LEARNING SELECTIVE REFERENCE DYNAMICS FOR PERCEPTUAL VIDEO COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recurrent video codecs must exploit decoded history while limiting the reuse of reference errors. We introduce SRDVC, which separates predictive reference evolution from reference-conditioned coding. A Latent Dynamical System factors reference prediction into dynamics estimation, State evolution, and Observation prediction. Dynamical Manifold Alignment uses current-step features to query the predicted reference, while residual paths convey content not captured by the prediction. A local perturbation model motivates direction-dependent propagation: under bounded total input, uniform contraction of an orthogonal component bounds its accumulation. This conditional analysis provides a design rationale, while temporal-subspace probes characterize the behavior learned by the codec. At matched compute on UVG, replacing LDS, DMA, or both degrades perceptual rate–distortion performance; replacing both increases LPIPS/DISTS BD-Rate by 52.41%/65.76%. Across these variants, stronger relative directional selectivity accompanies better compression. On UVG, MCL-JCV, and HEVC-B, SRDVC achieves mean BD-Rate reductions of 72.67% under LPIPS and 63.97% under DISTS relative to GLC-video, with efficient steady-state 1080p P-frame coding and favorable viewer preferences over the tested subjective-study baselines.

1 INTRODUCTION

Perceptual neural video compression optimizes reconstructed videos for visual quality as well as coding efficiency (Qi et al., 2025; Zheng et al., 2026). Its temporal coding context is built from decoded information carried from earlier frames (Lu et al., 2019; Li et al., 2023; 2024; Jia et al., 2025; Qi et al., 2025; Zheng et al., 2026). This recurrent reference supplies useful predictive content, but can also propagate deviations introduced by quantization, reconstruction, and imperfect prediction.

We study how a codec should organize the evolution and use of this reference. Preserving all variation does not distinguish predictive temporal changes from perturbations, while uniform attenuation may also suppress useful variation. We formulate this design objective as selective reference dynamics: retain responsiveness to temporal evolution while limiting the propagation of other deviations. A local manifold view provides an idealized directional interpretation of this objective (Fefferman et al., 2016). The analysis in Sec. 3 describes sufficient conditions for bounded orthogonal perturbation accumulation; it provides a design rationale rather than a stability certificate for the implemented codec.

SRDVC separates predictive reference evolution from reference-conditioned coding. Its Latent Dynamical System (LDS) estimates dynamics, advances a compact State, and predicts an Observation using only decoded history, following latent world models (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; Wu et al., 2025). Its Dynamical Manifold Alignment (DMA) assigns current-step features to Queries and the predicted reference to Keys and Values. This source assignment makes the attention branch reference-conditioned, while residual paths carry content not represented by the prediction. The reconstructed Observation returns to the reference buffer for the next step.

Our strongest controlled evidence comes from structural replacements at matched compute. Replacing LDS with a ConvGRU update, DMA with concatenation and convolution, or both increases perceptual BD-Rate while retaining the specified interfaces and training protocol. Encoding and decoding MAC counts differ by less than 1% across these variants. Complementary temporal-subspace

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.

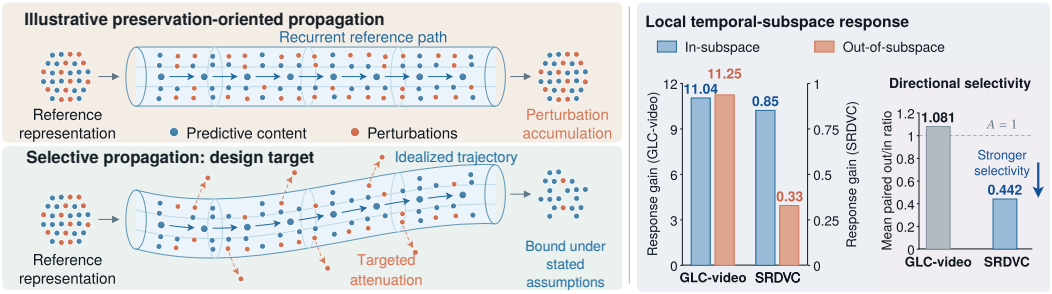


Figure 1: Selective reference dynamics: design target and local diagnostics. Left: schematic comparison of preservation-oriented propagation and the desired selective response; bounded accumulation requires the assumptions in Sec. 3. Right: sampled temporal-subspace responses measured in each model's own representation, with model-specific gain scales. These measurements describe local directional responses rather than certify closed-loop stability.

probes show that the full model has stronger relative directional selectivity in the reconstructed Observation space, and that the ordering of the diagnostic matches the ordering of RD penalties. We use this correspondence to characterize the learned behavior, without identifying selectivity as the cause of the compression gains.

Across three benchmarks, SRDVC achieves strong perceptual rate–distortion performance together with efficient P-frame coding. Bitrate-matched subjective comparisons provide additional evidence of viewer preference over the two tested baselines. These results support the explicit organization of reference prediction and use as a practical design dimension for perceptual video compression.

Our contributions are:

1. **A structured predictive-reference architecture.** We formulate reference propagation as selective dynamics and introduce SRDVC, combining factorized latent reference evolution with an asymmetric reference-conditioning interface.
2. **Compute-controlled evidence of architectural value.** Matched-compute replacements demonstrate the value of LDS and DMA within the tested codec. Evaluation across three benchmarks shows strong perceptual RD performance, efficient P-frame coding, and viewer preference over the tested subjective-study baselines.
3. **Conditional analysis and learned-response diagnostics.** A local perturbation model motivates selective propagation under explicit assumptions. Temporal-subspace probes characterize learned directional responses and their association with the structural-ablation results.

