

Generative Semantic Coding for Ultra-Low Bitrate Visual Communication and Analysis

Weiming Chen^{1*} Yijia Wang^{1*} Zhihan Zhu¹ Zhihai He^{1,2†}

¹Southern University of Science and Technology, Shenzhen, China

²Pengcheng Laboratory, Shenzhen, China

{chenwm2023, wangyj2022, 12312326}@mail.sustech.edu.cn

hezh@sustech.edu.cn

Abstract

We consider the problem of ultra-low bit rate visual communication for remote vision analysis, human interactions and control in challenging scenarios with very low communication bandwidth, such as deep space exploration, battlefield intelligence, and robot navigation in complex environments. In this paper, we ask the following important question: **can we accurately reconstruct the visual scene using only a very small portion of the bit rate in existing coding methods while not sacrificing the accuracy of vision analysis and performance of human interactions?** Existing text-to-image generation models offer a new approach for ultra-low bitrate image description. However, they can only achieve a semantic-level approximation of the visual scene, which is far insufficient for the purpose of visual communication and remote vision analysis and human interactions. To address this important issue, we propose to seamlessly integrate image generation with deep image compression, using joint text and coding latent to guide the rectified flow models for precise generation of the visual scene. The semantic text description and coding latent are both encoded and transmitted to the decoder at a very small bit rate. Experimental results demonstrate that our method can achieve the same image reconstruction quality and vision analysis accuracy as existing methods while using much less bandwidth. The code will be released upon paper acceptance.

1. Introduction

In this paper, we consider the problem of ultra-low bit rate visual communication for remote vision analysis, human interactions and control in challenging scenarios such as deep space exploration, battlefield intelligence, and robot navigation in complex environments. In these scenarios, the

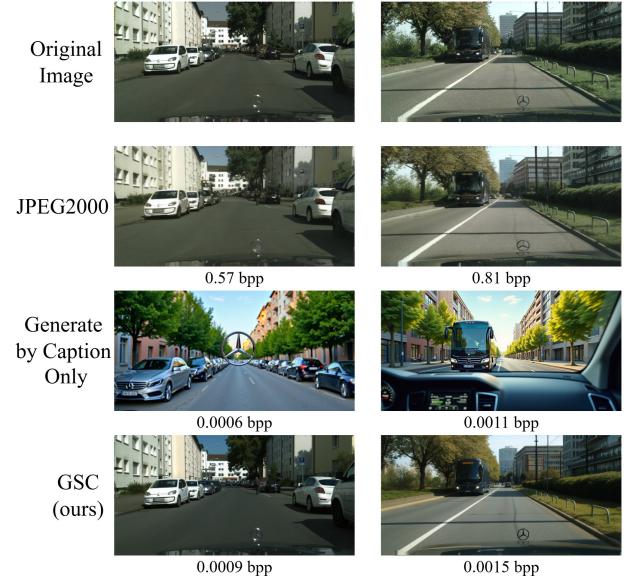


Figure 1. Example of GSC result compared with JPEG2000, one of the compression standards, and the result generated only guided by the caption.

sender and the receiver often have abundant computational power and resources. For example, the exploration robot on the moon or Mars, as well as the receiving station, is equipped with high-end GPUs and a sufficient power supply. However, the communication bandwidth between the sender and receiver is a very scarce resource due to the long transmission distance or strong interference. In these scenarios, we need to accurately reconstruct the visual scene for vision analysis, decision making, human interactions and control.

Existing image and video compression methods, such as JPEG2000 [1] and H.265 [2] excel in pixel-level reconstruction, but require high bandwidth. For instance, H.265 en-

*Equal contributions

†Corresponding author

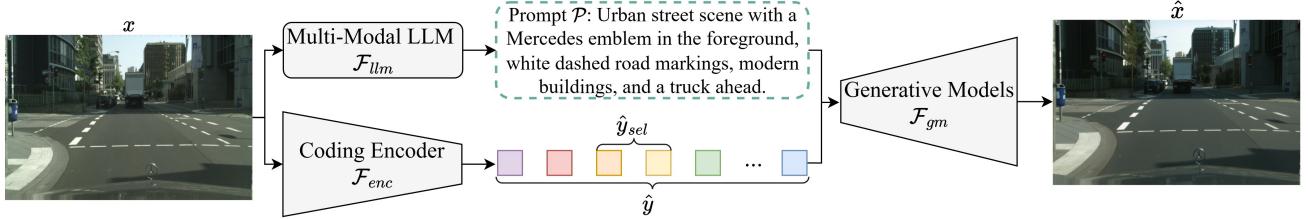


Figure 2. Overview of the proposed Generative Semantic Coding (GSC) framework.

coded standard-definition video often requires bandwidth ranging from 1 Mbps to 2 Mbps, which far exceeds the bandwidth available in these scenarios discussed above. Note that, in these scenarios, the purpose of the visual communication is to support accurate remote vision analysis, human interactions, control, and decisions. Thus, pixel-perfect reconstruction is not necessary. It only needs to reconstruct the image and visual scene such that subsequent vision analysis performance is consistent with that using the original images. An important question to ask is: *can we accurately reconstruct the visual scene using only a very small portion of the bit rate in existing coding methods while not sacrificing the accuracy of vision analysis and performance of human interactions?*

Recent advances in text-to-image generation models [9–12, 25, 28, 29, 31, 32] offer a novel approach for scene description and reconstruction. With this method, we only need to transmit the text descriptions of the scene to the receiver end, allowing the reconstruction of the visual scene. Unfortunately, the text description is often very subjective. With texts, they can only reconstruct and approximate the visual scene semantically at a very coarse level. Recently, researchers have studied using extra visual information, such as contours and sketches, to guide the text-image generation process [16, 31, 35]. They still suffer from inaccurate reconstruction of image details and high bit rate cost.

