On the Choice of Perception Loss Function for Learned Video Compression

Buu Phan [*1] **Sadaf Salehkalaibar** [*1] **Jun Chen** [2] **Wei Yu** [1] **Ashish Khisti** [1]

Abstract

We study causal, low-latency, sequential video compression when the output is subjected to both a mean squared-error (MSE) distortion loss as well as a perception loss to target realism. Motivated by prior approaches, we consider two different perception loss functions (PLFs). The first, PLF-JD, considers the joint distribution (JD) of all the video frames up to the current one, while the second metric, PLF-FMD, considers the framewise marginal distributions (FMD) between the source and reconstruction. Using deep-learning based experiments, we demonstrate that the choice of PLF can have a significant effect on the reconstruction, especially at low-bit rates. In particular, while the reconstruction based on PLF-JD can better preserve the temporal correlation across frames, it also imposes a significant penalty in distortion compared to PLF-FMD and further makes it more difficult to recover from errors made in the earlier output frames. We also demonstrate that encoded representations generated by training a system to minimize the MSE (without requiring either PLF) can be transformed to a reconstruction satisfying the perfect perceptual quality based on FMD by changing the distortion at most with a factor of two. A similar argument holds for the PLF-JD for a class of encoders operating at low-rate regime. We validate our results using information-theoretic analysis and deep-learning based experiments on moving MNIST and KTH datasets.

[*]Equal contribution [1]Department of Electrical & Computer Engineering, University of Toronto, Ontario, Canada [2]Department of Electrical& Computer Engineering, McMaster University, Ontario, Canada. Correspondence to: Buu Phan <truong.phan@mail.utoronto.ca>, Sadaf Salehkalaibar <sadafs@ece.utoronto.ca>.

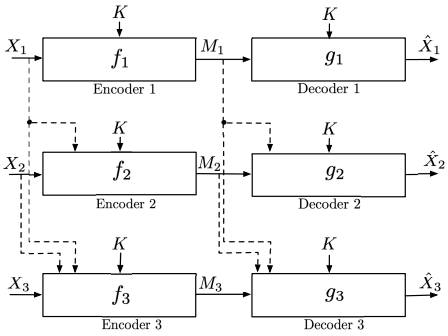
Accepted in Neural Compression Workshop in 40^{th} International Conference on Machine Learning, 2023.

1. Introduction

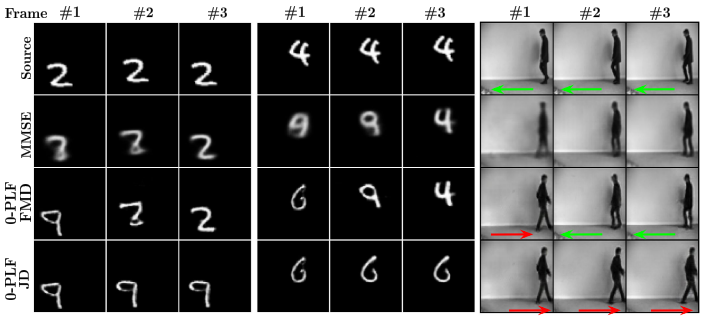
There is an increasing demand for video compression algorithms that are able to generate visually pleasing videos at low bitrates. Most of the current video codecs use distortion measures such as PSNR (Agustsson et al., 2020a; Yang et al., 2020; Rippel et al., 2021; Li et al., 2021a), MSE and MS-SSIM (Golinski et al., 2020; Rippel et al., 2021; Li et al., 2021a) to generate reconstructions which tend to be blurry at extremely low bitrates. In recent years, there has been a growing interest (see e.g., (Zhang et al., 2021b; Mentzer et al., 2022; Yang et al., 2021; Veerabadran et al., 2021; Wang et al., 2020)) in using deep generative models to make the reconstructions look more realistic. Such techniques introduce an additional perception loss function that measures a distance between distributions of the source and reconstruction, with *perfect* perception corresponding requiring that the two distributions be identical.

In compression systems, improving realism comes at the price of increasing distortion. The work of Blau and Michaeli (Blau & Michaeli, 2019) establishes the theoretical rate-distortion-perception (RDP) tradeoff which has also been validated in (Agustsson et al., 2019; Ballé et al., 2017; Theis et al., 2017; Mentzer et al., 2018). Furthermore *universal* encoded representations were proposed in (Zhang et al., 2021a) where the representation is fixed at the encoder and the decoder is adapted to achieve a performance near the optimal RDP tradeoff curve. The extension of these works to video compression involves many challenges. First, the compression system must not only account for spatial redundancy as in image compression, but also exploit the temporal redundancy across video frames, making the system design more complex. Secondly, unlike the case of image compression, there may be no clear choice of the perception loss function (PLF). Indeed, some prior works (Mentzer et al., 2022) consider PLF that preserves framewise marginal distribution (PLF-FMD) between the source and reconstruction, while other works consider joint distribution (PLF-JD) across multiple frames (Veerabadran et al., 2021).

As illustrated in Fig. 1a, we study causal, low-latency, sequential video compression when the output is subjected to both a mean squared-error (MSE) distortion loss and either a PLF-JD or PLF-FMD metric for perception loss.



(a) Encoders and decoders for T=3 frames.



(b) Effects of different PLFs on reconstructions for Moving MNIST and KTH datasets (best view in the monitor).

Figure 1. (a) Proposed System Model (b) Error permanence phenomenon under different PLF. High fidelity but incorrect I-frame reconstruction propagates the error to subsequent P-frames in 0-PLF-JD reconstructions. The MMSE and 0-PLF-FMD reconstructions do not have this problem.

Our main results are as follows. For general sources, we show that when using PLF-FMD, the MMSE reconstruction can be transformed to a reconstruction satisfying perfect perceptual quality by increasing the distortion at most by a *factor of two*. While a similar result does not hold for PLF-JD in general, it is satisfied for a special class of encoders which operate in the low-rate regime. On the experimental side, we demonstrate that while PLF-JD preserves better temporal consistency across video frames, it suffers from the *permanence of error* phenomenon in which the mistakes in reconstructions propagate to future frames [1]. On the other hand, the PLF-FMD metric shows more capability in correcting mistakes across frames (see Fig. 1b for visualizations involving three-frame videos).

