Supplementary materials

A Trainable parameters

As shown in Tab. 1, our proposed methods use far fewer training parameters than the full fine-tuning method. This is mainly because our methods do not have parameters in the encoder and hyperprior parts that need training, allowing for a faster learning process. Additionally, our methods only need to update a relatively small set of parameters, increasing their overall efficiency. Specifically, the method that repeats parameters has a much smaller number of parameters compared to the other methods. This points to the high efficiency and practicality of our methods and strengthens their potential for effective use in various real-world situations.

B Adapter on hyperprior model

We conducted experiments on integrating our adapter structure with hyperprior models. As Fig. 1 shows, this did not lead to improvements in PSNR and MS-SSIM values. Furthermore, the compression performance got worse at lower bitrates. This can be attributed to the fact that the added structure makes more bits need to be transferred.

C GoP size variation

As demonstrated in prior researches [3,4], the commonly used practical Group of Pictures (GoP) size is close to 32. Thus, We also evaluated the performance with the GoP size set to 32, using the same Rate-distortion (RD) curve. The datasets for this evaluation remained the same, including UVG [5], MCL-JCV [6], and HEVC class B and C [2]. As shown in Fig. 2, the RD curves of PSNR and MS-SSIM displayed largely similar performance to the previous result with a smaller GoP size. Although there was a slight decrease in performance at lower bitrates, the overall performance remained consistent, demonstrating the robustness and applicability of our proposed methods to larger GoP sizes. This suggests that our proposed methods can be effectively applied even when the GoP size is increased, further enhancing the versatility of our method.

D Qualitative results

As indicated in Sec. 4.3, some datasets, especially those with cartoon-style or complex movements, pose challenges in reconstructing images. Therefore, we present the qualitative results for each dataset in Fig. 3, Fig. 4, and Fig. 5. The comparison is made at similar bpp settings, revealing that SSF [1] exhibits motion blur in complex domains. Moreover, full fine-tuning results in distortion

	Total params. (M)	Train params. (M)			
		Encoder	Hyperprior	Decoder	all
Full fine-tuning	34.24	12.70	16.59	4.95	34.24
Ours	35.03	0	0	0.79	0.79
Ours(repeat)	34.27	0	0	0.03	0.03

Table 1: Number of training parameters for video sequence instance-adaptation.

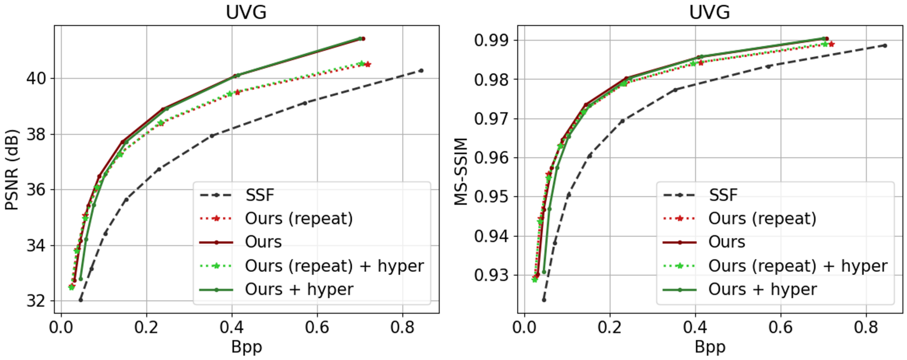

Fig. 1: RD-curve when apply adapter on hyperprior model. Comparison conducted on UVG dataset.

from the original, failing to accurately represent finer details. In contrast, both of our methods can effectively represent their respective areas without motion blur, even in the cartoon domain. These qualitative results highlight the superior overfitting mitigation capability of our methods.

E PSNR per frame

To assess the detailed performance of our method, we measure the PSNRs for each frame. As depicted in Fig. 6, the comparison is made between the baseline and our method with no duplication, focusing on the some of UVG dataset sequence. Both the baseline and our method show an increasing trend in PSNR. However, our method exhibits an overall improvement in PSNRs of approximately 1 dB, with smaller spikes, even though bpp is lower than baseline. This suggests that our method may be prone to overfitting the input video sequences with saving the number of bits.

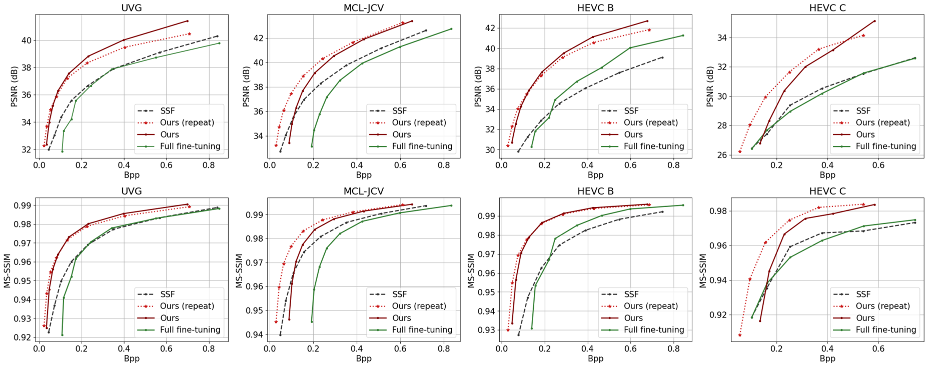

Fig. 2: RD-curve with GoP set to 32. Comparison conducted on UVG, MCL-JCV, HEVC class B, and C datasets.

