GIViC: Generative Implicit Video Compression

Supplementary Material

A. Additional Details

A.1. Data formatting in the YUV colorspace

With the YUV 4:2:0 format, the two chroma components are sampled at half the sample rate of luma. Hence, we perform trilinear interpolation to upsample the chroma components and yield the input of the GIViC network where all channels are of the same spatial resolution. We respectively adopted and experimented with the following design choices:

- (Current design) The denoiser ϵ_θ generates a two-channel output, where one of the channels could be de-multiplexed into U and V channels of the original frame. In this case, the diffusion denoising loss is calculated in the YUV 4:2:0 colorspace.
- (V6.1) The denoiser ϵ_θ generates a three-channel output that is matched against the upsampled input to GIViC, where the diffusion denoising loss is calculated in the YUV 4:4:4 colorspace. The predicted frame's U and V channels are downsampled via trilinear interpolation to yield the final YUV 4:2:0 reconstruction.
- (V6.2) The denoiser ϵ_θ generates a three-channel output that is matched against the upsampled input to GIViC, where the diffusion denoising loss is calculated in the YUV 4:4:4 colorspace. We then fix the main model, attach and fine-tune another downsampling layer which is optimize to reduce the distortion loss in the YUV 4:2:0 colorspace.

The above options were ablated on the UVG and JVET-B datasets and the resulting average BD-rate figures, with the original GIViC as the anchor, are reported in Table S1. It could be seen that although the resulting difference is trivial, the current design offers the best overall performance.

ablation option	UVG	JVET-B
(V6.1)	+0.3	+0.5
(V6.2)	+0.2	+0.4
GIViC	+0.0	+0.0

Table S1. Ablation results (BD-rate, %) for different diffusion loss calculations in the YUV 4:2:0 colorspace.

A.2. Baselines

All baseline models used for comparison in the present paper are open-source and were acquired from their official repositories. Specifically, we re-trained VCT [5], C3 [4], and PNVC [2] with the default training configurations in the YUV colorspace. For C3, we did not use its best performing variant (i.e., *adaptive* configuration) as it requires sweeping over nine hyperparameters per patch for the entire video.

B. Additional Experiments

B.1. Compression Performance

We further report the compression performance measured by VMAF (Video Multimethod Assessment Fusion) and LPIPS (Learned Perceptual Image Patch Similarity), as shown in Table S2). VMAF and LPIPS are perceptual metrics designed to better approximate human perception of quality. Unlike traditional metrics such as PSNR and MS-SSIM, which can often misalign with subjective human experience, VMAF and LPIPS are known to provide scores that more accurately reflect perceived visual quality. This closer alignment with human vision has led to significant shifts in the industry; for instance, the optimization of modern codecs like AV1 is increasingly oriented towards VMAF. Even when evaluated against these more challenging and perceptually-focused metrics, our proposed model demonstrates preferable performance. It could be seen from Table S2 that GIViC still outperforms the benchmark codecs in terms of LPIPS and VMAF, which further demonstrates its advanced compression efficiency.

BD-rate (%)	UVG		MCL-JCV		JVET-B	
Codec	LPIPS	VMAF	LPIPS	VMAF	LPIPS	VMAF
VTM 20.0 (RA) [1]	-16.65	-17.29	-13.31	-13.32	-10.98	-11.00
AV1 libaom v3.0.2 (RA) [3]	-24.46	-19.71	-19.39	18.97	-12.01	-7.33
C3 [4]	-63.33	-62.31	-	-	-56.32	-58.99
GIViC w/o Overfit.	-4.59	-4.87	-3.21	-3.41	-3.33	-3.35

Table S2. Compression performance results of the proposed GIViC framework. Here each BD-rate value is calculated when the corresponding benchmark codec is used as the anchor.