2 Related Work

2.1 Neural Video Compression

Conditional neural video codecs recurrently reuse decoded information as the reference for the next frame, typically strengthening temporal contexts or longer-range dependencies (Lu et al., 2019; Li et al., 2023; 2024; Jia et al., 2025; Qi et al., 2025). Our work studies an explicit factorization of reference prediction into dynamics estimation, State evolution, and Observation prediction before current-frame coding. The distinction is this organization of the reference path, rather than the availability of decoded history or the causality of prediction, which are also shared by the compared recurrent codecs.

2.2 Perceptual Video Compression

Most neural video codecs are optimized for pixel-wise rate-distortion performance (Li et al., 2023; 2024; Jia et al., 2025), yet pixel-wise fidelity does not fully capture perceptual quality, which has motivated the use of LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022) for evaluation. For

video in particular, perceptual reconstruction depends on the temporal behavior of the propagated reference as well as on single-frame quality (Qi et al., 2025; Zheng et al., 2026). Perceptual objectives and generative priors are important ingredients in this line of work (Qi et al., 2025; Zheng et al., 2026). We complement them by studying how reference evolution and reference-conditioned coding are organized; SRDVC itself is also trained with a perceptual rate–distortion objective.

2.3 Latent Dynamics and Manifold Structure

World models and recurrent state-space models learn structured latent dynamics and predict observations from a compact State (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; Wu et al., 2025). Manifold-based representation learning describes structured data as concentrated near low-dimensional regions of the ambient space (Fefferman et al., 2016), which suggests separating variation along a temporal trajectory from deviations orthogonal to it. We use the manifold view as a local geometric abstraction for a directional design target. The empirical temporal subspaces are estimated from memory differences and are not identified tangent spaces of an underlying data manifold.

3 A Local Perturbation Model for Selective Reference Propagation

This section develops a local perturbation model that motivates the design target of SRDVC. It distinguishes a preservation-oriented reference case from a sufficient condition for bounded accumulation of an orthogonal component.

3.1 The Reference Carries Perturbations Forward

Neural video codecs reuse history through recurrent reference propagation (Li et al., 2023; 2024; Jia et al., 2025; Qi et al., 2025; Zheng et al., 2026). Let M_t denote the propagated reference representation at step t; its concrete instantiation in SRDVC and in the compared codecs is summarized in Appendix C.6. A useful M_t predicts the observation about to be coded. Alongside that predictive content, M_t carries perturbations from quantization error, reconstruction error, and prediction mismatch.

A local perturbation recursion describes how carried-over deviations respond to an update. Appendix D.1 gives the assumptions and derivation. Appendix D.2 discusses an illustrative connection to coding cost under additional additive-Gaussian and high-rate quantization assumptions; this surrogate is not a quantitative model of the measured perceptual BD-Rate.

3.2 Perturbation Accumulation under Preservation-Oriented Updates

As a reference case, consider a local update that approximately preserves perturbation energy. Assume zero-mean perturbations, uncorrelated carried-over and incoming terms, and a constant incoming variance σ_δ^2. Under these assumptions, the linearized recursion in Appendix D.1 gives

$$\sigma_{\epsilon,t}^2 \approx \sigma_{\epsilon,0}^2 + t\sigma_\delta^2. \tag{1}$$

This conditional example yields linear variance accumulation. It is not asserted to describe every recurrent codec. Uniform preservation does not distinguish predictive variation from perturbations, while uniform attenuation may also suppress useful temporal variation. This motivates studying direction-dependent propagation.

3.3 A Directional Target and a Conditional Accumulation Bound

We use a local manifold view as an idealization (Fefferman et al., 2016) in which some directions follow temporal evolution and others are orthogonal to it. Let J_S denote the local Jacobian of the reference update with its other inputs held fixed, and let $v^{\parallel}$ and $v^{\perp}$ denote idealized trajectory-aligned and orthogonal directions. The desired response is

$$\frac{\|J_S v^{\parallel}\|_2}{\|v^{\parallel}\|_2} \approx 1, \qquad \frac{\|J_S v^{\perp}\|_2}{\|v^{\perp}\|_2} \approx \alpha, \quad 0 < \alpha \ll 1. \tag{2}$$

This is a directional design target. In the diagnostics, temporal directions are approximated from recent memory differences rather than obtained from a known data manifold.

For the accumulated orthogonal perturbation, Appendix D.1 considers the local recursion $\epsilon_t^{\perp} = B_t \epsilon_{t-1}^{\perp} + \delta_t^{\perp}$, where B_t is the projected propagation operator. The total input $\delta_t^{\perp}$ includes newly introduced errors, cross-direction coupling, changes in other update inputs, and the linearization remainder. Assume along the trajectory that $\|B_t\|_2 \le \alpha < 1$, that $(\mathbb{E}\|\delta_t^{\perp}\|_2^2)^{1/2} \le d$, and that the initial second moment is finite. Defining $m_t = (\mathbb{E}\|\epsilon_t^{\perp}\|_2^2)^{1/2}$ gives

$$m_t \le \alpha^t m_0 + d\,\frac{1-\alpha^t}{1-\alpha}, \qquad \limsup_{t\to\infty} m_t \le \frac{d}{1-\alpha}. \tag{3}$$

Appendix D.1 gives the derivation and the constant-gain, uncorrelated-input variance model as a special case. Appendix D.2 relates these models to an additional coding penalty only under its stated residual-model assumptions.

Scope. The sufficient condition concerns the local perturbation model. Bounded total input includes the combined effect of feedback through the other update inputs. Neither a uniform operator bound nor a bounded closed-loop input condition is verified for the implemented codec. The measurements in Sec. 5.4 characterize sampled directional responses rather than certify this condition.

Section 4 describes the architecture motivated by this target. Section 5.3 evaluates its structural choices, and Sec. 5.4 characterizes their learned directional responses.

4 Proposed Method

4.1 Overview: What to Keep, What to Use

SRDVC separates two parts of the recurrent loop: reference evolution and reference-conditioned coding (Fig. 2). LDS advances a compact State using decoded history and predicts the reference $\hat{O}_t$. DMA uses current-step features as Queries and the predicted reference as Keys and Values, while residual paths carry content not represented by the prediction. The reconstructed Observation O_t is stored for use by LDS at the next step.

