Steering One-Step Diffusion Model with Fidelity-Rich Decoder for Fast Image Compression

Zheng Chen[1*], **Mingde Zhou**[1*], **Jinpei Guo**[1,2], **Jiale Yuan**[1], **Yifei Ji**[1], **Yulun Zhang**[1†]

[1]Shanghai Jiao Tong University, [2]Carnegie Mellon University

LoRA Rank	LPIPS ↓	DISTS ↓	MS-SSIM ↑
8	0.351	0.164	0.839
16	0.343	0.160	0.852
32	**0.336**	**0.158**	**0.857**
64	0.339	0.160	0.849

Table 1: Ablation study on the LoRA rank. A rank of 32 achieves the best trade-off across all metrics.

Implementation Details

In this section, we provide more details on implementation to complement the main paper. We use HiFiC (Mentzer et al. 2020) as the VAE compression module and Stable Diffusion 2.1 (Rombach et al. 2022) as the one-step diffusion baseline. We set timestep t as 999. **In stage 1**, we set k_M=1 and k_P=2, and vary the Lagrange multiplier $\lambda \in \{0.5, 1, 2, 4, 8\}$ to obtain compression models with different bitrates (bpp). **In stage 2,** we adopt the same k_M and k_P values as in Stage 1 for the distortion function $d(\cdot)$. **In stage 3**, we set the alignment coefficient α to 30, and $\lambda \in \{1, 2, 4, 8, 16\}$ to increase the rate penalty. For the final fine-tuning stage, we set the generator loss weight β to 2×10^{-1}.

More Ablation Studies

Ablation on LoRA Rank. We conduct an ablation study to determine the optimal LoRA rank for fine-tuning the UNet in our diffusion model. We experiment with ranks of 8, 16, 32, and 64, with the results summarized in Tab. 1. We observe a consistent performance improvement across all metrics as the rank increases from 8 to 32. However, there is a slight degradation in performance when the rank is increased further. This suggests that a rank of 32 provides sufficient capacity for the model to adapt to the compression task, while a higher rank may lead to minor overfitting. Notably, we also found that full parameter fine-tuning yields inferior results than any of the LoRA variants, likely due to optimization instability. Therefore, to achieve the best trade-off between parameter efficiency and performance, we select a LoRA rank of 32 for our final model.

Ablation on GANs Hyperparameters. In the final fine-tuning stage, we apply a GAN-based objective to enhance realism. We introduce a GAN loss $\mathcal{L}_g$ to fine-tune the SODEC model, $\mathcal{L}_g = -\mathbb{E}_{\hat{z}_0 \sim p_g}[\log(\mathcal{D}(\hat{z}_0))]$. So the overall loss for this final fine-tuning stage can be written as:

$$\mathcal{L}_{\text{finetune}} = d(x, \hat{x}) + \lambda \cdot r(\hat{y}, \hat{z}) + \alpha \cdot \mathcal{L}_{align} + \beta \cdot \mathcal{L}_g, \quad (1)$$

and the discriminator model $\mathcal{D}_\phi$ is optimized for:

$$\begin{aligned} \mathcal{L}_d = &- \mathbb{E}_{z_{\text{real}} \sim p_{data}}[\log(\mathcal{D}_\phi(z_{\text{real}}))] \\ &- \mathbb{E}_{\hat{z}_0 \sim p_g}[\log(1 - \mathcal{D}_\phi(\text{sg}(\hat{z}_0)))]. \end{aligned} \quad (2)$$

GAN Loss Weight (β)	LPIPS ↓	DISTS ↓	MS-SSIM ↑	CLIPIQA ↑
w/o GANs	0.351	0.161	0.840	0.660
0.01	**0.341**	0.156	0.845	0.729
0.05	0.352	0.154	0.825	0.711
0.1	0.356	0.156	0.822	0.730
0.2	0.346	**0.151**	**0.847**	**0.735**
0.5	0.365	0.159	0.823	0.712

Table 2: Ablation study on the GAN loss weight (β) in the final stage. A weight of 0.2 achieves the best performance.