2. System Model

Let $(X_1, \ldots, X_T) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_T$ be T frames in a video (with each $\mathcal{X}_i \subseteq \mathbb{R}^d$) distributed according to $P_{X_1 \ldots X_T}$. The frames are available for encoding sequentially; X_1 is available first, then X_2 arrives, followed by X_3 and so on. There is a shared randomness $K \in \mathcal{K}$ which is available at all encoders and decoders. The following (possibly stochastic) mappings define the encoding and decoding functions:

$$f_j \colon \mathcal{X}_1 \times \ldots \times \mathcal{X}_j \times \mathcal{K} \to \mathcal{M}_j, \qquad j = 1, \ldots, T, \tag{1}$$

$$g_j \colon \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_j \times \mathcal{K} \to \hat{\mathcal{X}}_j, \tag{2}$$

where $\mathcal{M}_j \in \{0,1\}^\star$ denotes the set of (variable-length) messages assigned by the jth encoder and $\hat{\mathcal{X}}_j \subseteq \mathbb{R}^d$ is the j-th reconstruction alphabet (see Fig. 1a). Let $P_{\hat{X}_1 \ldots \hat{X}_T | X_1 \ldots X_T}$ be the conditional distribution of the reconstructed video given the original video which is basically determined by the mappings $\{f_j\}_{j=1}^T$ and $\{g_j\}_{j=1}^T$. The above setting is a *one-shot* setup as only a single source sample is compressed at a time. For each frame j, a distortion metric is imposed on the output, which we assume throughout is the mean squared-error (MSE) function i.e. $d(x_j, \hat{x}_j) = ||x - \hat{x}_j||^2$, which is commonly used in many applications. The compression rate of the jth frame is defined to be $\mathbb{E}[\ell(M_j)]$ where $\ell(.)$ denotes the length of the message M_j. From a perceptual point of view, for given probability distributions $P_{X_1 \ldots X_j}$ and $P_{\hat{X}_1 \ldots \hat{X}_j}$ on the original and reconstructed frame j, let $\phi_j(P_{X_1 \ldots X_j}, P_{\hat{X}_1 \ldots \hat{X}_j})$ be the perception function capturing the difference between them. Note that the function ϕ_j is defined based on the joint distribution of all first j frames. We call this metric as *perception loss function based on joint distribution (PLF-JD)*. Note that when $\phi_j(P_{X_1 \ldots X_j}, P_{\hat{X}_1 \ldots \hat{X}_j}) = 0$, we have:

$$P_{X_1 \ldots X_j} = P_{\hat{X}_1 \ldots \hat{X}_j}, \qquad j = 1, \ldots, T. \tag{3}$$

We refer to this case as *zero-perception loss function based on joint distribution (*0*-PLF-JD)*. Alternatively, the *perception loss function based on framewise marginal distribution (PLF-FMD)* is denoted by $\xi_j(P_{X_j}, P_{\hat{X}_j})$ and is based on only the marginal distribution of the j-th frame. In particular, note that 0-PLF-FMD implies that $P_{X_j} = P_{\hat{X}_j}$ for each j.

In most of the paper, for simplicity of presentation, we provide some of our results for $T = 3$ frames. In that case, we use the shorthand notation X to denote the tuple (X_1, X_2, X_3), e.g., $\mathsf{M} := (M_1, M_2, M_3)$, $\mathsf{D} := (D_1, D_2, D_3)$, $\mathsf{f} := (f_1, f_2, f_3)$.

3. Distortion Analysis for a Fixed Encoder and Zero-perception Loss

In this section, we assume that the encoding functions f are fixed, but the decoding functions g can be optimized to

[1] Unlike the error propagation issue (Mentzer et al., 2020; Lu et al., 2020), the *permanence of error* phenomenon cannot be resolved by increasing the code rate assigned to the P frames.

generate different reconstructions. Equivalently, the distribution $P_{\mathsf{M}|\mathsf{X}K}{:=}\mathbb{1}\{\mathsf{M} = \mathsf{f}(\mathsf{X}, K)\}$ is fixed, while by varying the reconstruction distribution $P_{\hat{\mathsf{X}}|\mathsf{M}K}{:=}\mathbb{1}\{\hat{\mathsf{X}} = \mathsf{g}(\mathsf{M}, K)\}$, one attains different reconstructions $\hat{\mathsf{X}}$, where $\mathbb{1}\{.\}$ denotes the indicator function. Furthermore defining $D_j{:=}\mathbb{E}_P[\|X_j - \hat{X}_j\|^2]$, we denote D as the achievable distortion tuple associated with $P_{\hat{\mathsf{X}}|\mathsf{M}K}$.

One natural choice of reconstructions is the minimum mean squared error (MMSE) reconstruction function. At step j, the reconstruction, which we denote in this case by $\tilde{X}_j$, is obtained by taking the conditional expectation of X_j given all information at the decoder up to time j i.e., $\tilde{X}_j{:=}\mathbb{E}_P[X_j|M_1 \dots M_j, K]$ for each $j = 1, 2, 3$. It is well known that the MMSE reconstruction functions minimize the reconstruction distortion i.e., if we define the set

$$\Phi_{\mathsf{D}^{\min}}(P_{\mathsf{M}|\mathsf{X}K}) = \{\mathsf{D} : D_j \geq \mathbb{E}_P[\|X_j - \tilde{X}_j\|^2],\ \ j = 1, 2, 3\}, \quad (4)$$

then the distortion tuple D associated with any reconstruction $P_{\hat{\mathsf{X}}|\mathsf{M}K}$ satisfies $\mathsf{D} \in \Phi_{\mathsf{D}^{\min}}(P_{\mathsf{M}|\mathsf{X}K})$.

The main result of this section is that assuming fixed encoder, the achievable distortions under 0-PLF-FMD is at most twice of that under the MMSE distortion loss alone. The same conclusion also holds for 0-PLF-JD for a class of encoders operating at low rate. We first consider the case of 0-PLF-FMD.

Definition 3.1 (0-PLF-FMD Distortion)**.** For an encoder $P_{\mathsf{M}|\mathsf{X}K}$, the set $\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$ denotes the set of all distortion tuples D for which there exists a reconstruction $P_{\hat{\mathsf{X}}|\mathsf{M}K}$ satisfying $P_{X_j} = P_{\hat{X}_j}$ for each $j \in \{1, 2, 3\}$.

Theorem 3.2. *The set $\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$ is characterized as follows:*

$$\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K}) = \{\mathsf{D} : D_j \geq \mathbb{E}_P[\|X_j - \tilde{X}_j\|^2] + W_2^2(P_{\tilde{X}_j}, P_{X_j}),\ j = 1, 2, 3\}, \quad (5)$$

where $W_2^2(P_{X_j}, P_{\hat{X}_j})$ denotes the Wasserstein-2 distance between the two distributions (Panaretos & Zemel, 2020). Furthermore, we also have that:

$$\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K}) \supseteq \{\mathsf{D} : D_j \geq 2\mathbb{E}_P[\|X_j - \tilde{X}_j\|^2],\quad j = 1, 2, 3\}, \quad (6)$$

i.e., minimum achievable distortion with 0-PLF-FMD is at most twice the MMSE distortion.

Proof: See Appendix A. ∎

We remark that the proof of Theorem 3.2, operationally demonstrates that the MMSE reconstruction can be converted to another reconstruction satisfying 0-PLF-FMD with at-most a factor of 2 increase in distortion, generalizing the result in (Zhang et al., 2021a) for the single frame scenario (see also (Blau & Michaeli, 2018)).