To overcome these limitations, we proposed a novel framework, called Generative Semantic Coding (GSC), as illustrated in Figure 2. We seamlessly integrate image generation with deep image compression, using joint text and coding latent to guide the rectified flow models for precise generation of the visual scene. We observe that the coded latents from the deep image compression system provide compact and high-quality guidance for the image generation. We dynamically select a tiny portion of the coding latents that contains the most significant information for preserving the structural consistency between the original and reconstructed images. The semantic text description and coding latent are both encoded and transmitted to the decoder at a very small bit rate. As shown in Figure 1, these selected coding latents only require less than 0.001

bpp, which is ultra-low but works well.

2. Related Work and Unique Contribution

In this section, we first review existing generative image compression methods related to our work. Then, we point out the necessity of adding conditional guidance. Finally, we summarize the unique contributions of this work.

2.1. Ultra-Low Bitrate Coding with Generative Models

Existing generative image compression methods typically for ultra-low bitrates operate within the bitrate range from 0.02 bpp to 0.10 bpp. For example, GLC [13] and HiFiC [20] employ GANs to learn image distributions for efficient compression but suffer from significant distortions and detail loss at extremely low bitrates. PerCo [3] trains a hyper-encoder and a codebook to extract image features, emphasizing perceptual quality via diffusion models; nevertheless, at extremely low bitrates, its perceptual quality still degrades. MS-ILLM [22] optimizes compression through multi-step iterations and language models to extract semantic information, but its image quality is severely compromised below 0.01 bpp. Recently, some methods [23, 41] transmit a quantized embedding as a conditional input to the diffusion-based decoder, while DiffC [35, 38] directly transmits pixels corrupted by noise in a diffusion process. But they don't focus on the semantic coding. Text-Sketch [16] adopts prompt inversion to maintain semantic consistency through CLIP [26], but struggles to keep spatial consistency and wastes lots of bits. These methods all face challenges at bitrates lower than 0.01 bpp, highlighting the need for more advanced techniques to address this issue. We leverage the inherent structural information embedded in the coded feature to ensure consistency under extremely low bitrate conditions.

2.2. Controllable Diffusion Models

One limitation of generative image compression methods is that textual descriptions alone cannot effectively control the image generation process. Therefore, it is necessary to incorporate additional conditional guidance mech-

anisms [21, 42, 44] to enhance controllability. ControlNet [44] augments diffusion models with additional conditional branches, enabling fine-grained control over the generation process using structural information, while preserving the original model’s generation fidelity. IP-Adapter [42] introduces a decoupled cross-attention mechanism by adding an additional cross-attention module to each existing cross-attention layer in the U-Net, facilitating more effective identity or style transfer in text-to-image generation. T2I-Adapter [21] introduces lightweight and composable adapters that align internal features of frozen text-to-image models with external control information. Inspired by ControlNet [44], we augment the FLUX model [15] with an additional module to inject encoded guidance, effectively controlling image generation and preserving both structural and semantic information.

2.3. Unique contributions

Our major unique contributions are as follows: (1) This paper considers an extreme scenario where transmission resources are severely limited while side resources are abundant. In this context, we discuss how to encode an image using minimal information, targeting bitrates below 0.01 bpp. (2) We develop a new approach, called generative semantic coding (GSC), which controls the image generation process to reconstruct images as precisely as possible. (3) Extensive experiments on three fundamental vision tasks demonstrate that our method achieves comparable performance to previous approaches while only utilizing less than 10% of their bpp, specifically less than 0.007 bpp.

3. The Proposed GSC Method

In this section, we begin with an overview of our proposed method (Section 3.1), followed by a detailed exposition of its two principal components (Section 3.2 and Section 3.3). Finally, we provide a theoretical analysis of the problem and our method (Section 3.4).

3.1. Method Overview

The architecture of our proposed GSC framework is shown in Figure 2. Given an input image x , we first extract its caption \mathcal{P} by a multi-model large language model (MMLLM). This caption encodes the semantic information of x . Structural and spatial details are extracted by a deep image encoder \mathcal{F}_{enc} that generates the latent representation $\hat{y} = \{\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \dots, \hat{Y}_n\}$, from which, we dynamically select a small subset $\hat{y}_{sel} = \{\hat{Y}_1^{sel}, \hat{Y}_2^{sel}, \hat{Y}_3^{sel}, \dots, \hat{Y}_C^{sel}\}$. Both \mathcal{P} and \hat{y}_{sel} are encoded and transmitted to the receiver. Guided by the \mathcal{P} and \hat{y}_{sel} , the receiver reconstructed image \hat{x} by the rectified flow (RF) [19]. P enforces the semantic consistency between the original and reconstructed image, while \hat{y}_{sel} ensures the structural consistency.