We conduct an ablation study to investigate the impact of the GAN loss weight β. The results, summarized in Tab. 2, compare our model trained without a GAN ($\beta = 0$) against variants with different β values. We observe that introducing the adversarial objective consistently improves most perceptual metrics (LPIPS, DISTS, CLIPIQA) compared to the baseline. Performance generally improves as β increases to 0.2, which achieves the best results on DISTS, MS-SSIM, and CLIPIQA. However, we note a trade-off, as a smaller weight of $\beta = 0.01$ yields the optimal LPIPS score. Increasing the weight further to 0.5 leads to a degradation across most metrics, suggesting potential training instability. Considering the strong results on multiple key metrics, we select $\beta = 0.2$ as the optimal setting for our final model.

More Visualizations

Comparison Between With/Without Alignment Loss. To visually demonstrate the necessity of the alignment loss during end-to-end fine-tuning, we present a visual comparison in Fig. 1. As the VAE encoder is updated, the latent representation $\hat{y}$ undergoes a distributional shift, which can severely degrade the output of the fixed, fidelity-oriented decoder $\mathcal{D}_a$. The following figure illustrates the output of the decoder with and without the proposed $\mathcal{L}_{align}$ to show how it counteracts this distortion and preserves fidelity.

*These authors contributed equally.

†Corresponding author: Yulun Zhang, yulun100@gmail.com

Figure 1: Visual comparison of the decoder output $\hat{x}_f$ with and without the alignment loss. The MSE-only alignment loss effectively counteracts distortions from the fine-tuned encoder and preserves high-fidelity details.

Reconstruction Quality of Different Training Strategies. To demonstrate the effectiveness of our proposed training methodology, we provide a visual comparison (Fig. 2) of the final reconstruction quality between our model trained with and without the rate annealing strategy. The strategy involves manually tuning the Lagrange multiplier λ to ensure the final bitrate is close to the original values.

Training Process with different Alignment Loss. We analyze the training dynamics of different alignment loss configurations during stage 3 (joint training with rate annealing), where we lift the rate penalty and perform a joint optimization. As presented in Fig. 3. The model trained without any alignment loss converges to the worst LPIPS and MS-SSIM scores, confirming the necessity of this constraint. A composite loss of MSE and LPIPS improves performance, and employing an MSE-only alignment loss yields the best results, achieving both the lowest LPIPS and the highest MS-SSIM scores. Furthermore, the training process with the MSE-only loss exhibits superior stability, as evidenced by the smoother curves. This is likely because optimizing with LPIPS becomes unreliable when the latent representation $\hat{y}$, distorted by the low-bitrate constraint, contains insufficient information for stable convergence.

More Qualitative Results

We provide more visual comparisons in Figs. 4, 5, 6, and 7. Compared to various existing methods, our SODEC reconstructs results with higher realism and fidelity, further demonstrating its effectiveness.

Limitations and Future Work

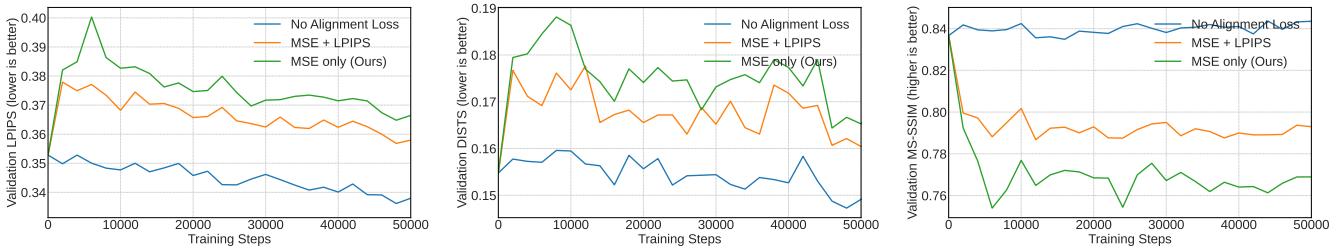
While our one-step design significantly accelerates decoding compared to multi-step methods, its latency is still higher than that of non-diffusion-based codecs. For future work, exploring model distillation could further enhance the performance and efficiency of the one-step diffusion model. Meanwhile, investigating more lightweight guidance mechanisms is another promising direction. In addition, model pruning is also a direction worth exploring.

References

Mentzer, F.; Toderici, G. D.; Tschannen, M.; and Agustsson, E. 2020. High-fidelity generative image compression. In *NeurIPS*. 1

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 1

Figure 2: Qualitative comparison of our rate annealing training strategy. Our proposed strategy (right column) consistently produces reconstructions with better perceptual quality and fidelity than those without rate annealing (middle column).