4.2 Latent Dynamical System: Predicting the Reference from the Past

LDS forms the reference for frame t from decoded history, without accessing frame t. Following the organization of latent world models (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; Wu et al., 2025), it estimates how the State should evolve, applies that evolution, and predicts an Observation from the evolved State:

$$D_t = f_D(S_{t-1}, O_{t-1}), \qquad S_t = f_S(S_{t-1}, D_t), \qquad \hat{O}_t = f_O(S_t, O_{t-1}), \tag{4}$$

where the Dynamic, State, and Observation Predictors f_D, f_S, and f_O carry out the three steps, and f_O anchors the rendering on O_{t-1}. Since no term depends on frame t, $\hat{O}_t$ must be extrapolated from the decoded past; content it does not predict reaches the coder through the residual paths of DMA. A generic recurrent update, such as the ConvGRU variant in Sec. 5.3, shares this property; what LDS adds is the factorization into evolution, State update, and rendering. The factorization makes State evolution an explicit component of the predictive reference path. Section 5.4 probes its local directional response; this component-level measurement is distinct from analyzing the complete recurrent buffer.

4.3 Dynamical Manifold Alignment: Reference-Conditioned Coding

DMA imposes an asymmetric source assignment in the fusion of current-step features and the predicted reference. At the encoder, current-frame features F_t supply the Queries, while $\hat{O}_t$ supplies

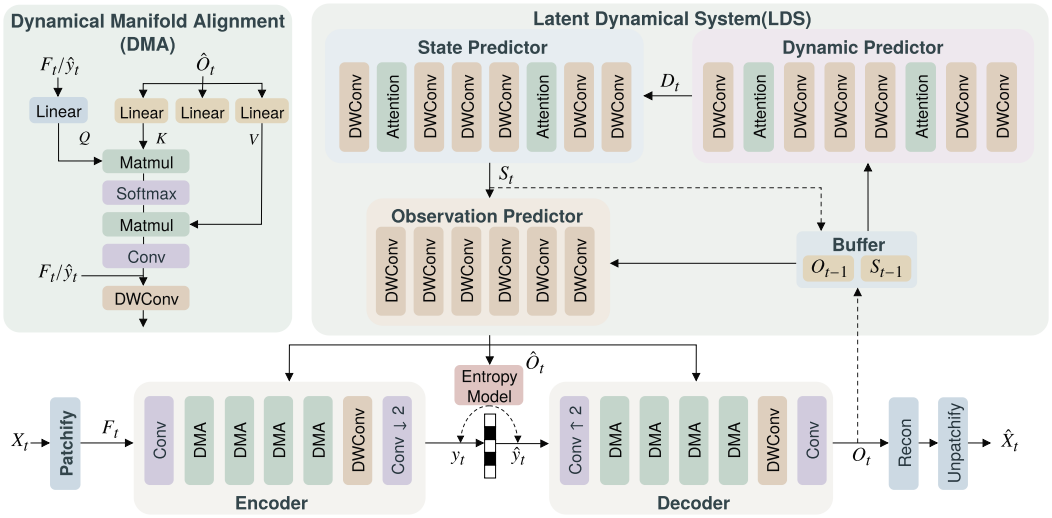


Figure 2: SRDVC. LDS advances the State and predicts $\hat{O}_t$ from decoded history. DMA conditions the coding paths on this prediction; the encoder uses current-frame features and the decoder uses decoded-latent features. The reconstructed Observation O_t is stored for the next reference update.

the Keys and Values. Suppressing residual connections, the attention-based content transformation (Vaswani et al., 2017) can be written as

$$Q_t = W_Q F_t, \qquad K_t = W_K \hat{O}_t, \qquad V_t = W_V \hat{O}_t, \qquad H_t^{\text{attn}} = \text{DWConv}(\text{Attn}(Q_t, K_t, V_t)) \,. \tag{5}$$

Here W_Q, W_K, and W_V are learnable projections. The expression specifies the source assignment in the attention-based path rather than the complete encoder latent: residual paths from the current-step features run around the attention and DWConv blocks, and the full encoder produces y_t for entropy coding.

Reference-sourced attention content. The retrieved Values come from the predicted reference. This gives the attention branch a reference-conditioned content source, while the residual paths can carry information not represented by the prediction. The source assignment is a structural bias toward using the reference, not a guarantee that the model cannot rely primarily on the current-step paths.

Query-conditioned relative weighting. Query–Key interactions determine relative weights among predicted features. These weights provide feature-dependent selection, but do not implement an explicit confidence gate that rejects an entirely mismatched reference. The residual paths provide a route for new content without requiring such content to be present in the prediction.

Explicit source separation. Unlike the tested concatenation–convolution replacement, DMA explicitly separates the features that determine the retrieval weights from the source of the retrieved Values. The structural ablation in Sec. 5.3 evaluates the overall effect of this fusion architecture while retaining both inputs; it does not separately identify the contribution of each design motivation.

At the decoder, Queries are computed from features of the decoded latent, not from the unavailable current frame. Keys and Values use the same past-derived prediction. The decoder reconstructs O_t, from which $\hat{X}_t$ is generated and the recurrent buffer is updated.

