Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng[1*] **Yunpeng Qi**[1*] **Jinming Liu**[2] **Yixin Gao**[1]
Xin Li[1†] **Xin Jin**[2] **Zhibo Chen**[1†]

[1]University of Science and Technology of China
[2]Eastern Institute of Technology, Ningbo

Abstract

Image compression methods are usually optimized isolatedly for human perception or machine analysis tasks. We reveal fundamental commonalities between these objectives: preserving accurate semantic information is paramount, as it directly dictates the integrity of critical information for intelligent tasks and aids human understanding. Concurrently, enhanced perceptual quality not only improves visual appeal but also, by ensuring realistic image distributions, benefits semantic feature extraction for machine tasks. Based on this insight, we propose *Diff-ICMH*, a generative image compression framework aiming for harmonizing machine and human vision in image compression. It ensures perceptual realism by leveraging generative priors and simultaneously guarantees semantic fidelity through the incorporation of *Semantic Consistency loss (SC loss)* during training. Additionally, we introduce the *Tag Guidance Module (TGM)* that leverages highly semantic image-level tags to stimulate the pre-trained diffusion model's generative capabilities, requiring minimal additional bit rates. Consequently, Diff-ICMH supports multiple intelligent tasks through a single codec and bitstream without any task-specific adaptation, while preserving high-quality visual experience for human perception. Extensive experimental results demonstrate Diff-ICMH's superiority and generalizability across diverse tasks, while maintaining visual appeal for human perception.

1 Introduction

The digital era has driven an explosive growth in network data. Compression algorithms are crucial in the storage and transmission. Image compression, a cornerstone of visual signal processing, is essential for both technological advancements and the operational efficiency of numerous digital systems. In terms of algorithmic development, traditional compression standards such as JPEG [1], JPEG2000 [2], H.264/AVC [3], H.265/HEVC [4], and H.266/VVC [5] have been widely applied. More recently, learned image compression methods [6–23] have emerged, demonstrating remarkable performance. However, these methods are primarily optimized for human visual perception. Concurrently, the rapid development of artificial intelligence technology means that the scale of visual data consumed by downstream intelligent tasks is growing substantially. This shift in application scenarios highlights the urgent need for developing compression methods tailored for intelligent tasks.

Representative existing approaches to image compression for machines (ICM) are depicted in Fig. 1 (a)-(c). Traditional codec-based methods optimize compression for machine tasks through quantization parameter tuning [24–29], strategic bit allocation [30–33], or by integrating neural network-based pre/post-processing modules [34–37]. While these approaches enable a single codec

[*]Equal contribution.
[†]Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

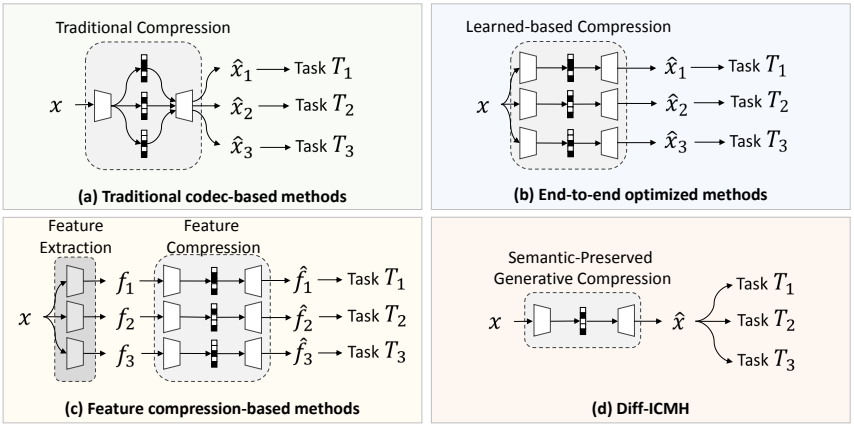


Figure 1: Comparison of compression and reconstruction workflows between different methods.

to support multiple tasks, their generalization and performance are often constrained due to the codec's inherent fidelity-oriented design and its non-differentiable nature. Task-driven end-to-end optimized methods [38–45] typically achieve superior performance on specific tasks but exhibit limited generalization at both the bitstream and codec levels. Feature compression-based methods [46–61] directly compress intermediate features from *specific* task models, often yielding a more favorable rate-distortion trade-off, while also reducing the computational burden on the server-side. However, such methods still face generalization challenges across different intelligent task models.

Optimizing image compression solely for specific intelligent tasks limits generalization across diverse tasks and human perception. To overcome this, we introduce Diff-ICMH (Fig. 1 (d)), a conditional generation approach targeting both versatile machine task support and human visual quality. From a novel perspective, we identify that performance degradation in intelligent tasks using compressed images primarily stems from two core information losses: (1) compromised *semantic integrity* (loss of core intelligible meaning), which directly impairs task analytics; and (2) reduced *perceptual realism* (textures and details deviating from natural distributions), leading to distribution mismatches that hinder feature extraction and cause inaccurate semantic analysis. Diff-ICMH tackles these issues by first employing a diffusion model-based generative framework, leveraging pre-trained models like Stable Diffusion [62], to ensure perceptual realism and mitigate domain shifts. Crucially, to preserve semantic integrity, we introduce Semantic Consistency loss (SC loss), which aligns features extracted by the pre-trained diffusion models. Furthermore, a Tag Guidance Module (TGM) utilizes efficiently coded, word-level image tags to activate generative priors, thereby enhancing both the subjective quality and semantic clarity of the reconstructed images.

The contributions of this paper are as follows:

- We introduce an innovative perspective for designing a versatile codec that jointly serves multiple intelligent tasks and human visual perception, identifying semantic fidelity and perceptual realism as critical determinants for this unification.
- Building on this insight, we introduce Diff-ICMH, which integrates generative priors with robust semantic information preservation, realized through proposed Semantic Consistency loss and Tag Guidance Module.
- Extensive experiments validate our approach, showcasing state-of-the-art performance across 10 diverse downstream intelligent tasks. All results are achieved without task-specific adaptation training, while concurrently delivering high-quality reconstructions for human visual perception.

2 Related work

2.1 Image compression for machines

Research in image compression for machines (ICM) has predominantly explored three primary directions. First, traditional codec-based methods adapt standardized formats like JPEG [1] and

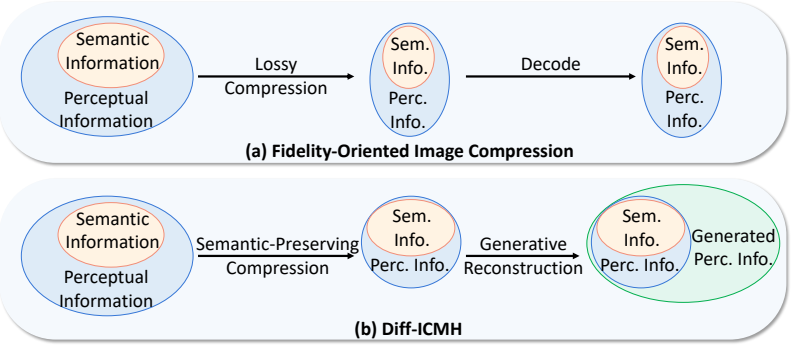


Figure 2: Illustration of the information transformation during compression processes in signal fidelity-oriented image compression and Diff-ICMH.

H.265 [4] by optimizing quantization parameters [24–29], employing strategic bit allocation [30–33], or applying neural network pre/post-processing [34–37]. While these methods offer the advantage of supporting multiple tasks with a single codec, their inherent signal fidelity-oriented design often limits their generalization capabilities and can lead to performance bottlenecks. Task-driven end-to-end optimized methods utilize learned codecs to directly optimize for specific machine tasks [38–43]. However, task-driven optimization often leads to poor generalization across different tasks. To mitigate this limitation, recent works [44, 45, 43] implement adaptation mechanisms aimed at enabling efficient task adaptation with minimal trainable parameters in the codec. Feature compression methods [46–61] directly encode intermediate neural network features instead of images, offering efficiency gains especially in cloud-edge scenarios, though they remain tightly coupled with specific feature extractors and cannot support human viewing.

2.2 Generative image compression

The basic objective of lossy compression is to optimize the trade-off between bit rate and quality. However, because traditional fidelity-oriented metrics like Peak Signal-to-Noise Ratio (PSNR) often correlate poorly with human visual perception, generative codecs have emerged as a promising direction for enhancing perceptual quality. Foundational theoretical work on rate-distortion-perception relationships [63–65] has catalyzed practical advances in this field. Two predominant approaches have gained prominence. First, GAN-based [66] methods [67–72] typically employ adversarial training to enhance perceptual quality. Second, diffusion model-based approaches [73–83] optimize for rate, distortion, and generation quality, either by training models from scratch or by leveraging large pre-trained models like Stable Diffusion [62]. Both approaches can be seen as using decoded features as conditional signals to guide perceptually realistic reconstructions. The ongoing advancement of generative models holds significant promise for their integration into image compression, paving the way for high-fidelity visual reconstructions at low bit rates.

3 Motivation

As shown in Figure 2 (a), traditional signal fidelity-oriented image compression methods [1–23] typically focus on indiscriminately minimizing pixel-wise errors, such as Mean Squared Error (MSE). This indiscriminate optimization, however, often incurs both semantic distortion and pixel-level perceptual mismatch. Semantic distortion directly impacts the information completeness crucial for downstream tasks. Furthermore, perceptual mismatch manifests as domain shifts, leading to error propagation during feature extraction and hindering the accurate mapping of inputs to semantic features. Generative image compression, by its very nature of reconstructing images to mimic natural distributions, is inherently well-suited to mitigating such domain shifts.

