VIDEO COMPRESSION UNDER LIMITED DELAY: THEORETICAL OPTIMALITY AND SYSTEM DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional neural video compression (NVC) methods predominantly follow strictly causal predictive-coding paradigm, treating zero-frame delay as a fixed constraint rather than delay as an explicit design parameter. From an information-theoretic perspective, allowing limited frame delay enables a codec to exploit correlations with future frames within the delay window, thereby reducing uncertainty that is inaccessible to strictly causal prediction and potentially improving the rate-distortion tradeoff. In this work, we formalize this intuition through a rigorous rate–distortion analysis builds on the clique factorization of Markov networks, and propose *Graph-Factored Sequential Coding* (GFSC), a principled framework for sequential video coding under limited delay. To our knowledge, we provide the first theoretical result showing that, for general K-order Markov sources with positive distributions, a K-frame delay is sufficient for GFSC to attain joint-coding rate-distortion optimality over a nontrivial distortion region, a guarantee unavailable to zero-delay causal coding in general. In practice, we instantiate GFSC through lightweight neural modules, allowing existing causal predictive codecs to be augmented without modifying their core architectures. Experiments across multiple NVC models and datasets show that GFSC consistently improves rate-distortion performance over the corresponding causal counterparts. Under matched delay budgets, GFSC also achieves competitive performance compared with representative bidirectional codecs.

1 INTRODUCTION

Video compression can be fundamentally viewed as a sequential lossy source coding problem (Viswanathan & Berger, 2002; Yang et al., 2011; Ma et al., 2007). Most existing neural video compression (NVC) models follow a strictly *causal predictive coding* paradigm (Agustsson et al., 2020; Lu et al., 2019; 2020; Li et al., 2021; Mentzer et al., 2022), where each frame is encoded using only previously reconstructed frames and therefore cannot exploit future correlations. For temporally correlated video, future context provides predictive information about the current frame and reduces its reconstruction uncertainty, potentially improving the rate-distortion tradeoff. In addition, allowing limited frame delay does not necessarily incur substantial additional latency. At 60 fps, a single-frame lookahead introduces only 16.67 ms of buffering latency, which is modest relative to the inference latency of modern NVC models and tolerable in many practical scenarios. Frame delay should therefore be treated as an explicit design parameter rather than fixed a priori to zero.

The information-theoretic basis of limited delay coding has been studied through *delayed sequential coding* (Ma et al., 2007). In particular, for K-order Gauss-Markov sources, a K-frame delay is sufficient to attain joint-coding rate-distortion optimality over a nonempty distortion region, a guarantee that zero-delay causal coding does not generally enjoy. However, it remains unclear whether this optimality extends beyond Gaussian sources, and the current theory does not provide a constructive architectural principle for translating a given delay budget into a practical coding system. Existing bidirectional codecs (Wu et al., 2018; Djelouah et al., 2019; Xu et al., 2024; Yang et al., 2020; Yılmaz & Tekalp, 2021) exploit future frames in practice, but typically rely on fixed anchor-based structures designed for relatively large delay budgets, thereby inefficient in the limited-delay regime of interest. Consequently, a unified framework that characterizes limited-delay optimality for general Markov sources while enabling flexible practical NVC remains largely underexplored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

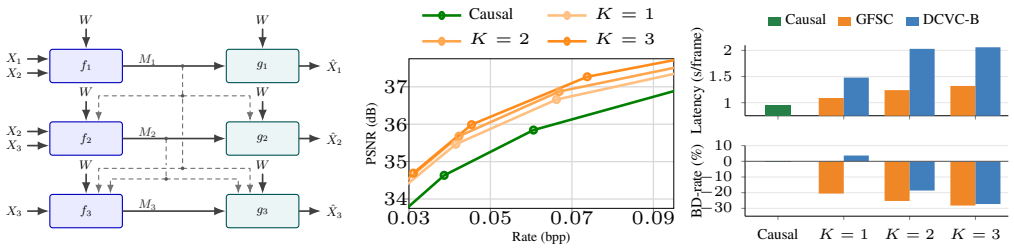


Figure 1: *Left*: GFSC-1 coding architecture for three frames. *Center*: Rate-PSNR comparison between zero-delay DCVC and GFSC with delays $K = 1, 2, 3$ on UVG. *Right*: Latency (top. Lower the better) and BD-rate (bottom. Lower the better) of GFSC and DCVC-B at matched delays, with zero-delay DCVC as the baseline. BD-rate is measured relative to zero-delay DCVC.

In this work, we propose *Graph Factored Sequential Coding* (GFSC), a principled framework for sequential coding under limited delay. To our knowledge, we are the first to prove that, for general K-order Markov sources with positive distributions, a K-frame delay is sufficient to attain joint-coding rate-distortion optimality over a nontrivial distortion region. Our analysis builds on probabilistic graphical models, where K-order Markov dependencies factorize into local clique-wise interactions. For K-order Gauss-Markov sources, the conditional independence structure further reduces to pairwise interactions with linear interpolation sufficient statistics. Figure 1 (left) illustrates the resulting GFSC architecture under one-frame delay. Although real videos are not assumed to follow an exact Markov model, their temporal dependencies are strongly local. As shown in Figure 2, the estimated inverse covariance of video features is approximately banded, with the strongest dependencies concentrated within a small temporal neighborhood. This empirical locality further motivates exploiting short, configurable delays.

We show that GFSC can be implemented using lightweight neural modules that fuse delayed latent representations and can be integrated into existing causal NVC models with minimal architectural changes. Experiments across multiple models, datasets, and delay settings show consistent rate-distortion improvements over causal baselines and competitive performance against representative B-frame codecs under matched delay budgets. The main contributions of this work are summarized as follows:

1. We propose GFSC, a flexible delayed sequential coding framework that uses clique-wise interactions architecture to systematically exploit a configurable frame-delay budget.
2. We establish rigorous rate-distortion guarantees for GFSC. For general K-order Markov sources, we prove that K-delay GFSC attains joint-coding rate-distortion optimality over a nontrivial distortion region. For K-order Gauss-Markov sources, the clique-wise construction further reduces to pairwise linear interpolation.
3. We realize GFSC using lightweight neural modules that can be integrated into existing causal NVC models with minimal architectural changes. Experiments demonstrate consistent rate-distortion gains over causal baselines and competitive performance against representative B-frame codecs under matched delay budgets.

2 Preliminaries

2.1 Joint coding

Consider an L-frame sequence $\mathbf{X} \triangleq (X_1, \cdots, X_L)$ where $X_\ell \in \mathcal{X}_\ell$ and $\mathbf{X} \sim P_{\mathbf{X}}$. In joint coding (JC), all frames are available prior to compression. We adopt a *one-shot* setup (Li & El Gamal, 2018), where a single source sequence $\mathbf{x}$ is compressed at one time. Given that encoder and decoder share the common randomness $W \in \mathcal{W}$ independent of source, a JC system is specified by

$$f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_L \times \mathcal{W} \to \mathcal{M}, \qquad g : \mathcal{M} \times \mathcal{W} \to \hat{\mathcal{X}}_1 \times \cdots \times \hat{\mathcal{X}}_L,$$

where the message $M \in \mathcal{M} = \{0,1\}^*$ is a binary string with expected length $\mathbb{E}[l(M)]$. The reconstruction $\hat{\mathbf{X}}$ follows $P_{\hat{\mathbf{X}}|\mathbf{X}}(\cdot|\mathbf{x})$, i.e., $\hat{\mathbf{X}} = g(f(\mathbf{x}, W), W) \sim P_{\hat{\mathbf{X}}|\mathbf{X}}(\cdot|\mathbf{x})$. The reconstruction

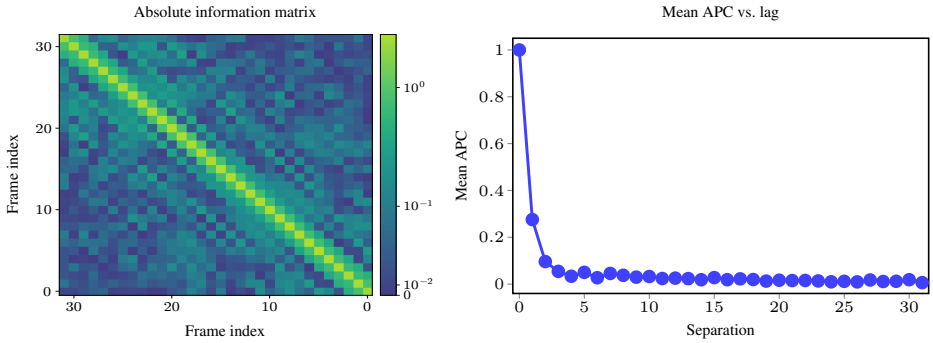


Figure 2: Temporal-dependence analysis for UVG dataset. *Left*: Estimated information matrix for 32-frame UVG video clips. *Right*: Mean APC vs. Separation curve. APC denotes Absolute partial correlation. Separation denotes the distance between two frame positions, i.e. $|j-k|$ for (X_j, X_k).

fidelity for each X_ℓ is measured by a distortion function $d_\ell : \mathcal{X}_\ell \times \mathcal{X}_\ell \to [0, \infty)$, with $d_\ell(x, x) = 0$ and $d_\ell(x, w) > 0$ for $w \neq x$. According to Li & El Gamal (2018), there exists a one-shot coding scheme such that minimum achievable $R^{\mathrm{JC},o}$ satisfies:

$$R^{\mathrm{JC},o} \leq I(\mathbf{X}; \hat{\mathbf{X}}) + \mathcal{O}(\log I(\mathbf{X}; \hat{\mathbf{X}})). \tag{1}$$

where the mutual information $I(\mathbf{X}; \hat{\mathbf{X}})$ is the theoretical optimal solution in the asymptotic setting (infinite blocklength), i.e., JC rate-distortion function.

Theorem 2.1 (JC rate-distortion function (Csiszár & Körner, 2011))**.** *The single-letter JC rate-distortion function is*

$$R^{\mathrm{JC}}(\mathbf{D}) = \min_{P_{\hat{\mathbf{X}}|\mathbf{X}} \in \mathcal{P}} I(\mathbf{X}; \hat{\mathbf{X}}), \tag{2}$$

where $\mathcal{P} \triangleq \{P_{\hat{\mathbf{X}}|\mathbf{X}} : \mathbb{E}[d_\ell(X_\ell, \hat{X}_\ell)] \leq D_\ell,\ \ell = 1, \cdots, L\}$.

2.2 Delayed Sequential Coding

In delayed sequential coding (DSC), source frames are observed sequentially with a limited look-ahead of K frames ($K < L$). Following Ma et al. (2007), there are three canonical setups. We focus on the noncausal-causal (NC-C) setting, which is most relevant to NVC, and defer the causal-noncausal (C-NC) and noncausal-noncausal (NC-NC) settings to Appendix A.1. An NC-C coding architecture is specified by:

$$f_\ell : \mathcal{X}_1 \times \cdots \times \mathcal{X}_{\min\{\ell+K, L\}} \times \mathcal{W} \to \mathcal{M}_\ell, \qquad g_\ell : \mathcal{M}_1 \times \cdots \times \mathcal{M}_\ell \times \mathcal{W} \to \hat{\mathcal{X}}_\ell,$$

where $M_\ell \in \mathcal{M}_\ell$ denotes ℓ-th encoder message. Under K-delay relaxation, the ℓ-th encoder has access up to $X_{\ell+K}$. The reconstruction $\hat{X}_\ell$ follows $P_{\hat{X}_\ell|X^{\ell+K}}(\cdot|x^{\ell+K})$ determined by $\{f_j, g_j\}_{j=1}^{\ell}$. A key result in Ma et al. (2007) shows that, for K-order Gauss-Markov sources, a K-frame delay is sufficient for DSC to asymptotically attain the joint-coding rate-distortion function in the high-resolution regime. However, this leaves two gaps. First, the result is established for the Gauss-Markov setting and does not characterize whether the same K-delay optimality extends to general K-order Markov sources. Second, DSC specifies the admissible information flow and its fundamental rate-distortion limits, but does not provide an architectural principle for realizing these limits in practical video coding systems.

In this work, we address both issues. We first establish the joint-coding optimality of K-delay sequential coding for a broader class of K-order Markov sources through a probabilistic graphical model analysis. We then introduce GFSC as a constructive realization of this structure.

3 Graph Factored Sequential Coding

We introduce *Graph Factored Sequential Coding* (GFSC), inspired by Markov networks. We adopt a one-shot setting and analyze the associated rate-distortion performance.

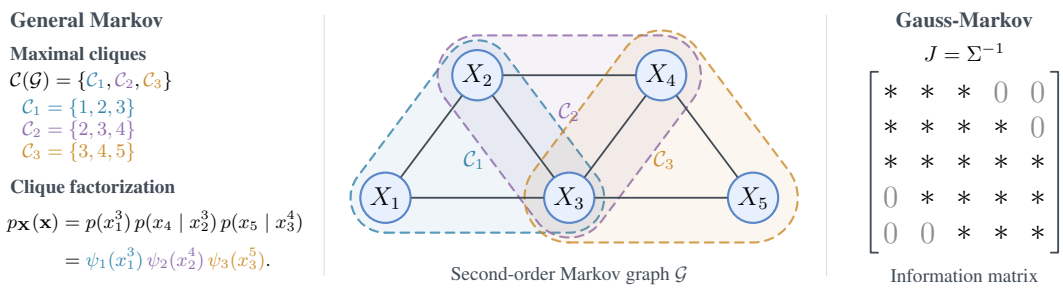


Figure 3: Second-order Markov structure illustration. *Left*: Maximal cliques and clique factorization. *Middle*: Second-order Markov network. *Right*: In the Gaussian case, the information matrix J satisfies $J_{ij} = 0$ for $|i - j| > 2$. "$*$" denotes potentially nonzero entries.

3.1 Markov Networks

Definition 3.1 (K-order Markov)**.** A K-order Markov source is a random vector $\mathbf{X}$ satisfying Markov chains $X_\ell \to (X_{\ell+1}, \cdots, X_{\ell+K}) \to (X_{\ell+K+1}, \cdots, X_L)$ for all ℓ.

Definition 3.2 (Clique)**.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. A nonempty subset $\mathcal{C} \subseteq \mathcal{V}$ is a *clique* if $\{i, j\} \in \mathcal{E}$ for every pair of distinct nodes $i, j \in \mathcal{C}$. A clique is *maximal* if it is not strictly contained in any other clique of $\mathcal{G}$.

Markov networks. A Markov network represents the conditional-independence structure of a joint distribution through an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node denotes a random variable (Koller & Friedman, 2009). Let $\mathcal{C}(\mathcal{G})$ be the set of maximal cliques of $\mathcal{G}$. For a strictly positive distribution $p_{\mathbf{X}}$, the Hammersley-Clifford theorem gives the clique factorization of probability density/mass function $p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{Z} \prod_{\mathcal{C}_j \in \mathcal{C}(\mathcal{G})} \psi_j(x_{\mathcal{C}_j})$, where ψ_j is a positive clique factor and Z a normalization constant. For the K-order Markov sources as in Definition 3.1, the graph has a simple temporal structure. Its maximal cliques are the consecutive $(K+1)$-variable windows, i.e. $\mathcal{C}_j \triangleq \{j, \ldots, j+K\}$. More importantly, the K-order Markov property itself directly yields a clique-wise factorization, without requiring the positivity condition:

$$p_{\mathbf{X}}(\mathbf{x}) = p(x^{K+1}) \prod_{j=2}^{L-K} p(x_{j+K} \mid x_j^{j+K-1}) \triangleq \prod_{j=1}^{L-K} \psi_j(x_{\mathcal{C}_j}), \tag{3}$$

where $\psi_1(x_{\mathcal{C}_1}) \triangleq p(x^{K+1})$ and $\psi_j(x_{\mathcal{C}_j}) \triangleq p(x_{j+K} | x_j^{j+K-1})$ for $j \geq 2$. Hence, the dependency structure of a K-order Markov source decomposes into local interactions over $(K+1)$-variable cliques. Figure 3 (*Left*) shows the clique factorization for a 2-order Markov example. This observation motivates the construction of GFSC introduced in Section 3.2.

Gaussian graphical models (GGMs). GGMs specialize Markov networks to Gaussian sources and have been exploited in multi-terminal source coding (Chen et al., 2017; 2020; Xie et al., 2020). For a Gaussian source $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, let $J \triangleq \Sigma^{-1}$ be its information matrix. The induced GGM $\mathcal{G}(J)$ has edge set $\mathcal{E} = \{\{i, j\} : i \neq j,\ J_{ij} \neq 0\}$. For a K-order Gauss-Markov source, J is K-banded, as illustrated in Figure 3 (*Right*). Moreover, the Gaussian density admits the pairwise factorization

$$p_{\mathbf{X}}(\mathbf{x}) \propto \prod_{\ell=1}^{L} \exp\left(-\frac{1}{2} J_{\ell\ell} x_\ell^2\right) \prod_{\{i,j\} \in \mathcal{E}} \exp\left(-J_{ij} x_i x_j\right). \tag{4}$$

This additional structure leads to the edge-wise and node-wise auxiliary construction and, ultimately, the pairwise interpolation form derived in Section 3.5.

3.2 Architecture of Coding Scheme

Let $\mathbf{X} = (X_1, \cdots, X_L)$ be a sequentially observed source and K the allowed delay relaxation ($K < L - 1$). We define the encoding and decoding mappings of GFSC-K architecture as follows.

Encoding / decoding. At each stage ℓ, the encoder observes at most K-delay source window $X_\ell^{\ell+K}$ and previous messages $M^{\ell-1}$. It then produces a new message M_ℓ. The ℓ-th decoder reconstructs X_ℓ using all messages available up to stage ℓ. The encoding and decoding mappings are given by:

$$f_\ell : \mathcal{X}_\ell^{\min\{\ell+K,L\}} \times \mathcal{M}^{\ell-1} \times \mathcal{W} \to \mathcal{M}_\ell, \qquad g_\ell : \mathcal{M}^\ell \times \mathcal{W} \to \hat{\mathcal{X}}_\ell. \tag{5}$$

There are L decoding stages for L source frames respectively. Compared with the NC-C DSC-K architecture, this formulation imposes a local-window structure. Figure 1 (*left*) and Figure 4 (*Right*) illustrate the structure for $K = 1, 2$. Now we characterize the operational rate-distortion region.

Definition 3.3 (Operational RD region)**.** A rate-distortion tuple $(\mathbf{R}, \mathbf{D})$ is achievable for one-shot GFSC-K system, if there exist encoder and decoder as defined above such that

$$R_\ell \geq \mathbb{E}[l(M_\ell)], \qquad D_\ell \geq \mathbb{E}[d_\ell(X_\ell, \hat{X}_\ell)], \tag{6}$$

where $M_\ell \in \mathcal{M}_\ell$ denotes the ℓ-th encoder message. The operational RD region $\mathcal{R}_K^o$ is the closure of the set of all achievable $(\mathbf{R}, \mathbf{D})$. For a given $\mathbf{D}$, the rate region $\mathcal{R}_K^o(\mathbf{D})$ consists of all $\mathbf{R}$ such that $(\mathbf{R}, \mathbf{D}) \in \mathcal{R}_K^o$, and the minimum sum-rate function is $R_{K,\text{sum}}^{\text{GFSC},o}(\mathbf{D}) = \min_{\mathbf{R}\in\mathcal{R}_K^o(\mathbf{D})} \sum R_\ell$.

3.3 Rate-distortion analysis

The operational region $\mathcal{R}_K^o$ cannot be directly computed as it involves all possible one-shot encoders and decoders. Here we provide a tractable approximation in terms of mutual information, which corresponds to the information RD region. We introduce auxiliary variable U_ℓ, which serves as a single-letter representation of the information conveyed by the ℓ-th encoder message.

Definition 3.4 (Information RD region)**.** For a given source $\mathbf{X}$, the single-letter information RD region of a GFSC-K system, denoted by $\mathcal{R}_K$, consists of all $(\mathbf{R}, \mathbf{D})$ satisfying

$$R_\ell \geq I(X_\ell^{\min\{\ell+K,L\}}; U_\ell | U^{\ell-1}), \qquad D_\ell \geq \mathbb{E}[d_\ell(X_\ell, \hat{X}_\ell)], \tag{7}$$

for some auxiliary random variables $\mathbf{U}$ and $\hat{\mathbf{X}}$ satisfying $\hat{X}_\ell = g_\ell(U^\ell)$ and the Markov conditions:

$$U_\ell \to (X_\ell^{\min\{\ell+K,L\}}, U^{\ell-1}) \to \mathbf{X}, \tag{8}$$

where $g_\ell(\cdot)$ are deterministic functions. For a given $\mathbf{D}$, the information rate region $\mathcal{R}_K(\mathbf{D})$ is the closure of all $\mathbf{R}$ such that $(\mathbf{R}, \mathbf{D}) \in \mathcal{R}_K$, and the information minimum sum-rate function is $R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) = \min_{\mathbf{R}\in\mathcal{R}_K(\mathbf{D})} \sum R_\ell$.

Remark 3.1 (Role of the auxiliary variables)**.** The auxiliary variable U_ℓ serves as a single-letter representation of the information generated at the ℓ-th GFSC stage. For a K-order Markov source, the local window observed by this stage is naturally associated with the local clique $\mathcal{C}_\ell$ of the underlying Markov network. Thus, U_ℓ can be viewed as a compressed representation of this clique, while $U^{\ell-1}$ carries information accumulated from preceding overlapping cliques. The Markov condition in equation 8 ensures that this representation respects the sequential information flow of GFSC.

Theorem 3.5 (Rate-distortion characterization)**.** *For a given distortion tuple* $\mathbf{D}$ *and any* $\mathbf{R} \in \mathcal{R}_K(\mathbf{D})$*, the operational rate region* $\mathcal{R}_K^o(\mathbf{D})$ *satisfies* $\mathbf{R} + \log(\mathbf{R}+1) + 5 \in \mathcal{R}_K^o(\mathbf{D}) \subseteq \mathcal{R}_K(\mathbf{D})$*. Moreover, the operational sum-rate function satisfies*

$$R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) \leq R_{K,\text{sum}}^{\text{GFSC},o}(\mathbf{D}) \leq R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) + \mathcal{O}\big(\log R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D})\big). \tag{9}$$

Proof. See Appendix B.1. □

Theorem 3.5 connects the information sum-rate function to the operational GFSC-K architecture up to a logarithmic one-shot overhead. We next identify a particular class of clique-wise auxiliary variables that satisfies the GFSC constraints together with stronger Markov conditions. For K-order Markov sources with positive distribution, this construction suffices to show that the GFSC information sum-rate coincides with the JC rate-distortion function over a nonempty distortion region.

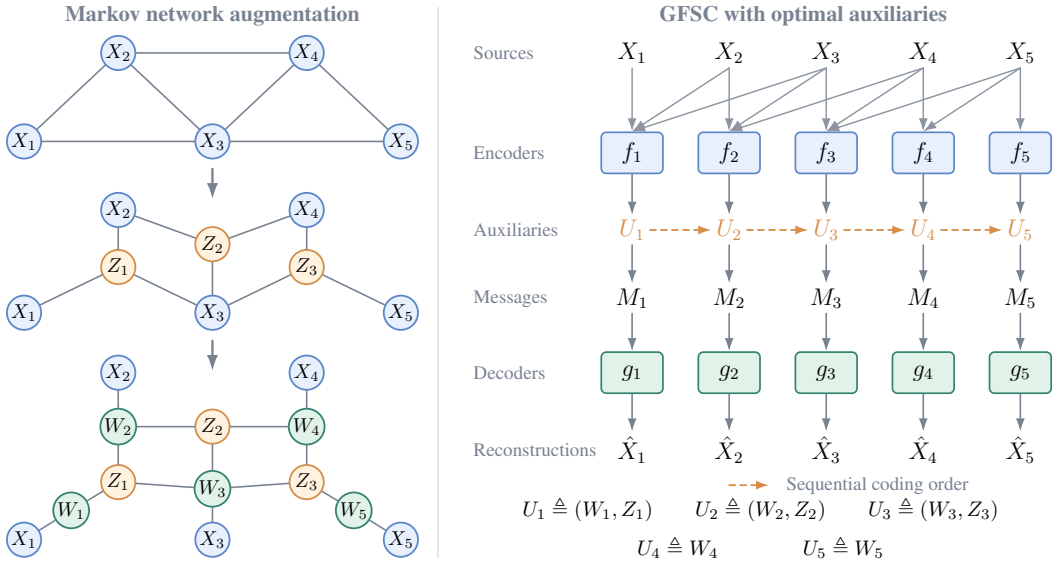


Figure 4: *Left*: A second-order Markov network (top), its augmentation by clique auxiliary variables Z_r (middle), and its further augmentation by node auxiliary variables W_j (bottom). *Right*: A 5-stage GFSC architecture with delay $K = 2$ and optimal auxiliary variables constructed in *Left*.

3.4 Joint-Coding Optimality for Markov Sources

3.4.1 General Markov sources

As discussed in Section 3.1, a K-order Markov source factorizes over consecutive cliques $\mathcal{C}_j = \{j, \dots, j+K\}$. This graphical structure motivates a clique-wise auxiliary construction: each auxiliary variable is generated from the source variables within one local clique, while the collection of overlapping auxiliaries captures the global source dependence through local interactions. For the optimality proof, we construct a particular class of such auxiliaries satisfying stronger Markov conditions than those required by GFSC. These conditions preserve the relevant dependency structure while remaining admissible under the GFSC-K information-flow constraints.

Definition 3.6 (K-order Markov source with positive distribution)**.** Let $\mathbf{X}$ be a K-order Markov source as in Definition 3.1. Moreover, each X_j has finite alphabet $2 \le |\mathcal{X}_j| < \infty$ and the joint distribution is positive, i.e. $p_{\mathbf{X}}(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_L$.