We next consider the case when zero perception loss is satisfied under the PLF-JD metric. Analogous to $\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$ in Definition 3.1, one can define $\Phi^{\text{joint}}_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$ to be the set of distortions associated with reconstruction functions that satisfy (3). The analysis of $\Phi^{\text{joint}}_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$ is discussed in Appendix B as it is more involved. In general, the *factor of two bound* as in Theorem 3.2 cannot be realized in this case as demonstrated by a counter-example in Appendix B. Nevertheless, for a special family of encoders we can obtain a counterpart of Theorem 3.2. In this family of encoders, the source X_j at time j is nearly independent of the encoder outputs up to and including time j, i.e., we can express:

$$P^{\text{noisy}}_{X_j|M_1 \dots M_j K} = (1 - \mu)P_{X_j} + \mu Q^{\text{noisy}}_{X_j|M_1 \dots M_j K}, \quad j = 1, 2, 3. \quad (7)$$

where μ is a sufficiently small constant and the distribution $Q^{\text{noisy}}(\cdot)$ could be arbitrary conditional distribution with same marginal as P_{X_j}. We note that such encoders are studied in a variety of problems in information theory (see e.g., (Makur, 2019)) that correspond to the low rate operating regime. The following result states that the factor-two bound holds approximately for such encoders.

Theorem 3.3. *For the class of encoders given by (7), we have*

$$\Phi^{\textit{joint}}_{\mathsf{D}^0}(P^{\textit{noisy}}_{\mathsf{M}|\mathsf{X}K}) \supseteq \{\mathsf{D} : D_j \geq 2\mathbb{E}_{P^{\textit{noisy}}}[\|X_j - \tilde{X}_j\|^2] + O(\mu),\ j = 1, 2, 3\}. \quad (8)$$

Proof: See Appendix C. ∎

We note that the low-rate operating regime is practically important, as at higher rates MMSE based reconstructions can suffice and the use of PLF metrics may be less relevant.

4. Experiment

We conduct experiments on the MovingMNIST dataset (Srivastava et al., 2015) (with 1 digit) using Wasserstein GAN (Gulrajani et al., 2017), to verify the implications of our theoretical claims to perceptual video compression. Additional results on the KTH dataset (Schuldt et al., 2004) are available in Appendix D.2. Our compression network is built on the scale-space flow model (Agustsson et al., 2020b) and conditional module (Li et al., 2021b). Details about the architecture and training objectives are available in the Appendix D.1. The experimental setup is focused on validating our theory, rather than proposing state-of-the-art neural network architectures. Accordingly, we (1) validate Theorems 3.2 and 3.3, which characterize the factor-of-two

bounds on the distortion of 0-PLF reconstructions (2) empirically demonstrate the *error permanence* phenomenon of the PLF-JD.

Table 1. Distortions of optimal reconstructions at different regimes (✓means factor of 2 holds and ✗means otherwise). Distortion is scaled by 10^{-2}.

(a) Case 1: $R_1{=}\infty$ bits.

R_2	MMSE	0-PLF-FMD	0-PLF-JD
1	1.08 ± 0.01	1.74 ± 0.02 ✓	2.05 ± 0.03 ✓
2	0.88 ± 0.01	1.39 ± 0.03✓	1.46 ± 0.02 ✓
3.17	0.53 ± 0.01	0.76 ± 0.01 ✓	0.79 ± 0.01 ✓

(b) Case 2: $R_1{=}12$ bits(ϵ).

R_2	MMSE	0-PLF-FMD	0-PLF-JD
4	1.23 ± 0.01	2.21 ± 0.04✓	2.36 ± 0.04 ✓
8	1.04 ± 0.01	1.78 ± 0.03 ✓	2.28 ± 0.03 ✗
12	0.89 ± 0.02	1.43 ± 0.02 ✓	2.26 ± 0.03 ✗
∞	0.0	0.0 ✓	2.18 ± 0.02 ✗

As our first experimental result in Table 1, we validate the *factor of two bounds* in Theorems 3.2 and 3.3. We consider the compression of two frames X_1 and X_2 at rates R_1 and R_2 respectively. The compression of X_1 is performed without any prior reference and corresponds to the compression of the "I-frame", while the compression of X_2 corresponds to the "P-frame", using X_1 as the reference. We consider the cases when either $R_1{=}\infty$ or $R_1{=}12$ bits, where the former corresponds to lossless compression of X_1 and the latter corresponds to the low rate regime. The average distortion for the first frame when $R_1 = 12$ is 0.0124 for the MMSE reconstruction and 0.0235 for the 0-PLF reconstruction, thus satisfying the factor of two bound. In compression of X_2, we systematically vary the value of the rate $R_2{\in}\{4, 8, 12, \infty\}$. Following Table 1b, for 0-PLF-JD reconstruction, only $R_2{=}4$ bits (low rate) satisfies the factor of two bounds as expected. Intuitively, even as more bits are acquired, the 0-PLF-JD criteria actively restricts improving the reconstructions, resulting in persistently higher distortion. Even in the case when $R_2 = \infty$, the distortion remains non-zero as the decoder is forced to maintain temporal consistency with $\hat{X}_1$. In contrast, for FMD, the factor of 2 bound holds at all rates, consistent with Theorem 3.2.

In Fig. 2 and Fig. 3, we present our experimental results with a group of pictures (GOP) of size 3 (i.e. one I-frame followed by two P-frames). In Fig. 2, we visualize sample reconstructions for MSE, 0-PLF-FMD and 0-PLF-JD cases when operating in the low-rate regime with $R_j = 12$ bits for $j = 1, 2, 3$. Note that given an incorrect digit reconstruction in $\hat{X}_1$, the decoder with 0-PLF-JD consistently produces incorrect digits (or content) while the 0-PLF-FMD gradually "corrects" it, which is called as error permanence phenomenon. We also plot the framewise distortion in Fig. 3 to show the difference in achievable distortion of the two perception metrics across different values for R_2 and R_3

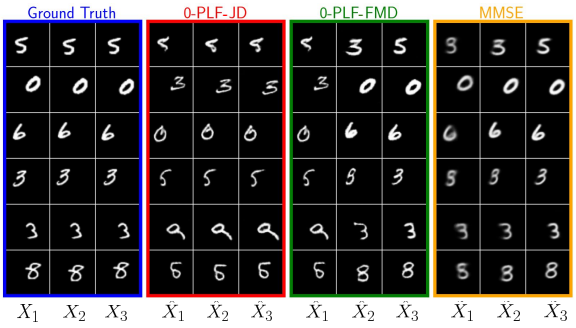


Figure 2. Permanence of Error Phenomenon. Ground-truth GOP and their optimal reconstructions with different PLFs for $R_1{=}R_2{=}R_3{=}12$ bits.