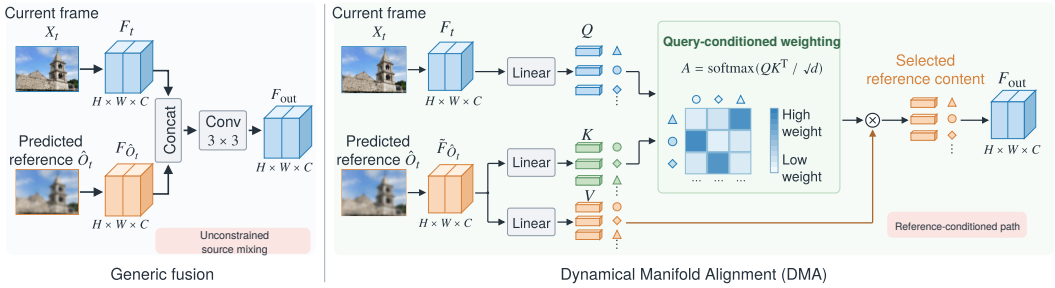


Figure 3: Source assignment in the tested fusion architectures. Left: concatenation and convolution combine current-step and predicted-reference features. Right: DMA uses current-step Queries with predicted-reference Keys and Values, making the attention branch reference-conditioned while retaining residual paths for new content. The diagram illustrates structural dependencies, not measured gradient magnitudes or a demonstrated elimination of shortcut learning.

4.4 Training

Training objective. We optimize the coding path with a perceptual rate–distortion objective, following the overall formulation of GVC1D (Zheng et al., 2026). For a training clip with T predicted frames,

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\left[R_t + \lambda(q) w_t \left(D_{\mathrm{RGB},t} + \beta D_{\mathrm{LPIPS},t}\right)\right],$$

where R_t is the estimated bitrate of the main latent and hyperlatent, $\lambda(q)$ controls the rate–distortion trade-off at base QP q, w_t is the temporal hierarchy weight, and $\beta = 1.2$ weights LPIPS relative to RGB reconstruction error. Appendix A.1 specifies the distortion normalization and training procedure.

Fine-tuning data. MASP is an auxiliary data-selection rule used only for 32-frame fine-tuning. It retains clips whose mean H.265 P-frame BPP at CRF=22 is below $\tau = 0.4$. Appendix A.4 specifies the rule, and the A/E comparison in Sec. 5.3 evaluates its effect within the tested setup.

5 Experiments

5.1 Experimental Setup

We train SRDVC on 7-frame Vimeo-90K sequences and fine-tune on 32-frame clips processed from raw Vimeo videos (Xue et al., 2019; Li et al., 2024; Jia et al., 2025; Zheng et al., 2026).

We evaluate the first 96 frames of each sequence on UVG, MCL-JCV, and HEVC-B (Mercat et al., 2020; Wang et al., 2016; Bossen, 2013). The comparison includes VTM-17.0, HM-16.25, ECM-5.0, DCVC-FM, DCVC-RT, GLC-video, and GVC1D. Available baseline implementations are evaluated using the settings documented in Appendix A. GVC1D entries are author-reported rather than rerun by us. LPIPS- and DISTS-based RD curves and BD-Rate (Zhang et al., 2018; Ding et al., 2022; Bjøntegaard, 2001) are reported, with PSNR results in Appendix B.1.

5.2 Perceptual Compression Performance

Figure 4 and Table 1 show that SRDVC has the best reported LPIPS BD-Rate on all three datasets and the lowest dataset-averaged DISTS BD-Rate among the listed methods. The average reductions relative to GLC-video are 72.67% under LPIPS and 63.97% under DISTS. The comparison is not a uniform dominance across metrics: GVC1D has better reported DISTS BD-Rate on MCL-JCV and HEVC-B, while SRDVC's larger improvement on UVG yields the lower average. Figure 5 and Appendix B.2 provide visual comparisons with DCVC-RT and GLC-video.

In the bitrate-matched 2AFC study, 60 participants evaluate 18 clips. SRDVC is preferred in 72.5% of judgments against DCVC-RT and 68.9% against GLC-video, with 1080 judgments per compar-

Table 1: Perceptual BD-Rate (%) relative to GLC-video; lower is better. GVC1D values are author-reported. SRDVC has the lowest LPIPS BD-Rate on each dataset and the lowest dataset-averaged DISTS BD-Rate, while GVC1D has lower DISTS BD-Rate on MCL-JCV and HEVC-B.

Method	UVG		MCL-JCV		HEVC-B		Average	
	LPIPS	DISTS	LPIPS	DISTS	LPIPS	DISTS	LPIPS	DISTS
GLC-video (Qi et al., 2025)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GVC1D (Zheng et al., 2026)	-66.97	-47.71	-62.27	**-61.57**	-65.06	**-68.72**	-64.77	-59.33
SRDVC (Ours)	**-79.61**	**-69.64**	**-64.22**	-57.21	**-74.19**	-65.05	**-72.67**	**-63.97**