Theorem 3.7 (JC optimality for Markov sources)**.** *Consider a K-order Markov source $\mathbf{X}$ as in Definition 3.6. There exists a non-empty distortion region $\mathcal{D}_\epsilon \triangleq \{\mathbf{D} : 0 < D_\ell \le \epsilon_\ell,\ \ell = 1, \cdots, L\}$, such that for all $\mathbf{D} \in \mathcal{D}_\epsilon$, $R^{\mathrm{GFSC}}_{K,\mathrm{sum}}(\mathbf{D})$ achieves the JC rate-distortion function in Theorem 2.1:*

$$R^{\mathrm{GFSC}}_{K,\mathrm{sum}}(\mathbf{D}) = R^{\mathrm{JC}}(\mathbf{D}), \quad \forall \mathbf{D} \in \mathcal{D}_\epsilon. \tag{10}$$

Proof. See Appendix B.2. □

Remark 3.2 (Graphical interpretation)**.** To our knowledge, Theorem 3.7 is the first rigorous result showing that K-delay coding can attain JC rate-distortion optimality for general K-order Markov sources. The key is the locality of the Markov network: clique-wise auxiliary variables compress local dependencies while preserving the conditional-independence structure required for optimal coding. Figure 4 explicitly illustrates this construction, where clique-wise auxiliary variables Z_ℓ induce the desired conditional independencies and node-wise auxiliary variables W_ℓ further refine individual reconstructions. This locality is also consistent with real video, whose estimated inverse covariance in Figure 2 is largely banded, with the strongest dependencies concentrated within about three off-diagonals. In practice, we show that these local clique-wise interactions can be efficiently realized in NVC using learnable neural modules.

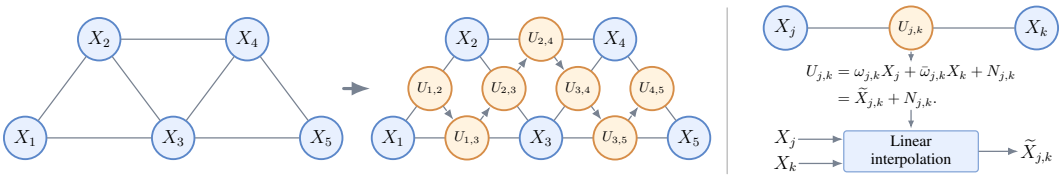


Figure 5: *Left*: A second-order GGM and its augmented representation within $\mathcal{D}_\epsilon^{\mathrm{G}}$; dashed arrows indicate sequential coding. *Right*: Edge-wise auxiliary construction by linear interpolation and additive noise, where $\bar{\omega}_{j,k} = 1 - \omega_{j,k}$.

3.5 Gauss-Markov sources under MSE

The general result above relies on clique-wise interactions of the underlying Markov network. Under the Gauss-Markov assumption and MSE distortion, the graphical structure further decomposes into edge-wise interactions as discussed in Section 3.1, allowing the clique-wise auxiliaries to be replaced by pairwise Gaussian auxiliaries.

Theorem 3.8 (Gauss-Markov specialization)**.** *Consider a K-order Gauss-Markov source $\mathbf{X}$ under MSE distortion. There exists a nonempty distortion region $\mathcal{D}_\epsilon^{\mathrm{G}}$ such that*

$$R_{K,\mathrm{sum}}^{\mathrm{GFSC-G}}(\mathbf{D}) = R^{\mathrm{JCG}}(\mathbf{D}), \qquad \forall\, \mathbf{D} \in \mathcal{D}_\epsilon^{\mathrm{G}}. \tag{11}$$

Moreover, the optimal clique-wise construction can be realized using pairwise Gaussian auxiliaries associated with the edges of the underlying GGM, with linear interpolations as sufficient statistics.

Proof. See Appendix C.4. □

Hence, Gaussianity is not required for GFSC, but provides a particularly simple interpretation of its optimal auxiliary structure: $(K+1)$-variable clique interactions reduce to ordered pairwise interactions within the K-frame neighborhood. Figure 5 illustrates this decomposition for a second-order Gauss-Markov source. Representative characterizations of $\mathcal{D}_\epsilon^{\mathrm{G}}$ for first-order Gauss-Markov example are provided in Appendix C.5.

4 Experimental Results

We conduct experiments on several NVC models based on causal predictive coding, including the DCVC (Li et al., 2021), video compression transformer (VCT) (Mentzer et al., 2022), and DHVC (Lu et al., 2024). We compare the rate-distortion performance of the GFSC extension and its causal counterparts. We further compare against both neural and conventional delayed coding baselines (bidirectional) under the same delay constraints. For quantitative comparison, we report BD-rate results on UVG (Mercat et al., 2020) and MCL-JCV (Wang et al., 2016).

Clique-wise interaction. For a K-delay setting, GFSC operates on the latent clique $(\mathbf{y}_\ell, \ldots, \mathbf{y}_{\ell+K})$, where $\mathbf{y}_\ell = f_\ell(\mathbf{x}_\ell)$ denotes the latent representation of frame $\mathbf{x}_\ell$. We implement this interaction using a window-based spatial-temporal attention module Liu et al. (2022), which jointly models the current and lookahead latent representations within each local spatial window. This enables configurable K-delay fusion without modifying the underlying analysis transforms and naturally handles boundary cases with fewer than K available future frames. Architectural details are provided in Appendix D.2.

4.1 Rate-distortion analysis

Causal counterparts. We evaluate GFSC on three causal NVC models. Figure 6 shows the rate-PSNR curves for the baselines and their GFSC-enhanced variants with delays $K = 1, 2, 3$. Across all models, incorporating the delay window interaction consistently improves rate-distortion performance, with gains increasing as K grows. Notably, even a single-frame delay yields significant improvement. Table 1 further quantifies the BD-rate improvements. Across all datasets and models, GFSC achieves consistent bitrate reductions relative to causal baselines, demonstrating the effectiveness of limited delay for improving rate-distortion performance. .

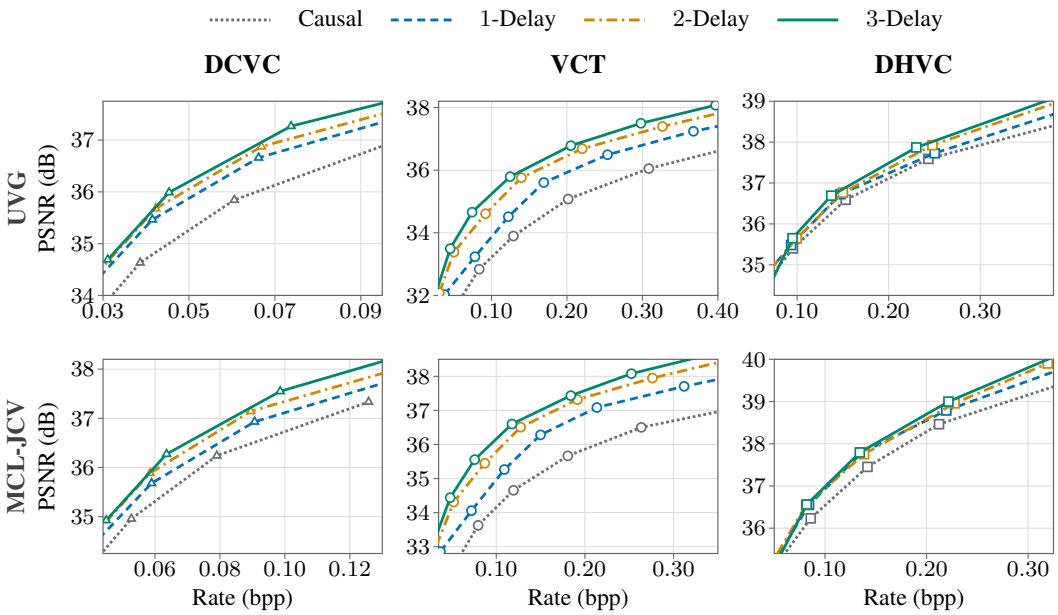


Figure 6: Rate-PSNR comparison between causal and GFSC-enhanced models on UVG and MCL-JCV dataset. Columns show DCVC, VCT, and DHVC, respectively.

Delayed baselines. In Fig. 7, we compare GFSC with delayed coding baselines, including DCVC-B (Sheng et al., 2025) and HM (*random access*), under delay budgets $K \leq 3$. For DCVC-B, we evaluate $K = 1, 2, 3$, corresponding to `I-B-I`, `I-B-B-I`, and `I-B-B-B-I`, respectively. For HM, due to its supported GOP structures, we evaluate $K = 1, 3$ using `I-B-I` and `I-B-B-B-I`. We also include DCVC-L, a widened DCVC baseline with a parameter count matched to DCVC-GFSC, to isolate the gain from delayed modeling and that due to increased model capacity. GFSC consistently achieves better rate-distortion performance than both the delayed and capacity-matched baselines. The corresponding BD-rate are reported in Table 2. We also provide 4K video dataset results in Appendix D.3.

Table 1: BD-rate (%) of GFSC-enhanced models relative to causal baselines.

	UVG (1080p)			MCL-JCV (1080p)		
Delay	DCVC	VCT	DHVC	DCVC	VCT	DHVC
$K = 1$	-20.57	-28.38	-7.39	-10.90	-32.35	-13.35
$K = 2$	-25.29	-48.00	-10.17	-17.32	-50.14	-14.65
$K = 3$	-28.22	-55.78	-11.91	-20.30	-56.70	-16.07

Complexity analysis. Table 2 compares DCVC-GFSC with causal DCVC and DCVC-B. GFSC introduces only modest additional latency over the causal baseline as the delay budget increases, while consistently improving rate-distortion performance. Under matched delay constraints, GFSC remains substantially faster than DCVC-B and also achieves better BD-rate performance. These results show that GFSC provides a more efficient use of limited delay, offering stronger compression gains with considerably lower latency than bidirectional coding.

5 Related Work

Neural video compression. Recent NVC methods have achieved rate-distortion performance competitive with conventional video codecs. Most existing approaches follow a strictly causal predictive coding paradigm, in which each frame is encoded using only previously reconstructed frames. Early methods adopt motion compensation followed by residual coding (Lu et al., 2019; Agustsson et al., 2020; Lin et al., 2020), whereas more recent approaches directly model temporal dependencies in the latent space through learned conditional coding (Li et al., 2021; Sheng et al., 2022; Li et al., 2023; 2024; Mentzer et al., 2022). Although these methods support zero-frame delay coding, they cannot exploit correlations with future frames. Bidirectional codecs address this limitation through interpolation-based prediction (Wu et al., 2018; Djelouah et al., 2019; Xu et al., 2024) or bidirectional conditional modeling (Yang et al., 2020; Yılmaz & Tekalp, 2021; Sheng et al., 2025). However, they typically rely on fixed anchor-frame structures and are designed for relatively large frame

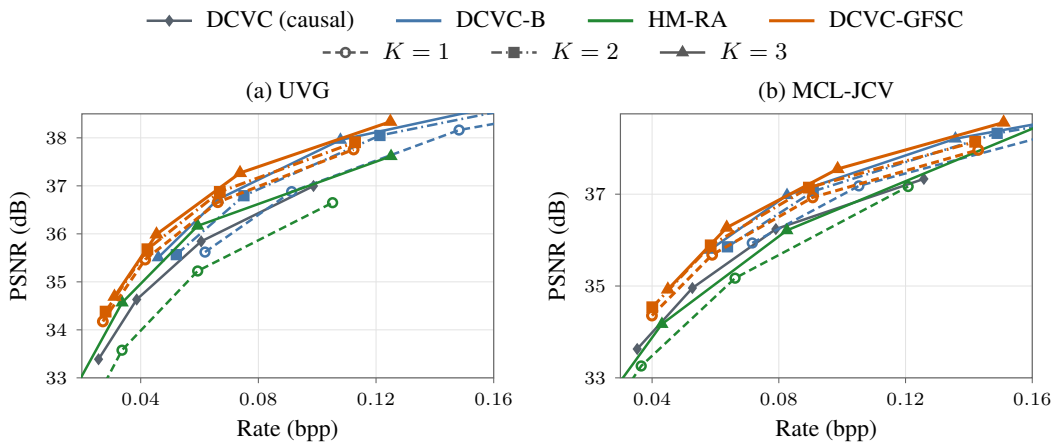


Figure 7: Rate-PSNR comparison under delay $K \leq 3$ on (a) UVG and (b) MCL-JCV. Colors identify model families. Line styles and markers distinguish delay settings.

Table 2: BD-rate (%) and computational complexity under matched delay budgets. Latency, memory, FLOPs, and model parameters are reported for DCVC, DCVC-GFSC, and DCVC-B.

Method	Delay K	BD-rate (%)	Enc. Latency (s/frame)	Dec. Latency (s/frame)	Memory (GB)	FLOPs (G/frame)	Params (M)
DCVC	0	0.0	0.52	0.44	5.54	4430.53	34.53
DCVC-L	0	-8.57	0.53	0.44	5.54	4447.51	39.21
DCVC-B	1	3.70	0.84	0.64	10.95	7153.37	55.38
	2	-18.64	1.14	0.89	10.95	8805.79	55.38
	3	-27.24	1.16	0.90	10.95	9632.01	55.38
DCVC-GFSC	1	-20.57	0.65	0.44	5.52	5916.28	39.21
	2	-25.29	0.80	0.44	5.52	7451.81	39.21
	3	-28.22	0.88	0.44	5.52	8987.58	39.21

delays, making them less flexible under small delay budgets. More recently, Xiang et al. (2025) incorporated 1-frame delay into a unified intra/inter coding framework. Their method, however, is specialized to the 1-delay setting and does not provide either a general construction for arbitrary delay budgets or an information-theoretic justification of the resulting rate-distortion gains.

Delayed sequential coding. The information-theoretic foundations of delayed sequential coding have been studied in Ma et al. (2007); Ma & Ishwar (2007), showing that limited delay can improve the rate-distortion performance of sequential coding and even a single-frame delay holds potential for matching the JC optimality in specific cases, e.g. first-order Gauss-Markov sources. However, these results do not establish whether such optimality extends to general K-order Markov sources, nor do they provide a constructive coding architecture for practical video compression.

6 Conclusion

We introduced GFSC, a general framework that treats frame delay as an explicit coding resource rather than fixing sequential video coding to the causal setting. By exploiting the clique structure of K-order Markov sources, we established the first rigorous result showing that a K-frame delay can attain joint-coding rate-distortion optimality for a broad class of general Markov sources. For Gauss-Markov sources, this construction further reduces to pairwise interactions with linear interpolation sufficient statistics, revealing a simple structure behind the optimal auxiliaries. GFSC can be implemented with lightweight neural modules and integrated into existing NVC models with minimal architectural changes. Experiments consistently demonstrate improved rate-distortion performance over causal coding and bidirectional codecs under matched delay budgets. Together, these results identify limited delay as a fundamental and practically useful design dimension for neural video compression, bridging information-theoretic optimality and implementable coding architectures. Limitations and future directions are discussed in Appendix E.

AI USE STATEMENT

No generative AI tools or large language models were used in this work. All research ideas, theoretical results, proofs, experimental designs, implementations, and writing were carried out by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All theoretical assumptions and complete proofs are provided in the appendix, together with additional implementation and experimental details. To ensure transparency and reproducibility, we will release all artifacts necessary to reproduce our results, including the source code, environment specifications, model configurations, and training/evaluation scripts.

REFERENCES

Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Ballé, Sung Jin Hwang, and George Toderici. 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.

Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

Toby Berger. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*, 2003.

Jun Chen, Farrokh Etezadi, and Ashish Khisti. Generalized gaussian multiterminal source coding and probabilistic graphical models. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 719–723. IEEE, 2017.

Jun Chen, Li Xie, Yameng Chang, Jia Wang, and Yizhong Wang. Generalized gaussian multiterminal source coding: The symmetric case. *IEEE Transactions on Information Theory*, 66(4): 2115–2128, 2020.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural inter-frame compression for video coding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6421–6429, 2019.

Robert M Gray and Thomas G Stockham. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 2002.

Daphne Koller and Nir Friedman. *Probabilistic graphical models*, volume 165. MIT press Cambridge, 2009.

Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018.

Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34:18114–18125, 2021.

Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22616–22626, 2023.

Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26099–26108, 2024.

Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-lvc: Multiple frames prediction for learned video compression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3546–3554, 2020.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3192–3201. IEEE, 2022.

Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11006–11015, 2019.

Guo Lu, Xiaoyun Zhang, Wanli Ouyang, Li Chen, Zhiyong Gao, and Dong Xu. An end-to-end learning framework for video compression. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3292–3308, 2020.

Ming Lu, Zhihao Duan, Fengqing Zhu, and Zhan Ma. Deep hierarchical video compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8859–8867, 2024.

Nan Ma and Prakash Ishwar. The value of frame-delays in the sequential coding of correlated sources. In *2007 IEEE International Symposium on Information Theory*, pp. 1496–1500. IEEE, 2007.

Nan Ma, Ye Wang, and Prakash Ishwar. Delayed sequential coding of correlated sources. In *2007 Information Theory and Applications Workshop*, pp. 214–222. IEEE, 2007.

Fabian Mentzer, George Toderici, David Minnen, Sung-Jin Hwang, Sergi Caelles, Mario Lucic, and Eirikur Agustsson. Vct: A video compression transformer. *arXiv preprint arXiv:2206.07307*, 2022.

Alexandre Mercat, Marko Viitanen, and Jarno Vanne. 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.

Zaixi Shang, Joshua P Ebenezer, Alan C Bovik, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Subjective assessment of high dynamic range videos under different ambient conditions. In *2022 IEEE international conference on image processing (ICIP)*, pp. 786–790. IEEE, 2022.

Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 25:7311–7322, 2022.

Xihua Sheng, Li Li, Dong Liu, and Shiqi Wang. Bi-directional deep contextual video compression. *IEEE Transactions on Multimedia*, 2025.

Lucas Theis and Eirikur Agustsson. On the advantages of stochastic encoders. *arXiv preprint arXiv:2102.09270*, 2021.

Harish Viswanathan and Toby Berger. Sequential coding of correlated sources. *IEEE Transactions on Information Theory*, 46(1):236–246, 2002.

Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. 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.

Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 416–431, 2018.

Hui Xiang, Yifan Bian, Li Li, Jingran Wu, Xianguo Zhang, and Dong Liu. Real-time neural video compression with unified intra and inter coding. *arXiv preprint arXiv:2510.14431*, 2025.

Li Xie, Xiaolan Tu, Siyao Zhou, and Jun Chen. Generalized gaussian multiterminal source coding in the high-resolution regime. *IEEE Transactions on Communications*, 68(6):3782–3791, 2020.

Chenming Xu, Meiqin Liu, Chao Yao, Weisi Lin, and Yao Zhao. Ibvc: Interpolation-driven b-frame video compression. *Pattern Recognition*, 153:110465, 2024.

Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

En-Hui Yang, Lin Zheng, Da-Ke He, and Zhen Zhang. Rate distortion theory for causal video coding: Characterization, computation algorithm, and comparison. *IEEE transactions on information theory*, 57(8):5258–5280, 2011.

Ren Yang, Fabian Mentzer, Luc Van Gool, and Radu Timofte. Learning for video compression with hierarchical quality and recurrent enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6628–6637, 2020.

M Akın Yılmaz and A Murat Tekalp. End-to-end rate-distortion optimized learned hierarchical bi-directional video compression. *IEEE Transactions on Image Processing*, 31:974–983, 2021.

Yichi Zhang and Fengqing Zhu. Opendcvcs: a pytorch open source implementation and performance evaluation of the dcvc series video codecs. *arXiv preprint arXiv:2508.04491*, 2025.

Jacob Ziv. On universal quantization. *IEEE Transactions on Information Theory*, 31(3):344–347, 2003.

A COMPARISON OF CODING STRUCTURES

A.1 DELAYED SEQUENTIAL CODING

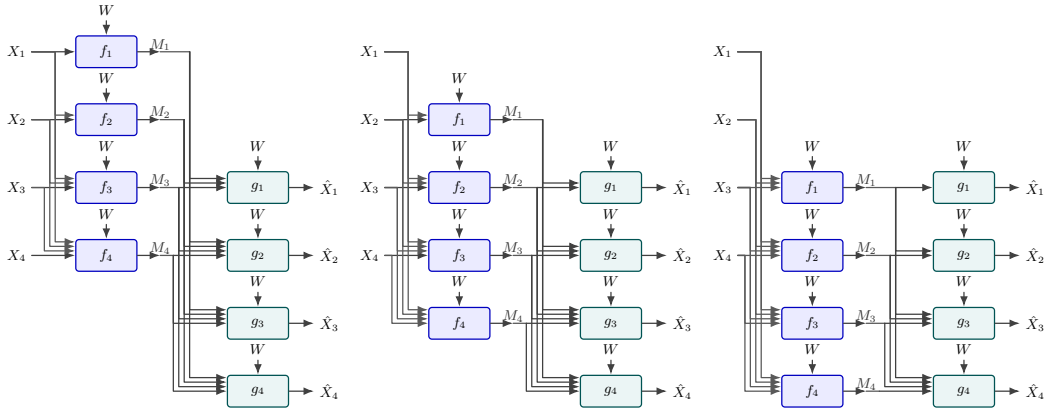


Figure 8: Illustration of three coding architectures for a four-frame source under 2-frame delay. *Left*: C-NC. *Middle*: NC-NC. *Right*: NC-C.

For completeness, we briefly review the three canonical setups of delayed sequential coding introduced in Ma et al. (2007). These include: (i) noncausal encoder and causal decoder (NC-C), (ii) causal encoder and noncausal decoder (C-NC), and (iii) noncausal encoder and noncausal decoder (NC-NC). The NC-C setting is defined in Section 2.2; here we provide the definitions of the remaining two setups.

C-NC (causal encoder, noncausal decoder). In the causal-noncausal setting, the encoder operates causally, while the decoder is allowed a delay of K frames. The encoding and decoding mappings are given by

$$f_\ell : \mathcal{X}_1 \times \cdots \times \mathcal{X}_\ell \times \mathcal{W} \to \mathcal{M}_\ell, \quad \ell = 1, \cdots, L, \tag{12}$$

$$g_\ell : \mathcal{M}_1 \times \cdots \times \mathcal{M}_{\min\{\ell+K,L\}} \times \mathcal{W} \to \hat{\mathcal{X}}_\ell, \quad \ell = 1, \cdots, L. \tag{13}$$

That is, the ℓ-th encoder has access only to past and current frames $(X_1, \cdots, X_\ell)$, while the decoder reconstructs $\hat{X}_\ell$ using messages up to time $\ell + K$.

NC-NC (noncausal encoder, noncausal decoder). In this setting, both the encoder and decoder are allowed limited lookahead. Let K_1 and K_2 denote the delays at the encoder and decoder, respectively, such that $K_1 + K_2 = K$. The encoding and decoding mappings are defined as

$$f_\ell : \mathcal{X}_1 \times \cdots \times \mathcal{X}_{\min\{\ell+K_1,L\}} \times \mathcal{W} \to \mathcal{M}_\ell, \quad \ell = 1, \cdots, L, \tag{14}$$

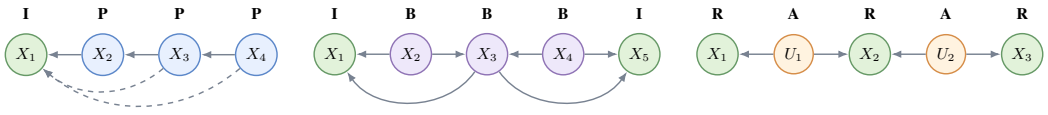
$$g_\ell : \mathcal{M}_1 \times \cdots \times \mathcal{M}_{\min\{\ell+K_2,L\}} \times \mathcal{W} \to \hat{\mathcal{X}}_\ell, \quad \ell = 1, \cdots, L. \tag{15}$$

Thus, the encoder can access future frames up to K_1 steps ahead, while the decoder reconstructs each frame with an additional delay of K_2.

These three setups characterize different ways of distributing the delay budget between the encoder and decoder, and together provide a unified framework for analyzing delayed sequential coding systems. We illustrate the corresponding system architectures for the three delayed setup in Figure 8. From a theoretical perspective, Corollary 6.1 in Ma et al. (2007) shows that all three setups achieve the same sum-rate performance for a given total delay K.

A.2 NVC PARADIGM COMPARISON

We briefly review common practical video coding paradigms and compare them with the proposed coding framework. Figure 9 provides a systematic comparison of these coding paradigms.



(a) Causal coding (b) Bidirectional coding (c) GFSC-1

Figure 9: *Left*: graphical illustration for causal predictive coding. *Middle*: graphical illustration for bidirectional coding. *Right*: graphical illustration for proposed GFSC-1.

P-frame coding. P-frame coding follows a causal predictive coding paradigm, where each frame is encoded conditioned only on previously reconstructed frames. This design enables low-latency processing and efficient implementation, and forms the foundation of most existing neural video compression methods. However, the strictly causal structure prevents the encoder from exploiting future frames, which limits the ability to capture longer-range temporal dependencies.

B-frame coding. B-frame coding extends the predictive paradigm by incorporating both past and future frames through bidirectional prediction. In practice, this is typically implemented via interpolation or bidirectional motion modeling between anchor frames. While such designs can improve compression efficiency, they usually rely on fixed temporal structures and are better suited to larger-delay settings. As a result, they are less flexible in scenarios with limited delay budgets.

GFSC. The proposed GFSC framework generalizes these paradigms by explicitly modeling a limited delay budget. Instead of relying on fixed bidirectional structures, GFSC constructs clique-wise dependencies within a delay window, enabling a flexible tradeoff between delay and compression efficiency. This formulation bridges the gap between strictly causal coding and joint coding, while remaining compatible with modern NVC architectures.

B Markov source models under general distortion

B.1 Operational rate-distortion analysis

Recall the information RD region for GFSC in Definition 3.4. We first derive the information sum-rate function.