(a) LPIPS (b) DISTS (c) MS-SSIM

Figure 3: Training dynamics of different alignment loss configurations. The LPIPS (a) and MS-SSIM (c) validation loss curves are shown. Using an MSE-only alignment loss leads to the most stable training.

Figure 4: Qualitative comparison with state-of-the-art methods on the DIV2K-Val dataset.

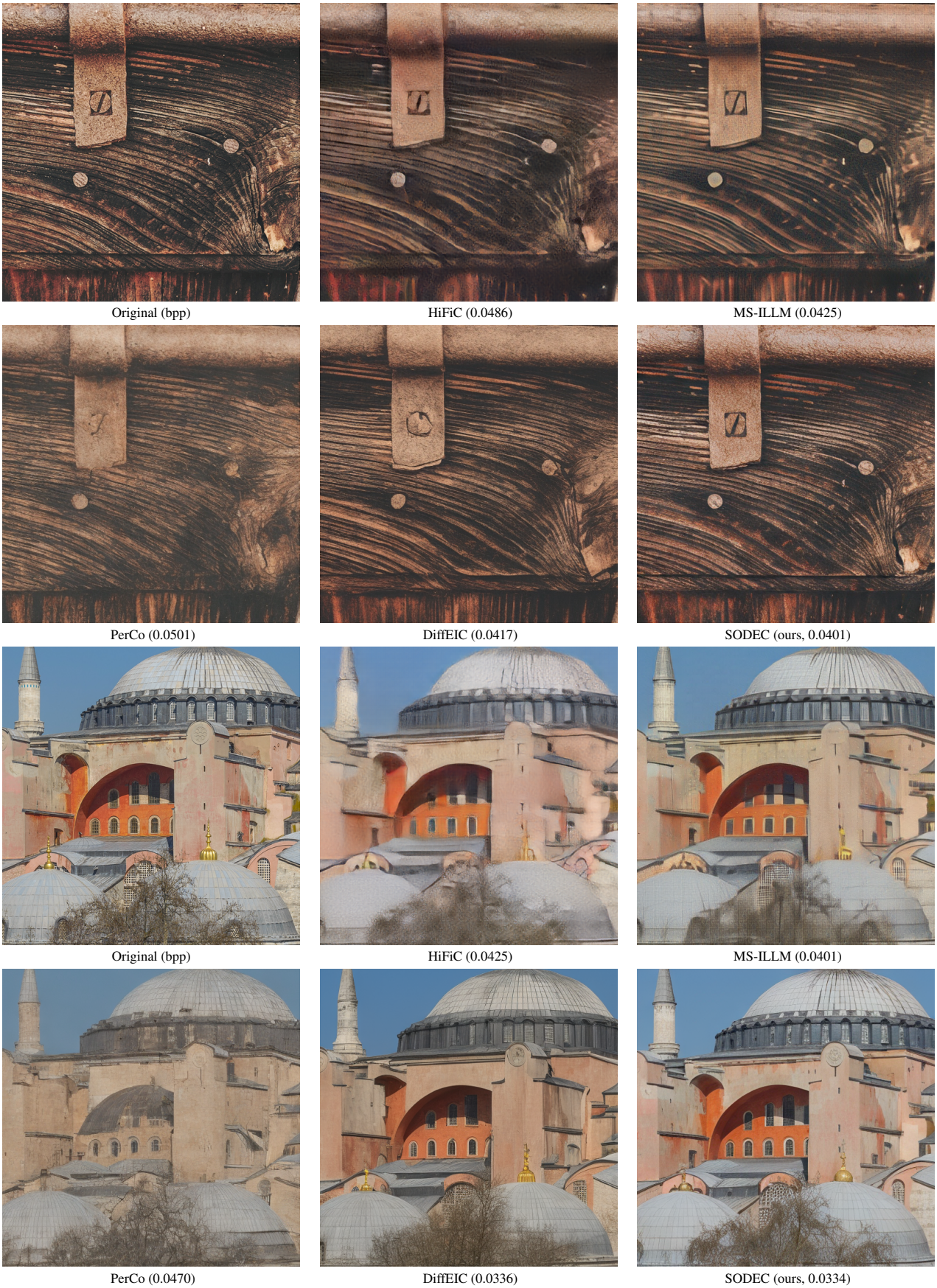


Figure 5: Qualitative comparison with state-of-the-art methods on the DIV2K-Val dataset.