To investigate this, we compared feature divergence (1 minus cosine similarity) between fidelity-oriented codecs (VTM-18.2, ELIC) and a GAN-based generative one (MS-ILLM [72]). As illustrated in Figure 3, while fidelity-oriented codecs exhibited smaller feature differences at shallow network layers (stem layer) due to superior signal fidelity (higher PSNR), they suffered significantly larger divergence at deeper layers (layer2 and layer4). This confirms that the realistic textures produced

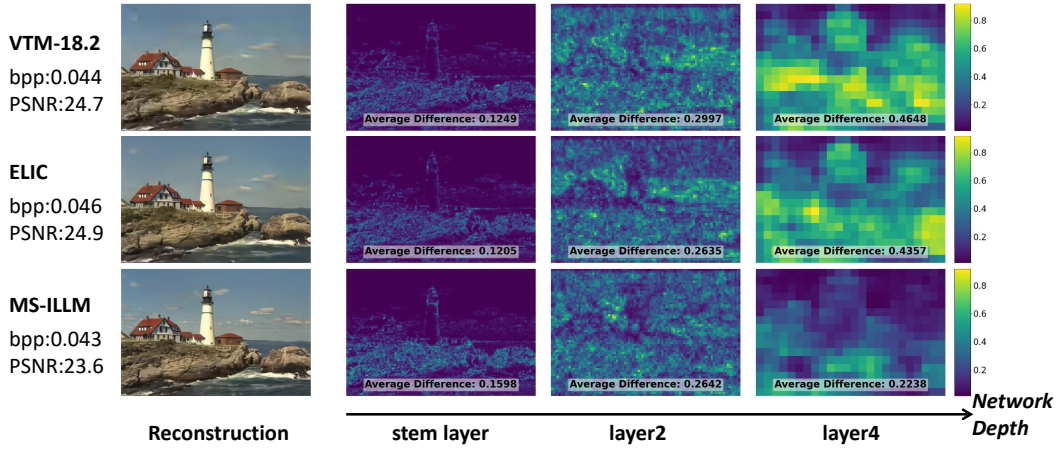


Figure 3: Variation in the difference (1 minus cosine similarity) between features extracted from reconstructed images by different codecs and those from original images when input to pre-trained ResNet50 [84], shown against increasing network depth.

by generative image compression effectively mitigate error accumulation arising from domain shift. However, realistic textures alone are insufficient for optimal intelligent task performance if semantic information integrity is compromised. Accurate extracted representations require *both* realistic textures and complete semantic information.

Diff-ICMH addresses these dual challenges based on a novel design philosophy: robust semantic information is prioritized and preserved during encoding, which then guides a generative reconstruction of perceptually realistic details during decoding, illustrated in Figure 2 (b). We realize this by developing a generative compression method leveraging a pre-trained Stable Diffusion model [62], further enhanced by our proposed Semantic Consistency loss (SC loss) and Tag Guidance Module (TGM). The subsequent sections will detail the overall framework of Diff-ICMH, along with the specifics of the SC loss and the TGM.

4 Diff-ICMH

4.1 Overall framework

The Diff-ICMH framework is depicted in Figure 4. The input image $\mathbf{x}$ is compressed and reconstructed as a latent feature $\hat{\mathbf{z}}$, targeting the VAE latent space of the pre-trained Stable Diffusion model. Concurrently, the tag extractor derives word-level tags $\mathbf{c}$ from $\mathbf{x}$ to capture coarse-grained semantics. In practical usage, the bitstream contains the compressed latent variables and the extracted tag IDs.

A key design choice in Diff-ICMH is the decoding of the bitstream directly into the VAE latent space of the pre-trained diffusion model instead of pixel space. This is motivated by several compelling properties of the latent space. It is inherently optimized for feature-level perceptual quality and realism, a result of its training objectives that include perceptual-oriented LPIPS loss [85] and adversarial losses [66]. Furthermore, this space provides a compact and perceptually rich representation (e.g., 8×8 spatial downsampling in Stable Diffusion) which effectively filters semantically irrelevant redundancy from the pixel domain. Consequently, optimizing for fidelity within this latent space enables the bitstream to prioritize perceptually salient and semantically coherent information, crucial for robust performance in downstream machine tasks.

Generative reconstruction is then conducted with $\hat{\mathbf{z}}$ as condition. Specifically, $\hat{\mathbf{z}}$ is fed to the control module, while the noisy latent $\mathbf{z}_t$ (at timestep t) is input to the diffusion model. The control module adapts ControlNet [86]-like architecture for conditioning on the downsampled latent feature $\hat{\mathbf{z}}$. Subsequently, generative reconstruction is performed using the control module and the pre-trained Stable Diffusion model. The reconstructed feature $\hat{\mathbf{z}}$ and the latent feature $\mathbf{z}_t$ at timestep t are input to the control module and diffusion model, respectively. The diffusion model then predicts the noise $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \hat{\mathbf{z}}, \mathbf{c}, t)$. Following the standard reverse diffusion process for T steps, we obtain the denoised

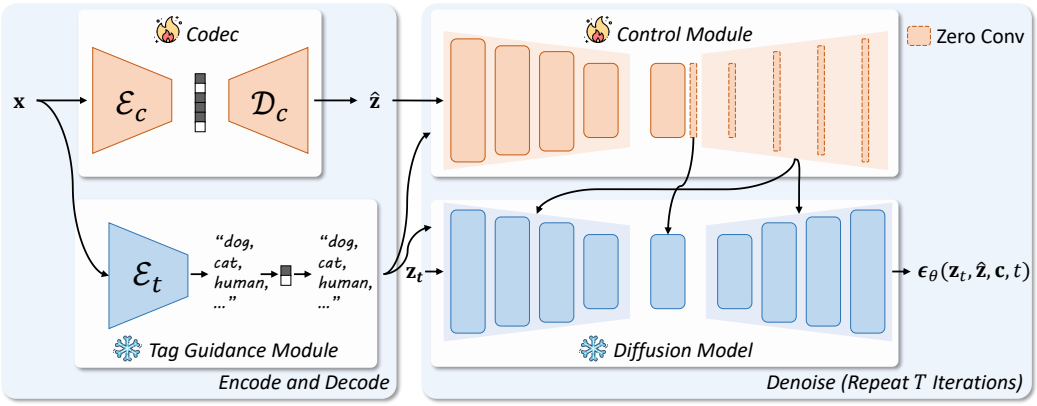


Figure 4: **Overview of Diff-ICMH.** Diff-ICMH consists of two parts: (left) image encoding/decoding and tag extraction; (right) condition-based image reconstruction. For simplicity, skip connections are omitted in the diagram.

latent $\mathbf{z}_0$. This $\mathbf{z}_0$ is then passed through the VAE decoder $\mathcal{D}(\cdot)$ to yield the final reconstructed image $\hat{\mathbf{x}}$. See Appendix for the detailed architecture of control module.

4.2 Semantic Consistency loss

To effectively preserve crucial semantic information during lossy coding, which is often compromised in traditional approaches, we propose the Semantic Consistency loss (SC loss). The core idea is to enforce semantic alignment between representations derived from the ground truth and the decoded signal. This is achieved by framing it as a pretext task: both signals are projected into a shared, high-quality semantic space where their representations are encouraged to be consistent.

The choice of this semantic space is critical for ensuring that the preserved semantics are both rich and generalizable for diverse downstream applications. Recent works [87–92] have demonstrated that large-scale pre-trained diffusion models possess strong inherent capabilities for image understanding and semantic feature extraction. Inspired by this, we leverage features extracted by these pre-trained diffusion models to instantiate our semantic space and guide the SC loss.

Specifically, as shown in Figure 5 (a), the ground truth latent variable $\mathbf{z} = \mathcal{E}(\mathbf{x})$ and the decoded feature $\hat{\mathbf{z}} = \mathcal{D}_c(\mathcal{E}_c(\mathbf{x}))$ are separately input into the pre-trained diffusion model, where $\mathcal{E}$ indicates the VAE's encoder and $\mathcal{E}_c$, $\mathcal{D}_c$ corresponds the the codec's encoder and decoder. We utilize the model's forward propagation $f(\cdot)$ as a bridge mapping from the original latent space to the semantic space, and align the two in this semantic space, optimizing the encoder-decoder through backpropagation.

This alignment is enforced by maximizing the similarity between the semantic representations $f(\mathbf{z})$ and $f(\hat{\mathbf{z}})$. The SC loss, $\mathcal{L}_{\text{sem}}$, is therefore formulated as:

$$\mathcal{L}_{\text{sem}} = -\mathbb{E}_{\mathbf{z},\hat{\mathbf{z}}}\left[\frac{1}{N}\sum_{n=1}^{N}\text{sim}(f(\mathbf{z})_n, f(\hat{\mathbf{z}})_n)\right], \tag{1}$$

where N represents the number of spatial positions in the feature, n indicates the spatial position index of the feature, and $\text{sim}(\cdot,\cdot)$ is a predefined similarity measurement function. In this paper, we instantiate the $\text{sim}(\cdot,\cdot)$ function using cosine similarity:

$$\text{sim}(\mathbf{z}, \hat{\mathbf{z}}) = \frac{\mathbf{z}^T\hat{\mathbf{z}}}{|\mathbf{z}|_2|\hat{\mathbf{z}}|_2}. \tag{2}$$

4.3 Tag Guidance Module

To activate generative priors in pre-trained diffusion models with minimal bitrate overhead, we introduce the Tag Guidance Module (TGM). As depicted in Figure 5 (b), a pre-trained tag extractor $\mathcal{E}_t$ (e.g., Recognize Anything [93] is used in this paper) first generates instance-level semantic tags for the