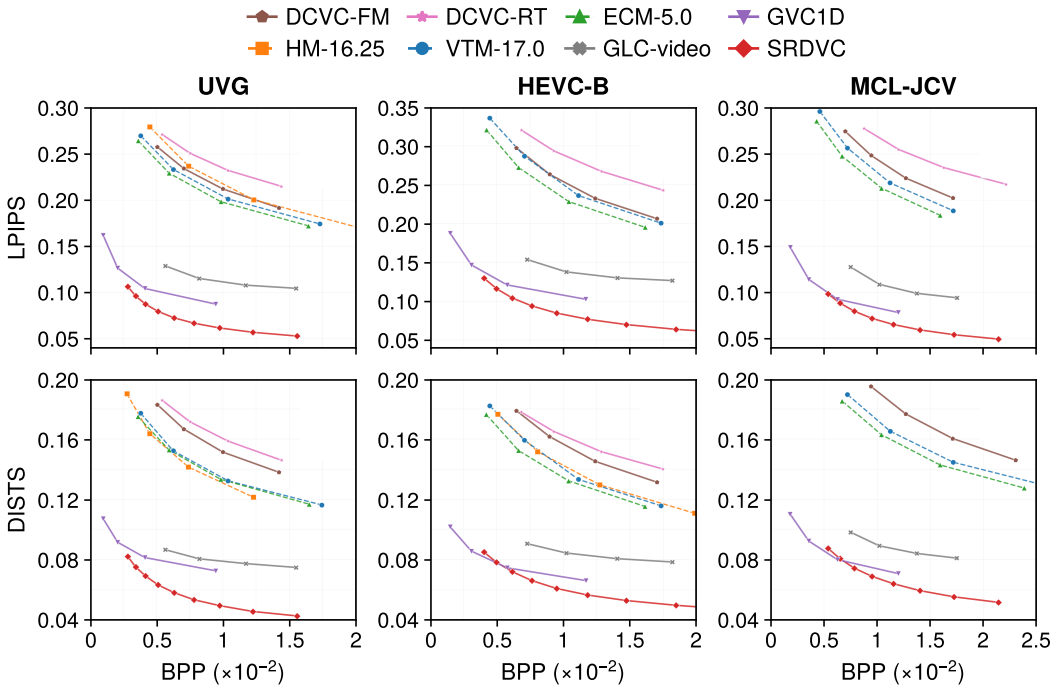


Figure 4: Perceptual RD curves on UVG, HEVC-B, and MCL-JCV. Lower LPIPS and DISTS are better. GVC1D curves are author-reported and were not independently rerun in this study.

ison. These are descriptive preference rates for the tested clips and baselines; Appendix C.4 gives the protocol.

5.3 Structural Ablations at Matched Compute

Table 2 evaluates the structural choices on UVG using five variants trained under the stated protocol. B and D replace LDS with a gated ConvGRU while retaining the observation extractor and the two output interfaces. C and D replace DMA with concatenation and convolution while retaining both the current-feature and predicted-reference inputs. All other components, including MASP, are unchanged across A–D; E omits only MASP.

Replacing LDS increases LPIPS/DISTS BD-Rate by 33.56%/46.77% relative to A. Replacing DMA increases it by 13.41%/12.63%, and replacing both increases it by 52.41%/65.76%. Encoding and decoding MAC counts differ from A by less than 1%, and the A–D P-frame parameter counts range from 31.52 to 32.94 million (Appendix A.2). These comparisons support the value of the proposed update and fusion structures within the tested codec, rather than attributing their benefit to a substantial compute increase.

The DMA replacement measures the overall effect of changing the fusion architecture while retaining both inputs; it does not separately test each motivation in Sec. 4.3. All BD-Rate penalties use



Method	BPP ↓	Relative BPP	LPIPS ↓	DISTS ↓
DCVC-RT	0.0066	1.20×	0.2718	0.2545
GLC-video	0.0064	1.16×	0.1356	0.1506
SRDVC (Ours)	**0.0055**	**1.00×**	**0.0912**	**0.1243**

Figure 5: Visual comparison with DCVC-RT and GLC-video. Relative BPP is normalized to SRDVC.

Table 2: Structural ablations on UVG. BD-Rate changes (%) are relative to A. Encoding and decoding MAC counts differ from A by less than 1% across A–D; parameter counts and interfaces are detailed in Appendix A.2. The post-training diagnostic A_{sel} is the mean paired out/in gain ratio in Observation space, not an absolute contraction measure. Dashes denote unreported results.

Model	Update	Fusion	MASP	BD-Rate (%)		Post-training
				LPIPS	DISTS	A_{sel} ↓
A	LDS	DMA	✓	**0.00**	**0.00**	**0.324**
C	LDS	Concat + conv	✓	+13.41	+12.63	0.502
B	ConvGRU	DMA	✓	+33.56	+46.77	0.703
D	ConvGRU	Concat + conv	✓	+52.41	+65.76	0.811
E	LDS	DMA	×	+5.9	+6.3	–

A as the anchor and are not additive. The post-training Observation-space diagnostic is reported alongside these results and interpreted in Sec. 5.4.

Removing MASP increases LPIPS/DISTS BD-Rate by 5.9%/6.3%. This is an auxiliary training-data effect; subset and threshold results are given in Appendix C.5.

5.4 Local Directional Responses of the Learned Reference Update

We use directional probes to characterize the learned reference update, complementing the structural RD comparisons rather than verifying the sufficient condition in Sec. 3.

Local temporal-subspace response. We estimate a temporal-subspace proxy from the leading r left singular vectors of K recent State differences. For the local update Φ with other update inputs held fixed, the one-step response is $G(v) = \|\Phi(S_t + \epsilon v) - \Phi(S_t)\|_2/(\epsilon\|v\|_2)$. At $S_{24} \to S_{25}$ with $K = 8$ and $r = 2$, the mean gains are $G_{\text{in}} = 0.852$ and $G_{\text{out}} = 0.329$, and the mean paired out/in ratio is 0.442. Gains and paired ratios are averaged equally over base QPs 0, 21, 42, and 63. These samples show a smaller out-of-subspace response while retaining a larger in-subspace response. The pattern persists across the tested histories, ranks, and update positions. GLC-video's reference latent has mean gains above one in both directions under the matched settings (Fig. 6; Appendix C.7).

Responses under structural replacements. The separate protocol in Appendix C.8 perturbs a stored memory and compares the next frame under normal propagation. For A–D, we probe the reconstructed Observation O_t using all seven UVG videos, the same four base QPs, $K = 8$, $r = 2$, and injection after frame 24. The paired statistic $A_{\text{sel}} = \mathbb{E}_{\text{pairs}}[G_\perp/G_\parallel]$ increases in the order A, C, B, D, matching the ordering of both perceptual BD-Rate penalties in Table 2. Both mean Observation-space gains exceed one in every variant, so the statistic describes relative directional balance, not absolute contraction.

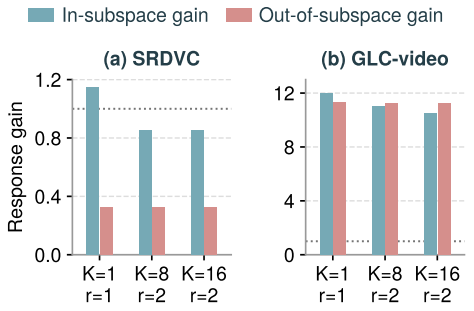


Figure 6: Sampled one-step temporal-subspace responses at frame $24 \rightarrow 25$. SRDVC's State has a smaller mean out-of-subspace response than in-subspace response, while both mean gains of GLC-video's reference latent exceed one. Panels use separate gain scales; unity denotes unchanged perturbation magnitude within each representation. Protocol and sensitivity results are given in Appendix C.7.

