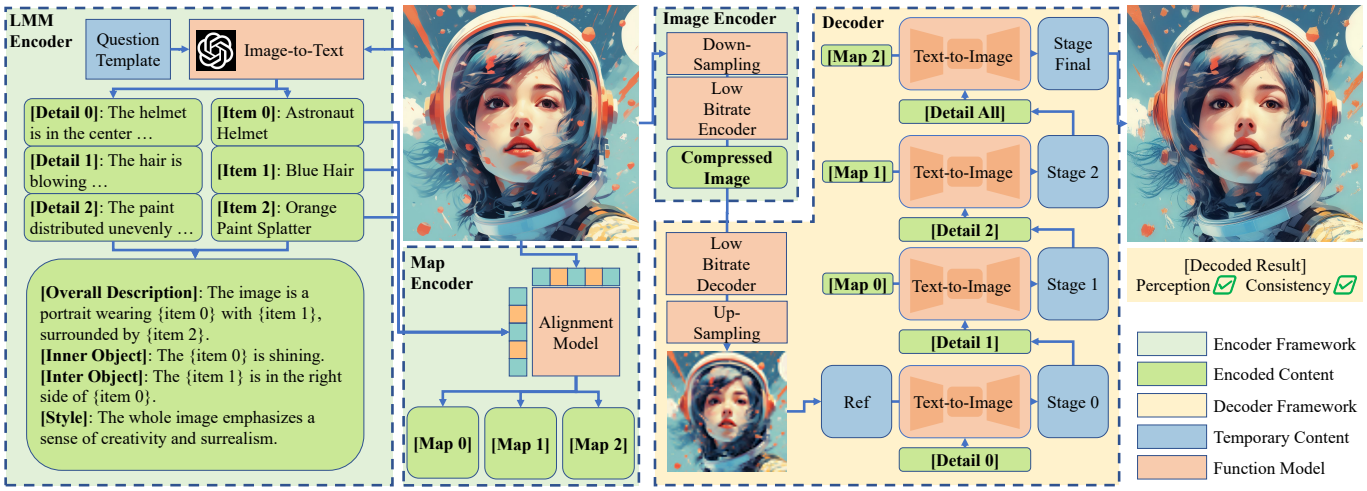


Fig. 2: The framework of the MISC model including LMM/map/image encoders and an LMM decoder. The compressed content (noted in **green**) has an extremely compressed image bitstream lower than 0.024 bpp, a detailed description of a whole image, and items' names, details, and supposed position maps. The decoder controls the diffusion process according to the above content to generate images that simultaneously satisfy high consistency and perceptual quality.

used, the resulting image quality is still low. Therefore, a high-quality AIGI database is needed to measure the performance of image compression algorithms.

C. Evaluation Criteria

Traditionally, the performance of image compression metrics is judged by pixel-level distortion, such as Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity (SSIM) [44]. That's because distortion is relatively small at normal bitrate, and the more consistent the compressed image is with the reference image, the higher the perceptual quality it has. However, as the bitrate decreases, the consistency perception becomes a dilemma [34], and high-quality images often contain details different from the reference. Therefore, to objectively evaluate the performance of a compression algorithm, both consistency and perception index need to be considered separately.

For consistency, at ultra-low bitrate, pixel-level fidelity metrics are poorly correlated with human subjective perception because they mainly focus on low-level details rather than high-level structures [34]; for perception, traditional LMM image generation works use Fréchet Inception Distance (FID) [45] to characterize human preference. However, research shows that subjective human perception does not depend on it, but on signal fidelity and aesthetics [46]–[50]. Therefore, the semantic similarity between the compressed image and the reference image, and the Image Quality / Aesethic Assessment (IQA/IAA) index [51]–[55] are more suitable as evaluation criteria of consistency and perception.

III. Proposed Compression Algorithm

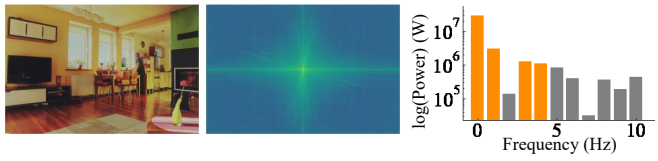
A. Framework

In this section, we propose the MISC framework for ultra-low bitrate image compression, as shown in Fig. 2. Specifically, the framework contains three encoding modules, including an LMM encoder for extracting semantic information of images, a map encoder for annotating regions corresponding to semantic information above, and an image encoder for extreme pixel-level compression, and a decoding module uses the text, map, and image obtained through the above process as constraints to reconstruct the image. Next, we will introduce the three encoders and one decoder module in detail.

B. LMM Encoder

The LMM encoder transforms images into multiple uncorrelated, sparse, weakly dependent semantic variables. It keeps the main variables and discards other variables. Then the decoder can perform an inverse transformation to obtain an image consistent with the original image and save a lot of data space. In recent years, the ability of LMMs to understand and generate images has contributed to more efficient semantic translation. By using Image-to-Text (I2T) and Text-to-Image (T2I) models as encoders and decoders, images can be compressed into more compact semantic information. Unlike Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT), semantically distinct items have lower correlations compared to the frequency domain. Under the same compression performance, the semantic domain can discard more information. Referring to the spatial-frequency domain, MISC designed the spatial-semantic domain mapping, as shown in Fig. 3. In traditional image compression, the base frequency is usually retained as the overall hue of the image, as well as the amplitude/phase of lower frequencies to represent important details, while other details at higher frequencies are discarded. Similarly, MISC will generate a long description of the image as a whole, as well as a description of some important items in the image, discarding other items. To accurately extract semantic information, MISC asked questions to the most advanced GPT-4 Vision [11] and obtained the following natural language feedback:

- $T_n[j]$: Item name ($\leq$ 3 words, $j \in \{0, 1, \cdots J-1\}$)
- $T_d[j]$: Item detail ($\leq$ 10 words, $j \in \{0, 1, \cdots J-1\}$)



Please list three most dominant frequency:
1Hz, 3Hz, 4Hz
Please give their magnitude and phase:
[**1Hz**]: 3.04 MW, 159.36°
[**3Hz**]: 1.29 MW, 92.29°
[**4Hz**]: 1.08 MW, 115.22°
Please give the magnitude of base frequency:
[**0Hz**]: 29.5 MW
Base and three other dominant frequency reserved for reconstructio n. Other abandoned.

(a) Frequency Domain Information (traditional compression)

LMM Question Template:
A. Image Resolution
B. Image Caption
C. Region Semantic
D. Dense Caption
…

Please list three most important items:
Television, Dining Table, Vase
Please give their detailed description:
[**Television**]: A large, flat-screen TV placed on a low stand…
[**Dining Table**]: A wooden dining table with chairs…
[**Vase**]: A tall, slim yellow vase sits on the floor…
Please generate an informative and nature paragraph from (A,B,C,D):
[**Overall**]: In a warmly lit room, a flat-screen television rests against the left wall on a dark stand … a lively and eclectic aesthetic.
Overall information and three other important items reserved for reconstructio n. Other abandoned.

(b) Semantic Domain Information (proposed MISC)

Fig. 3: Comparison of mapping spatial domain into frequency or semantic domain. Both methods compress images by retaining important information and discarding other information.

- T_{all}: Detail all (≈ 50 words)

where J stands for the number of items. **[Item name]** is the index of the item and is not directly used for image reconstruction, so it is only represented by no more than three words. **[Item detail]** includes the shape, color, status, or other attributes. Considering that the description of an object in existing visual question answering [46] tasks usually does not exceed 10 words, we set it as an upper limit. **[Detail all]** is a description of the image as a whole. Past GIC [31] shows that as the number of description words increases, the compression performance gradually increases and reaches a maximum of about 50 words. Considering the data scale of text format, for a 512×512 image, a word usually occupies $1 \sim 2 \times 10^{-4}$ bpp. To avoid unnecessary overhead, we set 50 as the benchmark. Assuming in the semantic domain, a few items take up most of the information in the image (just as the low frequencies take up most of the energy), MISC can set the threshold of the items S_{th} referring to the frequency threshold f_{th} of the Macro Block (MB) in H.264 [35] as:

$$S_{th} = \frac{f_{th}}{N_{pix}} \mathrm{E}(N_{item}), \tag{1}$$

where (N_{pix}, N_{item}) refer to the number of pixels in an MB, and the number of items in an image with Expectation $\mathrm{E}(\cdot)$.