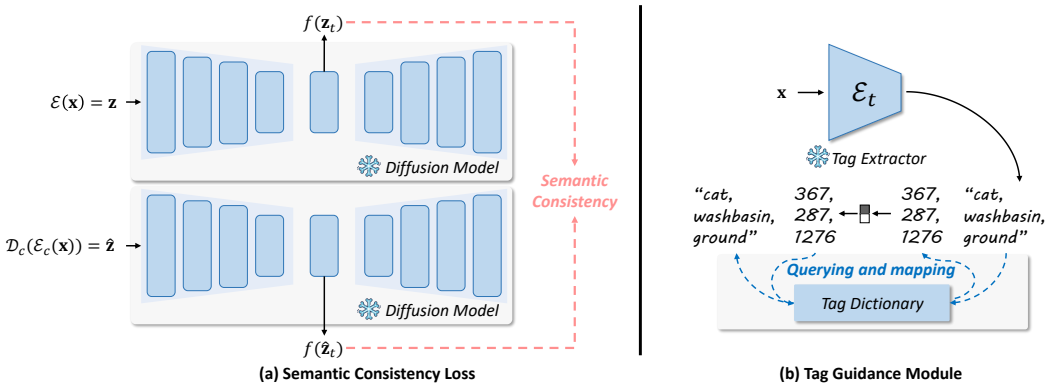


Figure 5: **(a) Semantic Consistency loss.** The latent variable $\mathbf{z}$ and the decoded latent variable $\hat{\mathbf{z}}$ are projected through the pre-trained diffusion model, resulting semantic representations that are then aligned. **(b) Tag Guidance Module.** Tags are extracted via tag extractor and losslessly compressed.

input image $\mathbf{x}$. These tags are subsequently mapped to numerical indices using a predefined dictionary and then losslessly encoded. At the decoder, these indices are converted back to their corresponding word-level tags via the dictionary. Finally, these tags are formatted as comma-separated text strings and input as conditioning to both the diffusion model and the control module.

This tag-based guidance mechanism incurs a very low bitrate overhead (approximately 100 bits per image) due to the typically small number of tags per image and a compact predefined dictionary. We therefore employ simple fixed-length coding for the tag indices. During the inference stage, Classifier-Free Guidance (CFG) [94] is utilized with these text tags to steer the diffusion model towards generating reconstructions with more distinct and accurate semantic content.

4.4 Loss function

Our final optimization objective, $\mathcal{L}_{\text{final}}$, combines four components: a rate loss $\mathcal{L}_{\text{rate}}$, a latent space reconstruction loss $\mathcal{L}_{\text{dist}}$, a diffusion model noise prediction loss $\mathcal{L}_{\text{diff}}$, and the SC loss $\mathcal{L}_{\text{sem}}$.

The rate loss $\mathcal{L}_{\text{rate}}$ penalizes the estimated entropy of the quantized primary latent variables $\hat{\mathbf{y}}$ and the quantized hyperprior latents $\hat{\mathbf{z}}_h$, following previous methods [7]:

$$\mathcal{L}_{\text{rate}} = \mathcal{R}(\hat{\mathbf{y}}) + \mathcal{R}(\hat{\mathbf{z}}_h). \tag{3}$$

During training, quantization is approximated by adding uniform noise [6] and we use the straight through estimator for the input of synthesizer.

The latent space reconstruction loss $\mathcal{L}_{\text{dist}}$ measures the distortion between the target VAE latent $\mathbf{z} = \mathcal{E}_{\text{VAE}}(\mathbf{x})$ and reconstructed latent $\hat{\mathbf{z}} = \mathcal{D}_c(\hat{\mathbf{y}})$. We define it as the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{dist}} = \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 = \|\mathcal{E}_{\text{VAE}}(\mathbf{x}) - \mathcal{D}_c(\hat{\mathbf{y}})\|_2^2, \tag{4}$$

where $\mathcal{E}_{\text{VAE}}$ is the encoder of the pre-trained VAE (associated with the diffusion model) and $\mathcal{D}_c$ is the codec's decoder operating on $\hat{\mathbf{y}}$.

The diffusion model's noise prediction loss $\mathcal{L}_{\text{diff}}$ follows the standard formulation:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z},t,\mathbf{c},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \hat{\mathbf{z}}, \mathbf{c}, t)\|_2^2\right], \tag{5}$$

where $\mathbf{z}$ is the clean VAE latent (serving as $\mathbf{z}_0$ for the diffusion process), $\mathbf{z}_t$ is its noised version at timestep t, $\boldsymbol{\epsilon}$ is the sampled Gaussian noise, $\boldsymbol{\epsilon}_\theta$ is the network's predicted noise, $\hat{\mathbf{z}}$ is the reconstructed latent from our codec acting as a condition, and $\mathbf{c}$ represents tags from TGM.

The final composite loss function is a weighted sum of these components:

$$\mathcal{L}_{\text{final}} = \lambda_{\text{rate}}\mathcal{L}_{\text{rate}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}, \tag{6}$$

where λ_{rate}, λ_{dist}, λ_{diff}, and λ_{sem} are scalar weights balancing each term.

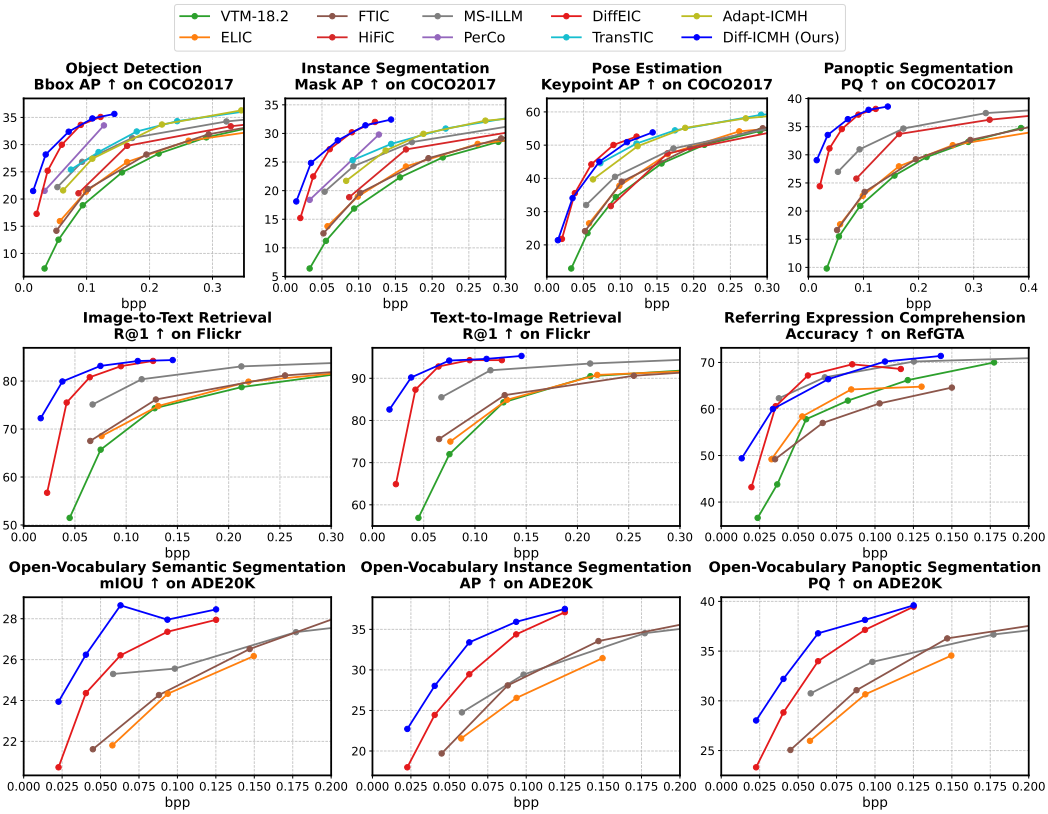


Figure 6: Performance comparison of Diff-ICMH with other methods across diverse intelligent tasks.

5 Experiments

5.1 Implementation details

Dataset. We use LSDIR [95] as our training dataset. During training, images from the dataset are randomly cropped to a size of 512×512.

Model setup. The experiments use Stable Diffusion 2.1 [1] as the pre-trained diffusion model. The control module is initialized in the same way as ControlNet [86], by copying parameters from the Stable Diffusion model for initialization, and initializing the weights of the Zero Convolution [86] to 0. During training, the pre-trained Stable Diffusion model and the tag extractor $\mathcal{E}_t$ remain frozen, while the encoder-decoder and control module parameters are learnable. For inference, we follow the standard DDPM process [96, 62] and start denoising from pure Gaussian noise.

Hyper-parameters. In equation (6), λ_{dist}, λ_{diff}, and λ_{sem} are empirically set to 1, 1, and 2 respectively, while λ_{rate} is set to $2, 4, 8, 16, 32$ to obtain codec models at different bit rates. We use the Adam optimizer [97] with β_1 and β_2 set to 0.9 and 0.999 respectively, and a training batch size of 16. Training is conducted in two stages. First, we train for 200K iterations with $\lambda_{\text{rate}} = 2$ and a learning rate of $1e-4$ to obtain a high bit rate model. Then, we fine-tune for another 200,000 iterations with a learning rate of $5e-5$ across all λ_{rate} values to obtain models for all bit rate points. For Classifier-Free Guidance (CFG), text tags are dropped with a probability of 0.1 during training. During inference, the CFG Scale is set to 5.0. Inference uses DDIM sampling [98] with 50 steps.

5.2 Evaluation protocol

We conduct a comprehensive evaluation of Diff-ICMH across two key dimensions: performance on a diverse set of downstream intelligent tasks and the perceptual quality of image reconstruction.