Theorem B.1 (Information minimum sum-rate)**.** *For a given distortion tuple* $\mathbf{D}$*, the minimum information sum-rate of GFSC-*K *as in Definition 3.4 is*

$$R^{\mathrm{GFSC}}_{K,\mathrm{sum}}(\mathbf{D}) = \min I(\mathbf{X};\mathbf{U}), \tag{16}$$

and the minimum is taken over all auxiliary variables $\mathbf{U}$ *satisfying Markov chains in equation 8 and the distortion constraints* $D_\ell \geq \mathbb{E}[d_\ell(X_\ell, \hat{X}_\ell)]$*,* $\ell = 1, \cdots, L$*.*

Proof. By Definition 3.4, for a given distortion tuple $\mathbf{D}$, the minimum sum-rate is given by

$$R^{\mathrm{GFSC}}_{K,\mathrm{sum}}(\mathbf{D}) = \min \sum_{\ell=1}^{L} I(X_\ell^{\min\{\ell+K,L\}}; U_\ell | U^{\ell-1}),$$

where the minimization is over all feasible auxiliary variables satisfying the constraints in equation 8 and the distortion constraints. For notational convenience, we denote each stage rate as $I(X_\ell^{\ell+K}; U_\ell | U^{\ell-1})$. Then by the Markov chain in equation 8, we apparently have:

$$\begin{aligned} I(\mathbf{X}; U_\ell | U^{\ell-1}) &= I(X_\ell^{\ell+K}; U_\ell | U^{\ell-1}) + I(\mathbf{X} \setminus X_\ell^{\ell+K}; U_\ell | X_\ell^{\ell+K}, U^{\ell-1}) \\ &= I(X_\ell^{\ell+K}; U_\ell | U^{\ell-1}). \end{aligned}$$

Substituting this into the sum-rate expression, we have

$$R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) = \min \sum_{\ell=1}^{L} I(X_\ell^{\min\{\ell+K,L\}}; U_\ell | U^{\ell-1}) \tag{17}$$

$$= \min \sum_{\ell=1}^{L} I(\mathbf{X}; U_\ell | U^{\ell-1}) \tag{18}$$

$$= \min I(\mathbf{X}; \mathbf{U}), \tag{19}$$

where equation 18 follows from the Markov chain in equation 8, and equation 19 follows from the chain rule. This completes the proof. □

Next, we show the proof for Theorem 3.5 to characterize the operational RD region.

Theorem B.2 (Rate-distortion characterization)**.** *For a given distortion tuple* $\mathbf{D}$*, the following holds:*

1. *For any* $\mathbf{R} \in \mathcal{R}_K(\mathbf{D})$*, the operational rate region* $\mathcal{R}_K^o(\mathbf{D})$ *satisfies*

$$\mathbf{R} + \log(\mathbf{R}+1) + 5 \in \mathcal{R}_K^o(\mathbf{D}), \tag{20}$$

and

$$\mathcal{R}_K^o(\mathbf{D}) \subseteq \mathcal{R}_K(\mathbf{D}). \tag{21}$$

2. *The operational sum-rate satisfies*

$$R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) \le R_{K,\text{sum}}^{\text{GFSC},o}(\mathbf{D}) \le R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) + \mathcal{O}\big(\log R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D})\big). \tag{22}$$

where $R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D})$ *is characterized in Theorem B.1.*

Proof. **Part 1. One-shot achievability**
By Strong Functional Representation Lemma (SFRL) in Li & El Gamal (2018), for jointly distributed random variables X and Y, there exists a random variable Z independent of X, and function f such that $Y = f(X, Z)$. Z is not necessarily unique. The SFRL further states that there exists a Z such that:

$$H(Y|Z) \le I(X;Y) + \log(I(X;Y)+1) + 4.$$

Notice that this can be extended to conditional distribution case. Given $P_{XY|W}$, there exists Z independent of (X, W) such that $Y = f(X, W, Z)$ and:

$$H(Y|W,Z) \le I(X;Y|W) + \log(I(X;Y|W)+1) + 4.$$

For an L-frame source $\mathbf{X} = (X_1, \cdots, X_L)$, we define the following notations for the proposed GFSC-K scheme:

$$\hat{\mathbf{X}} \triangleq (\hat{X}_1, \cdots, \hat{X}_L), \quad \mathbf{U} \triangleq (U_1, \cdots, U_L),$$
$$\mathbf{R} \triangleq (R_1, \cdots, R_L), \quad \mathbf{D} \triangleq (D_1, \cdots, D_L).$$

Proof of equation 20: $\mathbf{R} + \log(\mathbf{R}+1) + 5 \in \mathcal{R}_K^o(\mathbf{D})$.
For a given $\mathbf{D}$ and $\mathbf{R} \in \mathcal{R}_K(\mathbf{D})$, let $\mathbf{U}$ be jointly distributed with $\mathbf{X}$ where the Markov chains in equation 8 hold, and the rate and distortion constraints in equation 7 are all satisfied. Denote the joint distribution of such $(\mathbf{X}, \mathbf{U}, \hat{\mathbf{X}})$ by $P_{\mathbf{X}\mathbf{U}\hat{\mathbf{X}}}$ and according to the Markov chains in equation 8, we have:

$$P_{\mathbf{X}\mathbf{U}\hat{\mathbf{X}}} = P_{\mathbf{X}} \cdot \prod_{\ell=1}^{L} P_{U_\ell | X_\ell^{\min\{\ell+K,L\}}, U^{\ell-1}} \cdot \prod_{\ell=1}^{L} \mathbf{1}\{\hat{X}_\ell = g_\ell(U^\ell)\}. \tag{23}$$

Now, fix the marginal distribution $P_{\mathbf{U}}$ according to equation 23 and use it to construct the codebook, i.e.,

$$P_{\mathbf{U}} = P_{U_L|U^{L-1}} \cdots P_{U_\ell|U^{\ell-1}} \cdots P_{U_1}. \tag{24}$$

By the SFRL, we have:

1. (*First stage*) There exists a random variable Z_1 independent of X^{K+1} and a deterministic function q_1 such that $U_1 = q_1(X^{K+1}, Z_1)$ and:

$$H(U_1|Z_1) \leq I(X^{K+1}; U_1) + \log(I(X^{K+1}; U_1) + 1) + 4.$$

Therefore, for the first stage encoder, it applies the deterministic function q_1 to get U_1 whose distribution needs to be preserved according to equation 24;

2. (*Intermediate stages*) By the conditional SFRL, there exists a random variable Z_ℓ independent of $(U^{\ell-1}, X_\ell^{\min\{\ell+K,L\}})$ and a deterministic function q_ℓ such that $U_\ell = q_\ell(U^{\ell-1}, X_\ell^{\min\{\ell+K,L\}}, Z_\ell)$ and:

$$H(U_\ell|U^{\ell-1}, Z_\ell) \leq I(X_\ell^{\min\{\ell+K,L\}}; U_\ell|U^{\ell-1}) + \log\Big(I(X_\ell^{\min\{\ell+K,L\}}; U_\ell|U^{\ell-1}) + 1\Big) + 4$$

At the intermediate ℓ-th stage (including the last stage), the ℓ-th encoder has access to $U^{\ell-1}$ from earlier stages. So upon observing the delayed source window $X_\ell^{\min\{\ell+K,L\}}$, it applies q_ℓ to get U_ℓ whose conditional distribution needs to be preserved according to equation 24.

Now the encoding and decoding schemes are as follows:

1. (*First stage encoder*) With Z_1 available at encoder and decoder, we can design prefix-free codes indexed by Z_1 to losslessly represent U_1 with the expected codeword length $\bar{R}_1$

$$\bar{R}_1 \leq H(U_1|Z_1) + 1 \leq I(X^{K+1}; U_1) + \log(I(X^{K+1}; U_1) + 1) + 5$$

2. (*Intermediate stage encoders*) With Z_ℓ available at encoder and decoder, we can design prefix-free codes indexed by $(Z_\ell, U^{\ell-1})$ to losslessly represent U_ℓ with the expected codeword length $\bar{R}_\ell$

$$\begin{aligned}\bar{R}_\ell &\leq H(U_\ell|U^{\ell-1}, Z_\ell) + 1 \\ &\leq I(X_\ell^{\min\{\ell+K,L\}}; U_\ell|U^{\ell-1}) + \log(I(X_\ell^{\min\{\ell+K,L\}}; U_\ell|U^{\ell-1}) + 1) + 5\end{aligned} \tag{25}$$

3. (*Decoders*) Upon receiving all U^ℓ, the ℓth decoder can generate the reconstruction via:

$$\hat{X}_\ell = g_\ell(U^\ell). \tag{26}$$

Then we have $\mathbf{R} + \log(\mathbf{R} + 1) + 5 \in \mathcal{R}_K^o(\mathbf{D})$.

Proof of equation 21: $\mathcal{R}_K^o(\mathbf{D}) \subseteq \mathcal{R}_K(\mathbf{D})$.
For any $\mathbf{D}$ and $\mathbf{R} \in \mathcal{R}_K^o(\mathbf{D})$, shared randomness $W \in \mathcal{W}$, encoding and decoding functions

$$\begin{aligned}&(\text{Enc}) : f_\ell : \mathcal{X}_\ell^{\min\{\ell+K,L\}} \times \mathcal{M}^{\ell-1} \times \mathcal{W} \to \mathcal{M}_\ell, \quad \ell = 1, \cdots, L, \\ &(\text{Dec}) : g_\ell : \mathcal{M}^\ell \times \mathcal{W} \to \hat{\mathcal{X}}_\ell, \quad \ell = 1, \cdots, L,\end{aligned}$$

such that:

$$R_\ell \geq \mathbb{E}[l(M_\ell)], \quad \ell = 1, \cdots, L, \tag{27}$$

$$D_\ell \geq \mathbb{E}[d_\ell(X_\ell, \hat{X}_\ell)], \quad \ell = 1, \cdots, L, \tag{28}$$

we need to lower bound the expected length of messages $l(M_\ell)$. Define:

$$U_1 \triangleq (M_1, W), \qquad U_\ell \triangleq (M^\ell, W),\ \ell = 1, \cdots, L. \tag{29}$$

Then for the first stage:

$$\begin{aligned}R_1 \geq \mathbb{E}[l(M_1)] &\geq H(M_1|W) \\ &= I(X^{K+1}; M_1|W) \\ &= I(X^{K+1}; M_1, W) \\ &= I(X^{K+1}; U_1)\end{aligned} \tag{30}$$

For the intermediate stages:

$$\begin{aligned} R_\ell \geq \mathbb{E}[l(M_\ell)] &\geq H(M_\ell|M^{\ell-1}, W) \\ &= I(X_\ell^{\min\{\ell+K,L\}}; M_\ell|M^{\ell-1}, W) \\ &= I(X_\ell^{\min\{\ell+K,L\}}; U_\ell|U^{\ell-1}) \end{aligned} \tag{31}$$

Notice that the definitions equation 29 imply the following Markov chains:

$$U_\ell \to (X_\ell^{\min\{\ell+K,L\}}, U^{\ell-1}) \to \mathbf{X}, \quad \ell = 1, \cdots, L. \tag{32}$$

On the other hand, by the definition of decoding functions and equation 29, we have that:

$$\hat{X}_\ell = g_\ell(M^\ell, W) \triangleq \eta_\ell(U^\ell) \tag{33}$$

where η_ℓ, $\ell = 1, \cdots, L$ are deterministic functions. Now consider the fact that the set of constraints in eqs. (27), (28), (30) and (31) with Markov chains in eq. (32) and deterministic functions in eq. (33) constitute an information rate-distortion region, denote by $\tilde{\mathcal{R}}_K$, which is the set of $(\mathbf{R}, \mathbf{D})$ such that:

$$\begin{aligned} R_\ell \geq I(X_\ell^{\min\{\ell+K,L\}}; U_\ell|U^{\ell-1}), \quad \ell = 1, \cdots, L, \\ D_\ell \geq \mathbb{E}[d_\ell(X_\ell, \hat{X}_\ell)], \quad \ell = 1, \cdots, L, \end{aligned}$$

for all auxiliary random variables $\mathbf{U}$ and $\hat{\mathbf{X}}$ satisfying:

$$\begin{aligned} \hat{X}_\ell = \eta_\ell(U^\ell), \\ U_\ell \to (X_\ell^{\min\{\ell+K,L\}}, U^{\ell-1}) \to \mathbf{X}, \end{aligned}$$

where η_j are some deterministic functions. By the definition of $\mathcal{R}_K$ in Definition 3.4, apparently we have $\tilde{\mathcal{R}}_K = \mathcal{R}_K$. Therefore, $\mathcal{R}_K^o(\mathbf{D}) \subseteq \tilde{\mathcal{R}}_K(\mathbf{D}) = \mathcal{R}_K(\mathbf{D})$. This completes the proof of Part 1.

Part 2. Sum-rate bounds

As we have already shown in Theorem B.1 that the information minimum sum-rate function of GFSC-K system is:

$$R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) = \min I(\mathbf{X}; \mathbf{U}),$$

subject to constraints in eqs. (7) and (8). We also have $\mathcal{R}_K^o(\mathbf{D}) \subseteq \mathcal{R}_K(\mathbf{D})$ in the proof of equation 21. Thus, we have the following lower bound:

$$R_{K,\text{sum}}^{\text{GFSC},o}(\mathbf{D}) = \min_{\mathbf{R} \in \mathcal{R}_K^o(\mathbf{D})} \sum_{\ell=1}^{L} R_\ell \geq \min_{\mathbf{R} \in \mathcal{R}_K(\mathbf{D})} \sum_{\ell=1}^{L} R_\ell = R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}). \tag{34}$$

For the sum-rate upper bound, based on the proof of equation 20, we have:

$$R_{K,\text{sum}}^{\text{GFSC},o}(\mathbf{D}) = \min_{\mathbf{R} \in \mathcal{R}_K^o(\mathbf{D})} \sum_{\ell=1}^{L} R_\ell \tag{35}$$

$$\leq \min_{\mathbf{R} \in \mathcal{R}_K(\mathbf{D})} \sum_{\ell=1}^{L} (R_\ell + \log(R_\ell + 1) + 5) \tag{36}$$

$$= \min_{\mathbf{R} \in \mathcal{R}_K(\mathbf{D})} \sum_{\ell=1}^{L} R_\ell + \sum_{\ell=1}^{L} \log(R_\ell + 1) + 5L \tag{37}$$

$$\leq \min_{\mathbf{R} \in \mathcal{R}_K(\mathbf{D})} I(\mathbf{X}; \mathbf{U}) + L \cdot \log\left(\frac{\sum_{\ell=1}^{L}(R_\ell + 1)}{L}\right) + 5L \tag{38}$$

$$\leq R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) + L \cdot \log\left(\frac{R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) + L}{L}\right) + 5L \tag{39}$$

$$= R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D}) + \mathcal{O}\big(\log R_{K,\text{sum}}^{\text{GFSC}}(\mathbf{D})\big), \tag{40}$$

where equation 36 follows equation 20, equation 38 follows the Jensen's inequality for logarithms. This completes the proof of Part 2. □

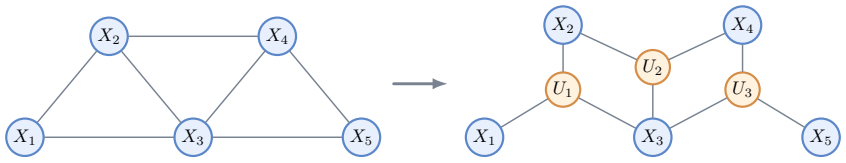


Figure 10: Clique-wise augmentation of a second-order Markov network, where U_j is associated with the maximal clique (X_j, X_{j+1}, X_{j+2}).

B.2 Joint-coding optimality for general Markov sources

We provide several definitions for a general distortion function.

Definition B.3 (Distortion function)**.** Define the general distortion function form as:

$$d : \mathcal{X} \times \mathcal{X} \to [0, \infty),$$

with $d(x, x) = 0$ and $d(x, w) > 0$ for $w, x \in \mathcal{X}$, $w \neq x$.

Definition B.4. Define $d_{\min} \triangleq \min_{x \neq w} d(x, w) > 0$ for all $x, w \in \mathcal{X}$.

Definition B.5. Define $d_{\max} \triangleq \max_{x \neq w} d(x, w) < \infty$ for all $x, w \in \mathcal{X}$.

For a general K-order Markov source as in Definition 3.6, its probability mass function $p_{\mathbf{X}}$ can be naturally factorized as $p_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{L-K} \psi_j(x_{\mathcal{C}_j})$ according to eq. (3), where $\mathcal{C}_j \triangleq \{j, \cdots, j+K\}$. We define ψ_j as the corresponding vectorized factor. Moreover, with the positive distribution assumption, $\psi_j(x_{\mathcal{C}_j})$ is strictly positive factor.

Definition B.6. Define $c_j \triangleq \min_{x_{\mathcal{C}_j} \in \mathcal{X}_{\mathcal{C}_j}} \psi_j(x_{\mathcal{C}_j}) > 0$ for $j = 1, \cdots, L-K$.

Definition B.7. For a positive distribution $p_{\mathbf{X}}$, define $p_{\min} \triangleq \min_{x \in \mathcal{X}} p_X(x) > 0$.

Definition B.8. Define the index set $\mathcal{I}_\ell \triangleq \{\max\{1, \ell - K\}, \cdots, \min\{\ell, L-K\}\}$ for $\ell = 1, \cdots, L$.

Now we provide the proof for Theorem 3.7.

Lemma B.9 (Nontrivial clique-wise auxiliary variables)**.** *Let* $\mathbf{X}$ *be a* K*-order Markov source as in Definition 3.6. Define a positive constant* δ_0 *as follow:*

$$\delta_0 \triangleq \min_{1 \leq j \leq L-K} \left\{ 1 - \left(1 + \frac{c_j}{\|\psi_j\|_2} \right)^{-\frac{1}{K+1}} \right\}. \tag{41}$$

Then for any $0 < \delta < \delta_0$*, there exist clique-wise auxiliary random variables* $\{U_j^{(\delta)}\}_{j=1}^{L-K}$*, jointly distributed with* $\mathbf{X}$*, such that:*

1. *Joint distribution* $p_{\mathbf{X}U^{(\delta),L-K}}$ *is also positive.*
2. *The following Markov chains hold:*

$$U_j^{(\delta)} \to X_j^{j+K} \to (\mathbf{X},\, U_{-j}^{(\delta)}), \quad j = 1, \ldots, L-K, \tag{42}$$

$$X_\ell \to U_{\mathcal{I}_\ell}^{(\delta)} \to X_{-\ell}, \quad \ell = 1, \ldots, L. \tag{43}$$

3. *Define* $\Delta_\ell(\delta) \triangleq \min_{g_\ell} \mathbb{E}[d_\ell(X_\ell, g_\ell(U_{\mathcal{I}_\ell}^{(\delta)}))]$ *over arbitrary deterministic mapping* g_ℓ *and* $\Delta_\ell(\delta)$ *is nontrivial, i.e.,*

$$0 < \Delta_\ell(\delta) \leq d_{\ell,\max}, \quad \ell = 1, \ldots, L.$$

Proof. We provide the proof in two parts. We first show that $\Delta_\ell(\delta) < d_{\ell,\max}$ is generally true based on the definition of distortion in Definition B.3. Then we provide a way to explicitly construct auxiliary variables $\{U_j\}_{j=1}^{L-K}$, which satisfy the Markov conditions in eqs. (42) and (43) and induce strictly positive distortion for each source variable, i.e. $\Delta_\ell(\delta) > 0$, for all $0 < \delta < \delta_0$ and $\ell = 1, \cdots, L$.

Part I. $\Delta_\ell(\delta) \le d_{\ell,\max}$**.**
We first show that $\Delta_\ell(\delta) \le d_{\ell,\max}$ is generally true based on the definition of distortion function. For arbitrary decoder input $U_{\mathcal{I}_\ell}$, we define $V_\ell \triangleq U_{\mathcal{I}_\ell}$ with $\mathcal{V}_\ell \triangleq \mathcal{U}_{\max\{1,\ell-K\}} \times \cdots \times \mathcal{U}_{\min\{\ell,L-K\}}$. Let $\hat{x}_\ell \triangleq g_\ell(v_\ell)$ denote the decoder reconstruction given observation $v_\ell \in \mathcal{V}_\ell$. By the definition of distortion $d_\ell : \mathcal{X}_\ell \times \mathcal{X}_\ell \to [0,\infty)$, we also have $\hat{x}_\ell \in \mathcal{X}_\ell$. Let $p_{X_\ell V_\ell}$ be the joint distribution of X_ℓ and V_ℓ, then we can bound the induced distortion:

$$\begin{aligned}
\Delta_\ell &= \min_{g_\ell} \mathbb{E}[d_\ell(X_\ell, g_\ell(V_\ell))] && (44)\\
&\le \mathbb{E}[d_\ell(X_\ell, g_\ell(V_\ell))] && (45)\\
&= \sum_{x_\ell, v_\ell} p_{X_\ell V_\ell}(x_\ell, v_\ell) d_\ell(x_\ell, \hat{x}_\ell) && (46)\\
&\le \sum_{x_\ell, v_\ell} p_{X_\ell V_\ell}(x_\ell, v_\ell) d_{\ell,\max} && (47)\\
&= d_{\ell,\max} < \infty, && (48)
\end{aligned}$$

where equation 45 is from the definition of Δ_ℓ, equation 47 is from Definition B.5. So $\Delta_\ell \le d_{\ell,\max}$ is generally true for arbitrary auxiliary variables V_ℓ.

Part II. $\Delta_\ell(\delta) > 0$**.**
For $\Delta_\ell(\delta) > 0$, we proceed in two steps. First, for the general distortion function in Definition B.3, we derive the necessary and sufficient condition to attain positive distortion. We then construct a set of auxiliary random variables $\{U_j\}_{j=1}^{L-K}$ and show that they satisfy both the required Markov conditions in eqs. (42) and (43) and the derived condition for positive distortion.

Positive distortion analysis. For a given observation $v_\ell = (u_1, \cdots, u_\ell) \in \mathcal{V}_\ell$ at decoder side, the goal of j-stage decoding is to minimize expected distortion:

$$\mathbb{E}[d_\ell(X_\ell, g_\ell(V_\ell))] = \sum_{v_\ell} p_{V_\ell}(v_\ell) \sum_{x_\ell} p_{X_\ell|V_\ell}(x_\ell|v_\ell) d_\ell(x_\ell, \hat{x}_\ell). \tag{49}$$

Here, we define the conditional risk $R_{v_\ell}(\hat{x}_\ell) \triangleq \sum_{x_\ell} p_{X_\ell|V_\ell}(x_\ell|v_\ell) d_\ell(x_\ell, \hat{x}_\ell)$. Apparently we have $R_{v_\ell}(\hat{x}_\ell) \ge \min_{\hat{x}_\ell \in \mathcal{X}_\ell} R_{v_\ell}(\hat{x}_\ell)$, thereby

$$\mathbb{E}[d_\ell(X_\ell, g_\ell(V_\ell))] \ge \sum_{v_\ell} p_{V_\ell}(v_\ell) \min_{\hat{x}_\ell \in \mathcal{X}_\ell} R_{v_\ell}(\hat{x}_\ell). \tag{50}$$

Then the optimal decoding function g_ℓ^* satisfies:

$$g_\ell^*(v_\ell) \in \arg\min_{\hat{x}_\ell \in \mathcal{X}_\ell} \sum_{x_\ell \in \mathcal{X}_\ell} p_{X_\ell|V_\ell}(x_\ell|v_\ell) d_\ell(x_\ell, \hat{x}_\ell). \tag{51}$$

In addition, as in Definition B.3, we have $d_\ell(x_\ell, x_\ell) = 0$ and $d_\ell(x_\ell, \hat{x}_\ell) > 0$ for any $x_\ell \ne \hat{x}_\ell$. Then for the minimum conditional risk $\min_{\hat{x}_\ell} R_{v_\ell}(\hat{x}_\ell)$, we have:

$$\begin{aligned}
\min_{\hat{x}_\ell} R_{v_\ell}(\hat{x}_\ell) &= \min_{\hat{x}_\ell} \sum_{x_\ell} p_{X_\ell|V_\ell}(x_\ell|v_\ell) d_\ell(x_\ell, \hat{x}_\ell) && (52)\\
&= \min_{\hat{x}_\ell} \sum_{x_\ell \ne \hat{x}_\ell} p_{X_\ell|V_\ell}(x_\ell|v_\ell) d_\ell(x_\ell, \hat{x}_\ell) && (53)\\
&\ge \min_{\hat{x}_\ell} \sum_{x_\ell \ne \hat{x}_\ell} p_{X_\ell|V_\ell}(x_\ell|v_\ell) d_{j,\min} && (54)\\
&= d_{j,\min} \left(1 - \max_{\hat{x}_\ell} p_{X_\ell|V_\ell}(\hat{x}_\ell|v_\ell)\right). && (55)
\end{aligned}$$

The minimum conditional risk is zero if and only if $p_{X_\ell|V_\ell}(\cdot|v_\ell)$ is Dirac delta function, i.e.,

$$p_{X_\ell|V_\ell}(x_\ell|v_\ell) = \delta(x_\ell - \hat{x}_\ell) = \begin{cases} 1, & x_\ell = \hat{x}_\ell, \\ 0, & x_\ell \ne \hat{x}_\ell. \end{cases} \tag{56}$$

Therefore, the induced distortion Δ_ℓ is zero if and only if $p_{X_\ell|V_\ell}(\cdot|v_\ell)$ is delta function for all v_ℓ with $p_{V_\ell}(v_\ell) > 0$. This is also equivalent to $H(X_\ell|V_\ell) = 0$. It is obvious that both distortion ϵ_ℓ and $p_{X_\ell|V_\ell}$ entries are non-negative. Therefore, it suffice to induce positive distortion $\Delta_\ell(\delta) > 0$ if we can construct auxiliary variables V_ℓ with positive distribution and positive backward test channel $p_{X_\ell|V_\ell}$ for all ℓ. In the following part, we provide an explicit construction.

Construction of auxiliary variables. In this part, we show that for K-order Markov source $\mathbf{X}$ as defined in Definition 3.6, we can always construct appropriate auxiliary variables, which satisfy all the conditions. For $\mathbf{X} \sim p_{\mathbf{X}}$, $p_{\mathbf{X}}$ naturally adopts the following clique factorization:

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{L-K} \psi_j(x_{\mathcal{C}_j}) \tag{57}$$

where $\mathcal{C}_j \triangleq \{j, \cdots, j+K\}$, and $\psi_j(x_{\mathcal{C}_j}) > 0$ is a positive clique factor since $\mathbf{X}$ has positive joint distribution. In addition, $\psi_j(x_{\mathcal{C}_j})$ can be further represented by a vectorized clique factor $\psi_j \in \mathbb{R}_+^{N_j}$ with $N_j \triangleq \prod_{r=j}^{j+K} |\mathcal{X}_r|$.