Table 3: 1080p runtime comparison in ms/frame. Locally measured entries use an NVIDIA A100 GPU with fp16; author-reported entries are marked † and are not re-run timing measurements. Bold identifies the fastest entry within each listed group.

Method	Enc.	Dec.
Pixel-wise RD optimized codecs		
DCVC-FM	183	190
DCVC-RT†	**7.99**	**8.87**
Perceptually optimized codecs		
GLC-video	290	183
GVC1D†	262	207
SRDVC	**15**	**14**

Interpretation. The correspondence between directional response and structural RD performance is descriptive. In particular, the single-impulse LPIPS BD-Rate changes in Appendix C.8 remain close to zero and show no consistent directional ordering at the tested perturbation scale. The probes therefore characterize learned dynamics but do not identify directional selectivity as the cause of the compression gains. The temporal subspaces are estimated proxies, not recovered tangent spaces, and sampled mean gains are not uniform operator bounds.

Appendices C.1–C.3 provide complementary finite-horizon observations: frame-group RD over 48 frames without intra refresh, sensitivity to input noise, and open-loop reconstruction behavior. These protocols characterize different aspects of reference use without isolating the same local perturbation mechanism.

5.5 Computational Efficiency

Table 3 reports the runtime comparison for 1080p coding. Our local timing protocol covers steady-state P-frame encoding and decoding, including entropy coding and file I/O, while excluding the I-frame and warm-up frame. DCVC-RT and GVC1D timings are author-reported. Appendix A.3 gives the measurement protocol and Fig. 7 shows the three recorded per-frame timing traces.

6 Conclusion

We introduced SRDVC, a perceptual video codec that separates predictive reference evolution from reference-conditioned coding. Its LDS factorizes dynamics estimation, State evolution, and Observation prediction, while DMA imposes an asymmetric source assignment for reference use. Matched-compute structural replacements demonstrate the value of these components within the tested codec. A conditional perturbation model motivates the design, and local temporal-subspace probes characterize the learned directional behavior and its association with the ordering of the structural RD results. Together with strong perceptual rate–distortion performance, efficient P-frame coding, and viewer preference over the tested subjective-study baselines, these results support the organization of recurrent reference prediction and use as a practical design dimension for perceptual video compression.

AI use statement

Generative AI tools were used to assist with drafting and copy-editing the manuscript text and with LaTeX formatting, working from an author-specified, evidence-bound outline. This assistance in-

cluded the exposition of theoretical arguments and diagnostic results, schematic editing, and plotting scripts. All technical content, including the method, experiments, numerical results, and their boundaries, was produced and verified by the authors from their own experimental assets. AI-assisted content was reviewed by the authors, who take full responsibility for the final content of this work. We did not use generative AI to generate experimental data or results.

Ethics statement

The subjective evaluation followed the applicable ethical requirements for research involving human participants. Participation was voluntary and based on informed consent, participants could withdraw, and participant privacy was protected. Appendix C.4 describes the study protocol and participant instructions. Appendix E.2 discusses the broader impacts of perceptual video compression.

Reproducibility statement

Section 4 specifies LDS, DMA, and the training objective. Appendix A.1 gives the distortion normalization, QP sampling, optimization, and data preprocessing; Appendix A.4 specifies the full MASP selection rule. Appendix A documents the evaluation datasets, codec configurations, and metric aggregation. Appendix A.2 reports the structural-ablation configurations and resource measurements, and Appendix A.3 details runtime measurement. Appendix C.7 gives the temporal-subspace construction and sensitivity results; Appendix C.8 specifies the structural-ablation diagnostic, perturbation normalization, direction seeds, quality settings, and paired-gain aggregation. The assumptions and derivations underlying the perturbation and rate analyses are provided in Appendices D.1–D.2. Code and model weights will be released after acceptance.

References

Gisle Bjøntegaard. Calculation of average PSNR differences between RD-curves. Technical Report VCEG-M33, VCEG, ITU-T SG16/Q6, Austin, Texas, USA, April 2001.

Frank Bossen. Common test conditions and software reference configurations. Technical Report JCTVC-L1100, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC 1/SC 29/WG 11, 2013.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006. doi: 10.1002/047174882X.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/852.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998. doi: 10.1109/18.720541.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2555–2565, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

JCT-VC and Fraunhofer HHI. HM-16.25. `https://vcgit.hhi.fraunhofer.de/jvet/HM/-/tags/HM-16.25`, 2022. Official HEVC reference software release tag.

Zhaoyang Jia, Bin Li, Jiahao Li, Wenxuan Xie, Linfeng Qi, Houqiang Li, and Yan Lu. Towards practical real-time neural video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12543–12552, June 2025.

JVET and Fraunhofer HHI. ECM-5.0. `https://vcgit.hhi.fraunhofer.de/ecm/ECM/-/tags/ECM-5.0`, 2022a. Official enhanced compression model release tag.

JVET and Fraunhofer HHI. VTM-17.0. `https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tags/VTM-17.0`, 2022b. Official VVC reference software release tag.

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 (CVPR)*, pp. 22616–22626, June 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 (CVPR)*, pp. 26099–26108, June 2024.

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 (CVPR)*, pp. 11006–11015, June 2019.

Alexandre Mercat, Marko Viitanen, and Jarno Vanne. UVG dataset: 50/120fps 4K sequences for video codec analysis and development. In *Proceedings of the ACM Multimedia Systems Conference (MMSys)*, 2020. doi: 10.1145/3339825.3394937.