[1] https://github.com/Stability-AI/stablediffusion

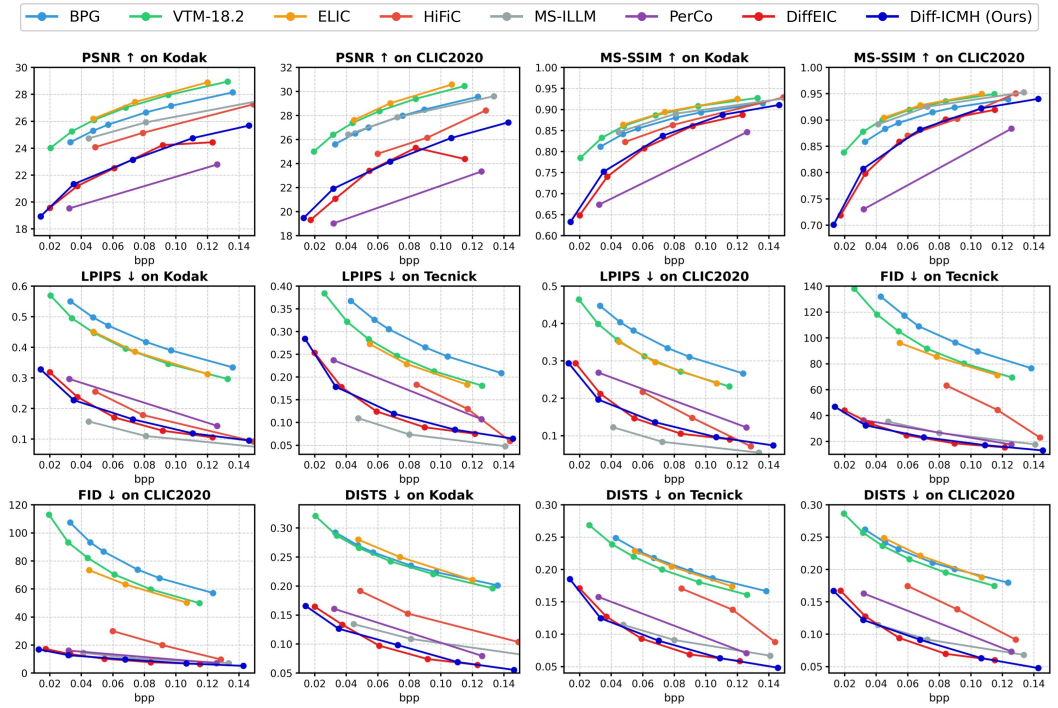


Figure 7: Comparison of Diff-ICMH and other methods on signal fidelity-oriented (PSNR↑, MS-SSIM↑) and perception-oriented (LPIPS↓, FID↓, DISTS↓) metrics.

Intelligent tasks. To evaluate the generalizability and effectiveness of our approach for machine vision, we conducted extensive experiments on a diverse range of intelligent tasks, encompassing a variety of task categories, model architectures, and backbone networks. These tasks includes traditional computer vision (e.g., detection and segmentation on COCO 2017 [99]), multimodal retrieval (Flickr30K [100]), and advanced Multimodal Large Language Model (MLLM) based understanding tasks (such as referring expression on RefGTA [101] and open-set segmentation on ADE20K [102]). Comprehensive details regarding these tasks are available in the Appendix.

Perceptual quality of reconstruction. Three public datasets are utilized: Kodak [103], Tecnick [104], and CLIC2020 [105]. We assess performance using metrics that measure signal fidelity (PSNR and MS-SSIM) as well as those that correlate with human perceptual quality (LPIPS, FID, and DISTS).

Compared methods for intelligent task support. We conduct comparative experiments with multiple high-performance codecs, including the powerful traditional image codec VVC [5](VTM-18.2), learned codecs ELIC [106] and FTIC [20], as well as codecs focusing on perceptual quality optimization, including HiFiC [71], PerCo [73, 107], MS-ILLM [72], DiffEIC [74], and task-specific optimized codecs TransTIC [44] and Adapter-ICMH [43] as comparative methods.

Compared methods for perceptual quality. We conduct comparisons with the traditional codec BPG, VVC [5] (VTM-18.2), the learned codec ELIC [106], and perceptual quality optimization codecs including HiFiC [71], PerCo [73, 107], MS-ILLM [72], and DiffEIC [74]. Numerical results of compared methods are obtained from the DiffEIC[2] repository.

5.3 Results

Multiple intelligent task supporting. Figure 6 illustrates Diff-ICMH's performance across diverse intelligent tasks on COCO, Flickr30K, RefGTA, and ADE20K datasets, compared against existing methods. Overall, Diff-ICMH consistently achieves state-of-the-art (SOTA) or highly competitive results. On COCO, it generally excels in object detection, instance, and panoptic segmentation, outperforming traditional, perception-oriented, and several task-specific codecs. For Flickr30K

[2] https://github.com/huai-chang/DiffEIC

cross-modal retrieval and ADE20K open-vocabulary segmentation tasks, Diff-ICMH demonstrates significant advantages, particularly at very low bitrates (e.g., 0.01-0.05 bpp on Flickr, 0.02-0.1 bpp on ADE20K), underscoring its effective semantic preservation capabilities attributed to our SC loss and TGM. This highlights strong adaptability to both traditional vision and advanced MLLM-based understanding tasks. While generally leading, performance is comparable to some methods like DiffEIC in specific scenarios, such as COCO pose estimation or RefGTA referring expression, potentially due to inherent VAE latent space limitations for fine details or domain shift from synthetic data, respectively. Despite these minor trade-offs, the extensive evaluations confirm Diff-ICMH's broad generalizability and excellent overall rate-distortion performance across a multitude of tasks without requiring any task-specific fine-tuning, emphasizing its practical value.

Perception-oriented reconstruction. Figure 7 presents the compared results of reconstruction quality. Consistent with its generative nature, Diff-ICMH's PSNR/MS-SSIM scores trail behind traditional fidelity-optimized codecs, a common characteristic when prioritizing perceptual realism. Besides, it is comparable to DiffEIC and outperforms PerCo in these fidelity metrics. Conversely, Diff-ICMH demonstrates substantial advantages in perceptual quality. It significantly surpasses fidelity-focused codecs (e.g., BPG, VTM-18.2) and other perception-oriented methods (e.g., HiFiC, PerCo) across LPIPS, FID, and DISTS. Notably, Diff-ICMH achieves SOTA performance in FID and DISTS scores among all compared methods, exhibiting particular strength at extremely low bitrates. These perceptual gains, especially at low BPP, are attributed to the effectiveness of our SC loss and TGM in preserving semantic integrity and guiding realistic reconstruction.

Additional results. For brevity, *visualizations*, *feature difference analysis*, and *computational complexity analysis* are provided in Section B.2, B.3, B.4.

5.4 Ablation study

We conducted ablation studies to validate the effectiveness of the proposed Semantic Consistency loss (SC loss) and Tag Guidance Module (TGM), along with an analysis of the SC loss setup. Models are trained on the LSDIR dataset and evaluated using object detection on COCO 2017. More details are provided in the Appendix.

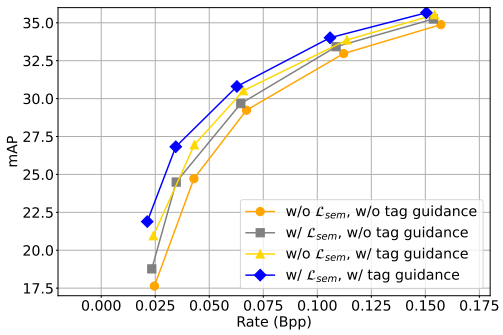
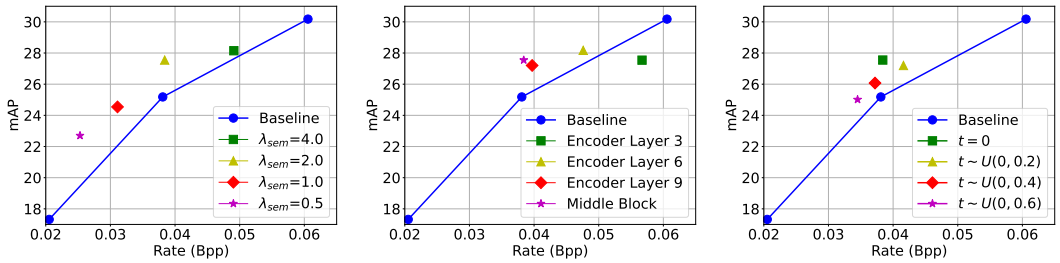


Figure 9: Ablation of Semantic Consistency loss and Tag Guidance Module.

Effectiveness of SC loss and TGM. Figure 9 presents ablation results for the SC loss and TGM. Both modules significantly improve the task performance, with their combination (blue curve) yielding the best performance (e.g., around 4 mAP gain over the baseline at approximately 0.025 Bpp). This confirms their individual contributions and synergistic effect. These results validate that SC loss maintains semantic similarity while TGM activates generative priors, complementarily boosting performance.

Setup of SC loss. Figure 8 details ablations for SC loss hyperparameters.

- **Weight λ_{sem}.** As shown in Figure 8 (a), $\lambda_{\text{sem}} = 2.0$ achieves the optimal balance, effectively preserving semantic information while maintaining low bitrate.
- **Position of SC loss application.** Applying SC loss to deeper features within the U-Net yields better rate-distortion performance (Figure 8 (b)). The middle block is found to be optimal, as features at this depth better capture abstract semantic content compared to shallower layers.
- **Input noise level for SC loss.** We compared using clean latent features $(\mathbf{z}, \hat{\mathbf{z}})$ versus noised versions $(\mathbf{z}_t, \hat{\mathbf{z}}_t)$ as inputs to the semantic feature extractor $f(\cdot)$ for the SC loss. Figure 8 (c) demonstrates that noise-free inputs ($t = 0$) result in the optimal rate-distortion performance. This suggests that while the diffusion model is trained with noise, its capability for semantic feature extraction (for our SC loss) is maximized with clean signals. Notably, all tested noise levels still improved performance over the baseline, confirming the robustness of the SC loss.