For j-th clique, we define the following auxiliary variable $U_j^{(\delta)}$ with $(K+1)$ components:

$$U_j^{(\delta)} \triangleq (U_{j,r}^{(\delta)} : r \in \mathcal{C}_j) \quad \text{with} \quad U_{j,r}^{(\delta)} \in \mathcal{X}_r, \tag{58}$$

which is also $U_j^{(\delta)} \triangleq (U_{j,j}^{(\delta)}, \cdots, U_{j,j+K}^{(\delta)})$. Now suppose a positive constant $0 < \delta < 1$, for each component $U_{j,r}^{(\delta)}$, we define a transition matrix $Q_{j,r}^{(\delta)} \in \mathbb{R}^{|\mathcal{X}_r| \times |\mathcal{X}_r|}$ from $U_{j,r}^{(\delta)}$ to X_r, with each entry as:

$$Q_{j,r}^{(\delta)}(x_r, u_{j,r}^{(\delta)}) = q_{X_r|U_{j,r}^{(\delta)}}(x_r|u_{j,r}^{(\delta)}) = (1-\delta)\mathbf{1}\{x_r = u_{j,r}^{(\delta)}\} + \delta \cdot \nu_r(x_r), \tag{59}$$

where ν_r denotes arbitrary positive distribution over $\mathcal{X}_r$. Here, we choose $\nu_r(x_r)$ to be the uniform distribution, therefore $\nu_r(x_r) = \frac{1}{|\mathcal{X}_r|}$ for all $x_r \in \mathcal{X}_r$. Then $Q_{j,r}^{(\delta)}$ can be further written as:

$$Q_{j,r}^{(\delta)} = (1-\delta) \cdot I + \frac{\delta}{|\mathcal{X}_r|} \cdot \mathbf{1}\mathbf{1}^\top = \underbrace{\begin{bmatrix} 1-\delta+\frac{\delta}{|\mathcal{X}_r|} & \cdots & \frac{\delta}{|\mathcal{X}_r|} \\ \vdots & \ddots & \vdots \\ \frac{\delta}{|\mathcal{X}_r|} & \cdots & 1-\delta+\frac{\delta}{|\mathcal{X}_r|} \end{bmatrix}}_{u_{j,r}^{(\delta)} \in \mathcal{X}_r} \Bigg\} x_r \in \mathcal{X}_r. \tag{60}$$

It can be verified that $Q_{j,r}^{(\delta)}$ has eigenvalues:

$$\lambda(Q_{j,r}^{(\delta)}) = \{1, 1-\delta\}.$$

Therefore as long as $0 < \delta < 1$, $Q_{j,r}^{(\delta)}$ is invertible and has all positive entries, which makes it a valid transition matrix. We also have the following form of the inverse of $Q_{j,r}^{(\delta)}$ is

$$\left(Q_{j,r}^{(\delta)}\right)^{-1} = \frac{1}{1-\delta}\left(I - \frac{\delta}{|\mathcal{X}_r|}\mathbf{1}\mathbf{1}^\top\right). \tag{61}$$

For the maximal clique $\mathcal{C}_j$, we further define the clique-wise transition matrix $Q_j^{(\delta)} \in \mathbb{R}_+^{N_j \times N_j}$ as:

$$Q_j^{(\delta)} \triangleq \bigotimes_{r \in \mathcal{C}_j} Q_{j,r}^{(\delta)}, \tag{62}$$

where $\bigotimes$ denotes the matrix Kronecker product. Each entry $Q_j^{(\delta)}(x_{\mathcal{C}_j}, u_j^{(\delta)})$ represents the backward test channel from $u_j^{(\delta)}$ to $x_{\mathcal{C}_j}$, i.e. $q_{X_{\mathcal{C}_j}|U_j^{(\delta)}}(x_{\mathcal{C}_j}|u_j^{(\delta)})$. For example, consider the 2-order Markov source $(X_1, \cdots, X_5)$ in Figure 10, we define the following constructions:

$$\begin{aligned} U_1^{(\delta)} &\triangleq (U_{1,1}^{(\delta)}, U_{1,2}^{(\delta)}, U_{1,3}^{(\delta)}) \xrightarrow{Q_1^{(\delta)} \triangleq Q_{1,1}^{(\delta)} \otimes Q_{1,2}^{(\delta)} \otimes Q_{1,3}^{(\delta)}} (X_1, X_2, X_3) \\ U_2^{(\delta)} &\triangleq (U_{2,2}^{(\delta)}, U_{2,3}^{(\delta)}, U_{2,4}^{(\delta)}) \xrightarrow{Q_2^{(\delta)} \triangleq Q_{2,2}^{(\delta)} \otimes Q_{2,3}^{(\delta)} \otimes Q_{2,4}^{(\delta)}} (X_2, X_3, X_4) \\ U_3^{(\delta)} &\triangleq (U_{3,3}^{(\delta)}, U_{3,4}^{(\delta)}, U_{3,5}^{(\delta)}) \xrightarrow{Q_3^{(\delta)} \triangleq Q_{3,3}^{(\delta)} \otimes Q_{3,4}^{(\delta)} \otimes Q_{3,5}^{(\delta)}} (X_3, X_4, X_5) \end{aligned} \tag{63}$$

For $0 < \delta < 1$, $Q_j^{(\delta)}$ is also invertible since the Kronecker product of invertible matrices is also invertible. It can be verified that $Q_j^{(\delta)}$ has the following set of eigenvalues:

$$\lambda(Q_j^{(\delta)}) = \{1, 1-\delta, \cdots, (1-\delta)^{K+1}\}. \tag{64}$$

Accordingly, the eigenvalues for its inverse $(Q_j^{(\delta)})^{-1}$ can be written as $(1-\delta)^{-k}$, $k = 0, \cdots, K+1$. Recall the positive vectorized clique factor ψ_j for clique $\mathcal{C}_j$, given $Q_j^{(\delta)}$ constructed above, we need to further find a positive vectorized factor $\phi_j^{(\delta)}$ satisfying:

$$\psi_j = Q_j^{(\delta)} \phi_j^{(\delta)}. \tag{65}$$

Since $Q_j^{(\delta)}$ is invertible, it is equivalent to find a valid positive factor such that:

$$\phi_j^{(\delta)} \triangleq (Q_j^{(\delta)})^{-1} \psi_j \in \mathbb{R}^{N_j}. \tag{66}$$

Now we will show that $Q_j^{(\delta)} \to I$ as $\delta \to 0$, thereby $\phi_j^{(\delta)} \to \psi_j$. Let $\|\cdot\|_2$ denotes the induced spectral norm of a matrix. It is known that the spectral norm of a positive semi-definite matrix is also its maximum eigenvalue. Notice that $(Q_j^{(\delta)})^{-1} - I$ is a positive semi-definite matrix, then we have:

$$\left\| (Q_j^{(\delta)})^{-1} - I \right\|_2 = \lambda_{\max}\left((Q_j^{(\delta)})^{-1} - I \right) = (1-\delta)^{-(K+1)} - 1. \tag{67}$$

By the definition of $\phi_j^{(\delta)}$ in equation 66, we further have:

$$\begin{aligned}
\|\phi_j^{(\delta)} - \psi_j\|_\infty &\le \|\phi_j^{(\delta)} - \psi_j\|_2 && (68)\\
&= \left\| (Q_j^{(\delta)})^{-1}\psi_j - \psi_j \right\|_2 && (69)\\
&= \left\| \left((Q_j^{(\delta)})^{-1} - I \right) \psi_j \right\|_2 && (70)\\
&\le \left\| (Q_j^{(\delta)})^{-1} - I \right\|_2 \|\psi_j\|_2 && (71)\\
&= \left((1-\delta)^{-(K+1)} - 1 \right) \|\psi_j\|_2 \triangleq F(\delta). && (72)
\end{aligned}$$

We also know that ψ_j is strictly positive factor due to positive joint distribution of source $\mathbf{X}$. Due to the definition of $U_j^{(\delta)}$ in equation 58, we have $\mathcal{U}_j = \mathcal{X}_{\mathcal{C}_j}$. Therefore, vectorized factors $\phi_j^{(\delta)}$ and ψ_j have the same coordinate ordering. Then we can have:

$$\begin{aligned}
\phi_j^{(\delta)}(u_j) &\ge \psi_j(u_j) - |\phi_j^{(\delta)}(u_j) - \psi_j(u_j)| && (73)\\
&\ge c_j - \|\phi_j^{(\delta)} - \psi_j\|_\infty && (74)\\
&\ge c_j - F(\delta). && (75)
\end{aligned}$$

Now we can set δ as following:

$$0 < \delta < 1 - \left(1 + \frac{c_j}{\|\psi_j\|_2} \right)^{-\frac{1}{K+1}}. \tag{76}$$

Then it can be verified that $F(\delta) < c_j$, thereby $\phi_j^{(\delta)}(u_j) > 0$ for every u_j. As we have $L-K$ maximal cliques for the K-order Markov source $\mathbf{X}$, we can define the following positive constant:

$$\delta_0 \triangleq \min_{1 \le j \le L-K} \left\{ 1 - \left(1 + \frac{c_j}{\|\psi_j\|_2} \right)^{-\frac{1}{K+1}} \right\} \tag{77}$$

Then for any $0 < \delta < \delta_0$ and all j, $\phi_j^{(\delta)}$ is a valid positive vectorized factor. By equation 66, we also have $\psi_j(x_{\mathcal{C}_j}) = \sum_{u_j} \phi_j^{(\delta)}(u_j^{(\delta)}) \prod_{r \in \mathcal{C}_j} Q_{j,r}^{(\delta)}(x_r, u_{j,r}^{(\delta)})$. Now we define the following augmented joint distribution:

$$q(\mathbf{x}, u^{(\delta),L-K}) \triangleq \prod_{j=1}^{L-K} \left(\phi_j^{(\delta)}(u_j^{(\delta)}) \prod_{r \in \mathcal{C}_j} Q_{j,r}^{(\delta)}(x_r, u_{j,r}^{(\delta)}) \right). \tag{78}$$

First, we already know that $Q_{j,r}^{(\delta)}$ has all positive entries and $\phi_j^{(\delta)}$ is a positive vectorized factor for $0 < \delta < \delta_0$. Then $q(\mathbf{x}, u^{(\delta),L-K})$ is also positive for $0 < \delta < \delta_0$. Second, we have:

$$\begin{aligned}
\sum_{\mathbf{x}} \sum_{u^{(\delta),L-K}} q(\mathbf{x}, u^{L-K}) &= \sum_{\mathbf{x}} \sum_{u_1^{(\delta)}} \cdots \sum_{u_{L-K}^{(\delta)}} \prod_{j=1}^{L-K} \left(\phi_j^{(\delta)}(u_j^{(\delta)}) \prod_{r \in \mathcal{C}_j} Q_{j,r}^{(\delta)}(x_r, u_{j,r}^{(\delta)}) \right) \\
&= \sum_{\mathbf{x}} \prod_{j=1}^{L-K} \sum_{u_j^{(\delta)}} \left(\phi_j^{(\delta)}(u_j^{(\delta)}) \prod_{r \in \mathcal{C}_j} Q_{j,r}^{(\delta)}(x_r, u_{j,r}^{(\delta)}) \right) \\
&= \sum_{\mathbf{x}} \prod_{j=1}^{L-K} \psi_j(x_{\mathcal{C}_j}) \\
&= \sum_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x}) = 1.
\end{aligned} \tag{79}$$

Therefore, this is a valid joint distribution for $(\mathbf{X}, U^{(\delta),L-K})$. Next we show that the above constructed auxiliary variables and the corresponding joint distribution satisfy all the Markov constraints and the positive induced distortion.

Markov conditions in equation 42. Now we verify the constructed joint distribution satisfy Markov conditions in equation 42. First, Markov conditions in equation 42 is equivalent to:

$$p(u^{L-K}|\mathbf{x}) = \prod_{j=1}^{L-K} p(u_j|x_j^{j+K}) = \prod_{j=1}^{L-K} p(u_j|x_{\mathcal{C}_j}) \tag{80}$$

Based on the joint distribution defined in equation 78, we can write:

$$\begin{aligned}
q(u^{(\delta),L-K}|\mathbf{x}) &= \frac{q(\mathbf{x}, u^{(\delta),L-K})}{p_{\mathbf{X}}(\mathbf{x})} \\
&= \frac{\prod_{j=1}^{L-K} \left(\phi_j^{(\delta)}(u_j^{(\delta)}) \prod_{r \in \mathcal{C}_j} Q_{j,r}^{(\delta)}(x_r, u_{j,r}^{(\delta)}) \right)}{\prod_{j=1}^{L-K} \psi_j(x_{\mathcal{C}_j})} \\
&= \prod_{j=1}^{L-K} \frac{\phi_j^{(\delta)}(u_j^{(\delta)}) \prod_{r \in \mathcal{C}_j} Q_{j,r}^{(\delta)}(x_r, u_{j,r}^{(\delta)})}{\psi_j(x_{\mathcal{C}_j})}
\end{aligned} \tag{81}$$

Then it can be verified that $q(u^{L-K}|\mathbf{x}) = \prod_{j=1}^{L-K} q(u_j|x_j^{j+K})$ is true, thereby the Markov conditions in equation 42 are satisfied.

Markov conditions in equation 43. Now we verify the constructed joint distribution satisfy Markov conditions in equation 43, which is equivalent to:

$$\begin{aligned}
p(\mathbf{x}|u^{L-K}) &= p(x_1|u_1)p(x_2|u^2) \cdots p(x_\ell|u_{\ell-K}^{\ell}) \cdots p(x_L|u_{L-K}) \\
&= \prod_{\ell=1}^{L} p(x_\ell|u_{\max\{1,\ell-K\}}^{\min\{\ell,L-K\}}) = \prod_{\ell=1}^{L} p(x_\ell|u_{\mathcal{I}_\ell}).
\end{aligned} \tag{82}$$

Now for the constructed joint distribution in equation 78, we can reorder the probability mass function as:

$$\begin{aligned}
q(\mathbf{x}, u^{(\delta),L-K}) &= \left[\prod_{j=1}^{L-K} \phi_j^{(\delta)}(u_j^{(\delta)}) \right] \left[\prod_{j=1}^{L-K} \prod_{r \in \mathcal{C}_j} Q_{j,r}^{(\delta)}(x_r, u_{j,r}^{(\delta)}) \right] \\
&= \left[\prod_{j=1}^{L-K} \phi_j^{(\delta)}(u_j^{(\delta)}) \right] \left[\prod_{\ell=1}^{L} \prod_{m \in \mathcal{I}_\ell} Q_{m,\ell}^{(\delta)}(x_\ell, u_{m,\ell}^{(\delta)}) \right] \\
&= \left[\prod_{j=1}^{L-K} \phi_j^{(\delta)}(u_j^{(\delta)}) \right] \left[\prod_{\ell=1}^{L} F_\ell(x_\ell, u_{\mathcal{I}_\ell}^{(\delta)}) \right]
\end{aligned} \tag{83}$$

It can be verified that we have the following conditional distribution form:

$$q(x_j|x_{-j}, u^{(\delta),L-K}) = \frac{F_j(x_j, u_{\mathcal{I}_j}^{(\delta)})}{\sum_{y\in\mathcal{X}_j} F_j(y, u_{\mathcal{I}_j}^{(\delta)})}, \tag{84}$$

so the factorized form of this conditional distribution is only dependent on $u_{\mathcal{I}_j}$. Therefore, $X_j \perp\!\!\!\perp (X_{-j}, \mathbf{U} \setminus U_{\mathcal{I}_j})|U_{\mathcal{I}_j}$, which is equivalent to the Markov conditions in equation 43.

Positive induced distortion. We have already shown that $\Delta_\ell(\delta) > 0$ if and only if $p_{X_\ell|V_\ell}$ is not point mass. For arbitrary $0 < \delta < \delta_0$, we can construct auxiliary variables $U_j \triangleq (U_j^{(j)}, \cdots, U_j^{(j+K)})$ and the corresponding joint distribution $q(\mathbf{x}, u^{L-K})$ with all positive entries. Therefore the induced distortion for each X_ℓ is also positive, i.e. $\Delta_\ell(\delta) > 0$ for $0 < \delta < \delta_0$. Specifically, for X_ℓ and a set of auxiliary variables $\{U_j^{(\delta)}\}_{j=1}^{L-K}$ constructed above, we have the induced distortion:

$$\begin{aligned}
\Delta_\ell(\delta) &\triangleq \min_{g_\ell} \mathbb{E}[d_\ell(X_\ell, g_\ell(U_{\mathcal{I}_\ell}^{(\delta)}))] && (85)\\
&= \sum_{u_{\mathcal{I}_\ell}^{(\delta)}} q(u_{\mathcal{I}_\ell}^{(\delta)}) \min_{\hat{x}_\ell\in\mathcal{X}_\ell} \sum_{x_\ell} Q^{(\delta)}(x_\ell, u_{\mathcal{I}_\ell}) d_\ell(x_\ell, \hat{x}_\ell) && (86)\\
&\geq \sum_{u_{\mathcal{I}_\ell}^{(\delta)}} q(u_{\mathcal{I}_\ell}^{(\delta)}) d_{\ell,\min} \min_{\hat{x}_\ell} \sum_{x_\ell\neq\hat{x}_\ell} Q^{(\delta)}(x_\ell, u_{\mathcal{I}_\ell}) > 0 && (87)
\end{aligned}$$

where $\min_{\hat{x}_\ell} \sum_{x_\ell\neq\hat{x}_\ell} Q^{(\delta)}(x_\ell, u_{\mathcal{I}_\ell})$ is a positive function of $\hat{x}_\ell$. Moreover, for arbitrary δ such that $0 < \delta < \delta_0$, we can define the following distortion region:

$$\mathcal{D}^{\mathrm{AUX}} \triangleq \{\mathbf{D} : 0 < D_\ell < \Delta_\ell(\delta),\ 0 < \delta < \delta_0,\ \ell = 1, \cdots, L\}. \tag{88}$$

Within this distortion region $\mathcal{D}^{\mathrm{AUX}}$, there always exist appropriate auxiliary variables $\{U_j^{(\delta)}\}_{j=1}^{L-K}$ constructed as above. □

Lemma B.9 shows that there always exist a class of clique-wise auxiliary variables for a K-order Markov source with positive distribution, such that the induced distortion for each source variable is also positive. Next we prove two Lemmas to characterize the RD optimal backward test channel for a positive distributed source. We first define the Neumann series which will be used in the proof.

Definition B.10 (Neumann series)**.** For a matrix $A \in \mathbb{R}^{n\times n}$, the series $\sum_{k=0}^{\infty} A^k$, with $A^0 = I_n$, is called the Neumann series associated with A. If $\|A\|_1 < 1$, where $\|\cdot\|_1$ denotes the induced matrix 1-norm, this series converges, and $I_n - A$ is invertible with

$$(I_n - A)^{-1} = \sum_{k=0}^{\infty} A^k. \tag{89}$$

Consequently,

$$\left\|(I_n - A)^{-1} - I_n\right\|_1 \leq \sum_{k=1}^{\infty} \|A\|_1^k = \frac{\|A\|_1}{1 - \|A\|_1}. \tag{90}$$

Lemma B.11. *Consider random variable $X \in \mathcal{X}$ with alphabet size $2 \leq |\mathcal{X}| < \infty$ and positive distribution. Let $W \in \mathcal{X}$ denote arbitrary reconstruction of X with $D \triangleq \mathbb{E}[d(X, W)]$ and T the backward test channel from W to X. Then for any joint distribution of (X, W) satisfying $0 < D < p_{\min}d_{\min}$, we have:*

1. *p_W is positive distribution, i.e. $p_W(w) > 0$ for all $w \in \mathcal{X}$.*
2. *T is arbitrarily close to but not exact identity matrix:*

$$\frac{2D}{d_{\max}} \leq \|T - I\|_1 \leq \frac{2D}{p_{\min}d_{\min} - D}. \tag{91}$$

3. *If additionally $D < \frac{p_{\min}d_{\min}}{3}$, then T is invertible and*

$$\|T^{-1} - I\|_1 \leq \frac{2D}{p_{\min}d_{\min} - 3D}. \tag{92}$$

Proof. **Part 1. Positive reconstruction distribution.**
Consider arbitrary joint distribution p_{XW} with $D = \mathbb{E}[d(X, W)]$. We have:

$$D = \mathbb{E}[d(X, W)] \tag{93}$$
$$= \sum_w \sum_{x \neq w} p_{XW}(x, w) d(x, w) \tag{94}$$
$$\geq d_{\min} \sum_w \sum_{x \neq w} p_{XW}(x, w) \tag{95}$$
$$= d_{\min} \cdot \Pr\{X \neq W\}. \tag{96}$$

Therefore $\Pr\{X \neq W\} \leq \frac{D}{d_{\min}}$. By the coupling inequality for total variation distance, for any $w \in \mathcal{X}$ we have:

$$\Pr\{X \neq W\} \geq d_{TV}(p_X, p_W) \tag{97}$$
$$= \sup_{\mathcal{A} \in \mathcal{X}} |p_X(\mathcal{A}) - p_W(\mathcal{A})| \tag{98}$$
$$\geq p_X(w) - p_W(w) \tag{99}$$
$$\geq p_{\min} - p_W(w). \tag{100}$$

Then for all $0 < D < d_{\min} p_{\min}$, we have $\Pr\{X \neq W\} \leq \frac{D}{d_{\min}} < p_{\min}$, thereby:

$$p_W(w) \geq p_{\min} - \Pr\{X \neq W\} > 0. \tag{101}$$

This completes the proof of Part 1.

Part 2. Transition matrix.
For all $w \in \mathcal{X}$, we have:

$$\sum_x |T(x, w) - I(x, w)| = 1 - T(w, w) + \sum_{x \neq w} T(x, w) \tag{102}$$
$$= 2\Pr\{X \neq W | W = w\} \tag{103}$$
$$= 2 \frac{\Pr\{X \neq W, W = w\}}{p_W(w)} \tag{104}$$
$$\leq 2 \frac{\Pr\{X \neq W\}}{p_W(w)} \tag{105}$$
$$\leq 2 \frac{\Pr\{X \neq W\}}{p_{\min} - \Pr\{X \neq W\}} \tag{106}$$
$$\leq \frac{2D}{p_{\min} d_{\min} - D}. \tag{107}$$

Therefore by the definition of induced matrix 1-norm, we have $\|T - I\|_1 \leq \frac{2D}{p_{\min} d_{\min} - D}$. This proves the upper bound in equation 91. In addition, we also have:

$$D = \sum_{x \neq w} p_{XW}(x, w) d(x, w) \tag{108}$$
$$\leq d_{\max} \sum_{x \neq w} p_{XW}(x, w) \tag{109}$$
$$= d_{\max} \sum_w p_W(w) \sum_{x \neq w} p_{X|W}(x|w) \tag{110}$$
$$= d_{\max} \sum_w p_W(w) \left(1 - T(w, w)\right) \tag{111}$$
$$\leq d_{\max} \max_w \left(1 - T(w, w)\right) \tag{112}$$
$$= \frac{d_{\max}}{2} \|T - I\|_1. \tag{113}$$

Therefore, we have $\|T - I\|_1 \geq \frac{2D}{d_{\max}} > 0$, which proves the lower bound in equation 91. This completes the proof of equation 91.

Part 3. Existence of inverse matrix.
Now we define $E \triangleq I - T$, then:

$$\|E\|_1 = \|T - I\|_1 \leq \frac{2D}{p_{\min} d_{\min} - D}.$$

If $D < \frac{p_{\min} d_{\min}}{3}$, then $\|E\|_1 < 1$. Therefore, consider $\sum_{k=0}^{\infty} E^k$ as the Neumann series, it converges and $T^{-1} = \sum_{k=0}^{\infty} E^k$ exists. By the definition of Neumann series in Definition B.10, we also have:

$$T^{-1} - I = E^0 + \sum_{k=1}^{\infty} E^k - I = \sum_{k=1}^{\infty} E^k. \tag{114}$$

Then:

$$\begin{aligned} \|T^{-1} - I\|_1 &= \left\| \sum_{k=1}^{\infty} E^k \right\|_1 \\ &\leq \sum_{k=1}^{\infty} \|E\|_1^k \\ &\leq \sum_{k=1}^{\infty} \left(\frac{2D}{p_{\min} d_{\min} - D} \right)^k \\ &= \frac{2D}{p_{\min} d_{\min} - 3D}. \end{aligned} \tag{115}$$

Then we have equation 92. This completes the proof. □

For arbitrary random vector, recall the characterization of joint-coding rate-distortion function in Theorem 2.1, i.e.,

$$R^{\mathrm{JC}}(\mathbf{D}) \triangleq \min_{\mathcal{P}} I(\mathbf{X}; \hat{\mathbf{X}}).$$

where $\mathcal{P} \triangleq \{P_{\hat{\mathbf{X}}|\mathbf{X}} : \mathbb{E}[d_\ell(X_\ell, \hat{X}_\ell)] \leq D_\ell,\ \ell = 1, \cdots, L\}$. For the per-letter distortion D_ℓ, we have the general form of Shannon lower bound:

$$I(\mathbf{X}; \hat{\mathbf{X}}) = H(\mathbf{X}) - H(\mathbf{X}|\hat{\mathbf{X}}) \tag{116}$$

$$\geq H(\mathbf{X}) - \sum_{\ell=1}^{L} H(X_\ell|\hat{X}_\ell) \tag{117}$$

$$= H(\mathbf{X}) - \sum_{\ell=1}^{L} H(X_\ell) + \sum_{\ell=1}^{L} I(X_\ell; \hat{X}_\ell) \tag{118}$$

$$\geq H(\mathbf{X}) - \sum_{\ell=1}^{L} H(X_\ell) + \sum_{\ell=1}^{L} R_\ell(D_\ell), \tag{119}$$

where $R_\ell(D_\ell)$ denotes the marginal rate-distortion function of source variable X_ℓ. Next we prove that for source with positive distribution, the Shannon lower bound can be achieved within a non-empty distortion region.