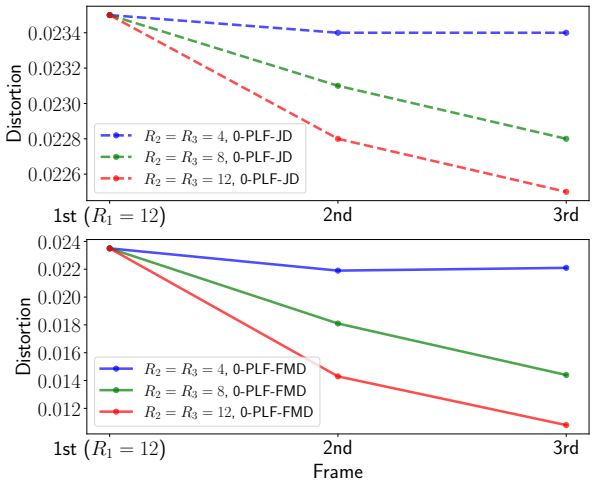


Figure 3. Permanence of Error Phenomenon. Distortion per frame $(X_i{-}\hat{X}_i)^2$ with 0-PLF-FMD and 0-PLF-JD reconstructions for different R_2, R_3 ($R_1{=}12$ bits for all the cases). When $R_2{=}R_3{=}\infty$, the distortions for the second and third frames are $2.18{\times}10^{-2}$ for 0-PLF JD and 0.0 for 0-PLF FMD.

as a function of the frame index. We note that the achievable distortion decreases much faster for 0-PLF-FMD than 0-PLF-JD for all selection of rates.

5. Conclusion

This work examines key theoretical properties of different perception loss functions, namely PLF-FMD and PLF-JD, for causal video coding. Our analysis highlights that while 0-PLF-JD reconstruction preserves temporal correlation, it is susceptible to the error permanence phenomenon. On the contrary, despite sacrificing temporal consistency, the 0-PLF-FMD reconstruction method effectively avoids this issue, ensuring that the reconstructed results are always confined within a factor of 2 from the MMSE reconstructions. We suggest future research directions such as exploring region-based perceptual metrics (Pergament et al., 2022), incorporating image-aware bits allocation, and leveraging conditional perception metric (Mentzer et al., 2020).

References

Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., and Van Gool, L. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 221–231, 2019.

Agustsson, E., Minnen, D., Johnston, N., Ballé, J., Hwang, S. J., and Toderici, G. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8503–8512, 2020a.

Agustsson, E., Minnen, D., Johnston, N., Balle, J., Hwang, S. J., and 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, 2020b.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. In *5th International Conference on Learning Representations*, 2017.

Blau, Y. and Michaeli, T. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

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

Golinski, A., Pourreza, R., Yang, Y., Sautiere, G., and Cohen, T. S. Feedback recurrent autoencoder for video compression. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, J., Li, B., and Lu, Y. Deep contextual video compression. In *Advances in Neural Information Processing Systems*, pp. 18114–18125, 2021a.

Li, J., Li, B., and Lu, Y. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34: 18114–18125, 2021b.

Lu, G., Cai, C., Zhang, X., Chen, L., Ouyang, W., Xu, D., and Gao, Z. Content adaptive and error propagation aware deep video compression. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 456–472. Springer, 2020.

Makur, A. *Information contraction and decomposition*. PhD Thesis, MIT, 2019.

Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., and Gool, L. V. Conditional probability models for deep image compression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Mentzer, F., Toderici, G., Tschannen, M., and Agustsson, E. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems*, 2020.

Mentzer, F., Agustsson, E., Ballé, J., Minnen, D., Johnston, N., and Toderici, G. Neural video compression using gans for detail synthesis and propagation. In *European Conference on Computer Vision*, 2022.

Panaretos, V. M. and Zemel, Y. *An invitation to statistics in Wasserstein space*. Springer, 2020.

Pergament, E., Tandon, P., Rippel, O., Bourdev, L., Anderson, A. G., Olshausen, B., Weissman, T., Katti, S., and Tatwawadi, K. Pim: Video coding using perceptual importance maps. *arXiv preprint arXiv:2212.10674*, 2022.

Rippel, O., Anderson, A. G., Tatwawadi, K., Nair, S., Lytle, C., and Bourdev, L. Elf-vc: Efficient learned flexible-rate video coding. 2021. URL https://arxiv.org/abs/2104.14335.

Schuldt, C., Laptev, I., and Caputo, B. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pp. 32–36. IEEE, 2004.

Srivastava, N., Mansimov, E., and Salakhudinov, R. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pp. 843–852. PMLR, 2015.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. Lossy image compression with compressive autoencoders. In *5th International Conference on Learning Representations*, 2017.

Veerabadran, V., Pourreza, R., Habibian, A., and Cohen, T. Adversarial distortion for learned video compression. 2021. URL https://arxiv.org/pdf/2004.09508.pdf.

Wang, Y., Bilinski, P., Bremond, F., and Dantcheva, A. G^3an: Disentangling appearance and motion for video generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Yang, R., Yang, Y., Marino, J., and Mandt, S. Hierarchical autoregressive modeling for neural video compression. 2020. URL https://arxiv.org/pdf/2010.10258.pdf.

Yang, R., Van Gool, L., and Timofte, R. Perceptual learned video compression with recurrent conditional gan. *arXiv preprint arXiv:2109.03082*, 1, 2021.

Zhang, G., Qian, J., Chen, J., and Khisti, A. Universal rate-distortion-perception representations for lossy compression. In *Advances in Neural Information Processing Systems*, pp. 11517–11529, 2021a.

Zhang, S., Mrak, M., Herranz, L., Blanch, M. G., Wan, S., and Yang, F. Dvc-p: Deep video compression with perceptual optimizations. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5. IEEE, 2021b.

A. Distortion Analysis for 0-PLF-FMD

Recall the definition of Wasserstein-2 distance (Panaretos & Zemel, 2020) as follows. For given distributions P_{X_j} and $P_{\tilde{X}_j}$, let

$$W_2^2(P_{\tilde{X}_j}, P_{X_j}) := \inf \mathbb{E}[\|X_j - \tilde{X}_j\|^2], \tag{9}$$

where the infimum is over all joint distributions of $(X_j, \tilde{X}_j)$ with marginals P_{X_j} and $P_{\tilde{X}_j}$.

Theorem A.1. *The set $\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$ is characterized as follows:*

$$\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K}) = \{\mathsf{D} : D_j \geq \mathbb{E}_P[\|X_j - \tilde{X}_j\|^2] + W_2^2(P_{\tilde{X}_j}, P_{X_j}),\ j = 1, 2, 3\}, \tag{10}$$

Furthermore, we also have that:

$$\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K}) \supseteq \{\mathsf{D} : D_j \geq 2\mathbb{E}_P[\|X_j - \tilde{X}_j\|^2], \quad j = 1, 2, 3\}, \tag{11}$$

i.e., minimum achievable distortion with 0-PLF-FMD is at most twice the MMSE distortion.