(a) Item 0: green bike (b) Item 1: wooden door

(c) Item 2: grass (d) Compressed without map (b)

(e) Compressed without map (c) (f) Compressed with all maps

Fig. 4: The positional map of three items, and the decompressed image with/without maps. When (b) and (c) are not used as constraints, 'wooden' and 'grass' will affect the 'bike' region respectively.

According to answers of GPT-4 Vision on Kodak24 [39] and CLIC2020 [41], a picture usually contains $10 \sim 15$ items. From (1), the optimal threshold $S_{th} = 2.81$, so MISC takes $J \leq 3$.

C. Map Encoder

The map encoder acts as an additional module for LMM encoder, by characterizing the spatial relationship between multiple items. This encoder can support a dynamic number of items to balance performance against bitrate. A three-item situation is shown in Fig. 4 as an example, each map including the following two aspects:

- The position of the object itself: For example, [Item 0] is located in the 'lower-right' of the entire picture. This word occupies 88 bits, but it is only a rough range; in contrast, an 8×8 map only occupies 64 bits and can more accurately specify its location.
- The relationship between objects: For example, [Item 2] is to the bottom of [Item 1], and it is difficult to describe the distance between them; in this case, only spatial information can complete this task.

Spatial information includes edge and map, of which edge occupies a larger space; considering our ultra-low bitrate requirements, we use maps to mark the corresponding position of each item. Using the text-image alignment capability of CLIP, we can map the name of the item to the corresponding

area of the image. First, multiply the image and text features to obtain a feature matrix F_T:

$$F_T[j] = \mathcal{C}_I(I) \odot \mathcal{C}_T(T_n[j]), \tag{2}$$

where I denotes the ground truth image and T_n means the name of item, $\mathcal{C}_{(I,T)}(\cdot)$ is the CLIP image/text encoder with the length (L_I, L_T). However, on the target detection task, CLIP has a large number of unactivated redundant features, which occupy a large part of the feature space F_T. Therefore, we adopt CLIP Surgery's [56] redundancy elimination method. Input an empty image to get the weight bias ω of the text encoder, and adjust the L_T dimension on F_T to get the redundant feature F_R:

$$\begin{cases} \omega[j] = \sum_{O} (\mathcal{C}_I(\text{null}) \odot \mathcal{C}_T(T_n[j])) \\ F_R[j] = F_T[j] \odot \omega[j], \end{cases} \tag{3}$$

where O is the number of $\mathcal{C}(\cdot)$ output channel. Subtracting them, we have the semantic map $M[i]$ for item $t_n[i]$:

$$\begin{cases} M[j] = \sum_{O} (F_T[j] - F_R[j]) \\ M[j] = \text{Bi}(\text{Pool}(M[j])), \end{cases} \tag{4}$$

where $\text{Bi}(\cdot)$ and $\text{Pool}(\cdot)$ represent binarization and pooling functions, which convert the original map into an 8×8 ~ 16×16 binary matrix to reduce its data size. The data size of maps is extremely small (< 10^{-3} bpp), but it greatly makes up for the shortcomings of the LMM encoder. As shown in Fig. 4, if the map of [Item 1] or [Item 2] is missing, [Item 0] 'bike' will become 'wooden' or grow 'grass' during the decoding process. Only by using map constraints for each item can the desired content be generated at the exact location.

D. Image Encoder

The image encoder provides a reference for the decoder through an extremely compressed image. By combining it with the two encoders above, the reconstruction result from the decoder can have a satisfying perceptual quality while avoiding major defects in consistency with the original image. This image only includes the rough outline of the original image, which can greatly improve the consistency score. At the same time, although the perceptual quality of this image is poor, the decoder reconstruction process will add certain details, and the perception score of the final decoded image is still acceptable. To achieve a balance between consistency and perception, we followed this idea under ultra-low bitrate and compressed the image into bitstream B:

$$B = \begin{cases} \text{En}(S_x^-(I)) & x \le \text{size}(I)/\text{En}_{th} \\ \text{Qu}(S_x^-(I)) & x > \text{size}(I)/\text{En}_{th}, \end{cases} \tag{5}$$

where $\text{En}(\cdot)$ and $\text{Qu}(\cdot)$ are the CNN encoder and the quantization function. $S_x^{(+,-)}$ stands for up/down-sampling. When the down-sampling rate x is low, we can directly use the existing low bitrate NIC/GIC encoder in TABLE I to output the compressed bit stream; when x becomes high, the image size is lower than the minimum input threshold En_{th} of the encoder. Thus, the value of each pixel is directly quantized, and the image is treated as a matrix and output as bitstream. Therefore, by inputting the bitstream together with maps and texts into the decoder, a high-consistency and high-perceptual quality reconstruction process can be achieved.

E. Decoder

The decoding process is based on the following information provided by the encoders, including a bitstream (representing an extremely compressed image), a detailed description of the overall image, and several name-detail-map (NDM) groups characterizing items in the image. Based on the semantic domain analysis in Sec. III-B, we take the number of group index j as (i) $J = 3$ to help image reconstruction, as reference information is limited at ultra-low bitrate; (ii) $J = 0$ to save bitrate, as relatively more information is contained at higher bitrate. In the first step, the decoder will perform the opposite operation of (5) to obtain the reference image I_{ref}:

$$I_{ref} = \begin{cases} S_x^+(\text{De}(B)) & x \le \text{size}(I)/\text{En}_{th} \\ S_x^+(\text{BIC}(B)) & x > \text{size}(I)/\text{En}_{th}, \end{cases} \tag{6}$$

where $\text{De}(\cdot)$ and $\text{BIC}(\cdot)$ are the CNN decoder and the bicubic interpolation. Then we extract the probability density z from text $T_d[j]$, and conduct diffusion progress referring to z with output $S[j]$:

$$\begin{cases} z = \text{QKV}(T_d[j]) \\ S[j] = \mathcal{D}_n^z(\mathcal{D}_{n-1}^z \cdots \mathcal{D}_1^z(S[j-1])) \\ S[j] = S[j]M[j] + S[j-1](1 - M[j]), \end{cases} \tag{7}$$

Where $\text{QKV}(\cdot)$ represents the multi-head attention and $\mathcal{D}_n^z$ denotes the diffusion operation at the n-th iteration. To reconstruct an image that aligns with both the reference image I_{ref} and the NDM groups, we utilize the weight of DiffBIR [57] in the diffusion model $\mathcal{D}$ to integrate the image reference and text description comprehensively. For the diffusion on [Item j], we only update the target region in $S[j]$ based on $M[j]$, while keeping other regions unchanged as the previous state $S[j-1]$. The original state is set as $S[-1] = I_{ref}$. After iterating this process for J times, referring to the descriptive text T_{all} of the entire image, we can enhance details in $S[J-1]$ and obtain the final decoding result $S[final]$:

$$\begin{cases} z = \text{QKV}(T_{all} + T_{aes}) \\ S[final] = \mathcal{D}_{\hat{n}}^z(\mathcal{D}_{\hat{n}-1}^z \cdots \mathcal{D}_1^z(S[2])), \end{cases} \tag{8}$$

where $\hat{n}$ is $4 \sim 8$ times the value of n and T_{aes} represent prompts that guide the high-quality generation (e.g., 'hyper detail', 'masterpiece', '4K'). This step plays a crucial role in the decoding process as it is responsible for ensuring consistency and enhancing perception scores. To prioritize this step, more iterations are allocated to it, and T_{aes} is fed into the QKV mechanism along with T_{all}. This allows the decoder to effectively balance these two conflicting objectives, particularly at ultra-low bitrates.