Lemma B.12 (SLB achievability)**.** *Let source $\mathbf{X}$ have finite component alphabets $2 \leq |\mathcal{X}_\ell| < \infty$ and positive joint distribution. For per-letter distortion function in Definition B.3, there exists a non-empty distortion region $\mathcal{D}^{\mathrm{SLB}} = \{\mathbf{D} : 0 < D_\ell \leq \varepsilon_\ell, \ell = 1, \cdots, L\}$, such that for all $\mathbf{D} \in \mathcal{D}^{\mathrm{SLB}}$, the Shannon lower bound can be attained.*

Proof. The main idea of the proof is to construct a valid joint reconstruction distribution from the marginal rate-distortion optimal backward test channels, and show that the SLB can be achieved.

First, we define the following positive constants:

$$p^*_{\min} \triangleq \min_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x}) > 0, \quad p_{\ell,\min} \triangleq \min_{x_\ell} p_{X_\ell}(x_\ell), \quad d_{\ell,\min} \triangleq \min_{x_\ell \neq w_\ell} d_\ell(x_\ell, w_\ell), \tag{120}$$

$$\tau \triangleq \frac{p^*_{\min}}{2 + p^*_{\min}}, \quad \varepsilon_\ell \triangleq \frac{p_{\ell,\min} d_{\ell,\min} \tau}{2L + \tau}. \tag{121}$$

Since $\tau < \frac{1}{2}$, we have $\varepsilon_\ell < \frac{p_{\ell,\min} d_{\ell,\min}}{3}$. Then we start the proof.

Step 1. Marginal RD problem. For each ℓ, select any W_ℓ that achieves the marginal RD function of X_ℓ. Such W_ℓ exists because X_ℓ has finite alphabet. Denote the vectorized reconstruction probability of p_{W_ℓ} as r_ℓ and the backward transition matrix as:

$$T_\ell(x_\ell, w_\ell) = p_{X_\ell|W_\ell}(x_\ell|w_\ell). \tag{122}$$

By Lemma B.11, for all $0 < D_\ell \leq \varepsilon_\ell < \frac{p_{\ell,\min} d_{\ell,\min}}{3}$, we have:

$$\|T_\ell - I\|_1 \leq \frac{2D_\ell}{p_{\ell,\min} d_{\ell,\min} - D_\ell} \leq \frac{\tau}{L}, \tag{123}$$

Apparently, T_ℓ^{-1} exists and we further have:

$$\|T_\ell^{-1} - I\|_1 \leq \frac{2D_\ell}{p_{\ell,\min} d_{\ell,\min} - 3D_\ell} \leq \frac{\tau}{L - \tau}. \tag{124}$$

Step 2. Joint backward test channel. In this part, we first provide an explicit way to construct a valid joint backward test channel based on marginal test channels. Then we show that this constructed joint backward test channel achieves Shannon lower bound within a non-empty distortion region. We first define the product backward transition matrix form:

$$S \triangleq \bigotimes_{\ell=1}^{L} T_\ell. \tag{125}$$

Then for all $\mathbf{D} \in \mathcal{D}^{\mathrm{SLB}}$, S^{-1} exists. Let p be the vectorized source distribution $p_{\mathbf{X}}$. Then we need to show that there exists a valid probability vector r satisfying:

$$r = S^{-1} p. \tag{126}$$

Since each T_ℓ is column transition matrix, we have $\|T_\ell\|_1 = 1$ for all ℓ. Let $N \triangleq \prod_{\ell=1}^{L} |\mathcal{X}_\ell|$. Using the telescoping expansion:

$$\bigotimes_{\ell=1}^{L} T_\ell - \bigotimes_{\ell=1}^{L} I_{|\mathcal{X}_\ell|} = \sum_{\ell=1}^{L} \left(\bigotimes_{j=1}^{\ell-1} T_j \right) \otimes (T_\ell - I_{|\mathcal{X}_\ell|}) \otimes \left(\bigotimes_{k=\ell+1}^{L} I_{|\mathcal{X}_k|} \right), \tag{127}$$

we have the following inequality:

$$\|S - I_N\|_1 = \left\| \bigotimes_{\ell=1}^{L} T_1 - \bigotimes_{\ell=1}^{L} I_{|\mathcal{X}_\ell|} \right\| \tag{128}$$

$$\leq \sum_{\ell=1}^{L} \left(\prod_{j=1}^{\ell-1} \|T_j\|_1 \right) \left\| T_\ell - I_{|\mathcal{X}_\ell|} \right\|_1 \left(\prod_{k=\ell+1}^{L} \|I_{|\mathcal{X}_k|}\|_1 \right) \tag{129}$$

$$= \sum_{\ell=1}^{L} \left\| T_\ell - I_{|\mathcal{X}_\ell|} \right\|_1 \leq \tau < 1. \tag{130}$$

By again applying the Neumann series, we have:

$$\|S^{-1} - I_N\| \leq \frac{\tau}{1 - \tau} = \frac{p^*_{\min}}{2}. \tag{131}$$

Then for the probability vectors, we have:

$$\max_{\mathbf{w}} |r(\mathbf{w}) - p_{\mathbf{X}}(\mathbf{w})| = \|r - p\|_\infty \tag{132}$$
$$\leq \|r - p\|_1 \tag{133}$$
$$= \|(S^{-1} - I)p\|_1 \tag{134}$$
$$\leq \|S^{-1} - I\|_1 \|p\|_1 \tag{135}$$
$$\leq \frac{p^*_{\min}}{2}. \tag{136}$$

Therefore, for the reconstruction probability vector r, we have:

1. $r(\mathbf{w}) \geq p_{\mathbf{X}}(\mathbf{w}) - \frac{p^*_{\min}}{2} \geq \frac{p^*_{\min}}{2} > 0$.
2. $\mathbf{1}^\top r = \mathbf{1}^\top S^{-1} p = \mathbf{1}^\top S S^{-1} p = \mathbf{1}^\top p = 1$.

Thus r is a valid positive probability distribution vector. Furthermore, by marginalizing the constructed joint distribution S, it can be verified that $p_{X_\ell} = T_\ell r_\ell$. Therefore, the marginal RD optimal backward test channels are preserved.

Step 3. Shannon lower bound achievability. For any feasible reconstruction $\hat{\mathbf{X}}$, the general form of Shannon lower bound (SLB) is shown in eq. (119). For the constructed auxiliary random variables $\mathbf{W}$, the product backward test channel and marginal RD optimality give:

$$I(\mathbf{X}; \mathbf{W}) = H(\mathbf{X}) - \sum_\ell H(X_\ell) + \sum_\ell R_\ell(D_\ell). \tag{137}$$

Therefore, the Shannon lower bound is attained by those constructed auxiliary variables $\mathbf{W}$ and the product backward test channel. Therefore, for all $0 < D_\ell < \varepsilon_\ell$, there always exist auxiliary variables $\mathbf{W}$ such that SLB is attained. This completes the proof. □

By Lemma B.9, for a K-order Markov source $\mathbf{X}$ as in Definition 3.6, we can construct nontrivial auxiliary random variables U^{L-K} satisfying the Markov constraints in eqs. (42) and (43). Those constructed auxiliary variables also satisfy the Markov constraints in Definition 3.4, thereby feasible for the GFSC architecture. For example, for first-order Markov, we can construct:

$$X_1 \to U_1 \to X_2 \to U_2 \to \cdots \to U_{L-1} \to X_L. \tag{138}$$

Apparently $X_\ell \to (U_{\ell-1}, U_\ell) \to X_{-\ell}$ holds. Next, to prove the JC rate-distortion optimality of GFSC architecture for general K-order Markov source, we introduce auxiliary random variables $\mathbf{W}$ satisfying the following properties:

1. Define $V_\ell \triangleq U_{\mathcal{I}_\ell}$ and $X_\ell \to W_\ell \to V_\ell$ holds.
2. The backward test channel from W_ℓ to X_ℓ is rate-distortion optimal.

Next, we show that such auxiliary variables can always be constructed within a nonempty distortion region for source with positive distribution.

Definition B.13. Let $Q_\ell \in \mathbb{R}^{|\mathcal{X}_\ell| \times |\mathcal{V}_\ell|}$ be the transition matrix constructed in Lemma B.9, representing the backward test channel maps P_{V_ℓ} to P_{X_ℓ}, i.e., $P_{X_\ell} = Q_\ell P_{V_\ell}$ with $Q_\ell(x, v) = p_{X_\ell|V_\ell}(x|v)$.

Definition B.14. Let $T_\ell \in \mathbb{R}^{|\mathcal{X}_\ell| \times |\mathcal{W}_\ell|}$ be the transition matrix constructed in Lemma B.12, representing the RD optimal backward test channel for X_ℓ, i.e., $P_{X_\ell} = T_\ell P_{W_\ell}$ with $T_\ell(x, w) = p_{X_\ell|W_\ell}(x|w)$.

Definition B.15. For Q_ℓ, define the positive term $q^{(\delta)}_{\ell,\min} \triangleq \min_{x,v} Q_\ell(x, v)$.

We also define the following positive constant:

$$\eta_\ell(\delta) \triangleq \frac{p_{\ell,\min} d_{\ell,\min} q^{(\delta)}_{\ell,\min}}{2 + 3q^{(\delta)}_{\ell,\min}}. \tag{139}$$

Lemma B.16 (RD-optimal refinement of clique auxiliary variables)**.** *Fix the clique-wise auxiliary variables constructed by equation 78 with $0 < \delta < \delta_0$. Assume the finite-alphabet and distortion conditions established above. Let $\varepsilon_\ell > 0$ be the distortion thresholds proved in Lemma B.12. Define $\epsilon_\ell(\delta) \triangleq \min\{\varepsilon_\ell,\ \eta_\ell(\delta)\}$ and the corresponding nonempty distortion region:*

$$\mathcal{D}_\epsilon \triangleq \{\mathbf{D} : 0 < D_\ell \le \epsilon_\ell(\delta),\ \ell = 1, \cdots, L\}. \tag{140}$$

Then for any distortion vector $\mathbf{D} \in \mathcal{D}_\epsilon$, there exist auxiliary variables $\mathbf{W} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_L$ such that:

1. *The Markov conditions hold:*
$$X_\ell \to W_\ell \to U_{\mathcal{I}_\ell}^{(\delta)} \to X_{-\ell}, \quad \ell = 1, \cdots, L. \tag{141}$$
2. *The reconstruction $\hat{X}_\ell \triangleq W_\ell$ and the corresponding backward test channel $p_{X_\ell|W_\ell}$ are jointly rate-distortion optimal.*

Proof. **Step 1. Construction.**
For a given δ such that $0 < \delta < \delta_0$, by the definition of $\eta_\ell(\delta)$ and $\Delta_\ell(\delta)$, we have:

$$\begin{aligned}
\Delta_\ell(\delta) &\triangleq \min_{g_\ell} \mathbb{E}[d_\ell(X_\ell, g_\ell(U_{\mathcal{I}_\ell}^{(\delta)}))] && (142)\\
&= \sum_{u_{\mathcal{I}_\ell}^{(\delta)}} q(u_{\mathcal{I}_\ell}^{(\delta)}) \min_{\hat{x}_\ell \in \mathcal{X}_\ell} \sum_{x_\ell} Q^{(\delta)}(x_\ell, u_{\mathcal{I}_\ell}) d_\ell(x_\ell, \hat{x}_\ell) && (143)\\
&\ge \sum_{u_{\mathcal{I}_\ell}^{(\delta)}} q(u_{\mathcal{I}_\ell}^{(\delta)}) d_{\ell,\min} \min_{\hat{x}_\ell} \sum_{x_\ell \ne \hat{x}_\ell} Q^{(\delta)}(x_\ell, u_{\mathcal{I}_\ell}) && (144)\\
&\ge \sum_{u_{\mathcal{I}_\ell}^{(\delta)}} q(u_{\mathcal{I}_\ell}^{(\delta)}) d_{\ell,\min} \min_{\hat{x}_\ell} \sum_{x_\ell \ne \hat{x}_\ell} q_{\ell,\min}^{(\delta)} && (145)\\
&= (|\mathcal{X}_\ell| - 1) d_{\ell,\min} q_{\ell,\min}^{(\delta)} \ge \eta_\ell(\delta). && (146)
\end{aligned}$$

Then for $\mathcal{D}_\epsilon$ defined in equation 140, we have:

$$\mathcal{D}_\epsilon \subseteq \mathcal{D}^{\mathrm{SLB}}, \quad \mathcal{D}_\epsilon \subseteq \mathcal{D}^{\mathrm{AUX}}. \tag{147}$$

Therefore, for all $\mathbf{D} \in \mathcal{D}_\epsilon$:

1. Lemma B.12 provides a backward test channel construction from $\mathbf{W}$ to $\mathbf{X}$:
$$p_{\mathbf{X}|\mathbf{W}}(\mathbf{x}|\mathbf{w}) = \prod_{\ell=1}^{L} T_\ell(x_\ell, w_\ell),$$
such that the Shannon lower bound can be attained. Thereby it is jointly rate-distortion optimal. T_ℓ denotes the marginal rate-distortion optimal test channel from W_ℓ to X_ℓ.
2. Lemma B.9 provides a backward test channel construction from U^{L-K} to $\mathbf{X}$ such that the induced distortion $\Delta_\ell(\delta)$ is positive and the following Markov conditions hold:
$$\begin{aligned}
&U_j^{(\delta)} \to X_j^{j+K} \to (\mathbf{X},\ U_{-j}^{(\delta)}), && j = 1, \ldots, L-K,\\
&X_\ell \to U_{\mathcal{I}_\ell}^{(\delta)} \to X_{-\ell}, && \ell = 1, \ldots, L.
\end{aligned}$$
Denote the backward test channel from $V_\ell \triangleq U_{\mathcal{I}_\ell}$ to X_ℓ as Q_ℓ.

Then it suffices to show that such constructed auxiliary variables W_ℓ and V_ℓ also satisfy the Markov conditions in equation 141.

Step 2. Markov conditions.
For constructed $T_\ell \in \mathbb{R}^{|\mathcal{X}_\ell| \times |\mathcal{W}_\ell|}$ and $Q_\ell \in \mathbb{R}^{|\mathcal{X}_\ell| \times |\mathcal{V}_\ell|}$, the Markov condition in equation 141 boils down to finding a valid transition matrix $A_\ell \in \mathbb{R}^{|\mathcal{W}_\ell| \times |\mathcal{V}_\ell|}$ such that

$$Q_\ell = T_\ell A_\ell. \tag{148}$$

As we already shown in Lemma B.11, T_ℓ is close to but not exactly an identity matrix and T_ℓ^{-1} exists. This is equivalent to showing that $A_\ell = T_\ell^{-1} Q_\ell$ is a valid transition matrix.

Non-negativity.

We first show that all the entries of $T_\ell^{-1} Q_\ell$ are non-negative. First, by Lemma B.11, we have

$$\begin{aligned} \|A_\ell - Q_\ell\|_1 &= \left\|(T_\ell^{-1} - I)Q_\ell\right\|_1 && (149)\\ &\le \left\|T_\ell^{-1} - I\right\|_1 \|Q_\ell\|_1 && (150)\\ &\le \frac{2D_\ell}{p_{\ell,\min} d_{\ell,\min} - 3D_\ell} && (151)\\ &\le \frac{2\eta_\ell(\delta)}{p_{\ell,\min} d_{\ell,\min} - 3\eta_\ell(\delta)} && (152)\\ &= q^{(\delta)}_{\ell,\min}. && (153) \end{aligned}$$

Then we have:

$$\begin{aligned} A_\ell(w_\ell, v_\ell) &= Q_\ell(w_\ell, v_\ell) - (Q_\ell(w_\ell, v_\ell) - A_\ell(w_\ell, v_\ell)) && (154)\\ &\ge Q_\ell(w_\ell, v_\ell) - \|A_\ell - Q_\ell\|_1 && (155)\\ &\ge q^{(\delta)}_{\ell,\min} - q^{(\delta)}_{\ell,\min} = 0. && (156) \end{aligned}$$

Therefore all entries of A_ℓ are non-negative by construction.

Column sum.

Moreover, for the column sum of A_ℓ, we directly have:

$$\mathbf{1}^\top A_\ell = \mathbf{1}^\top T_\ell^{-1} Q_\ell = \mathbf{1}^\top T_\ell T_\ell^{-1} Q_\ell = \mathbf{1}^\top Q_\ell = \mathbf{1}^\top. \tag{157}$$

Therefore, eqs. (156) and (157) prove that A_ℓ is a valid transition matrix. □

Theorem B.17 (JC optimality of GFSC)**.** *Consider a K-order Markov source $\mathbf{X}$ with positive distribution. There exist a non-empty distortion region $\mathcal{D}_\epsilon \triangleq \{\mathbf{D} : 0 < D_\ell \le \epsilon_\ell,\ \ell = 1, \cdots, L\}$, such that for all $\mathbf{D} \in \mathcal{D}_\epsilon$, the minimum information sum-rate $R^{\mathrm{GFSC}}_{K,\mathrm{sum}}(\mathbf{D})$ achieves the JC rate-distortion function:*

$$R^{\mathrm{GFSC}}_{K,\mathrm{sum}}(\mathbf{D}) = R^{\mathrm{JC}}(\mathbf{D}), \quad \forall \mathbf{D} \in \mathcal{D}_\epsilon. \tag{158}$$

Proof. **Step 1. Achievability.**

Fix an admissible δ in the clique-wise auxiliary variable construction such that $0 < \delta < \delta_0$. Choose the positive threshold $\epsilon_\ell(\delta)$ supplied by the Lemma B.16, and fix any $\mathbf{D} \in \mathcal{D}_\epsilon$. Then Lemma B.16 provides jointly distributed variables:

$$(\mathbf{X}, U^{L-K}, \mathbf{W}) \tag{159}$$

with the following properties:

1. The source marginal is still the prescribed $p_{\mathbf{X}}$.
2. The Markov conditions hold:
$$X_\ell \to W_\ell \to U_{\mathcal{I}_\ell}, \quad , \ell = 1, \cdots, L. \tag{160}$$
3. The refinement auxiliary variables $\mathbf{W}$ achieve the joint-coding rate-distortion function $R^{\mathrm{JC}}(\mathbf{D})$ for all $\mathbf{D} \in \mathcal{D}_\epsilon$.

We now construct augmented auxiliary variables satisfying the GFSC architecture. Specifically, we define:

$$Z_\ell \triangleq \begin{cases} (U^{(\delta)}_\ell, W_\ell), & 1 \le \ell \le L - K, \\ W_\ell, & L - K < \ell \le L. \end{cases} \tag{161}$$

Apparently, auxiliary variables $\{Z_\ell\}_{\ell=1}^{L}$ are just regrouping of constructed $\{U_j^{(\delta)}\}_{j=1}^{L-K}$ and $\mathbf{W}$. By equation 161, the previously received messages $Z^{\ell-1}$ contains $(U^{\min\{\ell-1,L-K\}}, W^{\ell-1})$. We first verify the required encoder Markov conditions in equation 8, i.e.,

$$U_\ell \to (X_\ell^{\min\{\ell+K,L\}}, U^{\ell-1}) \to \mathbf{X}. \tag{162}$$

For $1 \le \ell \le L-K$, we have the following:

$$\left.\begin{aligned} U_\ell^{(\delta)} \to X_\ell^{\ell+K} \to (\mathbf{X}, U_{-\ell}^{(\delta)}) \\ W_\ell \to (X_\ell, U_{\mathcal{I}_\ell}^{(\delta)}) \to X_{-\ell} \end{aligned}\right\} \Rightarrow Z_\ell \to (X_\ell^{\ell+K}, U_{\mathcal{I}_\ell}^{(\delta)}) \to \mathbf{X}. \tag{163}$$

For $L-K < \ell \le L$, we have $Z_\ell = W_\ell$. Then:

$$Z_\ell \to (X_\ell, U_{\mathcal{I}_\ell}^{(\delta)}) \to X_{-\ell} \tag{164}$$

Therefore, the encoder Markov conditions in equation 8 hold for every stage. Next, define the deterministic reconstruction as

$$\hat{X}_\ell = W_\ell \tag{165}$$

For $\ell \le L-K$, the decoder extracts the second component of Z_ℓ. For $\ell > L-K$, the decoder directly keep Z_ℓ itself. Thus this is admissible for the defined stronger decoder $g_\ell(Z^\ell)$. For ℓ-th encoding stage and augmented auxiliary variables Z_ℓ, by the conditional source coding theorem, the associate rate $\bar{R}_\ell$ is:

$$\bar{R}_\ell = I(X_\ell^{\min\{\ell+K,L\}}; Z_\ell | Z^{\ell-1}) \tag{166}$$

which also belongs to the information RD region in Definition 3.4. By the Markov condition in eq. (8), we have

$$I(X_\ell^{\min\{\ell+K,L\}}; Z_\ell | Z^{\ell-1}) = I(\mathbf{X}; Z_\ell | Z^{\ell-1}) \tag{167}$$

Consequently, we have the following sum-rate form by chain rule:

$$\sum_{\ell=1}^{L} \bar{R}_\ell = \sum_{\ell=1}^{L} I(X_\ell^{\min\{\ell+K,L\}}; Z_\ell | Z^{\ell-1}) \tag{168}$$

$$= \sum_{\ell=1}^{L} I(\mathbf{X}; Z_\ell | Z^{\ell-1}) \tag{169}$$

$$= I(\mathbf{X}; \mathbf{Z}) \tag{170}$$

$$= I(\mathbf{X}; U^{L-K}, \mathbf{W}) \tag{171}$$

$$= I(\mathbf{X}; \mathbf{W}) = R^{\mathrm{JC}}(\mathbf{D}). \tag{172}$$

Since this is just one possible realization of GFSC architecture, then we have:

$$R_{K,\mathrm{sum}}^{\mathrm{GFSC}}(\mathbf{D}) \le R^{\mathrm{JC}}(\mathbf{D}), \quad \forall \mathbf{D} \in \mathcal{D}_\epsilon. \tag{173}$$

Step 2. Converse.
By the Shannon lower bound, we have:

$$R_{K,\mathrm{sum}}^{\mathrm{GFSC}}(\mathbf{D}) \ge R^{\mathrm{SLB}}(\mathbf{D}) = R^{\mathrm{JC}}(\mathbf{D}), \quad \forall \mathbf{D} \in \mathcal{D}_\epsilon. \tag{174}$$

Combine with eq. (173), we have $R_{K,\mathrm{sum}}^{\mathrm{GFSC}}(\mathbf{D}) = R^{\mathrm{JC}}(\mathbf{D})$ for all $\mathbf{D} \in \mathcal{D}_\epsilon$. This completes the proof. □

C GAUSS-MARKOV SOURCE MODELS UNDER MSE DISTORTION

First, we review joint coding rate-distortion function for Gaussian source. Consider Gaussian source $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$. Let $\Sigma_e \triangleq \mathbb{E}[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^\top]$ be the error covariance, the JC rate-distortion function in Theorem 2.1 admits a closed-form characterization

$$R^{\mathrm{JCG}}(\mathbf{D}) = \min_{\Sigma_e} \frac{1}{2} \log \frac{|\Sigma|}{|\Sigma_e|} \quad \text{s.t.} \quad 0 \prec \Sigma_e \preceq \Sigma,\ (\Sigma_e)_{\ell\ell} \le D_\ell. \tag{175}$$

Moreover, in the distortion region $\mathcal{D}^{\mathrm{JC}} \triangleq \{\mathbf{D} : \mathrm{diag}(D_1, \cdots, D_L) \preceq \Sigma\}$, it simplifies to

$$R^{\mathrm{JCG}}(\mathbf{D}) = \frac{1}{2} \log \frac{|\Sigma|}{\prod_{\ell=1}^{L} D_\ell}. \tag{176}$$

This coincides with the Shannon lower bound (SLB) (Cover, 1999), which is tight over $\mathcal{D}^{\mathrm{JC}}$. Next, we briefly review Gaussian graphical models (GGMs).

C.1 GAUSSIAN GRAPHICAL MODEL

For Gaussian sources, as previously mentioned in Section 3.1, its graphical structure can be obtained directly from its inverse covariance matrix, i.e. information matrix or precision matrix. Let $J \triangleq \Sigma^{-1}$ denote the information matrix. The corresponding Gaussian graphical model is the undirected graph $\mathcal{G}(J) = (\mathcal{V}, \mathcal{E}(J))$, with node set $\mathcal{V} \triangleq \{1, \dots, L\}$ and edge set $\mathcal{E}(J) \triangleq \{\{j,k\} : J_{jk} \neq 0\}$, where J_{jk} denotes the (j,k)-th entry. For any $\mathcal{A} \subseteq \mathcal{V}$, write $X_{\mathcal{A}} \triangleq (X_i)_{i\in\mathcal{A}}$. The neighborhood of source node ℓ is

$$\mathcal{N}(\ell) \triangleq \{i \in \mathcal{V} \setminus \{\ell\} : \{\ell, i\} \in \mathcal{E}(J)\}.$$

For Gaussian sources, a missing edge has an exact conditional-independence interpretation:

$$J_{jk} = 0 \quad \Longleftrightarrow \quad X_j \perp\!\!\!\perp X_k | X_{\mathcal{V}\setminus\{j,k\}}.$$

Consequently, each variable is conditionally independent of all non-neighboring variables given its local neighborhood, i.e. $X_\ell \perp\!\!\!\perp X_{\mathcal{V}\setminus(\mathcal{N}(\ell)\cup\{\ell\})} | X_{\mathcal{N}(\ell)}$. For a K-order Gauss-Markov source, J is K-banded, i.e. $J_{jk} = 0$ for $|j-k| > K$. Therefore,

$$\mathcal{N}(\ell) \subseteq \{\ell - K, \dots, \ell - 1, \ell + 1, \dots, \ell + K\} \cap \mathcal{V}.$$

Thus, conditioned on at most $2K$ neighboring frames, X_ℓ is independent of all remaining source variables. Figure 3 illustrates this structure for a second-order Gauss-Markov source, whose edges connect only frames separated by at most two time steps.