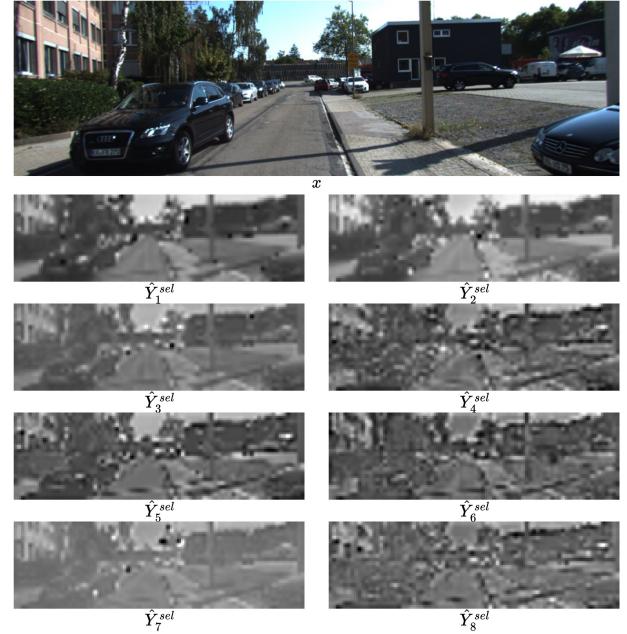


Figure 3. An example visualization of 8 selected channels.

3.2. Latent Construction and Channel Selection

As stated in the above section, we obtain the latent representation \hat{y} by encoding the original image x with a pre-trained image coding encoder. Instead of transmitting all n channels of \hat{y} , we focus on selecting C channels of \hat{y} to guide the generation process. Here, C is a very small number.

We first use the deep image encoder g_a to analyze the input image x to generate the latent representation $y = g_a(x; \phi)$, where ϕ is the learned parameters of g_a . Then, y is quantized into \hat{y} using a quantizer Q , $\hat{y} = Q(y)$. The entropy model is used to estimate the probability distribution Φ of \hat{y} to optimize bit allocation in encoding and decoding processes. This process can be written as:

$$\hat{y} = Q(g_a(x, \phi), \Phi). \quad (1)$$

From \hat{y} , we select a very small subset of channels \hat{y}_{sel} to guide the image generation process. In this work, we recognize that the task of \hat{y}_{sel} is to maintain the spatial and structural consistency between the reconstructed image and the original input. Therefore, we propose to use the SSIM (Structural Similarity Index) to dynamically select \hat{y}_{sel} . In our design, we select C channels with the largest SSIM value computed from the \hat{Y}_i ($i = 1, 2, 3, \dots, 320$) and an example of the gray-scale representation of the selected channels, i.e., \hat{Y}_i^{sel} ($i = 1, 2, 3, \dots, 8$), is shown in Figure 3.

It should be noted that, if more channels are selected to construct the \hat{y}_{sel} , higher accuracy can be achieved; however, more bits are required to encode them. The represents

a tradeoff between the visual analysis performance and encoding bit rate

$$\min_{\hat{y}_{sel}} \alpha |V(\hat{x}) - V(x)| + \beta B(\hat{y}_{sel}, \mathcal{P}), \quad (2)$$

where $B(\hat{y}_{sel}, \mathcal{P})$ represents the bits required to transmit the \hat{y}_{sel} and \mathcal{P} , $V(x)$ represents the visual analysis results of x , and α and β are weight parameters to control the trade-off between them.

3.3. Joint Text-Latent Guided Image Generation

As stated in the above section, guided by the image description \mathcal{P} and its coding latent \hat{y}_{sel} , we generate the reconstructed image \hat{x} using the FLUX text-to-image generation model [15]. As shown in Figure 4, a noise latent z_{t_N} is randomly sampled from the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. It is denoised under the guidance of \mathcal{P} and \hat{y}_{sel} . The \mathcal{P} is directly input into the T5 text encoder [27] to become the text embedding \mathcal{P}_{emb} to be used in the following Diffusion Transformer (DiT) [24] blocks. To incorporate the guidance of \hat{y}_{sel} , we create a trainable copy of the M multi-stream DiT blocks and S single-stream DiT blocks. Its initial inputs contain two parts: one is \mathcal{P}_{emb} , and the other is the sum of z_{t_N} and \hat{y}_{sel} . The outputs of each corresponding DiT block, after passing through a zero linear layer, are added to the first M multi-stream DiT blocks and S single-stream DiT blocks, respectively. As there are M_f Multi-stream DiT blocks and M_s Single-stream DiT blocks in the original FLUX [15], the rest $(M_f - M)$ multi-stream DiT blocks and $(S_f - S)$ single-stream DiT blocks remain the same as the original ones. After that, it performs denoising over N discrete timesteps $t = \{t_N, \dots, t_0\}$ by the following equation:

$$z_{t_{i-1}} = z_{t_i} + (t_{i-1} - t_i) v_\theta(z_{t_i}, t_i, \mathcal{P}_{emb}, \hat{y}_{sel}), \quad (3)$$

where $i = N, N-1, N-2, \dots, 1$ and v_θ is the predicted vector field obtained from the DiT blocks, parameterized by θ . After z_0 is obtained, it serves as an input to the VAE decoder to obtain the final output image. After T steps, we finally obtain the \hat{x} .