Original (bpp) | HiFiC (0.0453) | MS-ILLM (0.0478)

PerCo (0.0611) | DiffEIC (0.0400) | SODEC (ours, 0.0391)

Original (bpp) | HiFiC (0.0473) | MS-ILLM (0.0428)

PerCo (0.0536) | DiffEIC (0.0434) | SODEC (ours, 0.0400)

Figure 6: Qualitative comparison with state-of-the-art methods on the Kodak dataset.

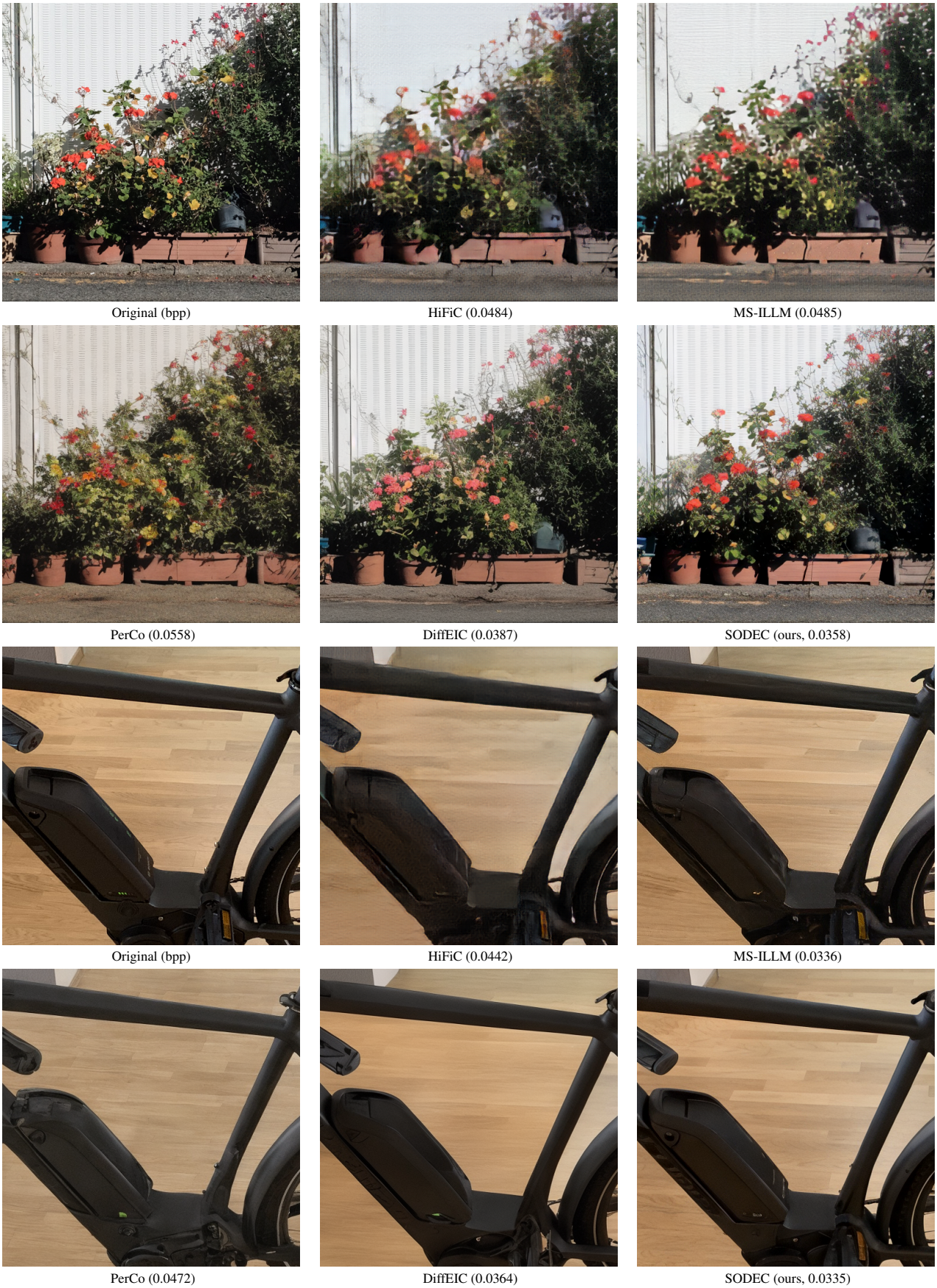


Figure 7: Qualitative comparison with state-of-the-art methods on the CLIC2020 dataset.