We introduce several definitions that can be used to specify a subset of edges in GGMs, and will be used later to define a sequential coding order for the Gaussian version coding system. For a K-order model, we first define

$$\mathcal{E}^{(K)} \triangleq \{(j,k) : 1 \le j < k \le L,\ k - j \le K\}. \tag{177}$$

Here, equation 177 contains all frame pairs within the K neighbors, which also represents the edge set of a K-order Gauss-Markov model. As illustrated in Figure 3, $\mathcal{E}^{(2)}$ includes all pairs with form $(j, j+1)$ and $(j, j+2)$, which are processed according to the ordering in equation 178. Then, we process the pairs in $\mathcal{E}^{(K)}$ lexicographically. Specifically, let

$$(i,\ell) \prec (j,k) \quad \Longleftrightarrow \quad i < j \text{ or } (i = j \text{ and } \ell < k). \tag{178}$$

For any pair-indexed collection $Z^N \triangleq (Z_{j,k})_{(j,k)\in\mathcal{E}^{(K)}}$ with $N \triangleq |\mathcal{E}^{(K)}|$, define

$$Z_{<(j,k)} \triangleq (Z_{i,\ell})_{(i,\ell)\in\mathcal{E}^{(K)},(i,\ell)\prec(j,k)}, \tag{179}$$

$$Z_{\le j} \triangleq (Z_{i,\ell})_{(i,\ell)\in\mathcal{E}^{(K)},i\le j}. \tag{180}$$

Here, $Z_{<(j,k)}$ contains all variables generated before stage (j,k), whereas $Z_{\le j}$ contains all variables available after completing stage j. We apply this convention to messages and auxiliary variables, yielding $M_{<(j,k)}$, $M_{\le j}$, $U_{<(j,k)}$, and $U_{\le j}$.

C.2 GFSC WITH PAIR-WISE FACTORIZATION

As discussed in Section 3.1, the clique factorization can be further decomposed as pair-wise factorization. In addition, as discussed in Section 3.5, the pair-wise interaction can be formalized as linear interpolation between frame pairs. We rename the further factorized system as *Interpolation based Sequential Coding* (ISC). Let $\mathbf{X} = (X_1, \cdots, X_L)$ be a sequentially observed source and K the allowed delay relaxation ($K < L-1$). We define the encoding and decoding mappings of ISC-K architecture as follows.

Encoding. The encoding proceeds sequentially over the set $\mathcal{E}^{(K)}$. For each stage $(j,k) \in \mathcal{E}^{(K)}$, the encoder first constructs a "virtual frame" $\tilde{X}_{j,k}$ with an pair-wise operator $h_{j,k} : \mathcal{X}_j \times \mathcal{X}_k \to \tilde{\mathcal{X}}_{j,k}$. $\tilde{X}_{j,k}$ is then encoded into message using a zero-delay sequential encoder $\phi_{j,k} : \tilde{\mathcal{X}}_{j,k} \times \mathcal{M}_{<(j,k)} \times \mathcal{W} \to \mathcal{M}_{j,k}$. The encoding mapping is given by $f_{j,k} = \phi_{j,k} \circ h_{j,k}$, i.e.,

$$f_{j,k} : \mathcal{X}_j \times \mathcal{X}_k \times \mathcal{M}_{<(j,k)} \times \mathcal{W} \to \mathcal{M}_{j,k}, \quad (j,k) \in \mathcal{E}^{(K)}. \tag{181}$$

This formulation makes explicit that delay is handled through interpolation between (X_j, X_k), while the coding process itself remains strictly sequential. There are generally $N \triangleq |\mathcal{E}^{(K)}|$ encoding stages. $N = L-1$ for 1-delay system and $N \ge L$ for $K > 1$ delay system.

Decoding. The ℓ-th decoder reconstructs X_ℓ using all messages available up to ℓ-stage decoding. The decoding mapping is given by:

$$g_\ell : \mathcal{M}_{\leq \ell} \times \mathcal{W} \to \hat{\mathcal{X}}_\ell, \quad \ell = 1, \cdots, L. \tag{182}$$

There are L decoding stages for L source frames respectively. So encoders and decoders are not necessarily matched one-by-one. Notice that we can combine and define a multi-stage encoder $\tilde{f}_\ell \triangleq f_{\ell,\ell+K} \circ \cdots \circ f_{\ell,\ell+1}$ with input $X_\ell^{\ell+K}$. This can be seen as a special case of GFSC architecture.

C.3 Rate-distortion analysis

For convenience, we only characterize the information RD region and sum-rate function of ISC-K architecture. The analysis imposes no assumptions on the source distribution and therefore applies to general sources. We first introduce the notation used throughout this section. For the length-L source and distortion sequences, define:

$$\mathbf{X} \triangleq (X_1, \cdots, X_L), \qquad \mathbf{D} \triangleq (D_1, \cdots, D_L),$$

For sequences indexed by the ordered set $\mathcal{E}^{(K)}$, define:

$$U^N \triangleq \{U_{j,k} : (j,k) \in \mathcal{E}^{(K)}\}, \quad R^N \triangleq \{R_{j,k} : (j,k) \in \mathcal{E}^{(K)}\},$$

where $N \triangleq |\mathcal{E}^{(K)}|$ and the entries follow the prescribed ordering of $\mathcal{E}^{(K)}$. With this notation in place, we begin with the definition of information rate-distortion region associated with the ISC-K system.

Definition C.1 (Information RD region of ISC-K)**.** For a given source $\mathbf{X}$, the single-letter information RD region of ISC-K, denoted by $\mathcal{R}_K$, consists of all $(R^N, \mathbf{D})$ satisfying

$$R_{j,k} \geq I(\tilde{X}_{j,k}; U_{j,k} | U_{<j,k}), \quad (j,k) \in \mathcal{E}^{(K)}, \tag{183}$$

$$D_\ell \geq \mathbb{E}[\|X_\ell - \hat{X}_\ell\|^2], \quad \ell = 1, \cdots, L, \tag{184}$$

for all auxiliary random variables $\{\tilde{X}_{j,k}, U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ and $\hat{\mathbf{X}}$ satisfying $\tilde{X}_{j,k} = h_{j,k}(X_j, X_k)$, $\hat{X}_\ell = g_\ell(U_{\leq \ell})$ and the Markov constraints:

$$U_{j,k} \to (\tilde{X}_{j,k}, U_{<j,k}) \to (X_j, X_k, U_{<j,k}) \to \mathbf{X}, \tag{185}$$

where $h_{j,k}(\cdot)$ and $g_\ell(\cdot)$ are deterministic functions. For a given $\mathbf{D}$, the information rate region $\mathcal{R}_K(\mathbf{D})$ is the closure of all R^N such that $(R^N, \mathbf{D}) \in \mathcal{R}_K$.

Theorem C.2 (Information minimum sum-rate)**.** *For a given distortion tuple* $\mathbf{D}$*, the information minimum sum-rate function of ISC-*K *is*

$$R_{K,\text{sum}}^{\text{ISC}}(\mathbf{D}) = \min I(\mathbf{X}; U^N). \tag{186}$$

The minimum is taken over all auxiliary variables $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ *satisfying Markov chains in equation 185 and the distortion constraints* $D_\ell \geq \mathbb{E}[\|X_\ell - g_\ell(U_{\leq\ell})\|^2]$ *for all* ℓ*.*

Proof. From Definition C.1, for a given distortion tuple $\mathbf{D}$, the minimum sum-rate is given by

$$R_{K,\text{sum}}^{\text{ISC}}(\mathbf{D}) = \min \sum_{(j,k)\in\mathcal{E}^{(K)}} I(\tilde{X}_{j,k}; U_{j,k} | U_{<j,k}), \tag{187}$$

where the minimization is over all feasible auxiliary variables satisfying the constraints in equation 185 and the distortion constraints. Since $\tilde{X}_{j,k} = h_{j,k}(X_j, X_k)$ is a deterministic function of (X_j, X_k), we have

$$I(\tilde{X}_{j,k}; U_{j,k} | U_{<j,k}) \leq I(X_j, X_k; U_{j,k} | U_{<j,k}).$$

On the other hand, by the Markov chain $U_{j,k} \to (\tilde{X}_{j,k}, U_{<j,k}) \to (X_j, X_k, U_{<j,k})$, the data processing inequality implies

$$I(\tilde{X}_{j,k}; U_{j,k} | U_{<j,k}) \geq I(X_j, X_k; U_{j,k} | U_{<j,k}).$$

Combining the above two inequalities gives $I(\tilde{X}_{j,k}; U_{j,k}|U_{<j,k}) = I(X_j, X_k; U_{j,k}|U_{<j,k})$. Substituting this into the sum-rate expression, we have

$$R_{K,\text{sum}}^{\text{ISC}}(\mathbf{D}) = \min \sum_{(j,k)\in\mathcal{E}^{(K)}} I(X_j, X_k; U_{j,k}|U_{<j,k}) \tag{188}$$

$$= \min \sum_{(j,k)\in\mathcal{E}^{(K)}} I(\mathbf{X}; U_{j,k}|U_{<j,k}) \tag{189}$$

$$= \min I(\mathbf{X}; U^N), \tag{190}$$

where equation 189 follows from the Markov chain $U_{j,k} \to (X_j, X_k, U_{<j,k}) \to \mathbf{X}$ and equation 190 follows from the chain rule of mutual information under the ordering induced by $\mathcal{E}^{(K)}$. This completes the proof. □

C.4 Joint-coding optimality for Gauss-Markov sources

In this section, we prove that the ISC architecture achieves JC rate-distortion optimality for Gauss-Markov sources within a nontrivial distortion region. Typically, the Gaussian graphical model (GGM) associated with a Gaussian source can be represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} = \{\{i, j\} : J_{ij} \neq 0\}$ denotes the set of undirected edges. For notational convenience, in this section, we identify each undirected edge $\{i, j\}$ with the unique ordered pair (i, j) satisfying $i < j$ and write $\mathcal{E} = \{(i, j) : J_{ij} \neq 0, i < j\}$. This convention merely provides a unique indexing of the edges. It neither assigns directions to edges nor changes the underlying graph.

Lemma C.3 (Existence of a nontrivial distortion region)**.** *Consider correlated Gaussian vector $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$ and the GGM $\mathcal{G}(J)$ induced by its information matrix $J \triangleq \Sigma^{-1}$. Let $\{U_{j,k}\}_{(j,k)\in\mathcal{E}}$ be a set of Gaussian random variables jointly distributed with $\mathbf{X}$ satisfying:*

$$U_{j,k} = \omega_{j,k} X_j + (1 - \omega_{j,k}) X_k + N_{j,k}, \quad (j, k) \in \mathcal{E}, \tag{191}$$

and the Markov chains:

$$X_j \to \left(X_{\mathcal{N}(j)\setminus\{k\}}, U_{j,k}\right) \to X_k, \quad (j, k) \in \mathcal{E}, \tag{192}$$

$$X_\ell \to U_{\ell,\mathcal{N}(\ell)} \to X_{-\ell}, \quad \ell = 1, \cdots, L, \tag{193}$$

where $N_{j,k} \sim \mathcal{N}(0, \sigma^2_{N_{j,k}})$ are Gaussian random variables independent of $\mathbf{X}$. Then for any $\omega_{j,k} \in \mathbb{R}\setminus\{0, 1\}$, the induced MSE distortion for each source variable X_ℓ, denoted as $\epsilon_\ell \triangleq \mathbb{E}[\|X_\ell - \mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)}]\|^2]$, is nontrivial, i.e.,

$$0 < \epsilon_\ell < \bar{D}_\ell, \tag{194}$$

where $\bar{D}_\ell \triangleq \mathbb{E}[\|X_\ell - \mathbb{E}[X_\ell|X_{\mathcal{N}(\ell)}]\|^2]$.

Proof. *Construction of appropriate auxiliary variables*:
First we show how to construct auxiliary random variables satisfying eqs. (191) to (193). For Gaussian source $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ with information matrix $J \triangleq \Sigma^{-1}$, its probability density function can be written as:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{L}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{x}^\top J \mathbf{x}\right) \tag{195}$$

$$= \frac{1}{(2\pi)^{\frac{L}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\sum_{\ell=1}^{L} J_{\ell\ell} x_\ell^2 - \sum_{1\le j<k\le L} J_{jk} x_j x_k\right) \tag{196}$$

$$= \frac{1}{(2\pi)^{\frac{L}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\sum_{\ell=1}^{L} J_{\ell\ell} x_\ell^2 - \sum_{(j,k)\in\mathcal{E}} J_{jk} x_j x_k\right), \tag{197}$$

where equation 196 follows from that J is a symmetric matrix, and equation 197 follows from that $J_{jk} = 0$ for $(j, k) \notin \mathcal{E}$. Thus, the dependency structure of a Gaussian source $\mathbf{X}$ is completely

characterized by the edge set $\mathcal{E}$ of its GGM. By equation 191, each auxiliary variable $U_{j,k}$ is associated with edge (j,k) in the GGM and independent additive Gaussian noise $N_{j,k}$. Therefore, its conditional density function given $(X_j, X_k) = (x_j, x_k)$ can be written as:

$$p(u_{j,k}|x_j, x_k) = \frac{1}{\sqrt{2\pi\sigma^2_{N_{j,k}}}} \exp\left(-\frac{(u_{j,k} - \omega_{j,k}x_j - \bar{\omega}_{j,k}x_k)^2}{2\sigma^2_{N_{j,k}}}\right), \tag{198}$$

where $\bar{\omega}_{j,k} \triangleq 1 - \omega_{j,k}$. In addition, since those noise items $\{N_{j,k}\}_{(j,k)\in\mathcal{E}}$ are also mutually independent with each other, $\{U_{j,k}\}_{(j,k)\in\mathcal{E}}$ are conditionally independent given $\mathbf{X}$, i.e., $p(\mathbf{u}|\mathbf{x}) = \prod_{(j,k)\in\mathcal{E}} p(u_{j,k}|x_j, x_k)$. Applying Bayes' rule, we have $p(\mathbf{x}|\mathbf{u}) \propto p(\mathbf{x})p(\mathbf{u}|\mathbf{x})$, which yields:

$$p(\mathbf{x}|\mathbf{u}) \propto \exp\left(-\frac{1}{2}\sum_{\ell=1}^{L} J_{\ell\ell}x_\ell^2 - \sum_{\{j,k\}\in\mathcal{E}} J_{jk}x_jx_k - \sum_{\{j,k\}\in\mathcal{E}} \frac{(u_{j,k} - \omega_{j,k}x_j - \bar{\omega}_{j,k}x_k)^2}{2\sigma^2_{N_{j,k}}}\right) \tag{199}$$

Expanding the quadratic terms and collecting coefficients gives

$$p(\mathbf{x}|\mathbf{u}) \propto \exp\left(-\frac{1}{2}\sum_{\ell=1}^{L} \lambda_\ell x_\ell^2 + \sum_{\ell=1}^{L} \gamma_\ell(\mathbf{u})x_\ell - \sum_{(j,k)\in\mathcal{E}} \left(J_{jk} + \frac{\omega_{j,k}\bar{\omega}_{j,k}}{\sigma^2_{N_{j,k}}}\right) x_jx_k\right), \tag{200}$$

where

$$\lambda_\ell \triangleq J_{\ell\ell} + \sum_{(\ell,k)\in\mathcal{E}} \frac{\omega^2_{\ell,k}}{\sigma^2_{N_{\ell,k}}} + \sum_{(j,\ell)\in\mathcal{E}} \frac{\bar{\omega}^2_{j,\ell}}{\sigma^2_{N_{j,\ell}}}, \tag{201}$$

$$\gamma_\ell(\mathbf{u}) \triangleq \sum_{(\ell,k)\in\mathcal{E}} \frac{\omega_{\ell,k}}{\sigma^2_{N_{\ell,k}}} u_{\ell,k} + \sum_{(j,\ell)\in\mathcal{E}} \frac{\bar{\omega}_{j,\ell}}{\sigma^2_{N_{j,\ell}}} u_{j,\ell}. \tag{202}$$

The Markov constraints in eqs. (192) and (193) imply that the source variables are conditionally independent given the adjacent auxiliary variables, i.e. $p(\mathbf{x} \mid \mathbf{u}) = \prod_{\ell=1}^{L} p(x_\ell|\mathbf{u})$. Therefore, all pairwise interaction terms x_jx_k in equation 200 must vanish, i.e.,

$$J_{jk} + \frac{\omega_{j,k}\bar{\omega}_{j,k}}{\sigma^2_{N_{j,k}}} = 0, \tag{203}$$

which yields $\sigma^2_{N_{j,k}} = -\frac{\omega_{j,k}\bar{\omega}_{j,k}}{J_{jk}}$. As $\sigma^2_{N_{j,k}}$ denotes the noise variance, to ensure $\sigma^2_{N_{j,k}} > 0$, we need to set $\omega_{j,k} \in (0,1)$ if $J_{jk} < 0$, and set $\omega_{j,k} > 1$ otherwise. With appropriate coefficients $\omega_{j,k}$, the constructed auxiliary variables apparently satisfy Markov constraints in eqs. (191) to (193). This completes the construction of the auxiliary random variables.

Nontrivial MSE distortion analysis:

We now establish the distortion bounds for the auxiliary random variables $\{U_{j,k}\}_{(j,k)\in\mathcal{E}}$ constructed above. Without loss of generality, we consider the source node $\ell = 1$. Suppose $\mathbf{X}$ is a pairwise correlated Gaussian vector, the set of neighbors for X_1 is apparently $\mathcal{N}(1) = \{2, \ldots, L\}$.

First, for the upper bound in equation 194, i.e.,

$$\epsilon_1 = \mathbb{E}[\|X_1 - \mathbb{E}[X_1|U_{1,\mathcal{N}(1)}]\|^2] < \bar{D}_1, \tag{204}$$

this follows directly from the Markov chain conditions in eqs. (191) to (193), which guarantee that the auxiliary random variables $U_{1,\mathcal{N}(1)}$ capture all the dependence of X_1 on the remaining source variables. Next, we prove the strict positivity in equation 194. Since the MSE is always nonnegative, it suffices to show that $\epsilon_1 \neq 0$. Equivalently, it suffices to show that X_1 cannot be recovered exactly from $U_{1,\mathcal{N}(1)}$. Suppose, for contradiction, that

$$X_1 = \mathbb{E}[X_1 \mid U_{1,\mathcal{N}(1)}] \quad \text{a.s.} \tag{205}$$

Since $(X_1, U_{1,\mathcal{N}(1)})$ are jointly Gaussian and zero mean, the conditional expectation is thereby a linear transformation. Therefore, equation 205 is equivalent to that there exist constants $a_2, \ldots, a_L$ such that

$$X_1 = \sum_{k=2}^{L} a_k U_{1,k} = \sum_{k=2}^{L} a_k \left(\omega_{1,k}X_1 + \bar{\omega}_{1,k}X_k + N_{1,k}\right) \quad \text{a.s.} \tag{206}$$

Rearranging terms gives

$$\left(1-\sum_{k=2}^{L} a_k\omega_{1,k}\right)X_1-\sum_{k=2}^{L} a_k\bar{\omega}_{1,k}X_k=\sum_{k=2}^{L} a_kN_{1,k} \quad \text{a.s.} \tag{207}$$

The left-hand side of equation 207 is a linear function of $\mathbf{X}$, whereas the right-hand side is a linear function of the independent Gaussian noises $\{N_{1,k}\}_{k=2}^{L}$. Hence the two sides are two independent Gaussian random variables. As they are equal almost surely, both sides must be almost surely constant. Because all variables have zero mean, this constant must be zero. Thus, we need

$$\left(1-\sum_{k=2}^{L} a_k\omega_{1,k}\right)X_1-\sum_{k=2}^{L} a_k\bar{\omega}_{1,k}X_k=0 \quad \text{a.s.}, \tag{208}$$

$$\sum_{k=2}^{L} a_kN_{1,k}=0 \quad \text{a.s.}, \tag{209}$$

where equation 208 implies that:

$$\begin{bmatrix} 1-\sum_{k=2}^{L} a_k\omega_{1,k} \\ -a_2\bar{\omega}_{1,2} \\ \vdots \\ -a_L\bar{\omega}_{1,L} \end{bmatrix} \in \ker(\Sigma), \tag{210}$$

Since $\Sigma \succ 0$, we have $\ker(\Sigma)=\{\mathbf{0}\}$. Therefore, equation 208 is equivalent to

$$1-\sum_{k=2}^{L} a_k\omega_{1,k}=0, \qquad a_k\bar{\omega}_{1,k}=0 \text{ for } k=2,\dots,L. \tag{211}$$

Because appropriate auxiliary variable construction requires $\omega_{1,k}\in\mathbb{R}\setminus\{0,1\}$, which implies $\omega_{1,k},\bar{\omega}_{1,k}\neq 0$. Therefore, $a_k\bar{\omega}_{1,k}=0$ requires $a_k=0$ for all $k=2,\dots,L$. This contradicts $1-\sum_{k=2}^{L} a_k\omega_{1,k}=0$. Therefore, equation 208 cannot hold for appropriate constructed auxiliary variables. Now for equation 209, since the noises $\{N_{1,k}\}_{k=2}^{L}$ are mutually independent Gaussian random variables with zero mean, equation 209 holds if and only if

$$\mathrm{Var}\left(\sum_{k=2}^{L} a_kN_{1,k}\right)=\sum_{k=2}^{L} a_k^2\sigma^2_{N_{1,k}}=0. \tag{212}$$

Therefore, we must have either $a_k=0$ or $\sigma^2_{N_{1,k}}=0$ for $k=2,\dots,L$. Recall that for appropriate constructed auxiliary variables, we have $\sigma^2_{N_{1,k}}=-\frac{\omega_{1,k}\bar{\omega}_{1,k}}{J_{1k}}$, where $J_{1k}\neq 0$. Since $\omega_{1,k}\in\mathbb{R}\setminus\{0,1\}$, we have $\omega_{1,k}\bar{\omega}_{1,k}\neq 0$ and thereby $\sigma^2_{N_{1,k}}>0$. Therefore equation 209 implies that $a_k=0$, for $k=2,\dots,L$. Substituting this into $X_1=\sum_{k=2}^{L} a_kU_{1,k}$ yields $X_1=0$ almost surely, which contradicts $\Sigma\succ 0$. Hence equation 209 also cannot hold for appropriate constructed auxiliary variables. Combining the above arguments, we conclude that ϵ_1 is strictly positive and finite.

This proof can be easily extended to all remaining source nodes X_ℓ, $\ell=2,\cdots,L$, with the same collection of auxiliary random variables $\{U_{j,k}\}_{(j,k)\in\mathcal{E}}$. Therefore, let $\epsilon_\ell\triangleq\mathbb{E}[\|X_\ell-\mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)}]\|^2]$ denotes the induced distortion for each source node X_ℓ, we have $0<\epsilon_\ell<\bar{D}_\ell$, for $\ell=1,\cdots,L$. This completes the proof. □

The above Lemma C.3 mainly shows that the existence of appropriate edge-wise auxiliary variables $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ for Gaussian source, which induce nontrivial MSE distortion ϵ_ℓ (strictly positive and bounded) for each source variable X_ℓ. A specific distortion region $\mathcal{D}_\epsilon$ can be thereby determined by those edge-wise auxiliary variables. The following Theorem C.5 further provide rate-distortion optimality analysis based on the result of Lemma C.3.

Berger-Tung scheme. Before proving Theorem C.5, we briefly review the Berger-Tung achievability scheme (Berger, 2003) through a simple example. Consider three correlated sources $\mathbf{X} = (X_1, X_2, X_3)$ and two encoders observing the overlapping source pairs (X_1, X_2) and (X_2, X_3), respectively. Choose auxiliary variables according to $P_{\mathbf{X},U_{1,2},U_{2,3}} = P_{\mathbf{X}} P_{U_{1,2}|X_1,X_2} P_{U_{2,3}|X_2,X_3}$, i.e.,

$$U_{1,2} \to (X_1, X_2) \to X_3, \quad U_{2,3} \to (X_2, X_3) \to X_1, \tag{213}$$

and let $\hat{X}_\ell = g_\ell(U_{1,2}, U_{2,3})$ satisfy the desired distortion constraint for each $\ell = 1, 2, 3$. Each encoder quantizes its observation into an auxiliary codeword and transmits its bin index. The decoder then jointly recovers both auxiliary codewords and reconstructs $\mathbf{X}$. The Berger-Tung inner bound guarantees that any rate pair satisfying

$$R_{1,2} \geq I(X_1, X_2; U_{1,2}|U_{2,3}), \tag{214}$$
$$R_{2,3} \geq I(X_2, X_3; U_{2,3}|U_{1,2}), \tag{215}$$
$$R_{1,2} + R_{2,3} \geq I(\mathbf{X}; U_{1,2}, U_{2,3}), \tag{216}$$

is achievable. Thus, for fixed auxiliary variables, the minimum achievable sum-rate is

$$R_{\text{sum}} = I(\mathbf{X}; U_{1,2}, U_{2,3}). \tag{217}$$

The same construction extends to an arbitrary collection of overlapping source subsets. In the proof below, the constructed pairwise and node-wise auxiliary variables play the roles of the Berger-Tung auxiliaries, yielding the desired achievable sum-rate upper bound.

Definition C.4. Define a non-empty distortion region under the Gaussian case as $\mathcal{D}_\epsilon^{\text{G}} \triangleq \{(D_1, \cdots, D_L) : 0 < D_\ell \leq \epsilon_\ell, \ell = 1, \cdots, L\}$ for some strictly positive $0 < \epsilon_j < \infty$.

Theorem C.5 (Gaussian optimality)**.** *Consider a K-order Gauss-Markov source $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$. There exists a non-empty distortion region $\mathcal{D}_\epsilon^{\text{G}}$, such that for all $\mathbf{D} \in \mathcal{D}_\epsilon^{\text{G}}$, the following sets of random variables can be constructed:*

i) (Gaussian auxiliary variables) It suffices to consider jointly Gaussian auxiliary random variables $\{\tilde{X}_{j,k}, U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$, $\{U_\ell\}_{\ell=1}^L$, and $\hat{\mathbf{X}}$. Moreover, the optimal $U_{j,k}$ and U_ℓ satisfy the Markov constraints:

$$\begin{aligned} X_j \to (X_{\mathcal{N}(j)\setminus\{k\}}, U_{j,k}) \to X_k, \quad (j,k) \in \mathcal{E}^{(K)}, \\ U_{j,k} \to (X_j, X_k) \to \mathbf{X}, \quad (j,k) \in \mathcal{E}^{(K)} \\ U_\ell \to X_\ell \to U_{\ell,\mathcal{N}(\ell)} \to X_{-\ell}, \quad \ell = 1, \cdots, L. \end{aligned} \tag{218}$$

ii) (Linear interpolation) Given (X_j, X_k) and $U_{j,k}$, there exists some coefficients $\omega_{j,k} \in \mathbb{R}\setminus\{0,1\}$ such that auxiliary variable $\tilde{X}_{j,k}$ admit linear deterministic functions for (X_j, X_k):

$$\tilde{X}_{j,k} = \omega_{j,k} X_j + (1 - \omega_{j,k}) X_k, \tag{219}$$

and satisfies the Markov chain $U_{j,k} \to \tilde{X}_{j,k} \to (X_j, X_k) \to \mathbf{X}$.

iii) (Linear reconstruction) The optimal $\hat{\mathbf{X}}$ adopt linear decoding mappings.