For training, we only activate and train M multi-stream DiT blocks and S single-stream DiT blocks, and freeze all the DiT blocks in the original FLUX [15]. The goal is to train a neural network to predict the v_θ . To this end, we couple samples from the target distribution with the samples from the Gaussian distribution via a linear path: $Z_t = tZ_1 + (1-t)Z_0$. Therefore, the marginal distribution of Z_t becomes:

$$p_t(z_t) = \mathbb{E}_{Z_1 \sim p_1} [p_t(z_t|Z_1)] = \int p_t(z_t|z_1)p_1(z_1) dz_1. \quad (4)$$

Given the initial state $Z_0 = z_0$ and the target state $Z_1 = z_1$, the linear path becomes $dZ_t = v_t(Z_t|z_1)dt = z_1 - z_0$. The

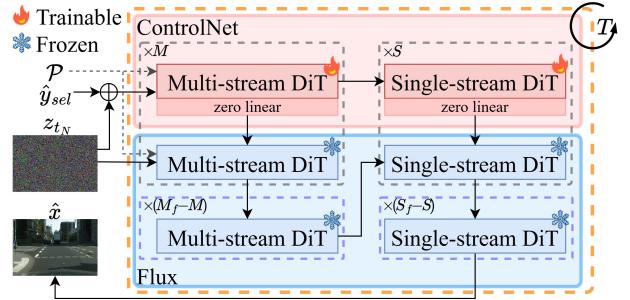


Figure 4. The details of the text-structural image generation process.

marginal vector field can be derived from the conditional vector field using the equation as follows,

$$\begin{aligned} v_t(z_t) &= \mathbb{E}_{Z_1 \sim p_1} \left[v_t(z_t|Z_1) \frac{p_t(z_t|Z_1)}{p_t(z_t)} \right] \\ &= \int v_t(z_t|z_1) \frac{p_t(z_t|z_1)}{p_t(z_t)} p_1(z_1) dz_1. \end{aligned} \quad (5)$$

After that, we use a neural network $v_\theta(z_t, t, \mathcal{P}, \hat{y}_{sel})$, parameterized by θ , to approximate the marginal vector field $v_t(z_t)$ through the conditional flow matching given by

$$\begin{aligned} \mathcal{L}_{CFM}(\varphi) &:= \mathbb{E}_{t \sim \mathcal{U}[0,1], Z_t \sim p_t(\cdot|Z_1), Z_1 \sim p_1} \\ &\quad [\|v_t(Z_t|Z_1) - v_\theta(Z_t, t, \mathcal{P}_{emb}, \hat{y}_{sel}; \varphi)\|_2^2]. \end{aligned} \quad (6)$$

3.4. Theoretical Analysis

In image compression with text and structural information, some guidance information might be useless or even misleading for the target image generation process. For example, different channels of \hat{y}_{sel} might contain similar information. Although it is difficult to accurately extract the useful guidance information, it is very important to understand its performance bound. Here, we present a theoretical analysis to characterize the lower bound of the coding bit rate.

We recognize that useful information is not uniformly distributed throughout the entire image, and only a subset of pixels contains important and useful information about the image. Motivated by this, we introduce a function $U(X)$ to quantize the information contained by pixel X in the image x has. We obtain the probability of quantized information by

$$P(X) = \frac{U(X)}{\sum_{X_i \in x} U(X_i)}, \quad P(E|X) = \frac{P(E \cap x)}{P(x)} \quad (7)$$

where E represents the information in the image x . The information entropy by the given image x is $H(E|x) = -\sum_{X_i \in x} P(E|x) \cdot \log P(E|x)$. As more proper vision

Table 1. Depth estimation results on KITTI and Hypersim.

Method	KITTI							Hypersim						
	bpp↓	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	AbsRel↓	RMSE↓	RMSE log↓	bpp↓	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	AbsRel↓	RMSE↓	RMSE log↓
Original	-	0.891	0.981	0.995	0.109	3.557	0.149	-	0.746	0.937	0.978	0.184	1.281	0.216
Directly Gen.	0.0063	0.255	0.538	0.720	0.756	15.906	0.611	0.0023	0.364	0.604	0.771	0.697	2.795	0.574
PIC	0.0011	0.381	0.628	0.773	0.644	14.371	0.581	0.0027	0.315	0.552	0.726	0.697	3.035	0.623
PICS	0.0235	0.703	0.877	0.948	0.208	7.263	0.292	0.0259	0.554	0.786	0.990	0.302	2.009	0.390
PerCo ₁₉	0.0037	0.619	0.816	0.910	0.300	8.631	0.361	0.0037	0.440	0.693	0.810	0.465	2.317	0.492
PerCo ₃₁₃	0.0329	0.808	0.943	0.979	0.153	5.468	0.215	0.0329	0.691	0.893	0.940	0.233	1.499	0.275
MS-ILLM ₂₀	0.0079	0.197	0.352	0.493	0.483	14.108	0.953	0.0049	0.276	0.501	0.663	0.484	2.936	0.674
MS-ILLM ₄₀	0.0124	0.506	0.699	0.799	0.283	10.594	0.529	0.0075	0.438	0.688	0.808	0.379	2.322	0.480
MS-ILLM ₃₅₀	0.0539	0.777	0.912	0.959	0.149	6.519	0.255	0.0327	0.650	0.891	0.944	0.223	1.522	0.270
Ours ($C = 1$)	0.0069	0.796	0.931	0.976	0.170	5.652	0.223	0.0026	0.615	0.837	0.907	0.288	1.763	0.334
Ours ($C = 2$)	0.0074	0.834	0.950	0.985	0.148	4.947	0.197	0.0028	0.627	0.847	0.907	0.254	1.667	0.304
Ours ($C = 4$)	0.0081	0.852	0.965	0.989	0.134	4.495	0.179	0.0032	0.663	0.896	0.934	0.227	1.492	0.269
Ours ($C = 8$)	0.0104	0.859	0.968	0.991	0.129	4.386	0.174	0.0043	0.668	0.892	0.952	0.225	1.447	0.261
Ours ($C = 16$)	0.0150	0.866	0.973	0.992	0.123	4.051	0.166	0.0064	0.711	0.912	0.959	0.211	1.375	0.242