IV. Proposed Database

A. Data Collection

Beyond NSIs, to compare the performance of image compression algorithms on AIGIs, we construct an AIGI Semantic Compression Database (AIGI-SCD). For a fair comparison, its

TABLE II: Existing Text-to-Image generative model. The database named AIGI-SCD is constructed by models with the best popularity (download times) and quality (overall IQA score, normalized average from ClipIQA↑ [58], DBCNN↑ [59], LIQE↑ [60], NIQE↓ [61]). The top three popularity/quality models are marked in **red**.

Metric	Popularity	Quality	
	Download	Score(↑, ↑, ↑, ↓)	Average
SDXL [14]	**4,400K**	(0.63, 0.64, 3.20, 4.27)	0.69
SD1.5 [13]	**3,780K**	(0.66, 0.55, 3.21, 4.56)	-0.06
SD1.4 [13]	**2,640K**	(0.66, 0.57, 3.24, 4.37)	0.83
SDXL-Turbo [62]	799K	(0.59, 0.64, 3.97, 4.41)	1.28
Midjourney [63]	248K	(0.71, 0.62, 4.09, 5.05)	1.29
DALLE-2 [15]	240K	(0.69, 0.48, 2.54, 5.10)	-2.55
SSD [64]	236K	(0.66, 0.66, 3.79, 5.02)	0.62
Playground [65]	132K	(0.70, 0.68, 3.66, 4.92)	**1.57**
Deramlike [66]	105K	(0.68, 0.61, 3.88, 4.62)	**1.68**
Pixart [67]	53K	(0.72, 0.62, 3.79, 4.47)	**2.75**
IF [68]	33K	(0.68, 0.54, 2.85, 5.00)	-1.31

data needs to represent most of the existing AIGIs, and should not have distortion (otherwise it will overlap with compression distortion, affecting the evaluation of the compression algorithm). Therefore, we conducted a detailed survey of today's mainstream T2I models and selected the most popular and highest-quality models for image generation. According to the huggingface website[1], we listed 11 common AIGI models, used their download times as popularity indicators, and generated 500 images for each; then, we used four IQA indicators to evaluate the images generated by each model. Quality scores S were collected four times, and each column in TABLE II is normalized to $-1 \sim 1$ as:

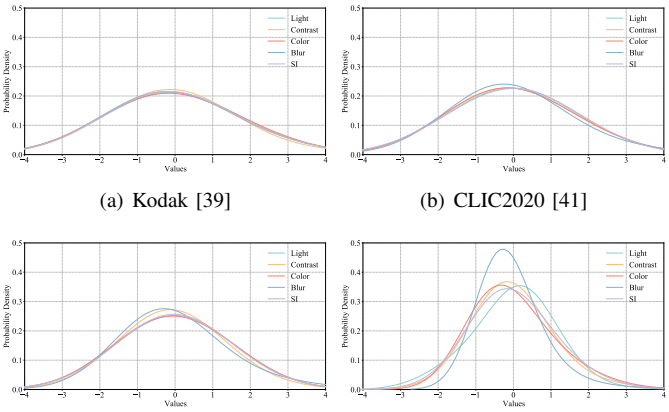
$$\overline{S_{p,q}} = 2 \cdot \frac{S_{p,q} - \min(S_{\cdot,q})}{\max(S_{\cdot,q}) - \min(S_{\cdot,q})} - 1, \tag{9}$$

where p represents the metric column in TABLE II while indicators $q \in$ (ClipIQA, DBCNN, LIQE, NIQE) [58]–[61] in the quality column. Then the average score A for each row's metric can be formulated below:

$$A_p = \sum_q (\pm 1) \cdot \overline{S_{p,q}}, \tag{10}$$

where we use $+1$ for upper-better and -1 for lower-better metric. According to Table II, the most widely used SD series models [13], [14] are representatives of AIGI and need to be included in this database. Meanwhile, the latest Pixart [67], Playground [65], and Dreamlike [66] models generate the highest quality results, thus suitable for compression tasks as the raw images. Considering the high similarity between SD1.5 and 1.4, we select SDXL and SD1.5 to generate mainstream AIGIs and use the above three high-quality generation models to characterize emerging AIGIs. Referring to the data scale and division of CLIC2020 [41], the AIGI-SCD contains 100 images from each of the five generative models, 450 are for training and 50 for testing.

[1]Data collected in January 2024. Considering that DALLE 2 and Midjourney are closed source models, their download times are replaced by OpenDALLE and Openjourney.



(a) Kodak [39] (b) CLIC2020 [41]

(c) Tecnick [69] (d) AIGI-SCD

Fig. 5: The normalized probability distributions of the low-level attributes. The distributions include NSIs in the Kodak24 [39], CLIC2020 [41], Tecnick [69], and the proposed AIGI-SCD database. The AIGIs have a sharper distribution and more common blur.

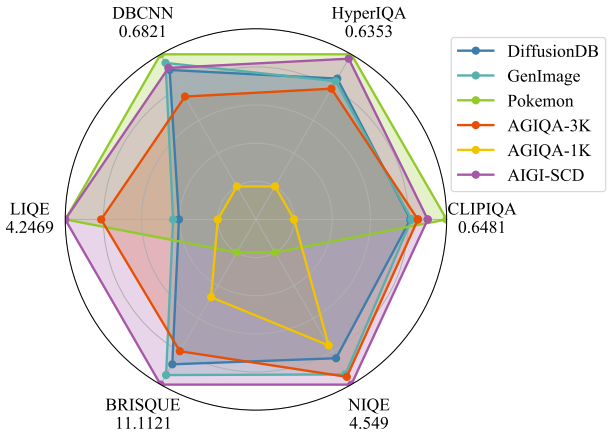


Fig. 6: Quality score comparison of the AIGI database. All existing AIGI databases have flaws in at least one quality indicator, while six quality scores of AIGI-SCD are all satisfactory, making them suitable for image compression tasks.

B. Data Analysis

This chapter will conduct a detailed analysis of the properties of the MISC-AIGI database. By comparing it with existing databases, we prove the importance of this database on AIGI compression tasks. For the existing NSI databases, to analyze their statistical differences against AIGIs, we use the distribution of five compression-related attributes for comparison, including light, contrast, color, blur, and Spatial Information (SI). Detailed descriptions of these attributes are defined in [70]. We selected three NSI compression databases, namely Kodak24 [39], CLIC2020 [41], and Tecnick [69] for comparison. As shown in Fig. 5, the attributes of NSI and AIGI are both normally distributed, indicating that AIGI-SCD has the same diversity as the traditional NSIs. However, compared to NSIs captured in the wild, the distribution curve of AIGIs is sharper. For example, AIGI is not affected by overexposure distortion in photography, so brightness and contrast are rel-

TABLE III: Performance of the state-of-art compression metrics and MISC-1/3 levels, validated on NSI database CLIC2020 [41], and AIGI database AIGI-SCD we constructed. The quality is evaluated by two consistency (LPIPS↓ [71], ClipSIM↑ [38]) and two perception (NIQE↓ [61], ClipIQA↑ [58]) indicators. A normalized average of those four indicators from equation (9) (10) is provided as the overall evaluation of consistency and perception together. [Key: **Best**; **Second Best**].