Proof: Define

$$\mathcal{D}^0 := \{\mathsf{D} : D_j \geq \mathbb{E}[\|X_j - \tilde{X}_j\|^2] + W_2^2(P_{\tilde{X}_j}, P_{X_j}),\ \ j = 1, 2, 3\}. \tag{12}$$

First, we show that $\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K}) \subseteq \mathcal{D}^0$. For any $\mathsf{D} \in \Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$, there exists $\hat{\mathsf{X}}_{\mathsf{D}^0} = (\hat{X}_{D_1^0}, \hat{X}_{D_2^0}, \hat{X}_{D_3^0})$ jointly distributed with $(\mathsf{M}, \mathsf{X}, K)$ such that

$$\mathbb{E}[\|X_j - \hat{X}_{D_j^0}\|^2] \leq D_j, \qquad j = 1, 2, 3, \tag{13}$$

$$P_{X_j} = P_{\hat{X}_{D_j^0}}. \tag{14}$$

Then, for example, the analysis for the second frame is as follows

$$\begin{aligned}
D_2 &\geq \mathbb{E}[\|X_2 - \hat{X}_{D_2^0}\|^2] && (15)\\
&= \mathbb{E}[\|(X_2 - \tilde{X}_2) - (\hat{X}_{D_2^0} - \tilde{X}_2)\|^2] && (16)\\
&= \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \mathbb{E}[\|\tilde{X}_2 - \hat{X}_{D_2^0}\|^2] && (17)\\
&\geq \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + W_2^2(P_{\tilde{X}_2}, P_{\hat{X}_{D_2^0}}) && (18)\\
&= \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + W_2^2(P_{\tilde{X}_2}, P_{X_2}), && (19)
\end{aligned}$$

where (17) holds because both $\tilde{X}_2$ and $\hat{X}_{D_2^0}$ are functions of (M_1, M_2, K) and thus the MMSE $(X_2 - \tilde{X}_2)$ is uncorrelated with $(\hat{X}_{D_2^0} - \tilde{X}_2)$; (19) follows because the 0-PLF-FMD implies that $P_{\hat{X}_{D_2^0}} = P_{X_2}$. Following similar steps for other frames, we get $\Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K}) \subseteq \mathcal{D}^0$.

Next, we show that $\mathcal{D}^0 \subseteq \Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$. Assume that $\mathsf{D} \in \mathcal{D}^0$. Let $\hat{X}_1^*$ be an auxiliary random variable jointly distributed with (M_1, K) such that it satisfies the following conditions

$$P_{\hat{X}_1^*} = P_{X_1}, \tag{20}$$

and

$$P_{\tilde{X}_1 \hat{X}_1^*} = \arg \inf_{\substack{\bar{P}_{\tilde{X}_1 \hat{X}_1^*}: \\ \bar{P}_{\tilde{X}_1} = P_{\tilde{X}_1} \\ \bar{P}_{\hat{X}_1^*} = P_{\hat{X}_1^*}}} \mathbb{E}_{\bar{P}}[\|\tilde{X}_1 - \hat{X}_1^*\|^2]. \tag{21}$$

Moreover, let $\hat{X}_2^*$ be an auxiliary random variable jointly distributed with (M_1, M_2, K) such that the following two conditions are satisfied

$$P_{\hat{X}_2^*} = P_{X_2}, \tag{22}$$

and

$$P_{\tilde{X}_2\hat{X}_2^*} = \arg \inf_{\substack{\bar{P}_{\tilde{X}_2\hat{X}_2^*}: \\ \bar{P}_{\tilde{X}_2}=P_{\tilde{X}_2} \\ \bar{P}_{\hat{X}_2^*}=P_{\hat{X}_2^*}}} \mathbb{E}_{\bar{P}}[\|\tilde{X}_2 - \hat{X}_2^*\|^2]. \tag{23}$$

Similarly, we define $\hat{X}_3^*$. Now, notice that since $\mathsf{D} \in \mathcal{D}^0$, we have:

$$D_2 \geq \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + W_2^2(P_{\tilde{X}_2}, P_{X_2}). \tag{24}$$

It then directly follows that

$$\begin{aligned}
\mathbb{E}[\|X_2 - \hat{X}_2^*\|^2] &= \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \mathbb{E}[\|\tilde{X}_2 - \hat{X}_2^*\|^2] && (25)\\
&= \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + W_2^2(P_{\tilde{X}_2}, P_{\hat{X}_2^*}) && (26)\\
&= \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + W_2^2(P_{\tilde{X}_2}, P_{X_2}) && (27)\\
&\leq D_2, && (28)
\end{aligned}$$

where

- (25) follows because $\tilde{X}_2$ and $\hat{X}_2^*$ are functions of (M_1, M_2, K) and thus the MMSE $(X_2 - \tilde{X}_2)$ is uncorrelated with $(\hat{X}_2^* - \tilde{X}_2)$;
- (26) follows from (23);
- (27) follows because $P_{\hat{X}_2^*} = P_{X_2}$.

Following similar steps for other frames, we get $\mathsf{D} \in \Phi_{\mathsf{D}^0}(P_{\mathsf{X_r}|\mathsf{X}})$.

Now, notice that $W_2^2(P_{\tilde{X}_2}, P_{X_2}) \leq \mathbb{E}[\|X_2 - \tilde{X}_2\|^2]$ since the Wasserstein-2 distance takes the infimum over all possible joint distributions $(X_2, \tilde{X}_2)$, but the expectation in $\mathbb{E}[\|X_2 - \tilde{X}_2\|^2]$ is taken over the given $P_{X_2\tilde{X}_2}$. Thus, we get

$$\mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + W_2^2(P_{\tilde{X}_2}, P_{X_2}) \leq 2\mathbb{E}[\|X_2 - \tilde{X}_2\|^2]. \tag{29}$$

This concludes the proof. ■

B. Distortion Analysis for 0-PLF-JD

Let $\hat{X}_1^*$ be defined as in (20)–(21). Moreover, let $\hat{X}_2^*$ be an auxiliary random variable jointly distributed with (M_1, M_2, K) such that the following conditions are satisfied

$$P_{\hat{X}_2^*|\hat{X}_1^*=x_1} = P_{X_2|X_1=x_1}, \qquad \forall x_1 \in \mathcal{X}_1, \tag{30}$$

and

$$P_{\tilde{X}_2\hat{X}_2^*|\hat{X}_1^*=x_1} = \arg \inf_{\substack{\bar{P}_{\tilde{X}_2\hat{X}_2^*|\hat{X}_1^*=x_1}: \\ \bar{P}_{\tilde{X}_2|\hat{X}_1^*=x_1}=P_{\tilde{X}_2|\hat{X}_1^*=x_1} \\ \bar{P}_{\hat{X}_2^*|\hat{X}_1^*=x_1}=P_{\hat{X}_2^*|\hat{X}_1^*=x_1}}} \mathbb{E}_{\bar{P}}[\|\tilde{X}_2 - \hat{X}_2^*\|^2|\hat{X}_1^* = x_1], \qquad \forall x_1 \in \mathcal{X}_1. \tag{31}$$

Then, the following result holds.