Moreover, with those constructed auxiliary variables, the sum-rate distortion of such system achieves the Shannon lower bound in equation 176 and hence the joint coding rate-distortion optimality.

Proof. Existence of $\mathcal{D}_\epsilon^{\text{G}}$:
First, as established in Lemma C.3, for the edge-wise auxiliary random variables $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ constructed according to eqs. (191) to (193), the induced distortion for each source variable X_ℓ, defined as $\epsilon_\ell \triangleq \mathbb{E}[\|X_\ell - \mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)}]\|^2]$, is strictly positive and bounded, i.e. $0 < \epsilon_\ell < \infty$. This establishes the existence of a nontrivial distortion region:

$$\mathcal{D}_\epsilon^{\text{G}} \triangleq \{(D_1, \cdots, D_L) : \ 0 < D_\ell \leq \epsilon_\ell, \ \ell = 1, \cdots, L\}. \tag{220}$$

Moreover, for each source variable X_ℓ in a K-order Gauss-Markov source, it has the K-banded conditional-dependence structure, i.e. $\mathcal{N}(\ell) \subset \{\ell - K, \cdots, \ell - 1, \ell + 1, \cdots, \ell +$

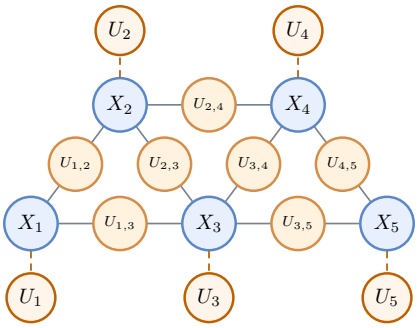


Figure 11: Augmented graphical model with node-wise and edge-wise auxiliary variables.

$K\}$. Therefore, by the Markov constraints in equation 218, $\mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)}]$ is a function of $(U_{\ell-K,\ell},\cdots,U_{\ell-1,\ell},U_{\ell,\ell+1},\cdots,U_{\ell,\ell+K})$.

Gaussian auxiliary variables:

Next, we show that arbitrary distortion tuple within the induced region, i.e. $\mathbf{D}\in\mathcal{D}_\epsilon^{\mathrm{G}}$, can be achieved by introducing a set of node-wise Gaussian auxiliary random variables $\{U_\ell\}_{\ell=1}^L$ satisfying the Markov chains:

$$U_\ell \to X_\ell \to (X_{-\ell},U_{\ell,\mathcal{N}(\ell)}), \tag{221}$$

and then the JC optimality is also achieved. Suppose the edge-wise auxiliary random variables $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ have been constructed according to eqs. (191) to (193), which are all Gaussian variables. Then each source node X_ℓ can be decomposed as

$$X_\ell = \mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)}] + N_\ell, \quad \ell=1,\cdots,L, \tag{222}$$

where $N_\ell\sim\mathcal{N}(0,\sigma_{N_\ell}^2)$ denotes Gaussian residual and is independent of $(U_{\ell,\mathcal{N}(\ell)},X_{\mathcal{N}(\ell)})$. To further refine the reconstruction of X_ℓ, we first introduce a Gaussian auxiliary random variable $\tilde{U}_\ell$ such that:

$$U_\ell = X_\ell + W_\ell, \quad \text{with } W_\ell\sim\mathcal{N}(0,\sigma_{W_\ell}^2),\ W_\ell\perp\!\!\!\perp X_\ell. \tag{223}$$

Apparently, we have $U_\ell\to X_\ell\to(X_{-\ell},U_{\ell,\mathcal{N}(\ell)})$. It can be verified that X_ℓ can be further decomposed as $X_\ell=\mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)},U_\ell]+Z_\ell$, where $Z_\ell\sim\mathcal{N}(0,\sigma_{Z_\ell}^2)$ denote an independent Gaussian noise. With $\epsilon_\ell\triangleq\mathbb{E}[\|X_\ell-\mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)}]\|^2]$ and introduced node-wise U_ℓ, then we further have the corresponding distortion term as $D_\ell=(\epsilon_\ell+\frac{1}{\sigma_{W_\ell}^2})^{-1}$. Observe that by setting the variance $\sigma_{W_\ell}^2$ arbitrarily small, the induced distortion of X_ℓ, i.e.,

$$D_\ell = \mathbb{E}[\|X_\ell-\mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)},U_\ell]\|^2], \tag{224}$$

can be made arbitrarily close to zero. Therefore, by introducing such auxiliary variables $\{U_\ell\}_{\ell=1}^L$ for each component, the achievable distortion can be continuously controlled over the range $(0,\epsilon_\ell]$. Consequently, any distortion tuple $\mathbf{D}\in\mathcal{D}_\epsilon^{\mathrm{G}}$ is achievable. For example, Figure 11 illustrates the construction of $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(2)}}$ and $\{U_\ell\}_{\ell=1}^5$ for a second-order Gauss-Markov source $(X_1,\cdots,X_5)$.

For the above designed coding system, by applying Lemma 1 in Chen et al. (2017) together with the Berger-Tung inner/achievability bound, we obtain that, for all distortion tuples $\mathbf{D}\in\mathcal{D}_\epsilon^{\mathrm{G}}$,

$$R_{K,\mathrm{sum}}^{\mathrm{ISC}}(\mathbf{D}) \le \frac{1}{2}\log\frac{|\Sigma|}{\prod_{\ell=1}^L D_\ell}. \tag{225}$$

Conversely, by the Shannon lower bound, we have for any $\mathbf{D}\in\mathcal{D}_\epsilon^{\mathrm{G}}\subseteq\mathcal{D}_{\mathrm{JC}}$,

$$R_{K,\mathrm{sum}}^{\mathrm{ISC}}(\mathbf{D}) \ge \frac{1}{2}\log\frac{|\Sigma|}{\prod_{\ell=1}^L D_\ell}. \tag{226}$$

Combining the upper and lower bounds, for any $\mathbf{D}\in\mathcal{D}_\epsilon$, it follows that

$$R_{K,\mathrm{sum}}^{\mathrm{ISC}}(\mathbf{D}) = \frac{1}{2}\log\frac{|\Sigma|}{\prod_{\ell=1}^L D_\ell}. \tag{227}$$

This coincides with the joint coding rate-distortion function in equation 176, and hence establishes joint coding optimality. This completes the proof.

Linear interpolation.
For jointly Gaussian $\mathbf{X}$ and $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$, the conditional expectation $\mathbb{E}[U_{j,k}|X_j, X_k]$ is linear on (X_j, X_k). Hence, $U_{j,k}$ admits the orthogonal decomposition

$$U_{j,k} = \gamma_{j,k}X_j + \theta_{j,k}X_k + \tilde{N}_{j,k}, \tag{228}$$

where $\tilde{N}_{j,k}$ is also Gaussian and independent of (X_j, X_k). Since scaling of $U_{j,k}$ does not affect the rate-distortion tradeoff, we can normalize the coefficients by $\gamma_{j,k} + \theta_{j,k}$ to obtain

$$U_{j,k} = \omega_{j,k}X_j + (1-\omega_{j,k})X_k + N_{j,k}, \tag{229}$$

where $\omega_{j,k} = \frac{\gamma_{j,k}}{\gamma_{j,k}+\theta_{j,k}}$ and $N_{j,k}$ is Gaussian and independent of (X_j, X_k). Furthermore, as shown in Lemma C.3, the noise variance satisfies

$$\sigma^2_{N_{j,k}} = -\frac{\omega_{j,k}(1-\omega_{j,k})}{J_{jk}}.$$

In particular, $\omega_{j,k} \in (0,1)$ when $J_{jk} < 0$, while $\omega_{j,k} > 1$ when $J_{jk} > 0$.

Linear reconstruction:
Given constructed edge-wise auxiliary variables $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ and node-wise auxiliary variables $\{U_\ell\}_{\ell=1}^L$, the final step is to reconstruct each source node X_ℓ. Due to the Markov constraints in equation 218, for each source variable X_ℓ, we have the following reconstruction:

$$\hat{X}_\ell = \mathbb{E}[X_\ell|\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}, \{U_\ell\}_{\ell=1}^L] = \mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)}, U_\ell].$$

Without loss of generality, we let $U_{\ell,\mathcal{N}(\ell)} = (U_{\ell-K,\ell}, \cdots, U_{\ell-1,\ell}, U_{\ell,\ell+1}, \cdots, U_{\ell,\ell+K})$. Then for jointly Gaussian auxiliary variables, we have the optimal reconstruction $\hat{X}_\ell$ with linear decoding function:

$$\begin{aligned}
\hat{X}_\ell &= \mathbf{a}^{(\ell)^\top}[U_{\ell-K,\ell}, \cdots, U_{\ell-1,\ell}, \cdots, U_{\ell,\ell+1}, \cdots, U_{\ell,\ell+K}, U_\ell]^\top \\
&= \sum_{k=0}^{K-1} a^{(\ell)}_{K-k} \cdot U_{\ell-1-k,\ell} + \sum_{k=K}^{2K-1} a^{(\ell)}_k \cdot U_{\ell,\ell-K+1+k} + a^{(\ell)}_{2K} \cdot U_\ell \\
&= \sum_{k=0}^{K-1} a^{(\ell)}_{K-k} \cdot (\omega_{\ell-1-k,\ell}X_{\ell-1-k} + \bar{\omega}_{\ell-1-k,\ell}X_\ell + N_{\ell-1-k,\ell}) \\
&\quad + \sum_{k=K}^{2K-1} a^{(\ell)}_k \cdot (\omega_{\ell,\ell-K+1+k}X_\ell + \bar{\omega}_{\ell,\ell-K+1+k}X_{\ell-K+1+k} + N_{\ell,\ell-K+1+k}) + a^{(\ell)}_{2K}\lambda_\ell \cdot X_\ell \\
&= \sum_{k=-K}^{K} \beta_{K+1+k}X_{\ell+k} + Z_\ell,
\end{aligned}$$

where $Z_\ell = \sum_{k=0}^{K-1} a^{(\ell)}_{K-k}N_{\ell-1-k,\ell} + \sum_{k=K}^{2K-1} a^{(\ell)}_k N_{\ell,\ell-K+1+k}$ is a Gaussian random variable independent with $\mathbf{X}$. This completes the proof. □

Theorem C.6 (JC rate-distortion optimality)**.** *Under the source model setting of Theorem C.5, there exists a non-empty distortion region $\mathcal{D}^{\mathrm{G}}_\epsilon$, such that for all $\mathbf{D} \in \mathcal{D}^{\mathrm{G}}_\epsilon$, $R^{\mathrm{ISC}}_{K,\mathrm{sum}}(\mathbf{D})$ achieves the joint-coding rate-distortion function:*

$$R^{\mathrm{ISC}}_{K,\mathrm{sum}}(\mathbf{D}) = R^{\mathrm{JCG}}(\mathbf{D}), \quad \forall \mathbf{D} \in \mathcal{D}^{\mathrm{G}}_\epsilon. \tag{230}$$

Proof. Theorem C.5 already establishes the existence of two sets of edge-wise auxiliary random variables $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ and node-wise auxiliary variables $\{U_j\}_{j=1}^L$, which satisfy the desired Markov constraints in equation 218 and achieve Shannon lower bound in the distortion region $\mathcal{D}_\epsilon$. In this proof, we show that these auxiliary variables are operationally compatible with the structure of the proposed ISC-K system.

Step 1. Achievability scheme for ISC:
In this part, we show those edge-wise and node-wise auxiliary variables can be incorporated into the proposed ISC-K system without violating its coding structure:

1. **Regrouped auxiliary variables.** As in the Definition C.1, there are only edge-wise auxiliary variable form $U_{j,k}$ involved in the ISC-K system. We first clarify how edge-wise auxiliary variables can be extended when node-wise auxiliary variables are present. Basically, the node-wise variables do not require a separate class of messages. They can be incorporated into the pair-indexed ISC auxiliaries through a deterministic augmentation. Given edge-wise and node-wise auxiliary variables as constructed in Theorem C.5, we consider a set of augmented auxiliary variables:

$$V_{j,k} \triangleq \begin{cases} U_{j,k}, & j < L-1,\ k < \min\{j+K, L\}, \\ (U_{j,k}, U_j), & j < L-1,\ k = \min\{j+K, L\}, \\ (U_{L-1,L}, U_{L-1}, U_L), & j = L-1,\ k = L. \end{cases} \tag{231}$$

   The idea is to simply augment the node-wise variable U_j into the last edge-wise auxiliary variable associated with source node X_j. The exception is the case when $j = L-1$, where we have to append both U_{L-1} and U_L as there are no future frames after frame X_L. Obviously, $\{V_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ is a regrouping of components of $\{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ and $\{U_\ell\}_{\ell=1}^{L}$. Therefore it represents the same collection of auxiliary variables.

2. **ISC Encoding.** As defined in Section 3.2, ISC-K encoding mapping $f_{j,k}$ can be decomposed as $f_{j,k} = \phi_{j,k} \circ h_{j,k}$, where $h_{j,k}$ denotes the deterministic interpolation function and $\phi_{j,k}$ denotes a sequential encoder. For deterministic functions $h_{j,k}$, we adopt the following realizations:

$$h_{j,k}(X_j, X_k) = \tilde{Y}_{j,k} \triangleq \begin{cases} \tilde{X}_{j,k}, & j < L-1,\ k < \min\{j+K, L\}, \\ (\tilde{X}_{j,k}, X_j), & j < L-1,\ k = \min\{j+K, L\}, \\ (\tilde{X}_{j,k}, X_j, X_k), & j = L-1,\ k = L. \end{cases} \tag{232}$$

   Here, $\tilde{X}_{j,k}$ is attained by performing a linear interpolation between (X_j, X_k) as stated in Theorem C.5. Apparently we have the following Markov chains:

$$\left.\begin{array}{r} U_{j,k} \to \tilde{X}_{j,k} \to (X_j, X_k) \to \mathbf{X} \\ U_\ell \to X_\ell \to \mathbf{X}_{-\ell} \end{array}\right\} \Rightarrow V_{j,k} \to \tilde{Y}_{j,k} \to (X_j, X_k) \to \mathbf{X}. \tag{233}$$

   Since the RHS Markov condition in equation 233 is stronger than the encoder-side Markov condition of ISC-K system in equation 8. Therefore, the constructed auxiliary variables $V_{j,k}$ and $\tilde{Y}_{j,k}$ are a valid instantiation of encoder test channels permitted of ISC-K system. Then we perform the ISC-K encoder on source pair (X_j, X_k) to generate $V_{j,k}$.

3. **ISC Decoding.** As in equation 8, given received $V_{\le\ell}$ at ℓ-stage ISC-K decoder, we adopt the deterministic decoding function $\hat{X}_\ell = g_\ell(V_{\le\ell})$.

For augmented auxiliary variables $\{V_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ and coding scheme constructed above, we have the following key observations:

1. Since each $V_{j,k}$ contains all information of $U_{j,k}$, it also follow the Markov constraints analogous to $U_{j,k}$ stated in eqs. (192) and (193).

2. Based on the definition of ISC-K encoder, for each (j,k)-encoding stage and auxiliary variable $V_{j,k}$, the associate rate $\bar{R}_{j,k}$ follows from the conditional source coding theorem.

$$\bar{R}_{j,k} = I(\tilde{Y}_{j,k}; V_{j,k} | V_{<j,k}). \tag{234}$$

   Note that this has the same mutual information form as in Definition C.1, while ISC-K coding is optimized over all admissible auxiliary variables.

Now for those constructed auxiliary variables, we can characterize the associate sum-rate and per-letter distortion. Specifically, for $j < L-1$ and $k < \min\{j+K, L\}$, the associate rate $\bar{R}_{j,k}$ follows by chain rule:

$$\bar{R}_{j,k} = I(\tilde{Y}_{j,k}; V_{j,k} | V_{<j,k}) \tag{235}$$

$$= I(\tilde{X}_{j,k}; V_{j,k} | V_{<j,k}) \tag{236}$$

$$= I(X_j, X_k; V_{j,k} | V_{<j,k}). \tag{237}$$

For $j < L-1$ and $k = \min\{j+K, L\}$, the associate rate $\bar{R}_{j,k}$ follows by chain rule:

$$\bar{R}_{j,k} = I(\tilde{Y}_{j,k}; V_{j,k}|V_{<j,k}) \tag{238}$$
$$= I(\tilde{X}_{j,k}, X_j; V_{j,k}|V_{<j,k}) \tag{239}$$
$$= I(X_j, X_k; V_{j,k}|V_{<j,k}). \tag{240}$$

For $j = L-1$ and $k = L$, the associate rate $R_{L-1,L}$:

$$\bar{R}_{L-1,L} = I(\tilde{Y}_{L-1,L}; V_{L-1,L}|V_{<L-1,L}) \tag{241}$$
$$= I(\tilde{X}_{L-1,L}, X_{L-1}, X_L; V_{L-1,L}|V_{<L-1,L}) \tag{242}$$
$$= I(X_{L-1}, X_L; V_{L-1,L}|V_{<L-1,L}). \tag{243}$$

Now we can directly apply the same steps in eqs. (188) to (190) in the proof of Theorem C.2 to augmented auxiliary variables $\{V_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$, i.e.,

$$\sum_{(j,k)\in\mathcal{E}^{(K)}} \bar{R}_{j,k} = \sum_{(j,k)\in\mathcal{E}^{(K)}} I(X_j, X_k; V_{j,k}|V_{<j,k}) \tag{244}$$
$$= \sum_{(j,k)\in\mathcal{E}^{(K)}} I(\mathbf{X}; V_{j,k}|V_{<j,k}) \tag{245}$$
$$= I(\mathbf{X}; \{V_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}) \tag{246}$$
$$= I(\mathbf{X}; \{U_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}, U_1, \cdots, U_L) \tag{247}$$

Now for decoding distortion of each source node X_ℓ, since $\mathbf{X}$ and $\{V_{j,k}\}_{(j,k)\in\mathcal{E}^{(K)}}$ are jointly Gaussian, we can adopt the decoding introduced in the proof of Theorem C.5, i.e. $\mathbb{E}[X_\ell|U_{\ell,\mathcal{N}(\ell)}, U_\ell]$, which is valid for the ISC architecture based on the definition of ℓ-th decoding function, i.e. $\hat{X}_\ell = g_\ell(V_{\leq\ell})$. Therefore we have the induced distortion $\mathbf{D} = (D_1, \cdots, D_L) \in \mathcal{D}_\epsilon^{\mathrm{G}}$ and the corresponding rate-distortion function in equation 227. Notice that given the induced distortion $\mathbf{D}$, this is only one possible realization of ISC-K coding system. Therefore we have the following upper bound:

$$R_{K,\mathrm{sum}}^{\mathrm{ISC}}(\mathbf{D}) \leq \bar{R}_{\mathrm{sum}}(\mathbf{D}) = \frac{1}{2}\log\frac{|\Sigma|}{\prod_{\ell=1}^{L} D_\ell}. \tag{248}$$

Step 2. Converse part
For the Gauss-Markov source $\mathbf{X}$ and arbitrary given distortion tuple $\mathbf{D} \in \mathcal{D}_\epsilon^{\mathrm{G}} \subseteq \mathcal{D}^{\mathrm{JC}}$, we can again use the Shannon lower bound:

$$R_{K,\mathrm{sum}}^{\mathrm{ISC}}(\mathbf{D}) \geq R^{\mathrm{JCG}}(\mathbf{D}) = \frac{1}{2}\log\frac{|\Sigma|}{\prod_{\ell=1}^{L} D_\ell}. \tag{249}$$

Then apparently, for K-order Gauss-Markov source and arbitrary distortion $\mathbf{D} \in \mathcal{D}_\epsilon^{\mathrm{G}}$, we can adopt the ISC-$K$ coding scheme alongside those auxiliary random variables constructed in Theorem C.5 and achieve the Shannon lower bound, thereby the joint-coding rate-distortion function. This completes the proof. □

C.5 Gaussian example

In this section, we provide the explicit numerical calculation for the two-frame Gaussian source $(X_1, X_2) \sim \mathcal{N}(0, \Sigma)$. We assume the symmetric setting with unit variances $\sigma_1^2 = \sigma_2^2 = 1$ and correlation coefficient $\rho > 0$, such that:

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{250}$$

C.5.1 Symmetric distortion case

We first consider the symmetric distortion constraint:

$$\mathbb{E}[\|X_j - \hat{X}_j\|^2] \leq D, \quad j = 1, 2 \tag{251}$$

Joint coding. Recall the Gaussian joint coding rate-distortion function in equation 175. Let Σ_e denote the reconstruction error covariance matrix. Under the symmetric distortion constraint, we parametrize:

$$\Sigma_e = \begin{bmatrix} d & \theta \\ \theta & d \end{bmatrix}. \tag{252}$$

where $d \triangleq (\Sigma_e)_{11} = (\Sigma_e)_{22}$. Then the joint coding rate-distortion optimization problem becomes:

$$\begin{aligned} R_{\mathrm{JC}}^G(\mathbf{D}) = \min_{\Sigma_e} \ & \frac{1}{2}\log\frac{1-\rho^2}{|\Sigma_e|} \\ \text{s.t. } & 0 \prec \Sigma_e \preceq \Sigma, \\ & (\Sigma_e)_{jj} \le D, \qquad j = 1, 2. \end{aligned} \tag{253}$$

Under the above parameterizations, this further reduces to

$$\begin{aligned} R_{\mathrm{JC}}^G(D) = \min_{\theta} \ & \frac{1}{2}\log\frac{1-\rho^2}{D^2-\theta^2} \\ \text{s.t. } & -D < \theta < D, \\ & \rho + D - 1 \le \theta \le \rho + 1 - D. \end{aligned} \tag{254}$$

Solving the above optimization yields:

$$R_{\mathrm{JC}}^G(D) = \begin{cases} \frac{1}{2}\log\frac{1-\rho^2}{D^2}, & 0 < D \le 1-\rho, \\ \frac{1}{2}\log\frac{1+\rho}{2D+\rho-1}, & 1-\rho < D \le 1, \\ 0, & D \ge 1. \end{cases} \tag{255}$$

The regime $0 < D \le 1 - \rho$ corresponds to the high-resolution region where the Shannon lower bound is tight.

Sequential coding. We next consider the strictly causal sequential coding setting. According to Corollary 1.2 in Ma et al. (2007), the sequential coding sum-rate function for the two-frame Gaussian source (X_1, X_2) under symmetric distortion constraint is given by

$$R_{\mathrm{seq,sum}}(D) = \begin{cases} \frac{1}{2}\log\frac{1}{D} + \frac{1}{2}\log\frac{\rho^2 D + (1-\rho^2)}{D}, & 0 < D \le 1, \\ 0, & D \ge 1. \end{cases} \tag{256}$$

The above expression can be interpreted as the sum of two terms corresponding to the sequential coding process:

$$R_1(D) = \frac{1}{2}\log\frac{1}{D}, \quad R_2(D) = \frac{1}{2}\log\frac{\rho^2 D + (1-\rho^2)}{D}. \tag{257}$$

The first term corresponds to encoding the first source X_1 without future information. The second term represents the additional rate required for encoding X_2 conditioned on the reconstruction of X_1. Compared with the joint-coding rate-distortion function in equation 255, the sequential coding system incurs a nonzero performance gap due to the causal constraint, which prevents the encoder from jointly exploiting the correlation between (X_1, X_2).

ISC. We now consider the proposed ISC-1 framework. For the symmetric distortion setting, we fix the interpolation coefficient to $\omega_{1,2} = \frac{1}{2}$, and construct the interpolation-based auxiliary variable:

$$\tilde{X}_{1,2} = \frac{1}{2}X_1 + \frac{1}{2}X_2, \quad U_{1,2} = \tilde{X}_{1,2} + N_{1,2}, \tag{258}$$

where $N_{1,2}$ is Gaussian noise independent of (X_1, X_2). The interpolated source has $\sigma^2_{\tilde{X}_{1,2}} = \frac{1+\rho}{2}$. In the high resolution regime corresponding to the Markov chain $X_1 \to U_{1,2} \to X_2$, the noise variance is given by $\sigma^2_{N_{1,2}} = -\frac{1}{4J_{12}} = \frac{1-\rho^2}{4\rho}$. This induces the distortion region:

$$\mathcal{D}_\epsilon = \{(D_1, D_2) : D_1, D_2 \le 1 - \rho\} \tag{259}$$

where the JC optimality is achieved through the introduction of additional auxiliary random variables U_1 and U_2.

Next, consider the low-resolution regime $D \in (1-\rho, 1]$. Under the symmetric distortion constraint, we have

$$\begin{aligned} D = D_1 &= \mathbb{E}[\|X_1 - \hat{X}_1\|^2] \\ &= \sigma_1^2 - \frac{\mathbb{E}[X_1 U_{1,2}]^2}{\sigma_{U_{1,2}}^2} \\ &= 1 - \frac{\frac{1}{4}(1+\rho)^2}{\frac{1+\rho}{2} + \sigma_{N_{1,2}}^2}. \end{aligned} \tag{260}$$

Solving for the noise variance yields:

$$\sigma_{N_{1,2}}^2 = \frac{(1+\rho)(2D+\rho-1)}{4(1-D)}. \tag{261}$$

The corresponding rate is

$$\begin{aligned} R &= I(\tilde{X}_{1,2}; U_{1,2}) \\ &= \frac{1}{2}\log\left(\frac{\sigma_{\tilde{X}_{1,2}}^2}{\sigma_{N_{1,2}}^2} + 1\right) \\ &= \frac{1}{2}\log\frac{1+\rho}{2D+\rho-1} \end{aligned} \tag{262}$$

This coincides exactly with the joint-coding rate-distortion function in equation 255 for the low-resolution regime. Combined with the high-resolution result above, ISC-1 therefore achieves the Gaussian joint-coding rate-distortion performance for the two-frame Gaussian source.