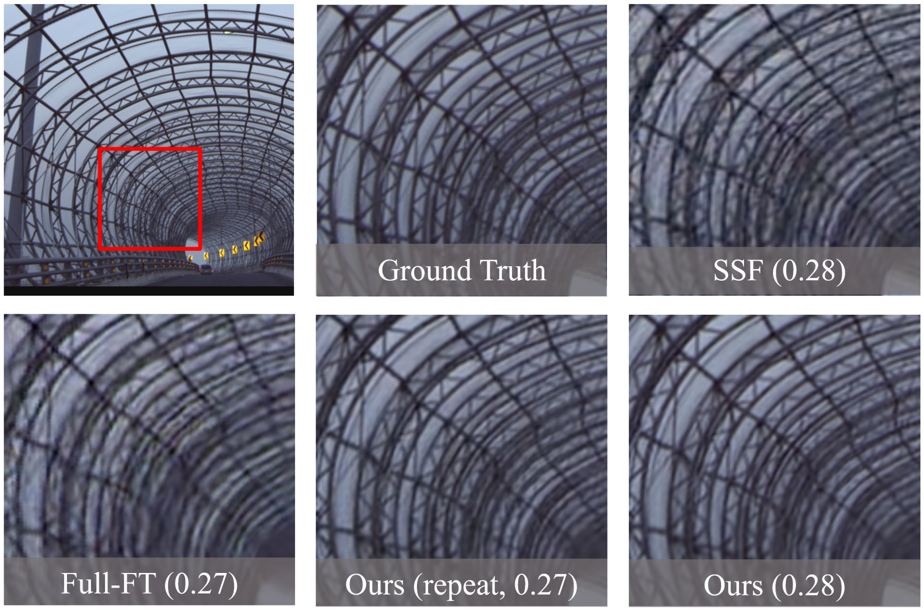

Fig. 3: Qualititive results of MCL-JCV 10 dataset.

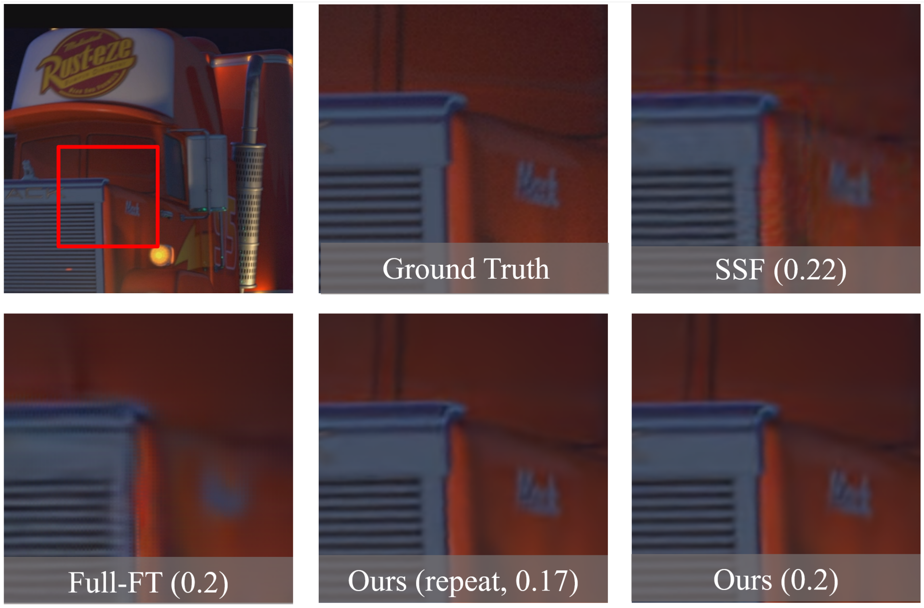

Fig. 4: Qualititive results of MCL-JCV 24 dataset.