B.2. Rate-Distortion-Complexity

One major limitation of GIViC is its non-trivial computational complexity due to the adoption of diffusion and transformer backbones, both of which are known to be computationally demanding. Its encoding time is particularly long: 1.78 hours per GOP of 32-frames at 1080P, which results in about 34 hours to train a 600-frame sequence on a single A100. However, we show in Table S3 that the encoding time could be drastically reduced by $2\times$ and $29\times$, respectively, when we allow the latent grids and hidden states to be initialized from the previous GOP and when we simply remove the overfitting steps. In Table S4 we further analyze the rate-distortion-complexity trade-off by modifying the number of diffusive sampling steps and the model size. The small (GIViC-S) and medium (GIViC-M) variants still demonstrate superior compression performance over VTM (RA) whilst achieving considerably improved decoding efficiency.

These results have also indicated the flexibility of GIViC in terms of supporting a diverse set of encoding and decoding complexities without much performance degradation, thanks to its pretrain-then-overfit scheme as well as the expressiveness of the proposed diffusion transformer backbone.

Encoding	GOP1	GOP2	GOP3	GOP4	Avg. Time	BD-rate (%)
w/ overfit.	1.78h	0.85h	0.84h	0.84h	16.56h	-15.89
w/o overfit.	0.06h	0.06h	0.06h	0.06h	1.15h	-12.27

Table S3. Per-GOP and whole sequence average convergence time on a single A100. BD-rate (%) is measured against VTM on UVG.

	BD-rate	Steps	Params (M)	Peak Memory (G)	Enc. FPS	Dec. FPS
Variants				Enc. / Dec.	A100 / 3090	A100 / 3090
GIViC-S	-3.77	4	79	4.5 / 1.9	3.51 / 1.62	18.90 / 12.88
GIViC-S	-4.00	8	79	4.5 / 1.9	3.51 / 1.62	14.86 / 8.71
GIViC-M	-8.01	4	135	11.5 / 4.3	1.01 / 0.41	15.55 / 10.94
GIViC-M	-8.99	8	135	11.5 / 4.3	1.01 / 0.41	12.21 / 5.99
Original	-15.94	8	226	22.8 / 8.5	0.03 / 0.01	9.79 / 3.45

Table S4. BD-rate (%, against VTM) and complexity results of GIViC variants on a single A100 and a single 3090 GPU, respectively.

C. Impact Statement

The development of neural video compression techniques can lead to significant economic, social, and environmental impacts. With video content comprising over 80% of total internet traffic, enhancing compression efficiency will substantially reduce storage demands and transmission costs. This, in turn, improves the scalability of video-based entertainment, real-time communication, and remote collaboration while also minimizing energy consumption and carbon emissions. By advancing video compression technologies, we can drive widespread economic savings, enhance digital connectivity, and contribute to environmental sustainability on a global scale.

References

[1] Adrian Browne, Yan Ye, and Seung Hwan Kim. Algorithm description for Versatile Video Coding and Test Model 19 (VTM 19). In *the JVET meeting*. ITU-T and ISO/IEC, 2023. 2

[2] Ge Gao, Ho Man Kwan, Fan Zhang, and David Bull. PNVC: Towards practical inr-based video compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3068–3076, 2025. 1

[3] Jingning Han, Bohan Li, Debargha Mukherjee, Ching-Han Chiang, Adrian Grange, Cheng Chen, Hui Su, Sarah Parker, Sai Deng, Urvang Joshi, et al. A technical overview of AV1. *Proceedings of the IEEE*, 109(9):1435–1462, 2021. 2

[4] Hyunjik Kim, Matthias Bauer, Lucas Theis, Jonathan Richard Schwarz, and Emilien Dupont. C3: High-performance and low-complexity neural compression from a single image or video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9347–9358, 2024. 1, 2

[5] Fabian Mentzer, George D Toderici, David Minnen, Sergi Caelles, Sung Jin Hwang, Mario Lucic, and Eirikur Agustsson. VCT: A video compression transformer. *Advances in Neural Information Processing Systems*, 35:13091–13103, 2022. 1