(a) Weight λ_{sem} of SC loss. (b) Position of SC loss application. (c) Noise level t for SC loss inputs.

Figure 8: **Ablation of hyper-parameters of SC loss.** "Baseline" indicates not including SC loss.

6 Limitation

The primary limitation of Diff-ICMH is computational complexity, as discussed in Section B.4. Diff-ICMH requires relatively significant decoding time due to the iterative nature of the diffusion denoising process, where each step necessitates a complete forward pass through the neural network. However, we emphasize that the primary focus of this paper is to demonstrate the generalizability and significant potential of Diff-ICMH for harmonizing diverse intelligent tasks with human perceptual quality. Therefore, computational complexity analysis is not the central emphasis of our current work. In future work, incorporating efficient sampling methods [108–110] and distillation techniques [111–113] to reduce sampling steps would make Diff-ICMH more practical for real-world applications.

7 Conclusion

In this paper, we introduced Diff-ICMH, a verstile generative image codec designed to effectively serve both intelligent tasks and human visual perception. The core design philosophy of Diff-ICMH departs from isolated optimization by revealing and leveraging fundamental commonalities between these two objectives: Preserving semantic information integrity is paramount for both machine analysis and human understanding. Concurrently, enhanced perceptual quality, which is built upon this semantic foundation, not only improves the visual experience but also benefits machine feature extraction. Diff-ICMH embodies this by integrating a diffusion model-based generative framework with a novel Semantic Consistency loss and an efficient Tag Guidance Module. Comprehensive evaluations demonstrate the effectiveness and generalization of our proposed method, which efficiently supports diverse tasks with a single codec and no task-specific adaptation, alongside excellent visual quality. Our work thus presents a promising pathway towards truly versatile image compression.

Acknowledgements This work was supported in part by NSFC under Grant 62371434, 623B2098, and 62021001, the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20252293, and the China Postdoctoral Science Foundation-Anhui Joint Support Program under Grant Number 2024T017AH. This work was also funded by Anhui Postdoctoral Scientific Research Program Foundation (No.2025A1015).

References

[1] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[2] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The jpeg2000 still image coding system: an overview," *IEEE Transactions on Consumer Electronics*, vol. 46, no. 4, pp. 1103–1127, 2000.

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

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

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

[6] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations*, 2017.

[7] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

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

[9] Z. Guo, Z. Zhang, R. Feng, and Z. Chen, "Causal contextual prediction for learned image compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2329–2341, 2021.

[10] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 771–14 780.

[11] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *IEEE International Conference on Image Processing*. IEEE, 2020, pp. 3339–3343.

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

[13] Y. Xie, K. L. Cheng, and Q. Chen, "Enhanced invertible encoding for learned image compression," in *ACM International Conference on Multimedia*, 2021, pp. 162–170.

[14] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 388–14 397.

[15] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2022.

[16] R. Zou, C. Song, and Z. Zhang, "The devil is in the details: Window-based attention for image compression," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 492–17 501.

[17] Y. Qian, X. Sun, M. Lin, Z. Tan, and R. Jin, "Entroformer: A transformer-based entropy model for learned image compression," in *International Conference on Learning Representations*, 2022.

[18] A. B. Koyuncu, H. Gao, A. Boev, G. Gaikov, E. Alshina, and E. Steinbach, "Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 447–463.

[19] J.-H. Kim, B. Heo, and J.-S. Lee, "Joint global and local hierarchical priors for learned image compression," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5992–6001.

[20] H. Li, S. Li, W. Dai, C. Li, J. Zou, and H. Xiong, "Frequency-aware transformer for learned image compression," in *International Conference on Learning Representations*, 2024.

[21] F. Zeng, H. Tang, Y. Shao, S. Chen, L. Shao, and Y. Wang, "Mambaic: State space models for high-performance learned image compression," *arXiv preprint arXiv:2503.12461*, 2025.

[22] S. Qin, J. Wang, Y. Zhou, B. Chen, T. Luo, B. An, T. Dai, S. Xia, and Y. Wang, "Mambavc: Learned visual compression with selective state spaces," *arXiv preprint arXiv:2405.15413*, 2024.

[23] S. Iwai, T. Miyazaki, and S. 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.

[24] Z. Liu, T. Liu, W. Wen, L. Jiang, J. Xu, Y. Wang, and G. Quan, "Deepn-jpeg: A deep neural network favorable jpeg-based image compression framework," in *Annual Design Automation Conference*, 2018, pp. 1–6.

[25] L. D. Chamain, S.-c. S. Cheung, and Z. Ding, "Quannet: Joint image compression and classification over channels with limited bandwidth," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2019, pp. 338–343.

[26] Z. D. S. Li, "Optimizing jpeg quantization for classification networks," *Resource-Constrained Machine Learning*, 2020.

[27] J. Choi and B. Han, "Task-aware quantization network for jpeg image compression," in *European Conference on Computer Vision*. Springer, 2020, pp. 309–324.

[28] L.-Y. Duan, X. Liu, J. Chen, T. Huang, and W. Gao, "Optimizing jpeg quantization table for low bit rate mobile visual search," in *Visual Communications and Image Processing*. IEEE, 2012, pp. 1–6.

[29] X. Luo, H. Talebi, F. Yang, M. Elad, and P. Milanfar, "The rate-distortion-accuracy tradeoff: Jpeg case study," 2020. [Online]. Available: https://arxiv.org/abs/2008.00605

[30] Q. Cai, Z. Chen, D. O. Wu, S. Liu, and X. Li, "A novel video coding strategy in hevc for object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4924–4937, 2021.

[31] Z. Huang, C. Jia, S. Wang, and S. Ma, "Visual analysis motivated rate-distortion model for image coding," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2021, pp. 1–6.

[32] X. Li, J. Shi, and Z. Chen, "Task-driven semantic coding via reinforcement learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 6307–6320, 2021.

[33] G. Xie, X. Li, S. Lin, Z. Chen, L. Zhang, K. Zhang, and Y. Li, "Hierarchical reinforcement learning based video semantic coding for segmentation," in *IEEE International Conference on Visual Communications and Image Processing*. IEEE, 2022, pp. 1–5.

[34] S. Suzuki, M. Takagi, K. Hayase, T. Onishi, and A. Shimizu, "Image pre-transformation for recognition-aware image compression," in *IEEE International Conference on Image Processing*. IEEE, 2019, pp. 2686–2690.

[35] G. Lu, X. Ge, T. Zhong, Q. Hu, and J. Geng, "Preprocessing enhanced image compression for machine vision," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[36] H. Zhang, J. I. Ahonen, N. Le, R. Yang, and F. Cricri, "Competitive learning for achieving content-specific filters in video coding for machines," in *IEEE International Conference on Image Processing*. IEEE, 2024, pp. 1877–1882.

[37] M. Yang, F. Yang, L. Murn, M. G. Blanch, J. Sock, S. Wan, F. Yang, and L. Herranz, "Task-switchable pre-processor for image compression for multiple machine vision tasks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6416–6429, 2024.

[38] S. Luo, Y. Yang, Y. Yin, C. Shen, Y. Zhao, and M. Song, "Deepsic: Deep semantic image compression," in *International Conference on Neural Information Processing*, 2018, pp. 96–106.

[39] N. Patwa, N. Ahuja, S. Somayazulu, O. Tickoo, S. Varadarajan, and S. Koolagudi, "Semantic-preserving image compression," in *IEEE International Conference on Image Processing*. IEEE, 2020, pp. 1281–1285.

[40] K. Fischer, F. Brand, and A. Kaup, "Boosting neural image compression for machines using latent space masking," 2021. [Online]. Available: https://arxiv.org/abs/2112.08168

[41] L. D. Chamain, F. Racapé, J. Bégaint, A. Pushparaja, and S. Feltman, "End-to-end optimized image compression for machines, a study," in *Data Compression Conference*. IEEE, 2021, pp. 163–172.

[42] N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, and E. Rahtu, "Image coding for machines: an end-to-end learned approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 1590–1594.

[43] H. Li, S. Li, S. Ding, W. Dai, M. Cao, C. Li, J. Zou, and H. Xiong, "Image compression for machine and human vision with spatial-frequency adaptation," in *European Conference on Computer Vision*. Springer, 2024, pp. 382–399.

[44] Y.-H. Chen, Y.-C. Weng, C.-H. Kao, C. Chien, W.-C. Chiu, and W.-H. Peng, "Transtic: Transferring transformer-based image compression from human perception to machine perception," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 297–23 307.

[45] J. Liu, X. Jin, R. Feng, Z. Chen, and W. Zeng, "Composable image coding for machine via task-oriented internal adaptor and external prior," in *IEEE International Conference on Visual Communications and Image Processing*. IEEE, 2023, pp. 1–5.

[46] H. Choi and I. V. Bajić, "Deep feature compression for collaborative object detection," in *IEEE International Conference on Image Processing*. IEEE, 2018, pp. 3743–3747.

[47] Z. Chen, K. Fan, S. Wang, L. Duan, W. Lin, and A. C. Kot, "Toward intelligent sensing: Intermediate deep feature compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 2230–2243, 2019.

[48] A. E. Eshratifar, A. Esmaili, and M. Pedram, "Bottlenet: A deep learning architecture for intelligent mobile cloud computing services," in *IEEE/ACM International Symposium on Low Power Electronics and Design*. IEEE, 2019, pp. 1–6.