Linfeng Qi, Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image and video compression. *arXiv preprint arXiv:2505.16177*, 2025.

Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012. doi: 10.1109/TCSVT.2012.2221191.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

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, 2016. doi: 10.1109/ICIP.2016.7532610.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

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. doi: 10.1007/s11263-018-01144-2.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, June 2018.

Zihan Zheng, Zhaoyang Jia, Naifu Xue, Jiahao Li, Bin Li, Zongyu Guo, Xiaoyi Zhang, Zhenghao Chen, Houqiang Li, and Yan Lu. Generative video compression with one-dimensional latent representation. *arXiv preprint arXiv:2603.15302*, 2026.

A Experimental Details

All main RD comparisons are conducted under low-delay settings in the RGB color space. Unless otherwise stated, we evaluate the first 96 frames of each sequence on UVG (Mercat et al., 2020), MCL-JCV (Wang et al., 2016), and HEVC-B (Bossen, 2013). LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), and PSNR are computed on the reconstructed RGB frames.

For traditional codecs, we use the official test models of VTM-17.0 (JVET and Fraunhofer HHI, 2022b), HM-16.25 (JCT-VC and Fraunhofer HHI, 2022), and ECM-5.0 (JVET and Fraunhofer HHI, 2022a) under low-delay settings. Since perceptual metrics are evaluated in RGB space, the input RGB frames are first converted to 10-bit YUV444 for internal codec processing. Specifically, HM uses `encoder_lowdelay_main_rext.cfg`, while VTM and ECM use `encoder_lowdelay_vtm.cfg` and `encoder_lowdelay_ecm.cfg`, respectively. We use the standard low-delay configuration with sequence-dependent frame rate and resolution, and convert the decoded outputs back to RGB before computing LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), and PSNR. For the learned baselines, we use the official inference implementations and recommended public configurations when available. In the current submission, all learned baselines except GVC1D are rerun by us under the same environment and evaluation protocol. For GVC1D, we report author-reported results that were not independently rerun in this study. BD-Rate (Bjøntegaard, 2001) is computed separately for each metric from its corresponding RD curves and then averaged over the sequences in each dataset. For runtime, our measurements use fp16 precision at 1080p resolution on a single NVIDIA A100 GPU; DCVC-RT and GVC1D runtimes are author-reported. The diagnostic analyses in the remaining appendix subsections use the same evaluation pipeline unless a different setup is explicitly stated.

A.1 Implementation Details

Objective and rate sampling. The training objective combines estimated bitrate with RGB reconstruction error and AlexNet-based LPIPS (Zhang et al., 2018). The RGB error is the sum of squared errors over all three channels divided by HW, where H and W are the frame dimensions. Frames are represented internally as normalized BT.709 YCbCr 4:4:4 and converted back to RGB for the distortion terms. This conversion is unclamped during loss computation; LPIPS receives RGB values transformed by $x \mapsto 2x-1$. The LPIPS network is frozen. The bitrate term is $R_t = R_{y,t} + R_{z,t}$, the estimated bits for the main latent and hyperlatent divided by HW.

A base QP q is sampled uniformly from $\{0, \dots, 63\}$ for each minibatch and shared across the clip. Its rate–distortion weight is interpolated logarithmically between $\lambda_{\min} = 0.07$ and $\lambda_{\max} = 1.5$:

$$\log \lambda(q) = \log \lambda_{\min} + \frac{q}{63} \left(\log \lambda_{\max} - \log \lambda_{\min} \right).$$

The temporal distortion weights repeat every eight predicted-frame positions:

$$(w_0, \dots, w_7) = (0.52, 1.33, 0.52, 0.83, 0.52, 0.83, 0.52, 0.83).$$

The weights are indexed by the predicted-frame position modulo eight. The first P-frame uses position one and weight 1.33. Cascaded training averages the loss over the predicted frames.

Optimization and preprocessing. Short-clip training uses AdamW with a global batch size of 40 across eight GPUs, or five clips per GPU. The optimizer uses its default $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 0.01; gradients are clipped to an ℓ_2 norm of 1.0. All P-frame model parameters are optimized. Training uses 256×384 random crops and random horizontal flips, with the same crop and flip applied to every frame of a clip. The spatial scale is fixed. Short-clip frame selection uses a maximum temporal stride of one. The Vimeo-90K training list contains 64,610 seven-frame sequences.

Short-clip training runs for 100 epochs. The curriculum begins with four epochs of two-frame distortion-only training. It then introduces the rate term and cascaded training with three, four, five, and six frames for one epoch each, followed by seven-frame clips. The learning rate is 10^{-4} for the first 18 epochs and 10^{-5} for the remaining 82 epochs. During cascaded training, a frozen pretrained I-frame codec reconstructs the first frame to initialize the reference buffer; the buffer is cleared between clips.

Table 4: Resource measurements for the UVG ablations in Table 2.

Model	Params (M)	Enc. GMAC	Dec. GMAC
A	32.65	653.7	635.6
B	31.52	658.6	640.6
C	32.94	654.6	636.1
D	31.81	659.5	641.1
E	32.65	653.7	635.6

Complexity is measured at 1920×1080, padded to 1920×1088, with batch size 1. Encoding/decoding GMAC are measured per frame along the actual `compress`/`decompress` paths for a P-frame following a P-frame ($P \to P$). Parameters count only the P-frame model, with shared weights counted once. E shares the architecture of A.

We subsequently fine-tune on the 32-frame dataset processed from raw Vimeo videos. MASP is used in this stage with default threshold $\tau = 0.4$; Section 4.4 summarizes the training setup, and Appendix A.4 gives the complete selection rule. For the learned baselines, we use the official code releases when available and follow their public evaluation pipelines as closely as possible.