Theorem B.1. *We have*

$$\begin{aligned}
\Phi_{\mathsf{D}^0}^{\textit{joint}}(P_{\mathsf{M}|\mathsf{X}K}) \supseteq \{\mathsf{D} : & D_1 \geq \mathbb{E}[\|X_1 - \tilde{X}_1\|^2] + W_2^2(P_{\tilde{X}_1}, P_{X_1}),\\
& D_2 \geq \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \sum_{x_1} P_{X_1}(x_1) W_2^2(P_{\tilde{X}_2|\hat{X}_1^*=x_1}, P_{X_2|X_1=x_1}),\\
& D_3 \geq \mathbb{E}[\|X_3 - \tilde{X}_3\|^2] + \sum_{x_1,x_2} P_{X_1X_2}(x_1,x_2) W_2^2(P_{\tilde{X}_3|\hat{X}_1^*=x_1,\hat{X}_2^*=x_2}, P_{X_3|X_1=x_1,X_2=x_2})\}.
\end{aligned} \tag{32}$$

Proof: Define

$$\begin{aligned}\mathcal{D}^0_{\text{joint}} := \{\mathsf{D} : D_1 &\geq \mathbb{E}[\|X_1 - \tilde{X}_1\|^2] + W_2^2(P_{\tilde{X}_1}, P_{X_1}),\\ D_2 &\geq \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \sum_{x_1} P_{X_1}(x_1) W_2^2(P_{\tilde{X}_2|\hat{X}_1^*=x_1}, P_{X_2|X_1=x_1}),\\ D_3 &\geq \mathbb{E}[\|X_3 - \tilde{X}_3\|^2] + \sum_{x_1,x_2} P_{X_1X_2}(x_1,x_2) W_2^2(P_{\tilde{X}_3|\hat{X}_1^*=x_1,\hat{X}_2^*=x_2}, P_{X_3|X_1=x_1,X_2=x_2})\}.\end{aligned} \tag{33}$$

Now, assume that $\mathsf{D} \in \mathcal{D}^0_{\text{joint}}$. For the first frame, recall that $\hat{X}_1^*$ is an auxiliary random variable jointly distributed with (M_1, K) such that it satisfies (20)–(21). From similar steps to (25)–(27), it then follows that

$$\mathbb{E}[\|X_1 - \hat{X}_1^*\|^2] = \mathbb{E}[\|X_1 - \tilde{X}_1\|^2] + W_2^2(P_{\tilde{X}_1}, P_{X_1}) \tag{34}$$

$$\leq D_1. \tag{35}$$

For the second frame, since $\mathsf{D} \in \mathcal{D}^0_{\text{joint}}$, we have:

$$D_2 \geq \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \sum_{x_1} P_{X_1}(x_1) W_2^2(P_{\tilde{X}_2|X_1=x_1}, P_{X_2|X_1=x_1}). \tag{36}$$

Recall that $\hat{X}_2^*$ is an auxiliary random variable jointly distributed with (M_1, M_2, K) such that (30)–(31) hold. It then directly follows that

$$\mathbb{E}[\|X_2 - \hat{X}_2^*\|^2] = \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \mathbb{E}[\|\tilde{X}_2 - \hat{X}_2^*\|^2] \tag{37}$$

$$= \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \sum_{x_1} P_{\hat{X}_1^*}(x_1) \mathbb{E}[\|\tilde{X}_2 - \hat{X}_2^*\|^2 | \hat{X}_1^* = x_1] \tag{38}$$

$$= \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \sum_{x_1} P_{\hat{X}_1^*}(x_1) W_2^2(P_{\tilde{X}_2|\hat{X}_1^*=x_1}, P_{\hat{X}_2^*|\hat{X}_1^*=x_1}) \tag{39}$$

$$= \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \sum_{x_1} P_{X_1}(x_1) W_2^2(P_{\tilde{X}_2|\hat{X}_1^*=x_1}, P_{X_2|X_1=x_1}), \tag{40}$$

where

- (37) follows because $\tilde{X}_2$ and $\hat{X}_2^*$ are functions of (M_1, M_2, K) and thus the MMSE $(X_2 - \tilde{X}_2)$ is uncorrelated with $(\hat{X}_2^* - \tilde{X}_2)$,
- (39) follows from (31),
- (40) follows because $P_{\hat{X}_1^*\hat{X}_2^*} = P_{X_1X_2}$.

Following similar steps for the third frame, we get $\mathsf{D} \in \Phi_{\mathsf{D}^0}(P_{\mathsf{M}|\mathsf{X}K})$. This concludes the proof. ■

B.1. A Counterexample for Factor-Two Bound in Case of 0-PLF-JD

Assume that we have only two frames, i.e., $D_3 \to \infty$. Let M_1 be independent of X_1 and $M_2 = X_2$. Then, we have $\tilde{X}_1 = \emptyset$ and $\tilde{X}_2 = X_2$. Consider the achievable distortion region of Theorem B.1. The distortion of the first step is given by the following

$$\mathbb{E}[\|X_1 - \tilde{X}_1\|^2] + W_2^2(P_{\tilde{X}_1}, P_{X_1}) = 2\mathbb{E}[X_1^2]. \tag{41}$$

For the second frame, we have

$$\begin{aligned}&\mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \sum_{x_1} P_{X_1}(x_1) W_2^2(P_{\tilde{X}_2|\hat{X}_1^*=x_1}, P_{X_2|X_1=x_1})\\ &= \sum_{x_1} P_{X_1}(x_1) W_2^2(P_{X_2|\hat{X}_1^*=x_1}, P_{X_2|X_1=x_1}) \qquad (42)\\ &= \sum_{x_1} P_{X_1}(x_1) W_2^2(P_{X_2}, P_{X_2|X_1=x_1}), \qquad (43)\end{aligned}$$

where (42) follows because $\tilde{X}_2 = X_2$ and (43) follows because X_2 is independent of $\hat{X}_1^*$ (M_1 is independent of X_1, then $\hat{X}_1^*$, which is a function of (M_1, K), would be independent of X_1 and hence independent of X_2).

Now, notice that the MMSE distortion of the second step is zero since $\tilde{X}_2 = X_2$. However, the achievable distortion of the second step for the reconstruction satisfying 0-PLF JD is given in (43) which clearly does not satisfy the factor-two bound.

C. Fixed Encoders Operating at Low rate regime

We consider the class of noisy encoders where the encoder distribution can be written as follows

$$P^{\text{noisy}}_{X_j|M_1\dots M_j K} = (1-\mu)P_{X_j} + \mu Q^{\text{noisy}}_{X_j|M_1\dots M_j K}, \qquad j = 1, 2, 3. \tag{44}$$

where μ is a sufficiently small constant and the distribution $Q^{\text{noisy}}(\cdot)$ could be arbitrary conditional distribution with same marginal as P_{X_j}.