[49] S. Suzuki, M. Takagi, S. Takeda, R. Tanida, and H. Kimata, "Deep feature compression with spatio-temporal arranging for collaborative intelligence," in *IEEE International Conference on Image Processing*. IEEE, 2020, pp. 3099–3103.

[50] S. Suzuki, S. Takeda, M. Takagi, R. Tanida, H. Kimata, and H. Shouno, "Deep feature compression using spatio-temporal arrangement toward collaborative intelligent world," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3934–3946, 2021.

[51] J. Shao and J. Zhang, "Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems," in *IEEE International Conference on Communications Workshops*. IEEE, 2020, pp. 1–6.

[52] S. Singh, S. Abu-El-Haija, N. Johnston, J. Ballé, A. Shrivastava, and G. Toderici, "End-to-end learning of compressible features," in *IEEE International Conference on Image Processing*. IEEE, 2020, pp. 3349–3353.

[53] S. R. Alvar and I. V. Bajić, "Bit allocation for multi-task collaborative intelligence," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 4342–4346.

[54] ——, "Pareto-optimal bit allocation for collaborative intelligence," *IEEE Transactions on Image Processing*, vol. 30, pp. 3348–3361, 2021.

[55] R. Henzel, K. Misra, and T. Ji, "Efficient feature compression for the object tracking task," in *IEEE International Conference on Image Processing*. IEEE, 2022, pp. 3505–3509.

[56] H. Tu, L. Li, W. Zhou, and H. Li, "Reconstruction-free image compression for machine vision via knowledge transfer," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 10, pp. 1–19, 2024.

[57] N. Ahuja, P. Datta, B. Kanzariya, V. S. Somayazulu, and O. Tickoo, "Neural rate estimator and unsupervised learning for efficient distributed image analytics in split-dnn models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2022–2030.

[58] Y. Kim, H. Jeong, J. Yu, Y. Kim, J. Lee, S. Y. Jeong, and H. Y. Kim, "End-to-end learnable multi-scale feature compression for vcm," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[59] D. Nazir, T. Bartels, J. Piewek, T. Bagdonat, and T. Fingscheidt, "Distributed semantic segmentation with efficient joint source and task decoding," in *European Conference on Computer Vision*. Springer, 2024, pp. 195–212.

[60] L. Xiong, X. Luo, Z. Wang, C. He, S. Zhu, and B. Zeng, "Texture-guided coding for deep features," 2024. [Online]. Available: https://arxiv.org/abs/2405.19669

[61] S. Guo, Z. Chen, Y. Zhao, N. Zhang, X. Li, and L. Duan, "Toward scalable image feature compression: a content-adaptive and diffusion-based approach," in *ACM International Conference on Multimedia*, 2023, pp. 1431–1442.

[62] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

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

[64] Z. Yan, F. Wen, R. Ying, C. Ma, and P. Liu, "On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 682–11 692.

[65] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 517–11 529, 2021.

[66] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[67] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, "Extreme learned image compression with gans," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2587–2590.

[68] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 221–231.

[69] S. Iwai, T. Miyazaki, Y. Sugaya, and S. Omachi, "Fidelity-controllable extreme image compression with generative adversarial networks," in *International Conference on Pattern Recognition*. IEEE, 2021, pp. 8235–8242.

[70] L. Wu, K. Huang, and H. Shen, "A gan-based tunable image compression system," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2334–2342.

[71] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 913–11 924, 2020.

[72] M. J. Muckley, A. El-Nouby, K. Ullrich, H. Jégou, and J. Verbeek, "Improving statistical fidelity for neural image compression with implicit local likelihood models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 25 426–25 443.

[73] M. Careil, M. J. Muckley, J. Verbeek, and S. Lathuilière, "Towards image compression with perfect realism at ultra-low bitrates," in *International Conference on Learning Representations*, 2023.

[74] Z. Li, Y. Zhou, H. Wei, C. Ge, and J. Jiang, "Towards extreme image compression with latent feature guidance and diffusion prior," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[75] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 971–64 995, 2023.

[76] L. Relic, R. Azevedo, M. Gross, and C. Schroers, "Lossy image compression with foundation diffusion models," in *European Conference on Computer Vision*. Springer, 2024, pp. 303–319.

[77] L. Theis, T. Salimans, M. D. Hoffman, and F. Mentzer, "Lossy compression with gaussian diffusion," 2022. [Online]. Available: https://arxiv.org/abs/2206.08889

[78] Z. Pan, X. Zhou, and H. Tian, "Extreme generative image compression by learning text embedding from diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2211.07793

[79] T. Xu, Z. Zhu, D. He, Y. Li, L. Guo, Y. Wang, Z. Wang, H. Qin, Y. Wang, J. Liu *et al.*, "Idempotence and perceptual image compression," in *International Conference on Learning Representations*, 2024.

[80] E. Hoogeboom, E. Agustsson, F. Mentzer, L. Versari, G. Toderici, and L. Theis, "High-fidelity image compression with score-based generative models," 2023. [Online]. Available: https://arxiv.org/abs/2305.18231

[81] N. F. Ghouse, J. Petersen, A. Wiggers, T. Xu, and G. Sautiere, "A residual diffusion model for high perceptual quality codec augmentation," 2023. [Online]. Available: https://arxiv.org/abs/2301.05489

[82] Y. Ma, W. Yang, and J. Liu, "Correcting diffusion-based perceptual image compression with privileged end-to-end decoder," in *International Conference on Machine Learning*. PMLR, 2024, pp. 34 075–34 093.

[83] H. Kuang, Y. Ma, W. Yang, Z. Guo, and J. Liu, "Consistency guided diffusion model with neural syntax for perceptual image compression," in *ACM International Conference on Multimedia*, 2024, pp. 1622–1631.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[85] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[86] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[87] S. Mukhopadhyay, M. Gwilliam, V. Agarwal, N. Padmanabhan, A. Swaminathan, S. Hegde, T. Zhou, and A. Shrivastava, "Diffusion models beat gans on image classification," 2023. [Online]. Available: https://arxiv.org/abs/2307.08702

[88] J. Zhang, C. Herrmann, J. Hur, L. Polania Cabrera, V. Jampani, D. Sun, and M.-H. Yang, "A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence," *Advances in Neural Information Processing Systems*, vol. 36, pp. 45 533–45 547, 2023.

[89] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie, "Representation alignment for generation: Training diffusion transformers is easier than you think," 2024. [Online]. Available: https://arxiv.org/abs/2410.06940

[90] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2206–2217.

[91] W. Xiang, H. Yang, D. Huang, and Y. Wang, "Denoising diffusion autoencoders are unified self-supervised learners," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 802–15 812.

[92] X. Chen, Z. Liu, S. Xie, and K. He, "Deconstructing denoising diffusion models for self-supervised learning," 2024. [Online]. Available: https://arxiv.org/abs/2401.14404

[93] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu *et al.*, "Recognize anything: A strong image tagging model," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1724–1732.

[94] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Advances in Neural Information Processing Systems Workshop on Deep Generative Models and Downstream Applications*, 2021.

[95] Y. Li, K. Zhang, J. Liang, J. Cao, C. Liu, R. Gong, Y. Zhang, H. Tang, Y. Liu, D. Demandolx *et al.*, "Lsdir: A large scale dataset for image restoration," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1775–1787.

[96] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[97] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[98] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[99] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[100] A. B. Plummer, L. Wang, M. C. Cervantes, C. J. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *International Journal of Computer Vision*, pp. 74–93, 2017.

[101] M. Tanaka, T. Itamochi, K. Narioka, I. Sato, Y. Ushiku, and T. Harada, "Generating easy-to-understand referring expressions for target identifications," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5794–5803.

[102] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[103] E. Kodak, "Kodak lossless true color image suite (photocd pcd0992)," http://r0k.us/graphics/kodak, 1993.

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

[105] G. Toderici, L. Theis, N. Johnston, E. Agustsson, F. Mentzer, J. Ballé, W. Shi, and R. Timofte, "Clic 2020: Challenge on learned image compression," *Retrieved March*, vol. 29, p. 2021, 2020.

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

[107] N. Körber, E. Kromer, A. Siebert, S. Hauke, D. Mueller-Gritschneder, and B. Schuller, "Perco (SD): Open perceptual compression," in *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024. [Online]. Available: https://openreview.net/forum?id=8xvygfdRWy

[108] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.

[109] ——, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *Machine Intelligence Research*, pp. 1–22, 2025.

[110] K. Zheng, C. Lu, J. Chen, and J. Zhu, "Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics," *Advances in Neural Information Processing Systems*, vol. 36, pp. 55 502–55 542, 2023.

[111] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *International Conference on Learning Representations*, 2023.

[112] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–103.

[113] A. Sauer, F. Boesel, T. Dockhorn, A. Blattmann, P. Esser, and R. Rombach, "Fast high-resolution image synthesis with latent adversarial diffusion distillation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[114] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[115] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som *et al.*, "Image as a foreign language: Beit pretraining for vision and vision-language tasks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 175–19 186.

[116] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 897–32 912, 2022.

[117] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[118] Y. Yuan, W. Li, J. Liu, D. Tang, X. Luo, C. Qin, L. Zhang, and J. Zhu, "Osprey: Pixel understanding with visual instruction tuning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 202–28 211.