C.5.2 Asymmetric distortion

We next consider the asymmetric distortion constraints:

$$\mathbb{E}[\|X_1 - \hat{X}_1\|^2] \le D_1, \quad \mathbb{E}[\|X_2 - \hat{X}_2\|^2] \le D_2, \tag{263}$$

where $D_1 \in (0, 1]$ is fixed. For convenience to compare within the high resolution regime, we further assume $D_1 \in (0, 1-\rho^2]$.

Joint coding. Under the asymmetric distortion setting, the reconstruction error covariance matrix takes the form

$$\Sigma_e = \begin{bmatrix} D_1 & \theta \\ \theta & d_2 \end{bmatrix} \tag{264}$$

where $d_2 \le D_2$. Then joint-coding rate-distortion optimization becomes:

$$\begin{aligned} R_{\mathrm{JC}}^G(\mathbf{D}) = \min_{\theta, d_2}\ & \frac{1}{2}\log\frac{1-\rho^2}{D_1 d_2 - \theta^2} \\ \text{s.t. } & D_1 d_2 - \theta^2 > 0, \\ & (1-D_1)(1-d_2) - (\rho-\theta)^2 \ge 0, \\ & d_2 \le D_2. \end{aligned} \tag{265}$$

Solving the above optimization yields the following rate-distortion function:

$$R_{\mathrm{JC}}^G(\mathbf{D}) = \begin{cases} \frac{1}{2}\log\frac{1-\rho^2}{D_1 D_2}, & 0 < D_2 \le \dfrac{1-\rho^2-D_1}{1-D_1}, \\ \frac{1}{2}\log\frac{1-\rho^2}{D_1 D_2 - \left(\rho - \sqrt{(1-D_1)(1-D_2)}\right)^2}, & \dfrac{1-\rho^2-D_1}{1-D_1} < D_2 < 1-(1-D_1)\rho^2, \\ \frac{1}{2}\log\frac{1}{D_1}, & D_2 \ge 1-(1-D_1)\rho^2. \end{cases} \tag{266}$$

The three regions correspond to different operating regimes of the Gaussian joint-coding solution. In particular, when D_2 is sufficiently small, the Shannon lower bound is tight and the optimal rate-distortion function reduces to the product form in the first case. As D_2 increases, the optimal reconstruction error covariance becomes correlated, leading to the middle regime. Finally, when D_2 is sufficiently large, only the distortion constraint on X_1 remains active, and the sum-rate reduces to the scalar Gaussian rate-distortion function of X_1 alone.

Sequential coding. For the strictly causal sequential coding setting, the sum-rate function under asymmetric distortion constraints is given by:

$$R_{\text{seq,sum}}(\mathbf{D}) = \begin{cases} \frac{1}{2}\log\frac{1}{D_1} + \frac{1}{2}\log\frac{\rho^2 D_1 + (1-\rho^2)}{D_2}, & D_2 < \rho^2 D_1 + (1-\rho^2), \\ \frac{1}{2}\log\frac{1}{D_1}, & \rho^2 D_1 + (1-\rho^2) \le D_2 \le 1, \\ 0, & D_1 \ge 1,\ D_2 \ge 1. \end{cases} \tag{267}$$

The above characterization still reflects the sequential nature of the coding process. When $D_2 \ge \rho^2 D_1 + (1-\rho^2)$, the distortion requirement on X_2 can already be satisfied using only the reconstruction of X_1, and therefore no additional rate is required for the second stage.

ISC. We now analyze the proposed ISC-1 under asymmetric distortion constraints. Unlike the symmetric setting, the interpolation coefficient $\omega_{1,2}$ is optimized jointly with the distortion constraint. From the proof of Lemma C.3, for $\rho > 0$, the optimal interpolation coefficient satisfies $\omega_{1,2} > 0$. We first characterize the high resolution regime where JC optimality is achieved through additional auxiliary random variables U_1 and U_2. Setting $\epsilon_1 \triangleq D_1$, the corresponding distortion threshold for the second source is

$$\epsilon_2 = \frac{1-\rho^2-\epsilon_1}{1-\epsilon_1}. \tag{268}$$

Therefore, for arbitrary $D_1 \in (0, 1-\rho^2]$ and $D_2 \le \frac{1-\rho^2-D_1}{1-D_1}$, ISC achieves JC rate-distortion function. Notably, the region coincides with the high-resolution regimes of the Gaussian joint-coding solution in equation 266. Outside this region, the sum-rate optimization reduces to:

$$\begin{aligned} R_{1,\text{sum}}(\mathbf{D}) = \min_{\omega_{1,2}} \frac{1}{2}\log\frac{\sigma^2_{U_{1,2}}}{\sigma^2_{N_{1,2}}} \\ \text{s.t. } \mathbb{E}[\|X_1 - \hat{X}_1\|^2] = D_1, \\ \mathbb{E}[\|X_2 - \hat{X}_2\|^2] \le D_2. \end{aligned} \tag{269}$$

The variance term of auxiliary variables are:

$$\sigma^2_{U_{1,2}} = \sigma^2_{\tilde{X}_{1,2}} + \sigma^2_{N_{1,2}} \tag{270}$$

$$\sigma^2_{\tilde{X}_{1,2}} = \omega^2_{1,2} + 2\omega_{1,2}(1-\omega_{1,2})\rho + (1-\omega_{1,2})^2. \tag{271}$$

We first consider the equality distortion constraint on the first source:

$$\begin{aligned} D_1 &= \mathbb{E}[\|X_1 - \hat{X}_1\|^2] \\ &= 1 - \frac{\left(\omega_{1,2} + (1-\omega_{1,2})\rho\right)^2}{\sigma^2_{U_{1,2}}}. \end{aligned} \tag{272}$$

Equivalently, we have the noise variance:

$$\sigma^2_{U_{1,2}} = \frac{\left(\omega_{1,2} + (1-\omega_{1,2})\rho\right)^2}{1-D_1}. \tag{273}$$

For the second source, we obtain:

$$\begin{aligned} D_2 &\ge \mathbb{E}[\|X_2 - \hat{X}_2\|^2] \\ &= 1 - \left(\frac{\omega_{1,2}\rho + (1-\omega_{1,2})}{\omega_{1,2} + (1-\omega_{1,2})\rho}\right)^2 (1-D_1). \end{aligned} \tag{274}$$

Define:

$$t \triangleq \frac{\omega_{1,2}\rho + (1-\omega_{1,2})}{\omega_{1,2} + (1-\omega_{1,2})\rho}. \tag{275}$$

Since $\omega_{1,2} \in (0,1)$, we have $t > 0$. The distortion constraint in equation 274 becomes

$$t^2 \ge \frac{1-D_2}{1-D_1}. \tag{276}$$

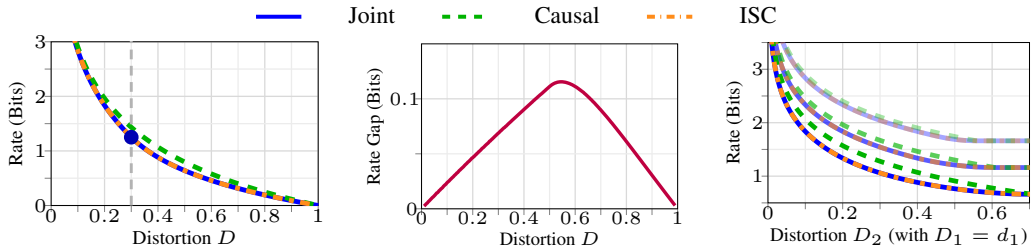


Figure 12: Gaussian source (X_1, X_2) with $\sigma_1^2 = \sigma_2^2 = 1$ and $\rho = 0.7$. *Left*: rate-distortion $R(D)$ comparisons for symmetric distortion constraint D. Interpolation coefficient $\omega_{1,2} = 0.5$ for ISC. *Middle*: Rate gap between joint coding (ISC) and causal coding for symmetric distortion case. *Right*: rate-distortion $R(D_2)$ comparisons for asymmetric distortion constraint with fixed $d_1 = 0.1, 0.2$, and 0.4 upside down.

which is equivalent to $t \geq \sqrt{\frac{1-D_2}{1-D_1}}$. Substituting this relation to the rate expression yields

$$\begin{aligned}\frac{1}{2}\log\frac{\sigma_{U_{1,2}}^2}{\sigma_{N_{1,2}}^2} &= \frac{1}{2}\log\frac{1}{1-\sigma_{\tilde{X}_{1,2}}^2/\sigma_{U_{1,2}}^2} \\ &= \frac{1}{2}\log\frac{1}{1-\frac{1-D_1}{(\omega_{1,2}+(1-\omega_{1,2})\rho)^2}\sigma_{\tilde{X}_{1,2}}^2} \\ &= \frac{1}{2}\log\frac{1}{1-\frac{1-D_1}{1-\rho^2}(1-2\rho t+t^2)}\end{aligned} \tag{277}$$

Therefore, the optimization reduces to minimizing $1 - 2\rho t + t^2$, whose unconstrained minimizer is $t^* = \rho$. Combining this with the distortion constraint in equation 274 gives:

$$t^* = \max\left\{\rho, \sqrt{\frac{1-D_2}{1-D_1}}\right\}. \tag{278}$$

When $t^* = \rho$, the coding degenerates to $\omega_{1,2} = 1$, corresponding to coding only the first source. Otherwise, we have the optimal interpolation coefficient:

$$\omega_{1,2}^* = \frac{1-\rho\sqrt{\frac{1-D_2}{1-D_1}}}{1-\rho+(1-\rho)\sqrt{\frac{1-D_2}{1-D_1}}}, \tag{279}$$

for $D_2 < 1 - \rho^2(1 - D_1)$. Finally, the ISC sum-rate-distortion function becomes:

$$R_{1,\text{sum}}(\mathbf{D}) = \begin{cases} \frac{1}{2}\log\frac{1-\rho^2}{D_1 D_2}, & 0 < D_2 \leq \dfrac{1-\rho^2-D_1}{1-D_1}, \\ \frac{1}{2}\log\frac{1-\rho^2}{D_1 D_2 - \left(\rho-\sqrt{(1-D_1)(1-D_2)}\right)^2}, & \dfrac{1-\rho^2-D_1}{1-D_1} < D_2 < 1-(1-D_1)\rho^2, \\ \frac{1}{2}\log\frac{1}{D_1}, & D_2 \geq 1-(1-D_1)\rho^2. \end{cases} \tag{280}$$

This coincides exactly with the Gaussian joint-coding rate-distortion function in equation 266, thereby establishing the JC optimality of ISC-1 for the asymmetric distortion setting.

C.5.3 Summary

We provide a simple two-frame Gaussian example to illustrate the behavior of the proposed ISC framework. Consider jointly Gaussian sources (X_1, X_2) with a quadratic distortion measure.

Symmetric distortion. We first consider the symmetric setting $D_1 = D_2 = D$. Figure 12 (left) shows the rate-distortion curves for joint coding, causal sequential coding, and ISC-1. In this case, the sum-rate is denoted by $R_{\text{sum}}(D^2)$. As shown, ISC-1 achieves the same rate-distortion performance as joint coding in the high-resolution regime by introducing auxiliary variables through interpolation. Moreover, it strictly outperforms causal coding across the entire rate region.

Asymmetric distortion. We next consider the case where D_1 is fixed and vary D_2. Figure 12 (right) shows the corresponding rate-distortion curves, denoted by $R_{D_1,\mathrm{sum}}(D_2)$. In this setting, the interpolation coefficients can be optimized according to the distortion constraint, and ISC achieves joint coding optimality across the entire feasible rate region.

Discussion. This example illustrates how interpolation-based auxiliary variables enable ISC to capture delayed dependencies efficiently. While the exact optimality result is specific to Gaussian sources, it provides a clear design principle for practical video coding: temporal dependencies can be exploited through local interpolation within a limited delay window without sacrificing coding efficiency.

D EXPERIMENT DETAILS

The proposed GFSC architecture is model-agnostic and can be integrated into different neural video compression (NVC) models. In this section, we provide more implementation details about model-training, evaluation, and baseline setups. We also explicitly show how GFSC can be incorporated into DCVC (Li et al., 2021) architecture, which serves as our primary baseline. The same integration principle can be applied to other NVC architectures, such as VCT and DHVC, with only minor modifications.

D.1 TRAINING SETUP AND OVERVIEW

Training and evaluation. All models are trained on Vimeo-90K (Xue et al., 2019) using 7-frame sequences ($\mathrm{GOP} = 7$), with the remaining training hyperparameters following the corresponding original implementations. In our implementation, we focus on evaluating the effectiveness of the proposed GFSC mechanism rather than optimizing for peak performance. Therefore, for VCT and DHVC, we adopt a simplified training procedure corresponding to the final stage of VCT, where all components (encoder, decoder, and entropy model) are trained jointly from scratch. Specifically, we train the models for 50 epochs using their original rate-distortion objectives and optimization settings. For DCVC, we adopt the full multi-stage training procedure of the original model as in Zhang & Zhu (2025) and introduce the delayed-frame modules only in the final training stage. For evaluation, we report results on UVG (Mercat et al., 2020) and MCL-JCV (Wang et al., 2016). Following standard evaluation protocols, we encode the first 96 frames of each sequence with $\mathrm{GOP} = 32$. We evaluate rate-distortion performance using bitrate (bits-per-pixel, bpp) and PSNR computed in the RGB space. For conventional video codecs, reconstructed `.YUV` videos are first converted to RGB before computing MSE and PSNR to ensure a consistent comparison.

Rate-distortion training. We follow the standard rate-distortion optimization used in DCVC Li et al. (2021). First, the rate-distortion loss function is defined as

$$\mathcal{L} = R + \lambda D,$$

where R denotes the bitrate and D denotes the distortion. Both are averaged across the video sequence batch during training. The trade-off hyperparameter λ is selected from $\lambda \in 0.01 \times \{2, 1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}\}$, covering both low- and high-rate regimes. For other settings, we use the Adam optimizer with a batch size of 8, an initial learning rate of 1×10^{-5} and aux learning rate of 1×10^{-3}. Training is performed for 30 epochs with a linear learning rate schedule: the learning rate is warmed up for the first 2k steps and then linearly decayed to 1×10^{-6}. All models are trained on a single NVIDIA RTX 5090 GPU using 2 days.

Delayed coding baselines. As discussed in Section 4, we additionally compare ISC with practical delayed coding frameworks, including Random Access (RA) coding (i.e., bidirectional coding), under matched delay constraints ($K \leq 3$). We first clarify the connection between sequential coding delay and RA coding configurations.

Consider an L-frame video sequence. In RA coding, the encoder first buffers a short frame group and performs bidirectional prediction within the group. For example, in an `I-B-B-B-I` structure, intermediate B-frames depend on both past and future anchor frames. As a result, the encoder must wait for the arrival of the future I-frame before encoding the intermediate frames, leading to an

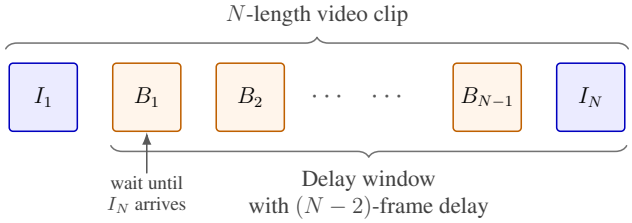


Figure 13: Illustration of the delay induced by B-frame coding over an N-length video clip.

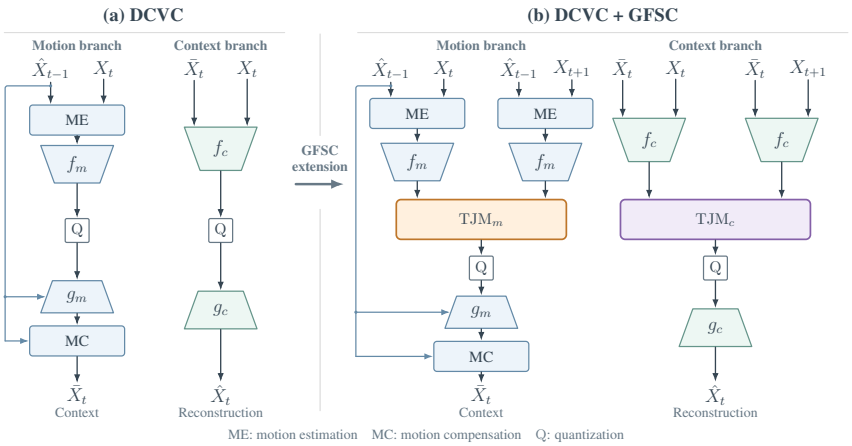


Figure 14: Simplified DCVC and GFSC architectures. GFSC introduces temporal joint modeling (TJM) modules to both motion and context branch, highlighted in *orange* and *purple*, respectively. One delayed frame case is shown. Additional delayed frames provide further inputs to these modules. The mappings f_m, f_c and g_m, g_c denote motion/context encoders and decoders, respectively.

effective practical delay determined by the distance between the first B-frame and the future anchor frame. The decoder then reconstructs frames in coding order and reorders them back to display order. Figure 13 illustrates the corresponding delay configurations.

We consider two representative delayed coding baselines: DCVC-B Sheng et al. (2025) and HM-RA. For DCVC-B, the neural architecture allows flexible extension to delay settings. Therefore, both $K = 1$ and $K = 2$ are implemented by customizing the reference frame lists for each frame. For HM-RA, the hierarchical GOP structure is designed primarily for even GOP sizes. Accordingly, we modify the standard `encoder_randomaccess_main.cfg` to implement 1-delay and 3-delay settings using `I-B-I` and `I-B-B-B-I` structures, respectively. The intermediate `I-B-B-I` configuration is not directly supported under the standard HM-RA hierarchical coding structure.

D.2 Network Architecture

We describe how GFSC is integrated into the causal DCVC architecture. The original DCVC codec contains separate I-frame and P-frame models. For P-frame coding, we distinguish two branches: a motion branch with analysis/synthesis transforms (f_m, g_m) and a context branch with transforms (f_c, g_c). Each branch retains its original entropy-coding module based on the CompressAI implementation (Bégaint et al., 2020). Generally, GFSC introduces temporal joint modeling (TJM) modules between the analysis transforms and the subsequent coding modules, while leaving the remaining DCVC pipeline unchanged. Figure 14 compares the original DCVC and DCVC-GFSC architectures for the one-frame-delay case.

One-frame delay ($K = 1$). For every stage $\ell < L$, GFSC jointly models the current representation and the available one-frame lookahead.

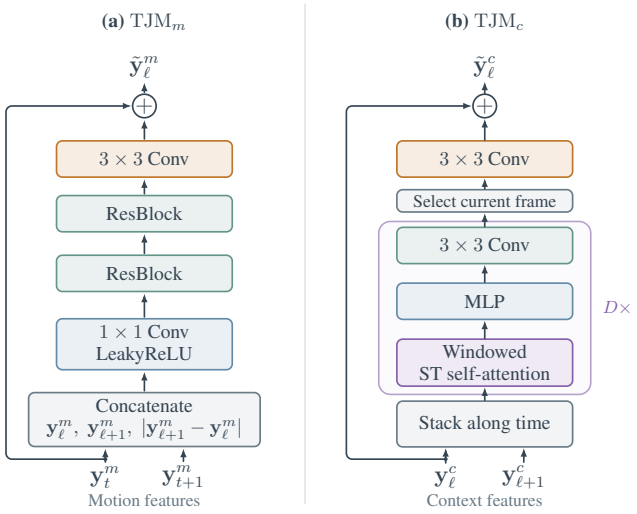


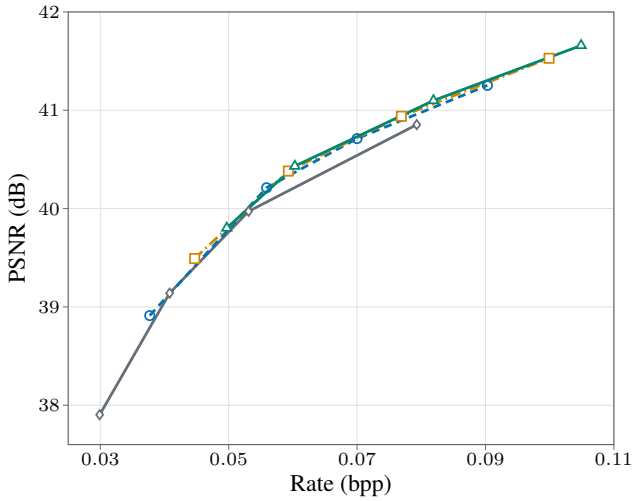
Figure 15: Temporal joint modeling modules for motion and context branches in DCVC architecture. *Left*: TJM_m concatenates motion features and their absolute difference as input. *Right*: TJM_c jointly updates the two context features using a spatial-temporal self attention module.

1. For an I-frame stage, let f_I and g_I denote the corresponding analysis and synthesis transforms. Given current and delayed frame pair $(\mathbf{x}_\ell, \mathbf{x}_{\ell+1})$, we first compute $\mathbf{y}_\ell = f_I(\mathbf{x}_\ell)$ and $\mathbf{y}_{\ell+1} = f_I(\mathbf{x}_{\ell+1})$. Then we have the following joint representation
$$\tilde{\mathbf{y}}_\ell = \mathrm{TJM}_c\left(\mathbf{y}_\ell, \mathbf{y}_{\ell+1}\right).$$
Here, TJM_c is implemented using the window-based spatial-temporal attention module described below.
2. For P-frames, DCVC contains separate motion and context representations. We therefore introduce two corresponding temporal modules, TJM_m and TJM_c. For each branch $b \in \{m, c\}$,
$$\tilde{\mathbf{y}}_\ell^b = \mathrm{TJM}_b\left(\mathbf{y}_\ell^b, \mathbf{y}_{\ell+1}^b\right), \qquad \mathbf{y}_\ell^b = f_b(\cdot).$$
The context module TJM_c uses spatial-temporal self-attention to jointly learn delayed latent representations, whereas TJM_m uses a lightweight convolutional module to jointly model the motion features. The resulting representations are then passed to the original entropy-coding and decoding pipeline. All P-frame stages share the same network.

Note that at the final stage $\ell = L$, no future frame is available. GFSC therefore bypasses temporal fusion and directly forwards the current latent representation $\tilde{\mathbf{y}}_L^b = \mathbf{y}_L^b$ for $b \in \{m, c\}$. Because the TJM outputs preserve the original latent tensor dimensions, the subsequent entropy models and synthesis transforms can be used without modification.

Extension to larger delays. The same construction extends directly to $K > 1$. For the context branch, the spatial-temporal attention module jointly processes all available latent representations within the delay window:
$$\tilde{\mathbf{y}}_\ell^c = \mathrm{TJM}_c\left(\mathbf{y}_\ell^c, \cdots, \mathbf{y}_{\min\{\ell+K, L\}}^c\right).$$
The attention formulation naturally supports a variable number of available future frames near the sequence boundary. For the motion branch, TJM_m uses a fixed-size input corresponding to the selected delay budget. For example, when $K = 2$,
$$\tilde{\mathbf{y}}_\ell^m = \mathrm{TJM}_m\left(\mathbf{y}_\ell^m, \mathbf{y}_{\ell+1}^m, \mathbf{y}_{\ell+2}^m\right).$$
If fewer than K future representations are available, the missing inputs are replaced by zero tensors of the same shape. The same padding rule is used during both training and inference.

Coding and quantization. After joint modeling within the K-delay window, the learned representation $\mathbf{z}_\ell$ is quantized and entropy-coded. Following the one-shot formulation in Section 3.2, we



BD-rate comparison.

Model	BD-rate (%)
DCVC	0.00
DCVC-GFSC-1	−3.78
DCVC-GFSC-2	−4.50
DCVC-GFSC-3	−5.11

Figure 16: LIVE-HDR rate-PSNR curves (left) and BD-rate relative to DCVC (right). Negative values indicate rate savings.

adopt universal quantization with shared randomness Gray & Stockham (2002); Ziv (2003); Theis & Agustsson (2021) to implement one-shot channel simulation. Specifically, the encoder computes

$$\mathbf{z}_\ell = \Delta \left\lfloor \frac{\tilde{\mathbf{y}}_\ell + \mathbf{w}}{\Delta} \right\rceil,$$

where $\mathbf{w} \sim \mathcal{U}(-\Delta/2, \Delta/2)$ denotes shared randomness. At the decoder side, the quantized latent is reconstructed as

$$\mathbf{u}_\ell = \mathbf{z}_\ell - \mathbf{w}, \tag{281}$$

since the randomness $\mathbf{w}$ is shared between the encoder and decoder. Let g_ℓ denote the decoder network. The reconstructed frame is then obtained by:

$$\hat{\mathbf{x}}_\ell = g_\ell(\mathbf{u}^\ell). \tag{282}$$

D.3 Additional experimental analysis

We further evaluate GFSC-enhanced DCVC on the 4K LIVE-HDR dataset (Shang et al., 2022). Compared with causal DCVC, GFSC achieves BD-rate reductions of 3.78%, 4.50%, and 5.11% for delays $K = 1, 2, 3$, respectively. These results show that exploiting information from delayed frames also improves rate-distortion performance for 4K video.

E Limitations

Despite the theoretical and empirical results established in this work, several questions remain open. First, our joint-coding optimality result is characterized over a nontrivial distortion region. Extending this characterization beyond the current region, and determining whether alternative auxiliary-variable constructions can enlarge the optimality regime, are important directions for future work. This limitation concerns only the extent of the characterized region and does not affect the established optimality guarantee within it.

Second, while the general Markov analysis establishes the existence and graphical structure of JC-optimal clique-wise auxiliary variables, it does not yield a simple closed-form parameterization for arbitrary source distributions. For Gauss-Markov sources, the additional Gaussian structure makes this construction explicit and reduces the clique-wise interactions to pairwise linear interpolation. Developing similarly tractable constructions for broader non-Gaussian source families is an interesting theoretical direction.

Finally, the computational cost of GFSC grows approximately with the delay budget because additional future-frame interactions must be modeled. This overhead is moderate in the limited-delay

regime studied in this work, but scaling GFSC to substantially longer delays while controlling computation and memory remains an important engineering direction.