Theorem C.1. *For the class of encoders given by (44), we have*

$$\Phi^{joint}_{\mathsf{D}^0}(P^{noisy}_{\mathsf{M}|\mathsf{X}K}) \supseteq \{\mathsf{D} : D_j \geq 2\mathbb{E}_{P^{noisy}}[\|X_j - \tilde{X}_j\|^2] + O(\mu), \quad j = 2, \dots, 3\}. \tag{45}$$

Proof: We analyze the distortion for the second frame. A similar argument holds for other frames.

Denote the reconstruction of the second step by $\hat{X}_2^*$ and consider the expected distortion. From a similar justification starting from (15) and leading to (17), we can write the distortion as follows

$$\mathbb{E}[\|X_2 - \hat{X}_2^*\|^2] = \mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + \mathbb{E}[\|\tilde{X}_2 - \hat{X}_2^*\|^2]. \tag{46}$$

Now, we study the expected term $\mathbb{E}[\|\tilde{X}_2 - \hat{X}_2^*\|^2]$ as follows

$$\mathbb{E}[\|\tilde{X}_2 - \hat{X}_2^*\|^2] = \sum_{x_1} P_{\hat{X}_1^*}(x_1)\mathbb{E}[\|\tilde{X}_2 - \hat{X}_2^*\|^2 | \hat{X}_1^* = x_1]. \tag{47}$$

In order to analyze the above expression, we first approximate the MMSE reconstruction $\tilde{X}_2$ as follows

$$\begin{aligned}
\tilde{X}_2 &= \mathbb{E}_{P^{\text{noisy}}}[X_2|M_1, M_2, K] && (48)\\
&= (1-\mu)\mathbb{E}_P[X_2] + \mu\mathbb{E}_{Q^{\text{noisy}}}[X_2|M_1, M_2, K] && (49)\\
&= \mathbb{E}[X_2] + O(\mu), && (50)
\end{aligned}$$

where (49) follows from (44). Moreover, notice that (50) implies that

$$\begin{aligned}
\mathbb{E}[\|X_2 - \tilde{X}_2\|^2] &= \mathbb{E}[\|X_2 - \mathbb{E}[X_2] + \mu(\mathbb{E}_{Q^{\text{noisy}}}[X_2|M_1, M_2, K] - \mathbb{E}[X_2])\|^2] && (51)\\
&= \mathbb{E}[\|X_2 - \mathbb{E}[X_2]\|^2] + O(\mu). && (52)
\end{aligned}$$

Next, consider the expected term in (47) as follows

$$\begin{aligned}
\sum_{x_1} P_{\hat{X}_1^*}(x_1)\mathbb{E}[\|\tilde{X}_2 - \hat{X}_2^*\|^2 | \hat{X}_1^* = x_1] &= \sum_{x_1} P_{\hat{X}_1^*}(x_1)\mathbb{E}[\|\mathbb{E}[X_2] - \hat{X}_2^*\|^2 | \hat{X}_1^* = x_1] + O(\mu) && (53)\\
&= \sum_{x_1} P_{\hat{X}_1^*}(x_1)\mathbb{E}[\|\mathbb{E}[X_2] - X_2\|^2 | X_1 = x_1] + O(\mu) && (54)\\
&= \sum_{x_1} P_{X_1}(x_1)\mathbb{E}[\|\mathbb{E}[X_2] - X_2\|^2 | X_1 = x_1] + O(\mu) && (55)\\
&= \mathbb{E}[\|\mathbb{E}[X_2] - X_2\|^2] + O(\mu) && (56)\\
&= \mathbb{E}[\|\tilde{X}_2 - X_2\|^2] + O(\mu), && (57)
\end{aligned}$$

where

- (53) follows from (50);
- (54) follows because the 0-PLF-JD implies that $P_{\hat{X}_2^*|\hat{X}_1^*} = P_{X_2|X_1}$ and $\mathbb{E}[X_2]$ is just a constant;
- (55) follows from 0-PLF-JD where $P_{\hat{X}_1^*} = P_{X_1}$;
- (57) follows from (52).

Considering (46) and (57), we get

$$\mathbb{E}[\|X_2 - \hat{X}_2^*\|^2] = 2\mathbb{E}[\|X_2 - \tilde{X}_2\|^2] + O(\mu). \tag{58}$$

The proof for the third frame follows similar steps. ■

D. Experiment Details

D.1. Training Setup and Overview

Our compression architecture is built on the scale-space flow model (Agustsson et al., 2020b), which allows end-to-end training without relying on pre-trained optical flow estimators. For better compression efficiency, we replace the residual compression module with the conditioning one (Li et al., 2021b). In the following, we will interchangeably refer X_1 as the I-frame and subsequent ones as P-frames. The annotation for the encoder, decoder, and critic (discriminator) will be referred to as f, g, and h respectively and their specific functionality (e.g motion compression, joint perception critic) will be described within context through a subscript/superscript.

Distortion and Perception Measurement: We follow the setup in prior works (Blau & Michaeli, 2018; Zhang et al., 2021a) for distortion and perception measurement. Specifically, we use MSE loss $\mathbb{E}[||X-\hat{X}||^2]$ as a distortion metric and Wasserstein-1 distance as a perception metric, which can be estimated through the WGAN critics (following the Kanotorovich-Rubinstein duality). For the marginal perception metric, we optimize our critics h_m to classify between original image X and synthetic ones $\hat{X}$. This will then allow us to measure $W_1(P_X, P_{\hat{X}})$ since:

$$W_1(P_X, P_{\hat{X}}) = \sup_{h_m \in \mathcal{F}} \mathbb{E}[h_m(X)] - \mathbb{E}[h_m(\hat{X})] \tag{59}$$

where $\mathcal{F}$ is a set of all bounded 1-Lipschitz functions. Similarly, the joint perception metric is realized through $W_1(P_{X_1...X_j}, P_{\hat{X}_1...\hat{X}_j})$ by training a critic h_j that classifies between synthetic and authentic sequences:

$$W_1(P_{X_1...X_j}, P_{\hat{X}_1...\hat{X}_j}) = \sup_{h_j \in \mathcal{F}} \mathbb{E}[h_j(X_1, ..., X_i)] - \mathbb{E}[h_j(\hat{X}_1, ..., \hat{X}_i)] \tag{60}$$

In practice, the set of 1-Lipschitz functions is limited by the neural network architecture. Also, although our analysis employs the Wasserstein-2 distance as a perception metric, it is worth noting that the ideal reconstructions (0-PLF) for this metric and the one used in our study should be identical.