analysis information an image contains, the larger its entropy value will be. So, $V(x) \propto H(x)$. According to the rate-distortion function [33], the compression rate $R = B(\hat{y}_{sel}, \mathcal{P})$ should be no less than the entropy of \hat{x} . Therefore, it can be formulated as the following optimization problem:

$$\begin{aligned} \min_{\hat{x}, \hat{y}_{sel}} \quad & \alpha(H(x) - H(\hat{x})) + \beta B(\hat{y}_{sel}, \mathcal{P}), \\ \text{s.t.} \quad & R \geq H(\hat{x}). \end{aligned} \quad (8)$$

As \hat{x} is obtained from denoising a sample from $\mathcal{N}(0, \mathbf{I})$, it follows with a normal distribution $\mathcal{N}(\mu, \Sigma)$. So, we can use the Lagrange multiplier method to find the solution even though it is not a convex problem. The constructed Lagrange function is,

$$\begin{aligned} L(\hat{x}, \hat{y}_{sel}, \lambda) = \alpha(H(x) - H(\hat{x})) + \beta B(\hat{y}_{sel}, \mathcal{P}) \\ + \lambda(H(\hat{x}) - R). \end{aligned} \quad (9)$$

And the theoretical optimal solution occurs when

$$\nabla_{\hat{x}, \hat{y}_{sel}, \lambda} L(\hat{x}, \hat{y}_{sel}, \lambda) = 0. \quad (10)$$

4. Experimental Results

In this section, we provide extensive experimental results to evaluate the proposed GSC method and ablation studies to understand its performance and evaluate its robustness.

4.1. Experimental Settings

(1) Datasets. For training the model, 20,000 images were constructed by randomly sampling 5,000 images from the training sets of KITTI [36], Flickr30k [43], COCO2017 [18], and iNaturalist [37], respectively, and combining them together. This enhances the diversity of the datasets and thus ensures the generalizability of the model. **(2) Implementation details.** Our model was implemented using PyTorch and trained on a single NVIDIA HGX H20-96G GPU. The number of multi-stream and single-stream

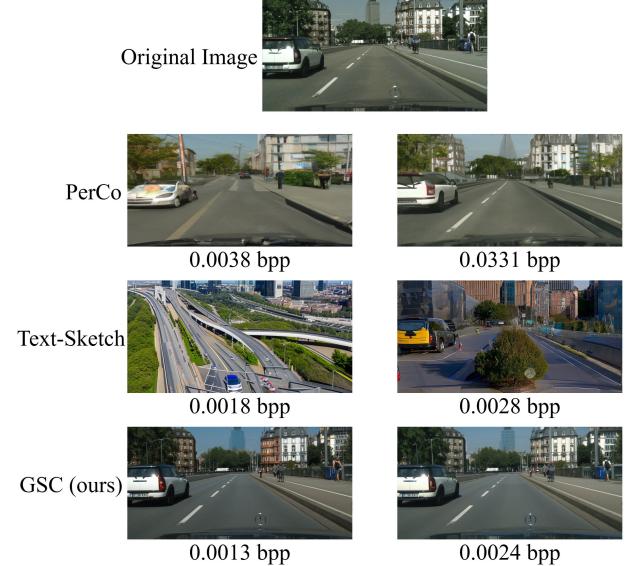


Figure 5. Qualitative results of CityScapes compared with other methods.

DiT blocks are set to $M = 4$ and $S = 2$, respectively. We trained the model for 15,000 steps using the AdamW optimizer with the learning rate and weight decay set to 4×10^{-5} and 0.01, respectively. The batch size is set to 1, and gradients are accumulated for 4 steps during the training. In our model, we trained 5 models with fixed channels of 1, 2, 4, 8, and 16. For getting textual descriptions of images, we use the Qwen2.5-vl-72b-Instruct [34].

4.2. Performance Comparisons

We compare our methods with other ultra-low bitrate methods, including Text-Sketch [16], Perco [3], and MS-ILLM [22]. These methods can still achieve the state-of-the-art (SOTA) when the bit rate is lower than 0.01 bpp.