Metric	Consistency(NSI) (LPIPS↓,ClipSIM↑)	Perception(NSI) (NIQE↓,ClipIQA↑)	Average Avg↑	Bitrate Bpp↓	Consistency(AIGI) (LPIPS↓,ClipSIM↑)	Perception(AIGI) (NIQE↓,ClipIQA↑)	Average Avg↑	Bitrate Bpp↓
{T}JPEG [17]	(0.5166, 0.7075)	(15.476, 0.2052)	-1.4036	0.1952	(0.5421, 0.8198)	(15.881, 0.2602)	-1.5949	0.1738
{T}WEBP [18]	(0.3780, 0.8374)	(7.5893, 0.1514)	0.8392	0.1296	(0.3578, 0.9160)	(6.9384, 0.1550)	0.9989	0.1065
{T}VVC [37]	(**0.3577**, 0.8969)	(6.6469, 0.2390)	1.7583	0.1035	(0.3088, 0.9561)	(6.5056, 0.3173)	2.0541	0.0951
{N}BMSHJ [72]	(0.7543, 0.6406)	(9.0865, 0.2224)	-1.4347	0.0581	(0.6859, 0.7739)	(9.1777, 0.2240)	-1.3768	0.0581
{N}MBT [19]	(0.8661, 0.5605)	(10.560, 0.2096)	-2.5149	0.0527	(0.7990, 0.6558)	(10.210, 0.2188)	-2.7817	0.0497
{N}CHENG [73]	(0.4625, 0.8198)	(6.4557, 0.2310)	0.9468	0.0623	(0.3723, 0.9316)	(6.7251, 0.2726)	1.4630	0.0565
{G}HiFiC [25]	(0.3980, **0.9194**)	(5.6601, 0.4209)	2.5088	0.0478	(**0.2876**, **0.9663**)	(5.7907, 0.4205)	**2.6619**	0.0444
{G}CDC [28]	(0.4692, 0.8564)	(6.2161, 0.2033)	1.0163	0.0469	(0.3416, 0.9414)	(7.3019, 0.2570)	1.4929	0.0451
{G}PICS [34]	(0.6080, 0.4668)	(**3.8469**, **0.6639**)	0.9735	0.0265	(0.6685, 0.7530)	(**4.6612**, 0.6484)	0.7417	0.0328
{G}MISC-1	(0.5142, 0.8252)	(4.3597, 0.6106)	**2.6162**	0.0225	(0.5026, 0.8921)	(5.0878, **0.7347**)	2.4924	0.0223
{G}MISC-3	(**0.3522**, **0.9106**)	(**3.8271**, **0.6612**)	**3.8957**	0.0470	(**0.3084**, **0.9570**)	(**4.6820**, **0.7701**)	**3.8588**	0.0446

atively stable. In addition, as AIGIs sometimes have limited iterations during the generation process, the image may suffer from some blurry regions. Thus, the center of the blur curve is biased to the left. These differences indicate some compression mechanisms for NSIs, such as the equalization of brightness and chrominance, and the enhancement of blurry regions may not be effective for AIGIs. In all, for a compression algorithm trained with the traditional *NSI database*, its compression performance on *AIGIs* is likely to be unsatisfactory due to their *different low-level attributes*.

For the existing AIGI database, we use six IQA methods[2] to comprehensively evaluate its quality in Fig. 6 including the four indicators in Section IV-A, as well as HyperIQA [75] and BRISQUE [74]. Higher quality indicates less distortion of the image. The comparison databases include: DiffusionDB [76] and GenImage [77], which contain millions of images generated by different models. They are the largest and most versatile. Pokemon [78] is the first AIGI database generated by GAN, which is suitable for early-stage generation task verification. The AGIQA-3K [42] and AGIQA-1K [43] databases cover three generation architectures: GAN, Auto-Regression, and Diffusion, and are mainly oriented to IQA tasks. In Fig. 6, compared to AIGI-SCD, all quality scores of AGIQA-3K/1K lag behind; although other remaining databases occasionally lead AIGI-SGD slightly in some IQA indicators, they all have obvious shortcomings in other indicators. Therefore, these databases from generative models of mixed results are far inferior to AIGI-SCD in overall quality. In all, the characteristics of AIGIs and NSIs are quite different and require targeted compression. Compared with the existing AIGI database, AIGI-SCD has higher quality, less distortion, and *more objective verification of compression algorithms*.

V. Expriment

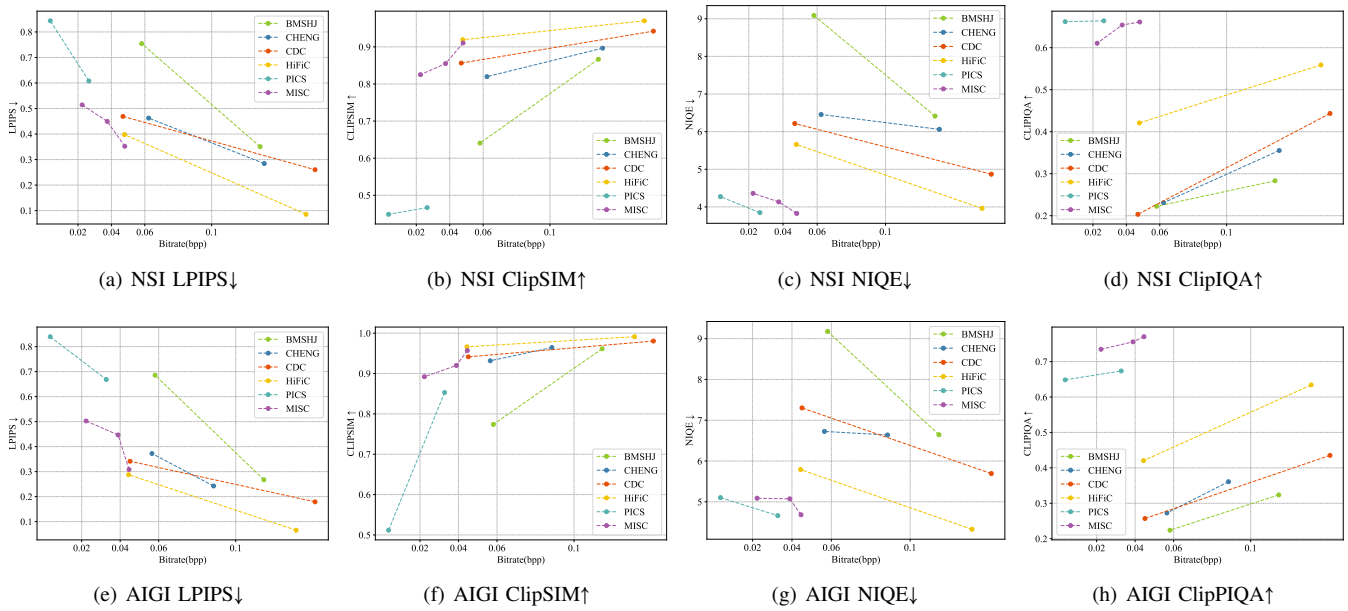
A. Experiment Settings

To assess the efficacy of the proposed MISC method across diverse image types, we conducted performance evaluations on two distinct databases: the commonly used CLIC2020[3] database for NSIs compression [41], and the AIGI-SCD database specifically designed for AIGIs. Following the standard partitioning, the training/test image distribution comprises 585/41 images for CLIC2020 and 450/50 images for AIGI-SCD. In our methodology, we kept the LMM and map encoders, as well as the T2I component of the decoder (utilizing default parameters of GPT-4 Vision [11], CLIP [38], and DiffBIR [57]) frozen, and focused on fine-tuning the low bitrate image encoder/decoder. While the encoder/decoder architecture aligns with existing NIC and GIC frameworks, the parameters underwent an intensive ultra-low fine-tuning process. Specifically, this process initiated the model at the lowest bitrate mode, followed by a tenfold increase in the bitrate weight of the loss function (λ reduced to $\frac{1}{10}$ of its original value), enabling training for extreme compression with a learning rate of 10^{-4}. MISC has three compression levels for dynamic adjustment. The first two levels (MISC-1/2) are for $0.02 \sim 0.03$ bpp, while we activate all encoders; the last level (MISC-3) is for $0.04 \sim 0.05$ bpp. This relatively higher bitrate can accommodate more image details, and using items to guide diffusion is not as important as before. To save bitrate, we discarded the information of these three NDM groups, by deactivating the item and detail questions in the LMM encoder, and the whole map encoder, then only used other modules for compression. In our comparative analysis, MISC is benchmarked against nine mainstream low-bitrate image compression methods, including traditional JPEG [17], WEBP [18], VVC (Intra frame mode) [37], NIC's BMSHJ [72], MBT [19], CHENG [73], and GIC's HiFiC [25], CDC [28], PICS [34]. As ultra-low bitrates are not their best scenario (excluding PICS), we applied the ultra-low fine-tuning approach above to all trainable models for a fair comparison. The experiments were conducted using NVIDIA RTX 4090 GPUs with the Adam optimizer [79].