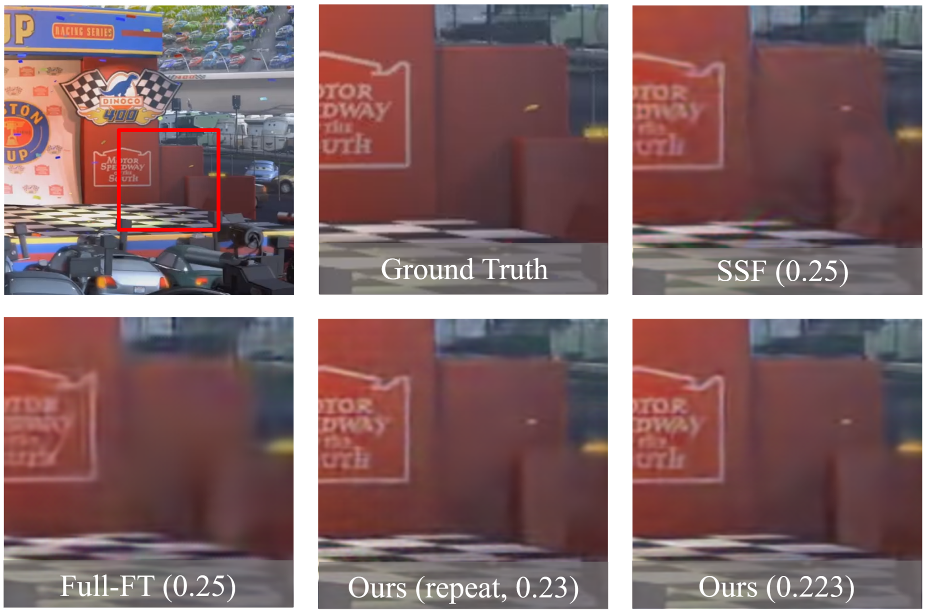

Fig. 5: Qualititive results of MCL-JCV 25 dataset.

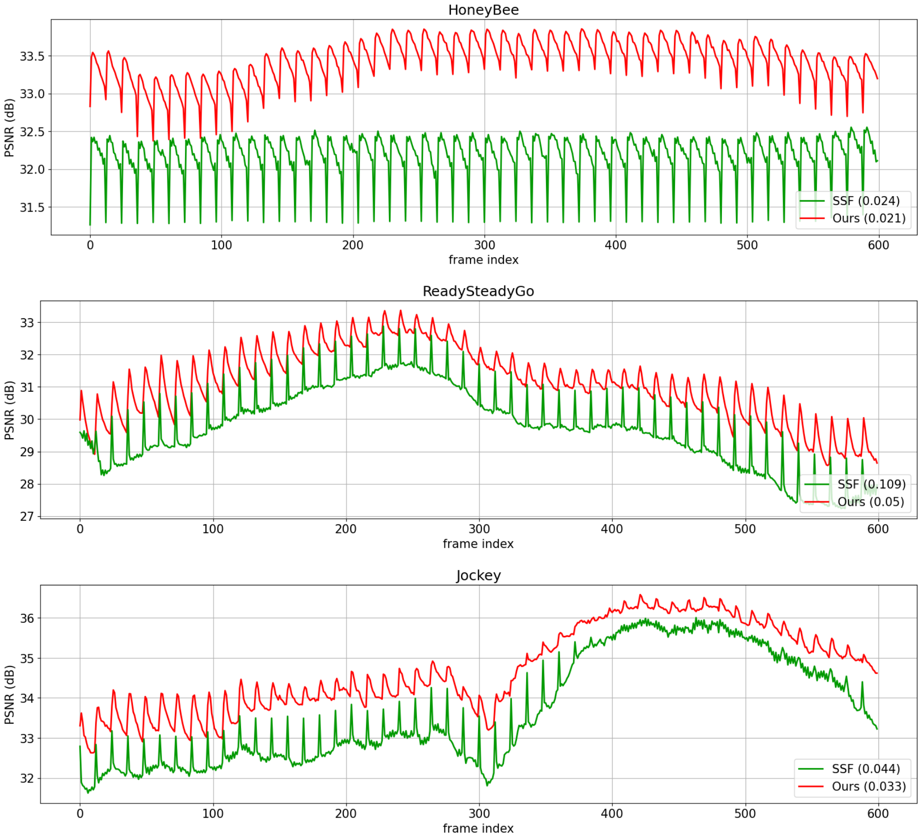

Fig. 6: PSNR for each frame using the same baseline model, tested on the 'HoneyBee', 'ReadySteadyGo', and 'Jockey' sequence. The Number in the legend represent bpp.

References

1. Agustsson, E., Minnen, D., Johnston, N., Balle, J., Hwang, S.J., Toderici, G.: Scale-space flow for end-to-end optimized video compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8503–8512 (2020)
2. Bossen, F., et al.: Common test conditions and software reference configurations. JCTVC-L1100 **12**(7), 1 (2013)
3. Li, J., Li, B., Lu, Y.: Hybrid spatial-temporal entropy modelling for neural video compression. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1503–1511 (2022)
4. Li, J., Li, B., Lu, Y.: Neural video compression with diverse contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22616–22626 (2023)
5. Mercat, A., Viitanen, M., Vanne, J.: Uvg dataset: 50/120fps 4k sequences for video codec analysis and development. In: Proceedings of the 11th ACM Multimedia Systems Conference. pp. 297–302 (2020)
6. Wang, H., Gan, W., Hu, S., Lin, J.Y., Jin, L., Song, L., Wang, P., Katsavounidis, I., Aaron, A., Kuo, C.C.J.: Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In: 2016 IEEE international conference on image processing (ICIP). pp. 1509–1513. IEEE (2016)