I-frame Compressor: We compress I-frames in a similar fashion as previous works (Blau & Michaeli, 2018; Zhang et al., 2021a). Our encoder f_I and decoder g_I contain a series of convolution operations and we control the rate R_1 by varying the dimension and quantization level in the bottleneck. The model utilizes common randomness through the dithered quantization operation. For a given rate R_1, we vary the amount of DP tradeoff by controlling the hyper-parameter $\lambda_i^{\text{marginal}}$ in the following minimization objective $\mathcal{L}_1$:

$$\mathcal{L}_1 = \mathbb{E}[||X_1 - \hat{X}_1||^2] + \lambda_i^{\text{marginal}} W_1(P_{X_1}, P_{\hat{X}_1}) \tag{61}$$

Following the results from Zhang et al.(Zhang et al., 2021a), we fix the encoder after optimizing the encoder-decoder pair for MSE representations. We then fix the encoder and train another decoder to obtain the optimal reconstruction with perfect perception, i.e, $W_1(P_X, P_{\hat{X}}) \approx 0$. This gives us the benefit of obtaining reconstructions at different DP tradeoffs (MMSE and 0-PLF FMD) for the I-frame.

P-frame Compressor: We describe the loss functions before explaining our architectures. Given previous reconstructions $\hat{X}_{[i-1]}:=\{\hat{X}_1, \hat{X}_2, ..., \hat{X}_{i-1}\}$, one can adjust the distortion-joint perception tradeoff by controlling the hyper-parameter λ_i^{joint} in the following objective $\mathcal{L}_i$.

$$\mathcal{L}_i^{\text{joint}} = \mathbb{E}[||X_i - \hat{X}_i||^2] + \lambda_i^{\text{joint}} W_1(P_{X_{[i]}}, P_{\hat{X}_{[i]}}) \tag{62}$$

Note that in order to achieve 0-PLF-JD, previous reconstructions $\hat{X}_{[i-1]}$ must also achieve 0-PLF-JD, since it is impossible to reconstruct such $\hat{X}_i$ if the previous $\hat{X}_{[i-1]}$ are not temporally consistent[2]. For the FMD metric, we use the loss function in (61).

D.2. Permanence of Error on KTH Datasets

The KTH dataset is a widely-used benchmark dataset in computer vision research, consisting of video sequences of human actions performed in various scenarios. We show more examples supporting our argument for the permanence of error on this realistic dataset. We use 16 bits for each frame. In general, the 0-PLF-JD decoder consistently outputs correlated but incorrect reconstructions due to the error induced by the first reconstructions, i.e., the P-frames will follow the wrong direction induced from the I-frame reconstruction. Besides the moving direction, we also notice that the type of actions (i.e. walking, jogging, and running) is also affected. On the other hand, while losing some temporal cohesion, MMSE and 0-PLF FMD decoders manage to fix the movement error.

[2]This follows from the inequality: $W_2^2(P_{X_1,X_2}, P_{\hat{X}_1,\hat{X}_2}) \geq W_2^2(P_{X_1}, P_{\hat{X}_1}) + W_2^2(P_{X_2}, P_{\hat{X}_2})$

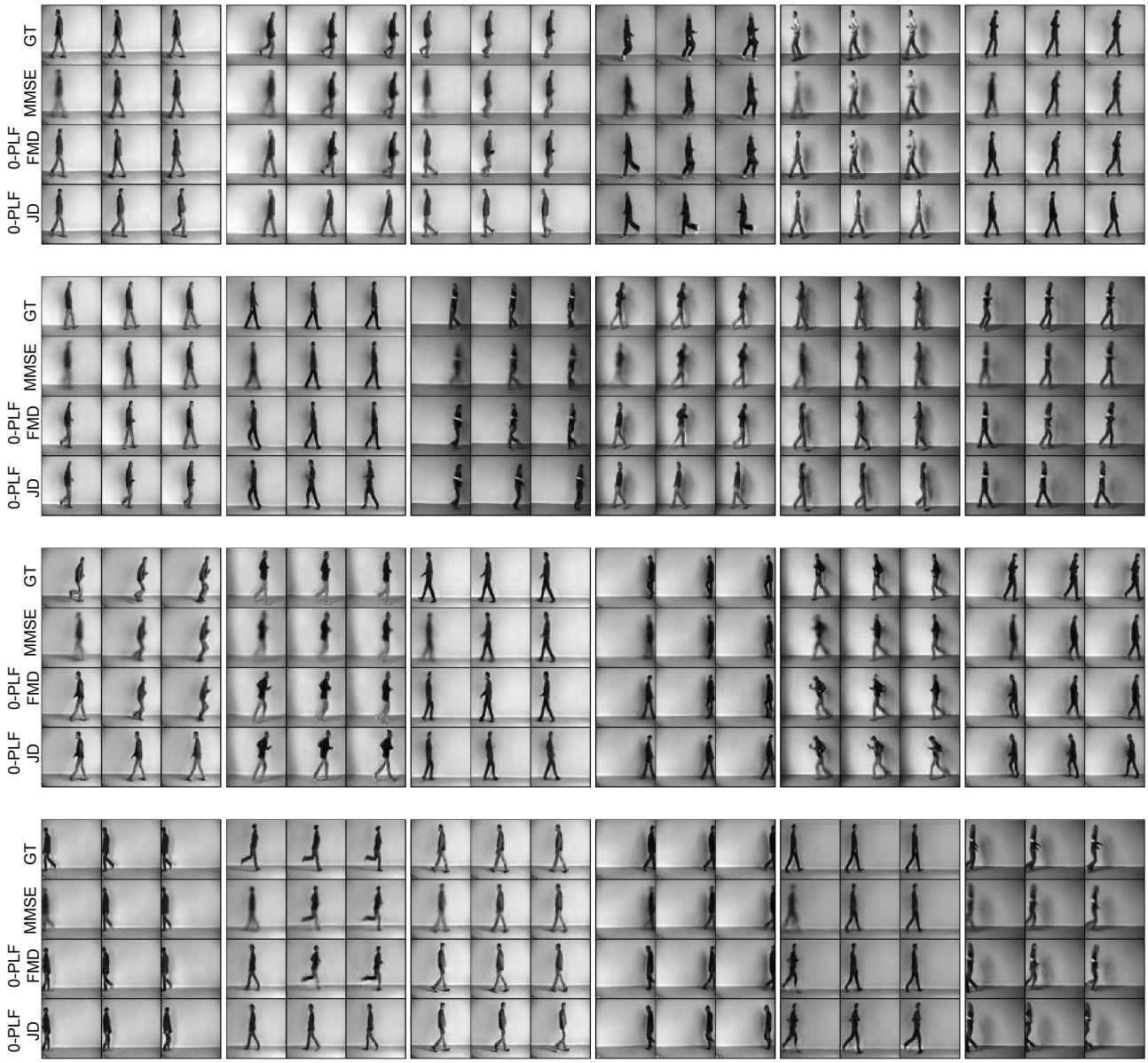


Figure 4. Additional Experimental Results for the Permanence of Error Phenomenon on KTH Dataset.