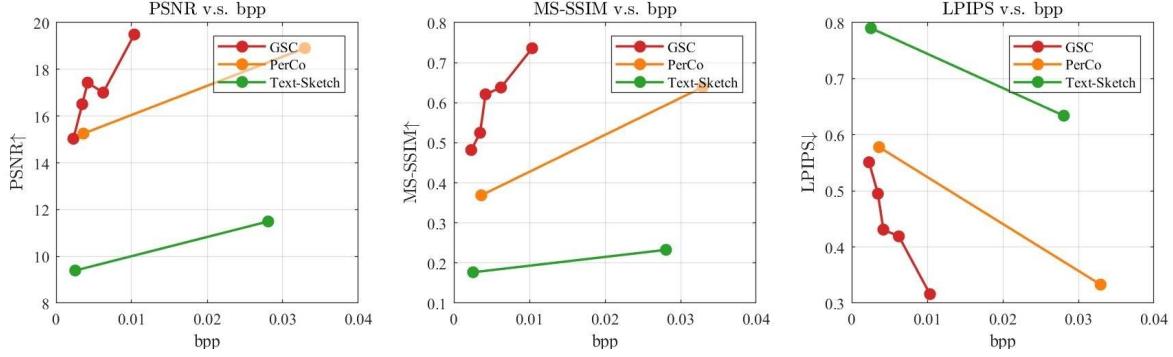


Figure 6. The rate-distortion performance comparison of different methods on the Kodak dataset.

Table 2. Pixel-level semantic segmentation result on the subset of CityScapes.

Method	bpp↓	aAcc↑	mIoU↑	mAcc↑
Original	-	96.370	82.280	88.600
Directly Gen.	0.0009	39.430	8.450	15.280
PIC	0.0020	36.200	6.450	11.860
PICS	0.0299	61.870	18.40	30.390
PerCo ₁₉	0.0037	67.260	21.700	31.170
PerCo ₃₁₃	0.0330	82.470	38.300	47.520
MS-ILLM ₂₀	0.0038	63.530	15.610	22.110
MS-ILLM ₄₀	0.0064	76.180	29.380	38.690
MS-ILLM ₃₅₀	0.0310	90.600	60.680	70.860
Ours ($C = 1$)	0.0011	85.250	47.840	58.360
Ours ($C = 2$)	0.0013	88.370	54.300	65.060
Ours ($C = 4$)	0.0015	90.540	61.090	71.510
Ours ($C = 8$)	0.0023	90.940	63.440	71.570
Ours ($C = 16$)	0.0039	93.440	70.730	78.090

Since our work focuses on the semantic coding at scenarios such as deep space exploration, with bitrates lower than 0.01 bpp, we assess the quality of reconstructed images by their downstream performance on fundamental vision tasks, and therefore adopt task-specific datasets. Specifically, we conduct evaluations across three vision tasks: depth estimation, semantic segmentation, and object detection. The goal is to evaluate whether the reconstructed images maintain sufficient information well needed for accurate vision analysis. In the following tables, the “Directly Gen.” means the result directly generated by FLUX [15] using only the prompt to guide the generation. The PIC and PICS are the methods from the paper [16]. The PerCo₁₉ and the PerCo₃₁₃ represent the pre-trained PerCo model [16] corresponding to 0.0019 bpp and 0.0313 bpp, respectively. The MS-ILLM₂₀, the MS-ILLM₄₀, and the MS-ILLM₃₅₀ represent the pre-trained MS-ILLM model corresponding to 0.0020 bpp, 0.0040 bpp, and 0.0350 bpp, respectively.

(1) Depth estimation. We evaluate the performance of depth estimation for the reconstructed images using the pre-trained Depth-Anything-V2-Large model of Depth Any-

Table 3. Object detection result on the subset of COCO2017.

Method	bpp↓	P↑	R↑	mAP50-95↑	mAP50↑	mAP75↑
Original	-	0.944	0.743	0.763	0.835	0.799
Directly Gen.	0.0037	0.481	0.350	0.092	0.327	0.022
PIC	0.0027	0.404	0.149	0.053	0.125	0.026
PICS	0.0221	0.815	0.656	0.584	0.708	0.631
PerCo ₁₉	0.0037	0.813	0.673	0.455	0.709	0.507
PerCo ₃₁₃	0.0329	0.900	0.720	0.697	0.785	0.760
MS-ILLM ₂₀	0.0086	0.405	0.217	0.123	0.179	0.132
MS-ILLM ₄₀	0.0122	0.406	0.350	0.273	0.344	0.274
MS-ILLM ₃₅₀	0.0496	0.857	0.536	0.600	0.679	0.635
Ours ($C = 1$)	0.0044	0.894	0.699	0.568	0.767	0.601
Ours ($C = 2$)	0.0051	0.838	0.710	0.623	0.765	0.685
Ours ($C = 4$)	0.0063	0.903	0.752	0.697	0.813	0.774
Ours ($C = 8$)	0.0094	0.882	0.739	0.701	0.817	0.746
Ours ($C = 16$)	0.0155	0.938	0.733	0.744	0.820	0.781

thing V2 [40] on the KITTI [36] depth validation set with the size 1216×352 . To demonstrate the generalization ability of our methods, we also test on an indoor scene dataset, that is Hypersim [30]. For evaluation metrics, $\delta_i = \text{percentage of } \max(d^*/d) < 1.25^i$, where $i = 1, 2, 3$, and d^* is the model prediction result and d is the ground truth. “AbsRel” represents the absolute relative error, given by $|d^* - d|/d$. “RMSE” is the root mean square error between the model prediction and the ground truth. “RMSE log” is the root mean square error of logarithms. As shown in Table 1, our method with $C = 1$ uses only 0.0069 bpp but achieves better performance than PICS using 0.0235 bpp and MS-ILLM₃₅₀ using 0.0539 bpp in the KITTI dataset. Furthermore, our method with $C = 2$ using 0.0074 bpp outperforms the PerCo₃₁₃ using 0.0329 bpp. On the Hypersim dataset, our method with $C = 8$ uses only 0.0043 bpp but achieves better performance than other methods except PerCo₃₁₃. And our methods with $C = 16$ use 0.0064 bpp to achieve better performance than PerCo₃₁₃ using 0.0313 bpp. **(2) Semantic segmentation.** We conduct semantic segmentation experiments on the Cityscapes [8] semantic segmentation validation set with the size 2048×1024 , using the Mask2Former [5] of open-mmsegmentation [7] with backbone Swin-L (in 22k). As shown in Table 2, our method