In our assessment of compression performance, we employ a comprehensive set of metrics to address both consistency and perception requirements. As discussed in Section II-C,

[2]BRISQUE [74] and NIQE [61] are lower-better values, so these two axes take the reciprocal in Fig. 6.

[3]As the maximum output of current GIC metrics is 1,024 pixels, we 2× downsampled images larger than this bound.



(a) NSI LPIPS↓ (b) NSI ClipSIM↑ (c) NSI NIQE↓ (d) NSI ClipIQA↑

(e) AIGI LPIPS↓ (f) AIGI ClipSIM↑ (g) AIGI NIQE↓ (h) AIGI ClipPIQA↑

Fig. 7: Variable bitrate of BMSHJ [72], CHENG [73], CDC [28], HiFiC [25], PICS [34], and MISC we proposed. The left four figures are for consistency while the right four are for perception. Existing methods can only achieve satisfactory results on either consistency or perception. In contrast, MISC provides the possibility for ultra-low bitrate compression at MISC-1 while simultaneously achieving the optimal of all indicators at MISC-3.

TABLE IV: Using different compressed content groups in MISC, including three Name-Detail-Map (NDM) groups, a description of all details of the image, and an extremely compressed image bitstream. [Key: **Best**; **Selected Content** ☑].

(a) MISC validation on CLIC2020 [41] database (Left, Right): (MISC-1, MISC-2)

Content			Consistency		Perception		Bitrate
NDM	Detail all	Bitstream	LPIPS↓	ClipSIM↑	NIQE↓	ClipIQA↑	Bpp↓
☑☑☑	☑	☑	(**0.5142**, **0.4494**)	(**0.8252**, **0.8550**)	(4.3597, 4.1336)	(0.6106, **0.6540**)	(0.0225, 0.0375)
☑☑☑		☑	(0.6244, 0.5648)	(0.7764, 0.8184)	(8.8310, 8.0048)	(0.2587, 0.3031)	(0.0187, 0.0337)
☑☑		☑	(0.6311, 0.5733)	(0.7725, 0.8149)	(8.6842, 7.9538)	(0.2372, 0.2753)	(0.0180, 0.0330)
☑		☑	(0.6387, 0.5856)	(0.7764, 0.8125)	(8.8231, 8.0993)	(0.2114, 0.2335)	(0.0172, 0.0323)
	☑	☑	(0.5222, 0.4690)	(0.8149, 0.8530)	(**4.0333**, **3.8690**)	(**0.6335**, 0.6368)	(0.0202, 0.0353)
☑☑☑	☑		0.7760	0.7613	4.5878	0.6315	0.0061

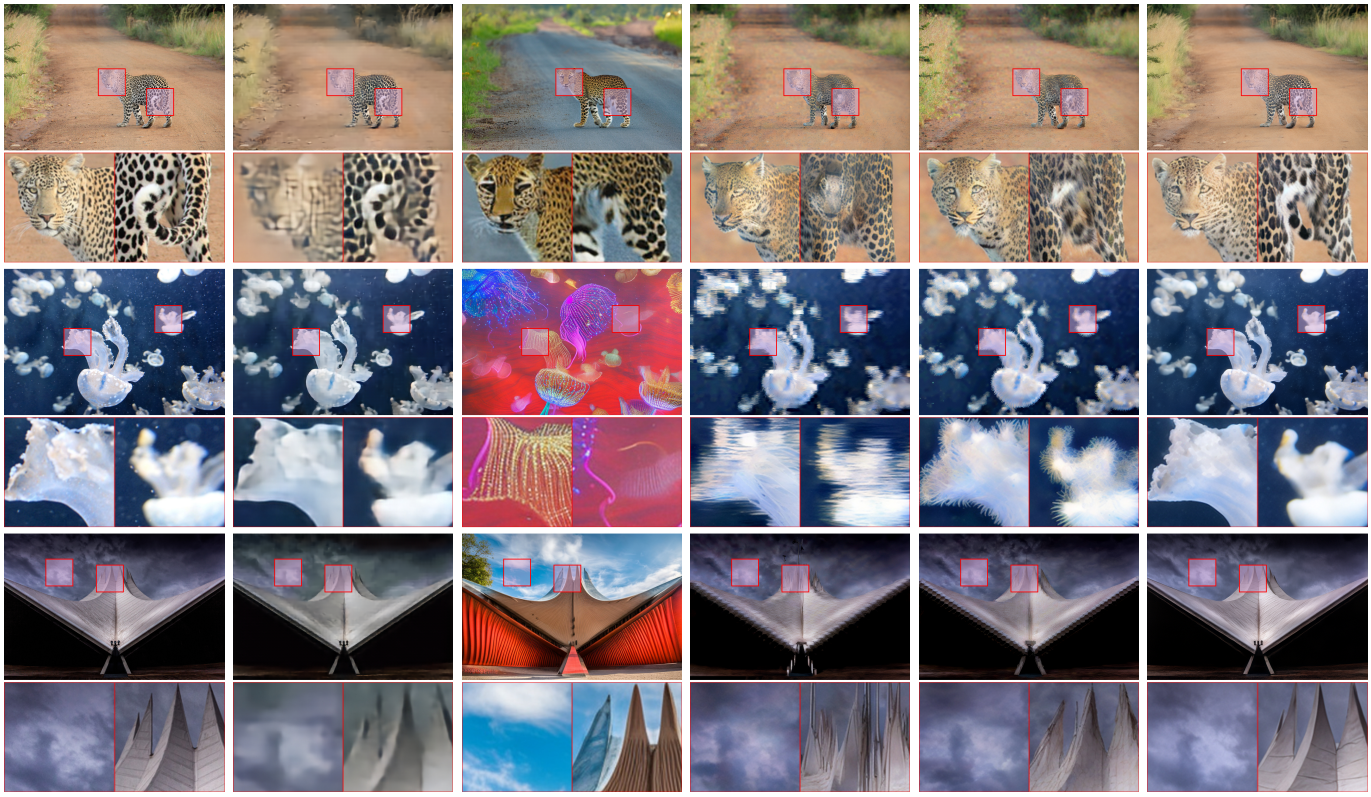
(b) MISC validation on AIGI-SCD database (Left, Right): (MISC-1, MISC-2)

Content			Consistency		Perception		Bitrate
NDM	Detail all	Bitstream	LPIPS↓	ClipSIM↑	NIQE↓	ClipIQA↑	Bpp↓
☑☑☑	☑	☑	(**0.5026**, **0.4468**)	(**0.8921**, **0.9204**)	(5.0878, 5.0734)	(**0.7347**, **0.7557**)	(0.0223, 0.0389)
☑☑☑		☑	(0.5299, 0.4651)	(0.8643, 0.9136)	(12.401, 11.784)	(0.3300, 0.4233)	(0.0188, 0.0354)
☑☑		☑	(0.5390, 0.4769)	(0.8638, 0.9082)	(12.780, 11.928)	(0.3051, 0.3705)	(0.0180, 0.0347)
☑		☑	(0.5469, 0.4862)	(0.8613, 0.9023)	(13.136, 12.391)	(0.2902, 0.3374)	(0.0173, 0.0340)
	☑	☑	(0.5480, 0.4671)	(0.8794, 0.9199)	(4.6725, 4.3572)	(0.7136, 0.7251)	(0.0200, 0.0367)
☑☑☑	☑		0.7416	0.7656	**4.2413**	0.6901	0.0058

semantic indicators become crucial at ultra-low bitrates for characterizing differences between the compressed image and the ground truth, surpassing the significance of pixel-level metrics such as PSNR and SSIM. Therefore, we utilize LPIPS [71], a widely used visual metric based on the Human Visual System (HVS), to quantify the distortion post-compression. Additionally, following prior work in LMM semantic compression [34], we utilize ClipSIM [38] to calculate the cosine distance of CLIP embeddings, assessing the similarity of semantic features between images. For perception evaluation, we utilize the IQA method NIQE [61] to evaluate low-level distortions like blur and noise. Furthermore, we incorporate the emerging IAA method ClipIQA [58] to measure human aesthetic satisfaction with the image. To provide a comprehensive representation of performance across all metrics, we calculate a normalized average of the four indicators mentioned above using equations (9) and (10) to offer an overall assessment of both consistency and perception aspects.

