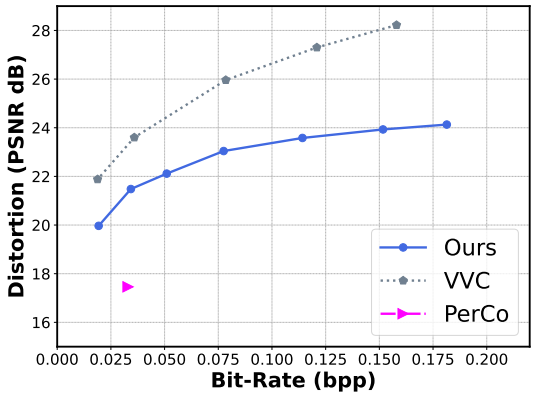
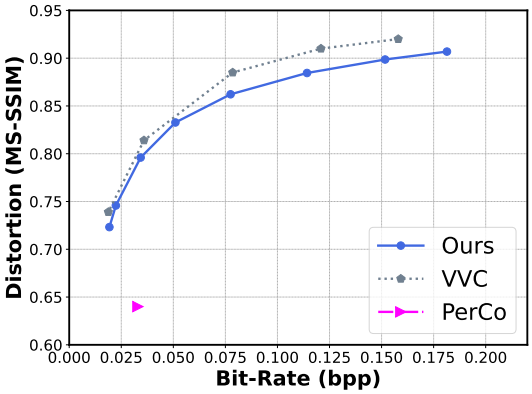
Supplementary

Jianhui Chang
China Telecom Cloud Computing Research Institute
changjh1@chinatelecom.cn

***Abstract*—This supplementary material provides additional quantitative comparisons between VVC, PerCo and our method on PSNR and MS-SSIM in Section I and presents ablation studies to evaluate the proposed attentive fusion and fidelity enhancement methods in Section II.**



(a) PSNR results.



(b) MS-SSIM results.

Fig. 1: The quantitative comparisons in terms of PSNR and MS-SSIM on the AIGC dataset, DiffusionDB. Higher scores indicate higher signal fidelity.

I. Additional Quantitative Comparisons on PSNR and MS-SSIM

Traditional coding methods typically prioritize signal fidelity and employ conventional objective quality assessment metrics, such as PSNR and MS-SSIM, to measure pixel-level distortions. In Fig. 1, we present additional quantitative comparison results for VVC, PerCo, and our method in terms of PSNR and MS-SSIM on the AIGC dataset. VVC, as the latest coding standard, primarily focuses on optimizing signal fidelity. In contrast, PerCo and our method are generative coding approaches based on pretrained latent diffusion models, with a focus on enhancing perceptual quality.

As shown in Fig. 1, generative coding methods generally underperform in PSNR and MS-SSIM metrics due to their probabilistic nature. Models such as GANs and diffusion models produce outputs aligned with the data distribution, resulting in perceptually plausible but not pixel-accurate reconstructions, which can negatively impact distortion metrics. Nevertheless, our method achieves MS-SSIM results comparable to those of VVC and significantly surpasses PerCo. This demonstrates that our approach effectively balances perceptual quality and reconstruction fidelity, validating its superior performance.

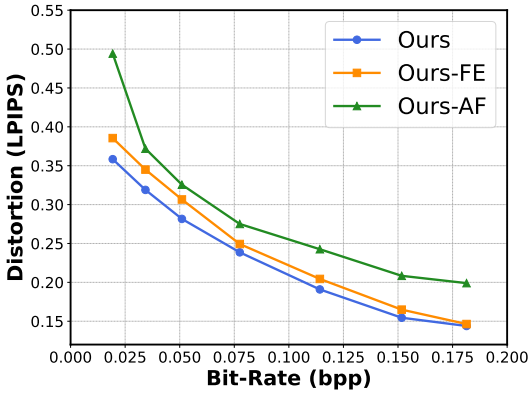
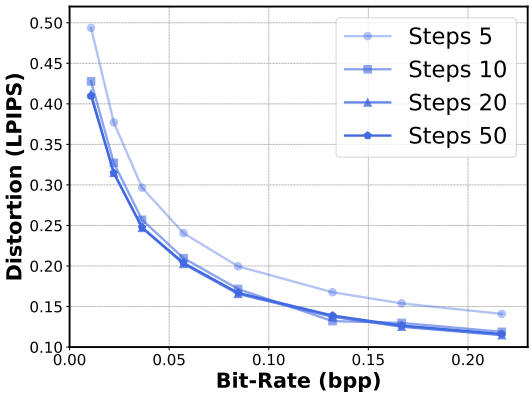


Fig. 2: Ablation results of attentive fusion and fidelity enhancement methods on the AIGC dataset.