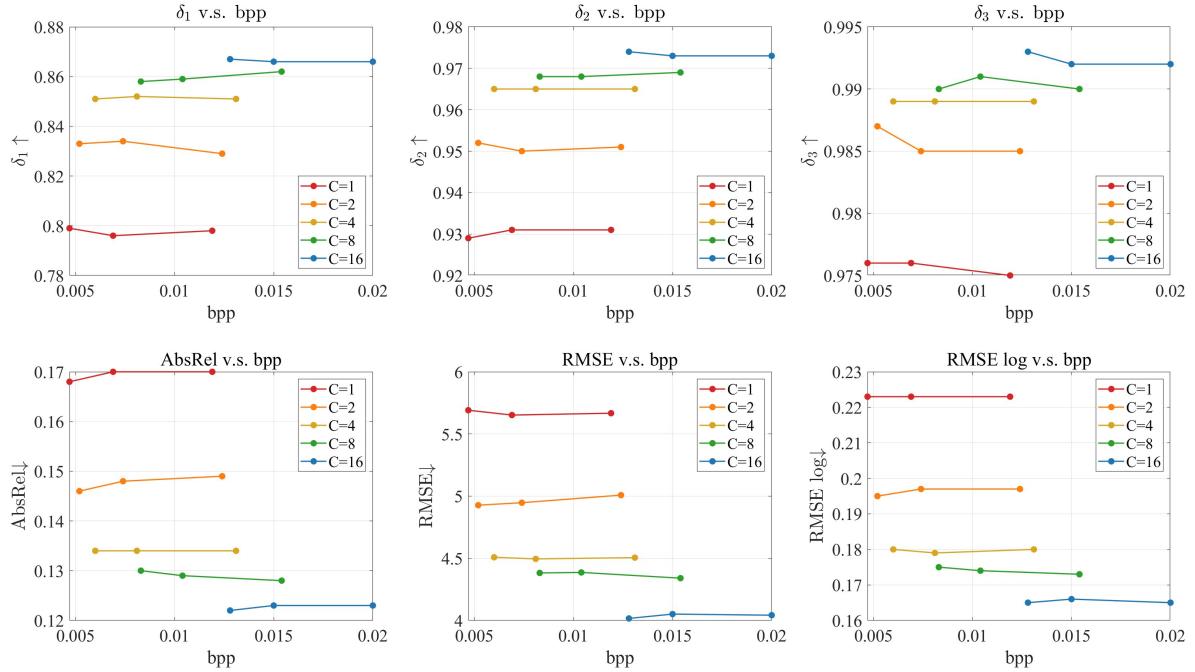


Figure 7. Ablations of the prompt with different lengths on the reconstruction quality in the depth estimation task using the KITTI sub test set.

with $C = 8$ uses only 0.0023 bpp, outperforming all the other methods with much higher bit rates. Although the PIC uses only 0.0020 bpp, its results are even worse than the results directly generated by FLUX [15], and our method with $C = 1$ uses only 0.0011 bpp to have better results. Figure 5 also demonstrates our method’s superior performance in preserving detailed structural information.

(3) Object detection. We evaluate object detection with the pre-trained YOLO11x of Ultralytics [14] on the COCO2017 [18] validation set. As shown in the Table. 3, our method with $C = 4$ uses less bpp but achieves better performance than other methods. Although PIC uses the least bpp among all methods, its performance is even worse than the performance of results directly generated by FLUX [15].

(4) Comparison on traditional compression performance. Our method not only performs very well on vision task-oriented image compression, but also achieves superior performance in conventional image compression. We conduct experiments on the Kodak [6]. Figure 6 shows the results of PSNR, MS-SSIM [39], and LPIPS [45], and our method achieves the best performance among all of them.

4.3. Ablation Studies

In the following, we provide detailed ablation studies to further understand our proposed method.

(1) Ablation studies on the number of channels. We change the selected structural guidance latent \hat{y}_{sel} in 1, 2, 4,

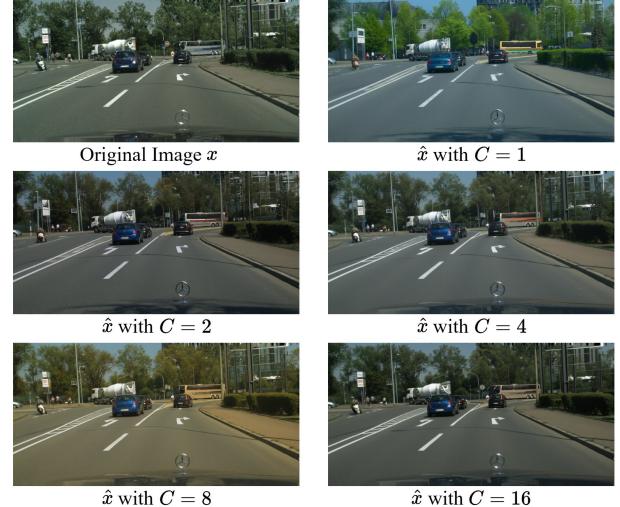


Figure 8. Reconstructed images with different numbers of channels.