B. Experiment Results and Discussion

At ultra-low/low bitrate, we compared other advanced methods against MISC from the lowest (MISC-1) to the



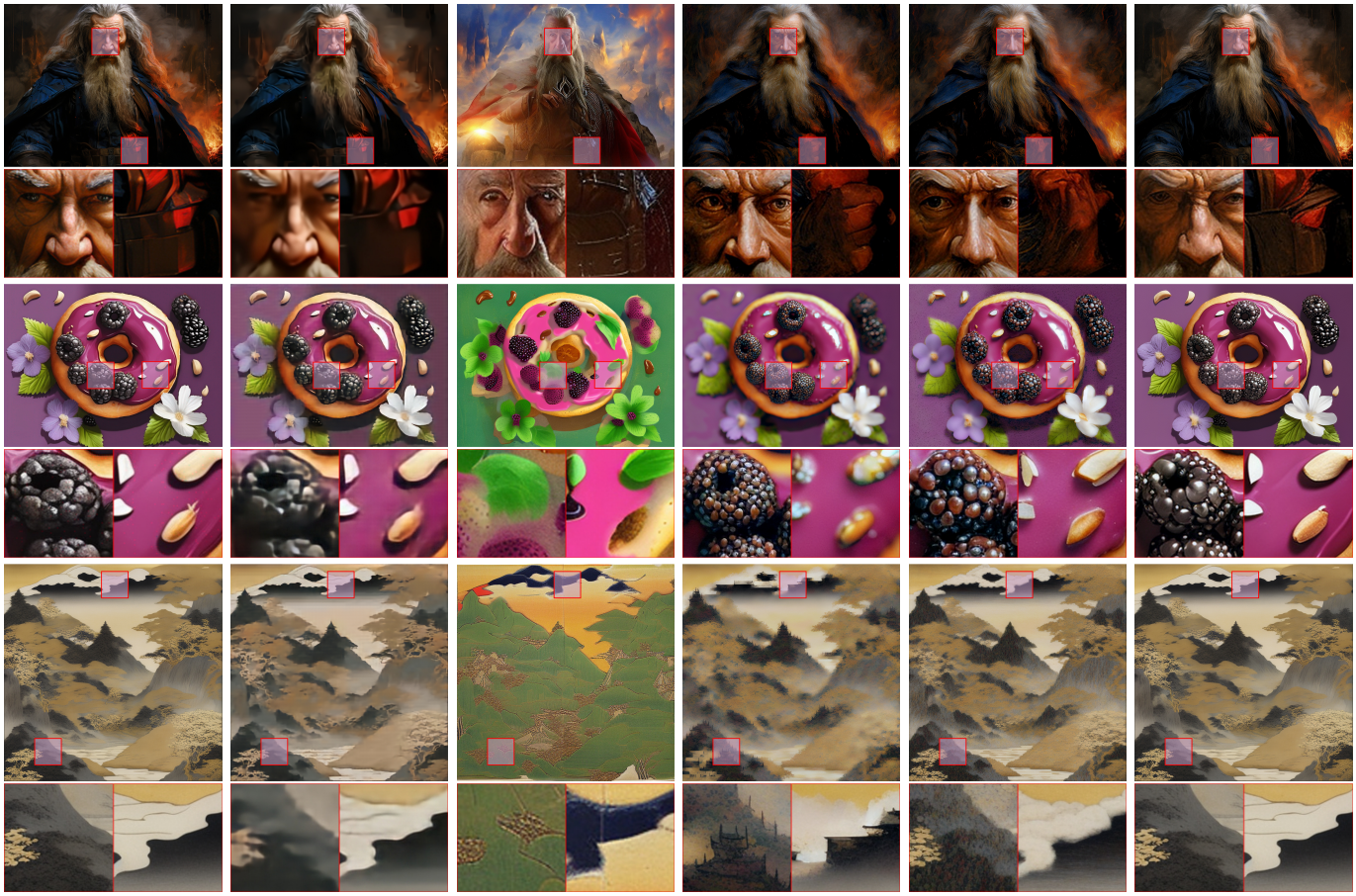
(a) Ground Truth (b) CHENG(0.0623 bpp) (c) PICS(0.0265 bpp) (d) MISC-1(0.0225 bpp) (e) MISC-2(0.0375 bpp) (f) MISC-3(0.0470 bpp)

Fig. 8: Qualitative experiment of NSI compression. CHENG [73] is suitable for only consistency, PICS [34] experts in only perception, and MISC performs the best in both indicators.

highest (MISC-3) bitrate levels respectively. As depicted in TABLE III, all the consistency and perception indicators of MISC-3 rank as the best or second best. Considering both consistency and perceptual quality collectively, MISC-3 achieves a normalized average score exceeding 3.8, significantly outperforming other methods. Even for MISC-1, it matches HiFiC in consistency while utilizing only 50% of its bitrate overhead. Examining each indicator independently, methods like VVC, HiFiC, and CDC can achieve high consistency with the original image but exhibit poor perceptual quality. For low-level vision, NIQE hovers around 6, while ClipIQA struggles to surpass 0.4 in aesthetics. In contrast, PICS offers higher perceptual quality but sacrifices consistency with the original image, ranking even the lowest in LPIPS and ClipIQA. Here, our MISC-3 matches HiFiC in consistency and surpasses NIQE/ClipIQA by approximately 1.8/0.24 in perception. Furthermore, it equals to PICS in perception and surpasses LPIPS/ClipSIM by around 0.35/0.45. In addition to its superior performance around 0.05 bpp, our approach excels in ultra-low bitrate compression below 0.024 bpp. Traditional compression methods have inherent bitrate limitations, and learning-based NIC/GIC methods struggle to train and converge at such low bitrates. In contrast, MISC-1 achieves acceptable consistency results for the first time, with slightly inferior LPIPS/ClipSIM compared to CHENG but nearly three times bitrate savings, while maintaining superior image perceptual quality. In summary, MISC has achieved a balance for the first time in addressing the conflicting optimization goals of extreme compression at low/ultra-low bitrates while reconstructing images with limited information, enhancing high-quality details consistent with the original image, and resolving the consistency-perception dilemma.

Fig. 7 further validates the performance of MISC and other methods across different bitrates. MISC not only leads at the 0.05 bpp bitrate but also maintains a significant advantage in perception compared to methods in the $0.1 \sim 0.2$ bpp range, with only a 30% bitrate overhead. Fig. 8 and 9 showcase visual compression examples of MISC on NSIs and AIGIs, encompassing characters, items, and backgrounds. Compared to traditional methods, MISC generates rich high-quality details and aligns more closely with the ground truth. Internally comparing different levels of MISC, MISC-1 produces outlines roughly consistent with the ground truth but exhibits some freedom in character expressions, object colors, and building shapes. MISC-2 improves on consistency with the ground truth but still lacks accuracy in rendering details like clothing, sky, and mountains. MISC-3 achieves near-complete consistency with the ground truth, with minimal texture differences upon zooming in that do not impact visual quality. In essence, MISC offers dynamic bitrate adjustment capabilities and demonstrates versatility in low/ultra-low bitrate scenarios.