[119] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

Appendix

Contents

A Implementation Details

A.1 Architecture of the control module

The control module described in Section 4.1 is modified from original ControlNet [86]. Key modifications include: (i) Replacing ControlNet's initial three stride-2 convolutions with a single stride-1 convolution to establish dimensional compatibility with $\hat{\mathbf{z}}$. (ii) Implementing bilateral feature injection through zero-convolutions from both encoder and decoder pathways of the diffusion model (unlike ControlNet's decoder-only approach) to establish more precise control mechanisms from $\hat{\mathbf{z}}$. Additionally, the extracted word-level tags $\mathbf{c}$ are strategically injected as text prompts into both the control module and the diffusion model to effectively activate relevant generative priors.

A.2 Implementation in ablation studies

Basic setup. As described in Section 5.4, we designed ablation experiments to verify the effectiveness of the SC loss and tag guidance module, along with the hyper-parameter configuration of SC loss. In the experimental setup, training was conducted on the LSDIR dataset with a batch size of 8, a total of $80,000$ iterations, and a learning rate of $1e-5$. The experimental results were rigorously evaluated on the COCO 2017 object detection benchmark dataset.

Semantic Consisitency loss with noisy input. In Section 4.2, for SC loss calculation, clean latent features $\mathbf{z}$ and $\hat{\mathbf{z}}$ are utilized as inputs to the diffusion model. Recognizing that diffusion models are typically trained on noisy signals, we conducted a series of comparative experiments (illustrated in Figure 10) to investigate the implications of using noisy latent features $\mathbf{z}_t$ and $\hat{\mathbf{z}}_t$ as inputs, corresponding to results in Section 5.4, "Input noise level for SC loss". Specifically, in this variant, the same sampled noise is first added to the latent variable $\mathbf{z}$ and the reconstructed latent variable $\hat{\mathbf{z}}$ using the same timestep t sampled from $U(0, t_{\text{max}})$, where t_{max} is a predefined threshold used to control the maximum noise intensity. A value of $t = 1$ corresponds to the complete 1000-step noising process. These are then separately fed into the trained diffusion model for feature mapping to obtain the corresponding semantic features, which are subsequently aligned.

Bits of Tag Guidance. We employ RAM++ from Recognize Anything [93], which has a maximum default vocabulary of 4585 tags. For simplicity, we utilize fixed-length encoding without entropy coding, requiring 13 bits per tag ID to accommodate a maximum of 8192 tags. Based on our analysis of 500 randomly sampled COCO images, RAM++ predicts an average of 8.7 tags per image under its

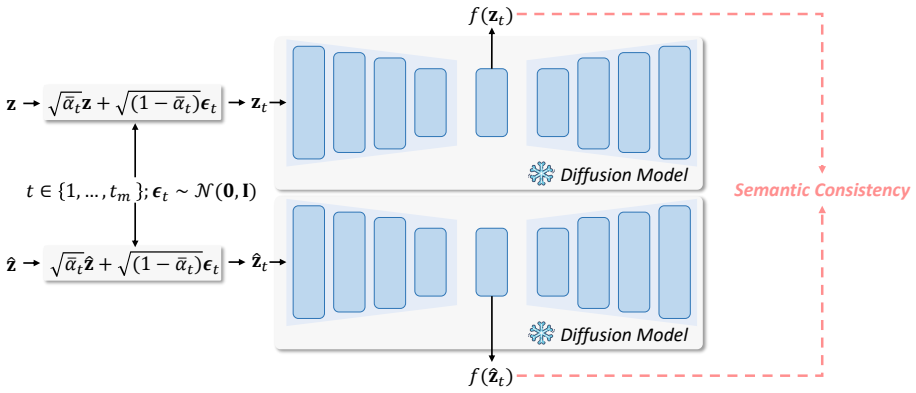


Figure 10: **Illustration of the variant of SC loss calculation.** Noisy latent variable $\mathbf{z}_t$ and $\hat{\mathbf{z}}_t$ are utilized as input into the pre-trained diffusion model.

default settings. Consequently, the average bit overhead for tag guidance amounts to $13 \times 8.7 = 113.1$ bits per image.

A.3 Hardware device

All training and inference experiments are conducted on 4 NVIDIA A100 Tensor Core GPUs.

B Experiments

B.1 Details of evaluation protocol

We conduct a comprehensive evaluation from two dimensions: the performance on various intelligent tasks and the perceptual quality of image reconstruction.

Intelligent tasks. As mentioned in Section 5.2, the efficacy and generalization capabilities of our proposed method are rigorously evaluated across a diverse spectrum of intelligent tasks spanning different domains, datasets, task-specific models, and vision backbones. To the best of our knowledge, this is the first comprehensive study to present such extensive experimental validation in the context of machine-oriented image coding. We believe this thorough empirical analysis not only substantiates our contributions but also establishes new benchmarks that may accelerate advancement in this emerging field.

The specific information about datasets and task models is shown in Table 1, where datasets include COCO 2017 [99], Flickr30K[3], RefGTA [101], and ADE20K[4]. The intelligent tasks cover three major categories: traditional perception-based computer vision tasks, multimodal retrieval tasks, and multimodal understanding tasks based on large language models. For task models, traditional perception tasks are evaluated through a series of models in the Detectron2 toolkit; multimodal retrieval tasks are evaluated through the BEiT-3 model; and for multimodal understanding capabilities, we specifically selected two multimodal large language models (MLLMs) based on different backbone networks, processing different data domains, and executing tasks of varying granularities.

This comprehensive experimental framework aims to thoroughly verify the generalizability and superior performance of the Diff-ICMH approach across different data types, datasets, intelligent tasks, and task models. For evaluation metrics, we employ standard measures appropriate to each task: mAP (mean Average Precision) of bounding boxes for object detection, mAP of segmentation masks for instance segmentation, AP of keypoints for pose estimation, PQ (Panoptic Quality) for panoptic segmentation, Recall@1 for multimodal retrieval tasks, and accuracy for referring expression comprehension. The open-set pixel-domain understanding tasks based on MLLMs are evaluated using mIoU for semantic segmentation, mAP for instance segmentation, and PQ for panoptic segmentation.

[3] https://hockenmaier.cs.illinois.edu/DenotationGraph/

[4] https://ade20k.csail.mit.edu/

Table 1: Overview of experimental datasets, tasks, and models.

Dataset	Task Type	Specific Task	Task Model	Backbone
COCO 2017	Traditional Perception	Object Detection Instance Segmentation Pose Estimation Panoptic Segmentation	Faster R-CNN Mask R-CNN Keypoint R-CNN Panoptic-FPN	R50-FPN [114]
Flickr30K	Multi-Modal Retrieval	Image-Text Retrieval Text-Image Retrieval	BEiT-3 [115]	MW-Transformer [116]
RefGTA ADE20K	MLLM Understanding	Referring Expression Open-Set Segmentation	Qwen2.5-VL[117] Ospray [118]	WA-ViT [117] ConvNext [119]

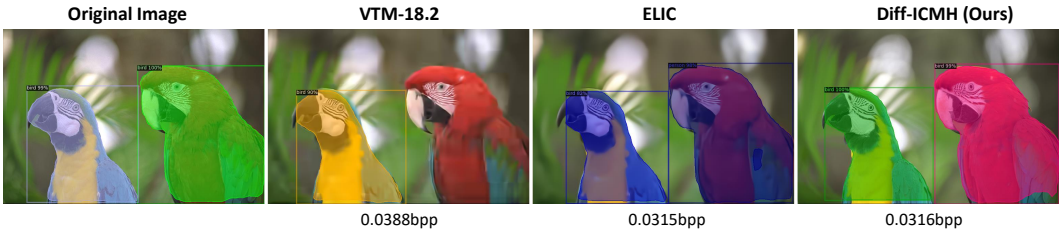


Figure 11: Visualized results on instance segmentation.

Image reconstruction. This paper uses three public datasets: Kodak [103], Tecnick [104], and CLIC2020 [105]. During the experiments, images from Tecnick and CLIC2020 datasets are rescaled to 768 pixels on the short side, and samples of size 768×768 are cropped from the center of the images for evaluation, following previous method [74]. Evaluation metrics include PSNR and MS-SSIM for measuring signal fidelity, as well as LPIPS, FID, and DISTS for evaluating perceptual quality.

B.2 Visualization results

Task supporting. Figure 11 presents visualization results comparing our method against VTM-18.2 [5], ELIC [106], and FTIC [20] on instance segmentation tasks. The comparison reveals that VTM-18.2, despite operating at its highest bit rate, fails to correctly identify the two distinct objects within the image. Similarly, ELIC erroneously classifies the "bird" object on the right as a "person," demonstrating significant recognition inaccuracy. In contrast, only the reconstruction output generated by our Diff-ICMH method successfully preserves the critical semantic information necessary for accurate object identification, thereby effectively supporting the completion of this intelligent task with precision.

Perception-oriented reconstruction. Figure 12 shows the visual comparison of reconstruction quality between Diff-ICMH and other compression methods. From the comparison, it can be observed that methods optimized for signal fidelity (VTM-18.2, ELIC, FTIC) produce reconstructed results with obvious blurring, significantly reducing visual quality. Although MS-ILLM shows some improvement in texture clarity, its reconstructed textures lack authenticity and are accompanied by obvious noise interference. In contrast, our proposed Diff-ICMH demonstrates the best visual realism in reconstruction results while maintaining clear boundary contour features.

B.3 Feature difference analysis

To further evaluate Diff-ICMH's performance advantages in semantic information protection and intelligent task support, we designed comparative experiments based on feature difference analysis. Specifically, we calculated the differences between feature representations of reconstructed images from various compression methods and those of original images at different layers of a ResNet50 feature extractor (using 1 minus cosine similarity as the metric, where lower values represent higher feature similarity), and plotted the curves of these differences across network depths.