8, 16 to examine the impact on compression performance. Figure 8 shows that the reconstructed image has more details aligning with the original one as more channels. Table 1, 2 and 3 also shows that more channels in the structural guidance latent usually lead to better performance, as more information has been used in the generation. How-

ever, these channels contain redundant and noisy information, so more channels don't always perform better than fewer channels, and adding more channels on top of one channel does not significantly improve the effect.

(2) Ablation studies on the length of the prompt. We conduct experiments with different lengths of prompts to evaluate the effect of prompts and the robustness of our method. Figure 7 shows the reconstruction quality in the depth estimation task on the KITTI sub test set. As longer prompts are used, higher bpp would be in the same number of channels, and it will give more detailed information about images. However, results show that our method achieves similar results using different lengths of the prompt, which means our method has great robustness.

In the Supplemental Materials, we have provided more experimental results to demonstrate the superior performance of our proposed GSC method.

5. Discussion

Semantic communication aims to interpret information at the semantic level and transmit representations that accurately convey the intended meaning, which is similar to the task in this paper. However, existing methods designed for semantic communication [4, 17, 46] primarily target bitrates above 0.1 bpp, making them unsuitable for the extremely low-bitrate scenarios considered in this paper.

6. Conclusion

We have developed Generative Semantic Coding (GSC), a new deep learning-based image compression method that uses multiple latent channels to guide the generation of images that preserve structural information as the original images while using less than 0.007 bpp. We developed new methods for constructing structural guidance and effectively utilizing it during the image generation process. This method will be very useful in scenarios where the communication channel conditions are very challenging and the bandwidth is very limited, however, both the sender and receiver have sufficient computational resources. Theoretical analysis is conducted to determine the lower bound of the compression. Future work includes eliminating redundant and noisy information in the latents to enhance compression and achieve a flexible balance between compression efficiency and visual analysis quality.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62331014) and Project 2021JC02X103. We acknowledge the computational support of the Center for Computational Science and Engineering at Southern University of Science and Technology.

References

- [1] Information technology – jpeg 2000 image coding system: Core coding system, 2000. 1
- [2] High efficiency video coding (hevc). <https://www.itu.int/rec/T-REC-H.265>, 2013. Accessed: 2024-01-10. 1
- [3] Marlene Careil, Matthew J. Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5
- [4] Yi-Hsin Chen, Ying-Chieh Weng, Chia-Hao Kao, Cheng Chien, Wei-Chen Chiu, and Wen-Hsiao Peng. Transtic: Transferring transformer-based image compression from human visualization to machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 8
- [5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 6
- [6] Eastman Kodak Company. Kodak image database. <https://r0k.us/graphics/kodak/>, 1994. Accessed: 2023-10-12. 7
- [7] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. <https://github.com/open-mmlab/mmsegmentation>, 2020. 6
- [8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6
- [9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021. 2
- [10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [12] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2
- [13] Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26088–26098, 2024. 2

- [14] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, 2023. 7
- [15] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 3, 4, 6, 7
- [16] 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. 2, 5, 6
- [17] Han Li, Shaohui Li, Shuangrui Ding, Wenrui Dai, Maida Cao, Chenglin Li, Junni Zou, and Hongkai Xiong. Image compression for machine and human vision with spatial-frequency adaptation. In *European Conference on Computer Vision*, 2024. 8
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 5, 7
- [19] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3
- [20] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems*, 33:11913–11924, 2020. 2
- [21] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 3
- [22] Matthew J Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, pages 25426–25443. PMLR, 2023. 2, 5
- [23] Zhihong Pan, Xin Zhou, and Hao Tian. Extreme generative image compression by learning text embedding from diffusion models. *arXiv preprint arXiv:2211.07793*, 2022. 2
- [24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4
- [25] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by re-description. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1505–1514, 2019. 2
- [26] 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*, pages 8748–8763. PMLR, 2021. 2
- [27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 4
- [28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2
- [29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2
- [30] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 6
- [31] 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*, pages 10684–10695, 2022. 2
- [32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2
- [33] Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142–163):1, 1959. 5
- [34] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 5
- [35] Lucas Theis, Tim Salimans, Matthew D Hoffman, and Fabian Mentzer. Lossy compression with gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022. 2
- [36] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 5, 6
- [37] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 5
- [38] Jeremy Vonderfecht and Feng Liu. Lossy compression with pretrained diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2
- [39] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirly-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. Ieee, 2003. 7
- [40] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jia Shi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 6
- [41] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36:64971–64995, 2023. 2

- [42] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3
- [43] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 5
- [44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3
- [45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7
- [46] Xu Zhang, Peiyao Guo, Ming Lu, and Zhan Ma. All-in-one image coding for joint human-machine vision with multi-path aggregation. In *Advances in Neural Information Processing Systems*, pages 71465–71503. Curran Associates, Inc., 2024. 8