Besides, regarding compressed image content, TABLE III and Fig. 7 illustrate that AIGI compression achieves higher consistency than NSI at equivalent bitrates, along with superior



(a) Ground Truth (b) CHENG(0.0565 bpp) (c) PICS(0.0328 bpp) (d) MISC-1(0.0223 bpp) (e) MISC-2(0.0389 bpp) (f) MISC-3(0.0446 bpp)

Fig. 9: Qualitative experiment of AIGI compression. CHENG [73] is suitable for only consistency, PICS [34] experts in only perception, and MISC performs the best in both indicators.

perceptual quality in aesthetics. However, existing compression algorithms fall short in low-level quality. Various methods on AIGI only reach 80% of NSI in terms of NIQE, attributed to both inherent low-level defects and the limited understanding of compression algorithms on AIGI. Notably, despite the low-level quality of NIQE, MISC excels in aesthetic quality, achieving a ClipIQA score above 0.7 even at the lowest MISC-1. In conclusion, beyond NSI, providing viewers with high perceptual quality and consistency for AIGI poses a significant challenge for future compression metrics.

C. Ablation Study

To validate the contributions of the different compressed content in MISC, an ablation study was conducted, and the results are presented in Table TABLE IV. The factors are specified as (i) Three Name-Detail-Map (NDM) group for items. (ii) A text description of all details in the image. (iii) A bitstream characterizing the extremely compressed image content. Considering that MISC-3 does not depend on NDM groups, we use MISC-1 and MISC-2, discard the above three contents during the compression process, and evaluate the results in terms of consistency and perception. The results show that [**Detail all**] is the most important content for both NSI and AIGI. Without its guidance, both consistency and perception performance decrease across the board. In addition, [**Bitstream**] also lays the foundation of MISC. Although its absence will not affect the Perception indicator, it will lead to a significant drop in Consistency. The only contentious issue lies in the utilization of [**NDM**] content. For perception, using NDM enhances aesthetic quality but diminishes low-level quality. Conversely, for consistency, omitting NDM results in an overall performance decline. As Fig. 7 illustrates MISC already has better perception beyond consistency. Given that incorporating three NDM groups only necessitates a bitrate overhead of 0.002 bpp, adding NDM groups can enhance consistency with minimal sacrifice in perception. In conclusion, eliminating any single content leads to performance deterioration, affirming their collective contribution to the final compression performance.

D. User Study

To verify the practicality of MISC in real-life scenarios, we conduct a subjective user study beyond the objective indicators, to analyze the human preference for the compressed image. We established an environment with standard lighting, displaying the ground truth centrally, and two compressed

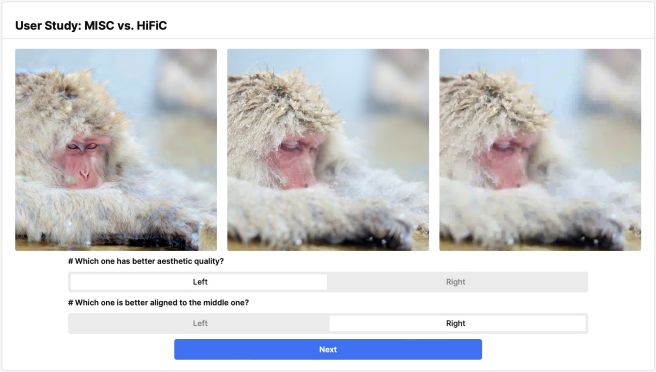
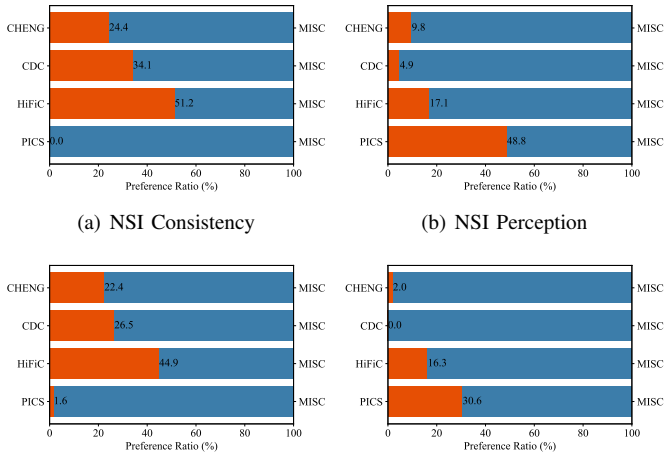


Fig. 10: User interface for choosing preference in terms of consistency/perception. The ground truth image in the middle is compressed by different metrics on the left/right.



(a) NSI Consistency (b) NSI Perception

(c) AIGI Consistency (d) AIGI Perception

Fig. 11: Statistical results of user study in the NSI database CLIC-2020 [41] and AIGI database AIGI-SCD. Humans subjectively believe that MISC is the best compression metric for both consistency and perception.

images on an iMac monitor with a resolution of up to 4,096 × 2,304 pixels. Viewers are required to select preferences between two images compressed by different algorithms, at both consistency and perception levels. The experiment involved 7 graduate students (4 males and 3 females) interacting with the interface in Fig. 10. The proposed MISC is compared with four state-of-art compression metrics, namely CHENG [73], CDC [28], HiFiC [25], and PICS [34]. A certain bitrate ($0.04 \sim 0.05$) is set for all those metrics for a fair comparison, by using MISC-3, and the bitrate in TABLE III for other metrics. The validation results illustrated in Fig. 11 demonstrate the superior performance of MISC across all evaluated criteria. Notably, MISC performs comparably to the PICS for consistency, and HiFiC for perception. Furthermore, compared to NSIs, AIGIs compressed by MISC were more preferred by human evaluators. It is worth mentioning that this subjective result is slightly different from the objective indicators in TABLE III. For example, in AIGI compression, the HiFiC [25] achieves the best objective indicators, but Fig. 11 shows that MISC is more subjectively preferred. Therefore, appropriate objective perceptual quality assessment measures should be developed to inspire image compression at ultra-low bitrates.

VI. Conclusion

In this paper, an image compression method called MISC is proposed at ultra-low bitrates. It can significantly reduce the storage space required and save bandwidth to transfer images. The framework of MISC consists of LMM/map/image encoders, and a general decoder. Experimental results on CLIC2020 and the AIGI-SCD database constructed in this paper show that MISC solves the trade-off problem between consistency and perception, as well as good scalability with bitrates. With the evolution of today's storage devices and communication protocols, we believe this LMM-driven methodology has the potential to facilitate a new paradigm for image compression.

References

[1] Qi Mao, Chongyu Wang, Meng Wang, Shiqi Wang, Ruijie Chen, Libiao Jin, and Siwei Ma, "Scalable face image coding via stylegan prior: Toward compression for human-machine collaborative vision," *IEEE Transactions on Image Processing*, vol. 33, pp. 408–422, 2024.

[2] Wenhong Duan, Zheng Chang, Chuanmin Jia, Shanshe Wang, Siwei Ma, Li Song, and Wen Gao, "Learned image compression using cross-component attention mechanism," *IEEE Transactions on Image Processing*, vol. 32, pp. 5478–5493, 2023.

[3] Juan Wang, Yiping Duan, Xiaoming Tao, Mai Xu, and Jianhua Lu, "Semantic perceptual image compression with a laplacian pyramid of convolutional networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 4225–4237, 2021.

[4] Chunyi Li, Haoyang Li, Ning Yang, and Dazhi He, "A pbch reception algorithm in 5g broadcasting," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2022.

[5] Zicheng Zhang, Yingjie Zhou, Long Teng, Wei Sun, Chunyi Li, Xiongkuo Min, Xiao-Ping Zhang, and Guangtao Zhai, "Quality-of-experience evaluation for digital twins in 6g network environments," *IEEE Transactions on Broadcasting*, 2024.

[6] Yochai Blau and Tomer Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6228–6237.

[7] Yochai Blau and Tomer Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*. PMLR, 2019, pp. 675–685.

[8] Bolin Chen, Shanzhi Yin, Peilin Chen, Shiqi Wang, and Yan Ye, "Generative visual compression: A review," arXiv preprint arXiv:2402.02140, 2024.

[9] Qi Mao, Tinghan Yang, Yinuo Zhang, Zijian Wang, Meng Wang, Shiqi Wang, and Siwei Ma, "Extreme image compression using fine-tuned vqgans," arXiv preprint arXiv:2307.08265, 2023.