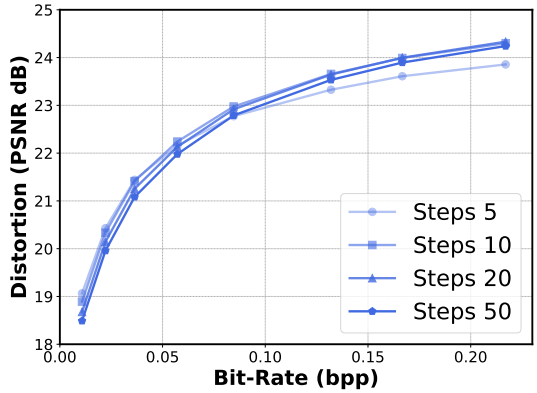
II. Ablation Studies

In this paper, as described in Sec. II-A and II-C, we propose an attentive latent fusion module to better integrate compressed latents with the internal features of pretrained diffusion models, as well as a fidelity enhancement method based on distribution renormalization. To evaluate the contributions of these components, we consider two model variants: (1) our model without the proposed fidelity enhancement method, referred to as "Ours-FE"; and (2) our model using direct additive fusion instead of the proposed attentive fusion method, referred to as "Ours-AE.". The "Ours-AE" variant is finetuned as described in Sec. II-B (2) under the same experimental conditions as our full model.

The ablation results on the AIGC dataset, evaluated in terms of the LPIPS metric, are shown in Fig. 2. These



(a) LPIPS results.



(b) PSNR results.

Fig. 3: The rate-distortion comparisons of different denoising steps in terms of LPIPS and PSNR on the AIGC dataset, DiffusionDB. Lower LPIPS indicates higher perceptual quality, and higher PSNR indicates higher signal fidelity.

results demonstrate that both the attentive fusion module and the fidelity enhancement method significantly improve reconstruction fidelity, leading to better compression performance. Quantitatively, based on the Bjontegaard metric for LPIPS, the attentive fusion method improves reconstruction quality by 20.06% at equal rate costs. This improvement is attributed to the method's ability to precisely capture spatial details from compressed latents and effectively leverage the correlations between the pretrained model's internal features and compressed latents. The fidelity enhancement method contributes an additional 6.05% improvement in reconstruction quality, primarily by enhancing color consistency with the original images. In terms of BD-rate, the attentive fusion and fidelity enhancement methods can achieve bitrate savings of 15.76% and 5.83%, respectively, while maintaining similar reconstruction quality. These findings validate the effectiveness of the proposed attentive fusion and fidelity enhancement methods in improving both reconstruction fidelity and compression efficiency.

III. Evaluation Effect of Denoising Steps

Diffusion models define a Markov chain that progressively adds random noise to the data and then learns to reverse this process to reconstruct the desired samples. Typically, the number of diffusion steps can range from a few hundred to over one thousand to ensure that the final forward diffusion samples approximate an isotropic Gaussian distribution. While one may generate data with fewer steps using strided sampling or other acceleration methods, the generation quality typically improves as the inference steps increase. To balance quality and speed, we adopt the DDIM sampling method, which introduces non-Markovian diffusion processes that share the same training objective but allow much faster sampling.

We conduct experiments to evaluate the impact of different denoising step configurations. Fig. 3 illustrates the reconstruction performance across various step counts. As shown in Fig. 3a, increasing the number of denoising steps improves both perceptual quality and reconstruction fidelity. For instance, raising the steps from 5 to 50 yields a 28.2% $\sim$ 33.3% performance gain in BD-Rate. However, once the number of steps exceeds 10, the improvement becomes marginal (up to around 7%). In Fig. 3b, at very low bitrates, increasing the number of steps leads to higher perceptual quality but lower PSNR, indicating a trade-off between signal fidelity and perceptual quality. However, when the bit rate exceeds 0.1 bpp, PSNR for 5 steps becomes significantly lower, suggesting that too few steps degrade overall quality. For over 10 steps, PSNR stabilizes, with little difference in quality observed across higher step counts. Overall, using 10 sampling steps strikes a favorable balance between decoding complexity and performance.