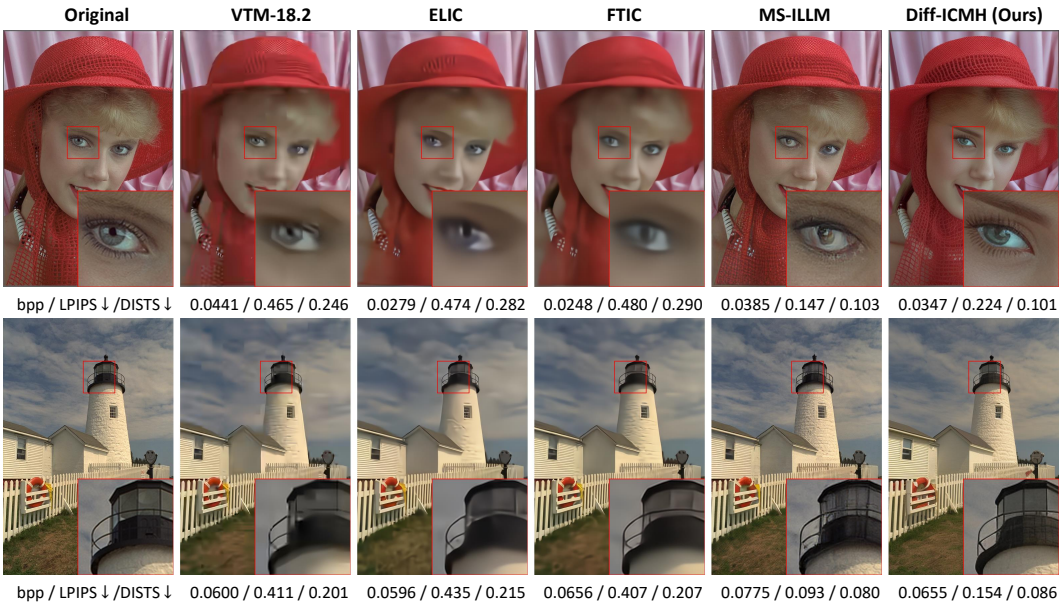


Figure 12: Reconstruction of Diff-ICMH and other compression methods.

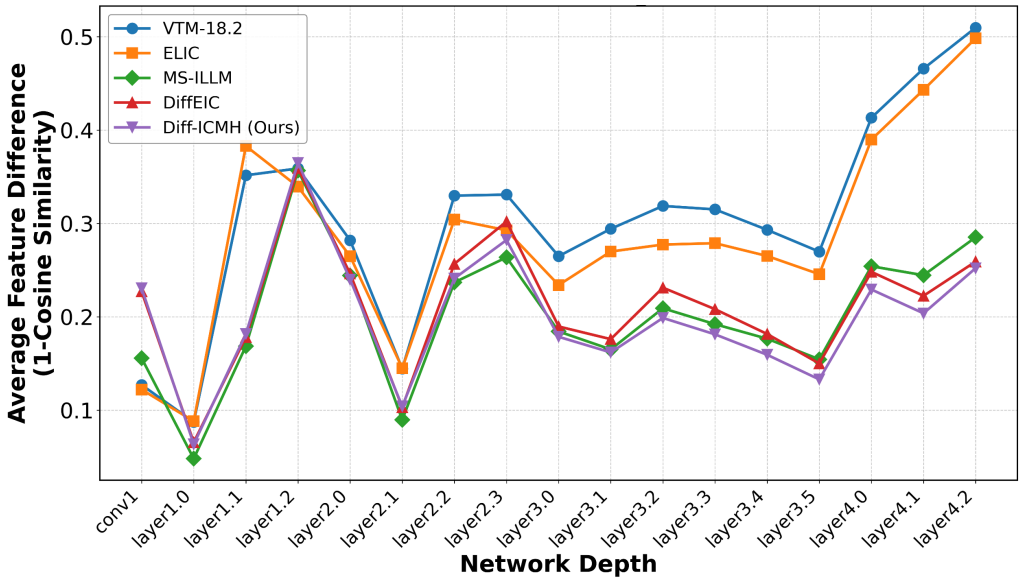


Figure 13: The evolution curves showing feature differences between reconstructed images from Diff-ICMH and other compression methods compared to original images across various layers of the ResNet50 feature extractor. Lower values represent higher feature similarity.

The experiments are conducted on the Kodak dataset, analyzing the average results of all images. As shown in Figure 13, the experimental results exhibit significant hierarchical characteristics: in shallow network layers, generative compression methods (MS-ILLM, DiffEIC, Diff-ICMH) show relatively higher difference values, while methods optimized for signal fidelity (VTM-18.2, ELIC) maintain minimal differences. However, as the network hierarchy and semantic level deepen, this pattern changes significantly—the latter methods' difference values increase rapidly, reaching a peak of approximately 0.5 at the deepest layer, far exceeding generative compression methods.

Among generative compression methods, Diff-ICMH's advantage is demonstrated in its performance from the middle network layers (layer3.0) to the deepest layers (layer4.2), consistently maintaining the

Table 2: Encoding and decoding time on Kodak dataset. (Second)

Method	NFE	Encoding Time	Decoding Time	Hardware
VVC	-	13.862	0.066	13th Core i9-13900K
ELIC	-	0.056	0.081	RTX4090
HiFiC	-	0.038	0.059	RTX4090
MS-ILLM	-	0.038	0.059	RTX4090
PerCo	5	0.080	0.665	A100
PerCo	20	0.080	2.551	A100
DiffEIC	20	0.128	1.964	RTX4090
DiffEIC	50	0.128	4.574	RTX4090
Diff-ICMH	20	0.232	5.456	A100
Diff-ICMH	50	0.232	13.14	A100

lowest feature difference values in this range. This result confirms the superiority of our approach in semantic information protection. More importantly, considering that the feature extractor (ResNet50) used in the experiment and the feature mapping (Stable Diffusion) used in the SC loss belong to different network architectures and different pre-training purpose, this performance advantage also highlights the excellent model generalization capability of our method.

B.4 Complexity analysis

Table 2 illustrates the encoding and decoding time of different methods. The data reveals that diffusion-based methods (such as DiffEIC and Diff-ICMH) perform well in terms of encoding speed, showing significant advantages compared to traditional VVC methods (13.862 seconds), with DiffEIC taking 0.128 seconds and Diff-ICMH taking 0.232 seconds.

However, there is room for improvement in our method's decoding time. Particularly, when the number of denoising steps increases, the decoding time increases significantly. For instance, Diff-ICMH requires 5.456 seconds and 13.14 seconds for decoding at 20 steps and 50 steps, respectively. The decoding time is considerably longer than traditional methods like VVC (0.066 seconds) and GAN-based methods such as HiFiC (0.038 seconds) and MS-ILLM (0.059 seconds).

The main reason for this phenomenon is that the progressive denoising process of diffusion models is inherently an iterative computational process, with each step requiring a complete network forward pass. When the number of steps increases, the computational cost increases linearly. Optimizing encoding and decoding speeds will be an important direction for future work: (i) Investigating more efficient sampling strategies, such as implementing samplers with less steps [108, 109]; (ii) exploring knowledge distillation techniques to distill multi-step denoising networks into lighter single-step or few-step networks [111]; (iii) and optimizing network structures to reduce the computational complexity of each denoising step.

While diffusion-based image methods currently face computational efficiency challenges, this work primarily serves to demonstrate the fundamental potential of this paradigm. We anticipate that ongoing advancements in sampling algorithm efficiency and targeted engineering optimizations will substantially reduce encoding and decoding latency, thereby enhancing the practical utility of diffusion models in this research field.

B.5 Ablation study of distortion loss calculation space

In Equation (4) of the main text, the distortion loss $\mathcal{L}_{\text{dist}}$ are calculated in the VAE latent space:

$$\mathcal{L}_{\text{dist}} = \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 = \|\mathcal{E}_{\text{VAE}}(\mathbf{x}) - \mathcal{D}_c(\hat{\mathbf{y}})\|_2^2, \quad (7)$$

Here we conduct ablation study of calculating loss in the pixel space:

$$\mathcal{L}_{\text{dist}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \|\mathbf{x} - \mathcal{D}'_c(\hat{\mathbf{y}})\|_2^2, \quad (8)$$

where $\mathcal{D}'_c$ maintains the same architectural foundation as $\mathcal{D}_c$ but incorporates additional upsampling blocks

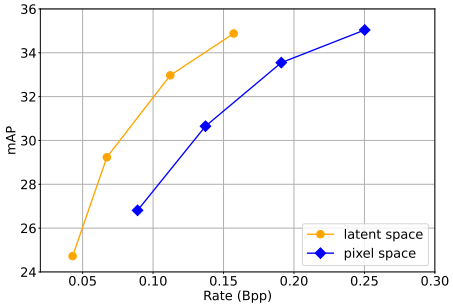


Figure 14: Ablation of distortion loss calculation space.

to reconstruct signals at pixel-level resolution. The reconstructed pixel-space output $\hat{\mathbf{x}}$ is subsequently fed into the control module for generative reconstruction. Figure 14 demonstrates the substantial performance advantage of calculating distortion in the latent space rather than pixel space. This finding confirms that the VAE latent space provides a more compact and perceptually meaningful representation that effectively filters semantically irrelevant information from the pixel domain. Through optimizing fidelity in the latent space, the compressed bitstream effectively prioritizes semantically salient information, thereby achieving superior rate-distortion performance across various downstream machine vision applications.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions of this paper are claimed in the abstract and the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation of this paper in the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the implementation details in the Section of Experiments. Besides, code will be released.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code will be released when published.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The full training and test details are described in Section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Rate-distortion curves are illustrated to verify the effectiveness of proposed method, which are commonly used in the area of compression.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computer resources used in our experiments in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This work is about the foundamental techniques in terms of image compression, with no societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all original paper and code used in this paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This submission does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.