[10] Jianhui Chang, Jian Zhang, Jiguo Li, Shiqi Wang, Qi Mao, Chuanmin Jia, Siwei Ma, and Wen Gao, "Semantic-aware visual decomposition for image coding," *International Journal of Computer Vision*, vol. 131, pp. 2333–2355, 2023.

[11] OpenAI, "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample, "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[14] Robin Rombach, Andreas Blattmann, and Björn Ommer, "Text-guided synthesis of artistic images with retrieval-augmented diffusion models," arXiv preprint arXiv:2207.13038, 2022.

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, 2022.

[16] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin, "Ai-generated content (aigc): A survey," arXiv preprint arXiv:2304.06632, 2023.

[17] Athanassios N. Skodras, Charilaos A. Christopoulos, and Touradj Ebrahimi, "The jpeg 2000 still image compression standard," *IEEE Signal Process. Mag.*, vol. 18, pp. 36–58, 2001.

[18] Oren Rippel and Lubomir Bourdev, "Real-time adaptive image compression," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2922–2930.

[19] David C. Minnen, Johannes Ballé, and George Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Neural Information Processing Systems*, 2018.

[20] Junqi Shi, Ming Lu, and Zhan Ma, "Rate-distortion optimized post-training quantization for learned image compression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[21] Sheng Cao, Chao-Yuan Wu, and Philipp Krähenbühl, "Lossless image compression through super-resolution," arXiv preprint arXiv:2004.02872, 2020.

[22] Wei Gao, Lvfang Tao, Linjie Zhou, Dinghao Yang, Xiaoyu Zhang, and Zixuan Guo, "Low-rate image compression with super-resolution learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 154–155.

[23] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 221–231.

[24] Fangyuan Gao, Xin Deng, Junpeng Jing, Xin Zou, and Mai Xu, "Extremely low bit-rate image compression via invertible image generation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[25] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11913–11924, 2020.

[26] Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer, "Multi-realism image compression with a conditional generator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22324–22333.

[27] Shoma Iwai, Tomo Miyazaki, and Shinichiro Omachi, "Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2900–2909.

[28] Ruihan Yang and Stephan Mandt, "Lossy image compression with conditional diffusion models," arXiv preprint arXiv:2209.06950, 2022.

[29] Zhihong Pan, Xin Zhou, and Hao Tian, "Extreme generative image compression by learning text embedding from diffusion models," arXiv preprint arXiv:2211.07793, 2022.

[30] Jiguo Li, Chuanmin Jia, Xinfeng Zhang, Siwei Ma, and Wen Gao, "Cross modal compression: Towards human-comprehensible semantic compression," in *ACM Multimedia*, 2021.

[31] Junlong Gao, Chuanmin Jia, Zhimeng Huang, Shanshe Wang, Siwei Ma, and Wen Gao, "Rate-distortion optimized cross modal compression with multiple domains," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[32] Tom Bordin and Thomas Maugey, "Semantic based generative compression of images for extremely low bitrates," in *2023 IEEE 25th International Workshop on Multimedia Signal Processing*. IEEE, 2023, pp. 1–6.

[33] Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis, "High-fidelity image compression with score-based generative models," arXiv preprint arXiv:2305.18231, 2023.

[34] Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti, "Text + sketch: Image compression at ultra low rates," arXiv preprint arXiv:2307.01944, 2023.

[35] Thomas Wiegand, Gary J. Sullivan, Gisle Bjøntegaard, and Ajay Luthra, "Overview of the h.264/avc video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 560–576, 2003.

[36] Gary J. Sullivan, Jens-Rainer Ohm, Woojin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 1649–1668, 2012.

[37] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 3736–3764, 2021.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[39] Gisle Bjontegaard, "Calculation of average psnr differences between rd-curves," *ITU SG16 Doc. VCEG-M33*, 2001.

[40] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, July 2017.

[41] Johannes Ballé, Philip A Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici, "Nonlinear transform coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 339–353, 2020.

[42] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[43] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai, "A perceptual quality assessment exploration for aigc images," in *IEEE International Conference on Multimedia and Expo Workshops*, 2023.

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[45] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[46] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al., "Q-bench: A benchmark for general-purpose foundation models on low-level vision," arXiv preprint arXiv:2309.14181, 2023.

[47] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al., "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," arXiv preprint arXiv:2311.06783, 2023.

[48] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui, et al., "Q-boost: On visual quality assessment ability of low-level multi-modality foundation models," arXiv preprint arXiv:2312.15300, 2023.

[49] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al., "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," arXiv preprint arXiv:2312.17090, 2023.

[50] Chunyi Li, Haoning Wu, Zicheng Zhang, Hongkun Hao, Kaiwei Zhang, Lei Bai, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai, "Q-refine: A perceptual quality refiner for ai-generated image," arXiv preprint arXiv:2401.01117, 2024.

[51] Zicheng Zhang, Wei Sun, Houning Wu, Yingjie Zhou, Chunyi Li, Xiongkuo Min, Guangtao Zhai, and Weisi Lin, "Gms-3dqa: Projection-based grid mini-patch sampling for 3d model quality assessment," arXiv preprint arXiv:2306.05658, 2023.

[52] Zicheng Zhang, Wei Sun, Yingjie Zhou, Haoning Wu, Chunyi Li, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin, "Advancing zero-shot digital human quality assessment through text-prompted evaluation," arXiv preprint arXiv:2307.02808, 2023.

[53] Chunyi Li, May Lim, Abdelhak Bentaleb, and Roger Zimmermann, "A real-time blind quality-of-experience assessment metric for http adaptive streaming," in *IEEE International Conference on Multimedia and Expo*, 2023.

[54] Xinhui Huang, Chunyi Li, Abdelhak Bentaleb, Roger Zimmermann, and Guangtao Zhai, "Xgc-vqa: A unified video quality assessment model for user, professionally, and occupationally-generated content," in *IEEE International Conference on Multimedia and Expo Workshops*, 2023.

[55] Chunyi Li, Zicheng Zhang, Wei Sun, Xiongkuo Min, and Guangtao Zhai, "A full-reference quality assessment metric for cartoon images," in *IEEE 24th International Workshop on Multimedia Signal Processing*, 2022.

[56] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li, "Clip surgery for better explainability with enhancement in open-vocabulary tasks," arXiv preprint arXiv:2304.05653, 2023.
[57] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong, "Diffbir: Towards blind image restoration with generative diffusion prior," arXiv preprint arXiv:2308.15070, 2023.
[58] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 2555–2563.
[59] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.
[60] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14071–14081.
[61] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.
[62] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach, "Adversarial diffusion distillation," arXiv preprint arXiv:2311.17042, 2023.
[63] David Holz, "Midjourney," https://www.midjourney.com, 2023.
[64] Yatharth Gupta, Vishnu V. Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen, "Progressive knowledge distillation of stable diffusion xl using layer level loss," arXiv preprint arXiv:2401.02677, 2024.
[65] PlaygroundAI, "playground-v2-1024px-aesthetic," https://playground.com, 2023.
[66] dreamlike art, "dreamlike-photoreal-2.0," https://dreamlike.art, 2023.
[67] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," arXiv preprint arXiv:2310.00426, 2023.
[68] DeepFloyd, "If-i-xl-v1.0," https://www.deepfloyd.ai, 2023.
[69] Nicola Asuni and Andrea Giachetti, "Testimages: a large-scale archive for testing visual devices and basic image processing algorithms.," in *Smart Tools and Applications in Graphics*, 2014, pp. 63–70.
[70] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth international conference on quality of multimedia experience*. IEEE, 2017, pp. 1–6.
[71] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," arXiv preprint arXiv:1801.03924, 2018.
[72] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.
[73] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7936–7945, 2020.
[74] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
[75] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.
[76] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau, "Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models," *arXiv preprint arXiv:2210.14896*, 2022.
[77] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang, "Genimage: A million-scale benchmark for detecting ai-generated image," arXiv preprint arXiv:2306.08571, 2023.
[78] Justin N. M. Pinkney, "Pokemon blip captions," https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.
[79] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.