A.2 Compute Resources and Reproducibility Notes

Training is conducted on 8 NVIDIA H200 GPUs. The runtime results in Table 3 are measured separately on a single NVIDIA A100 GPU at 1080p with fp16 inference. All runtime entries except DCVC-RT and GVC1D are measured by us using the official implementations under the same hardware and measurement protocol; DCVC-RT and GVC1D use author-reported runtimes. For SRDVC, the reported runtime is the actual end-to-end multi-frame average wall-clock time of the complete codec pipeline, including entropy coding, bitstream packing, and file I/O. To improve the auditability of the comparisons, we use official baseline implementations whenever they are publicly available. Unless otherwise stated, the reported SRDVC RD numbers in the main paper and appendix are obtained from a single deterministic evaluation run under the test protocol described above. The code and model weights for SRDVC will be released after acceptance.

Structural-ablation configurations and resource matching. In B and D, a standard ConvGRU replaces the LDS dynamics and State update, taking the previous decoded observation and its own hidden state. Both retain the same `ObsExtractor` architecture and the two output interfaces, `pred_O` for fusion and `ctx_t` for the entropy model. In C and D, concatenation followed by convolution replaces DMA while keeping both current features and the predicted reference as inputs. All remaining components, including MASP, are unchanged across A–D; E omits only MASP. Table 4 reports the associated resource measurements.

A.3 Inference-Time Measurement Protocol

Figure 7 shows three recorded per-frame timing traces for SRDVC at 1080p on a single NVIDIA A100 GPU. Each panel corresponds to a separate run under the stated measurement setup. The first frame is an I-frame and the second is a warm-up frame; both are excluded from the within-run means shown by dashed lines. The local protocol covers steady-state P-frame coding, including entropy coding, bitstream packing, and file I/O.

A.4 Codec-guided Fine-tuning with MASP

Manifold-Agnostic Sample Pruning (MASP) is a complementary training-side data-selection strategy that changes the fine-tuning distribution. It measures coding cost with a fixed H.265 encoder (Sullivan et al., 2012), independently of the learned model, and prunes the high-cost tail.

MASP is applied only during 32-frame fine-tuning. For each sequence $\mathcal{V}$, we encode it with H.265 at CRF=22 and compute its average P-frame bits-per-pixel,

$$\mathrm{BPP}_{H.265}(\mathcal{V}). \tag{6}$$

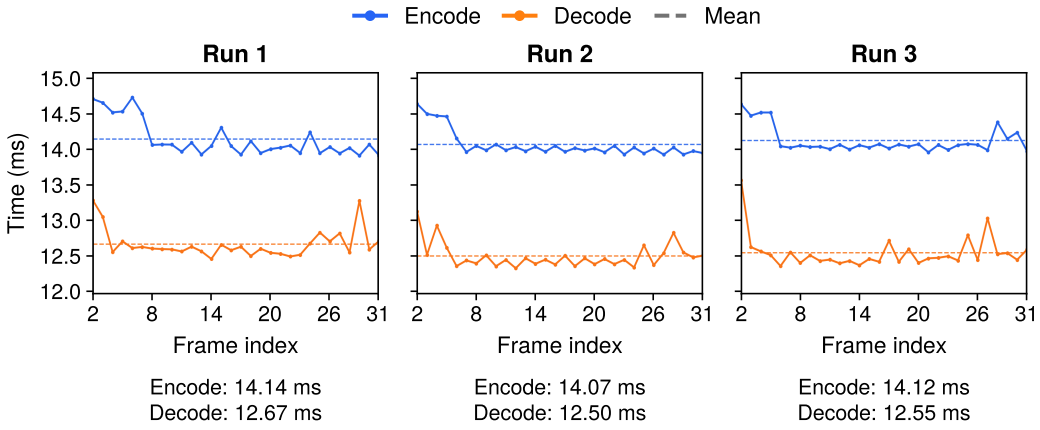


Figure 7: Per-frame encoding and decoding times in three recorded 1080p inference runs on a single NVIDIA A100 GPU. Dashed lines indicate the within-run means under the stated exclusion of the initial I-frame and warm-up frame.

We retain sequences whose score falls below τ:

$$\mathcal{D}_{\text{selected}} = \{\mathcal{V} \in \mathcal{D} \mid \text{BPP}_{H.265}(\mathcal{V}) < \tau\}, \tag{7}$$

We set $\tau = 0.4$. The score mixes temporal prediction difficulty with spatial texture and detail, occlusion, and noise under H.265 rate control, and a fixed CRF specifies an encoding setting rather than exactly matched distortion. The name records that selection uses this external coding statistic independently of the learned representation. Because selection depends only on the external codec, it can be computed once for a dataset. The A/E comparison in Table 2 reports its effect; Appendix C.5 gives the subset and threshold analyses.

B Additional Compression Results

B.1 PSNR Results

Although SRDVC is optimized for perceptual compression, Fig. 8 and Table 5 also report PSNR-based RD results. Relative to GLC-video, the reported PSNR BD-Rate reductions are 76.37% on UVG and 74.70% on HEVC-B. The MCL-JCV value is not reported because the compared PSNR RD curves do not overlap. We report per-dataset values without an aggregate average over unequal sets of defined comparisons.

Table 5: PSNR BD-Rate (%) relative to GLC-video. N/A denotes the MCL-JCV comparison whose PSNR RD curves do not overlap under the stated evaluation. No cross-dataset average is reported because the set of defined comparisons differs across methods.

Method	UVG	MCL-JCV	HEVC-B
GLC-video (Qi et al., 2025)	0.00	0.00	0.00
GVC1D (Zheng et al., 2026)	-62.89	-55.24	-54.27
SRDVC (Ours)	**-76.37**	N/A	**-74.70**

B.2 Visual Comparisons

Fig. 9 compares the visual results of DCVC-RT (Jia et al., 2025), GLC-video (Qi et al., 2025), and SRDVC. In all shown rows, SRDVC achieves the lowest BPP while also obtaining lower LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022) than the other two methods. Qualitatively, the first row shows that DCVC-RT loses the water details, while GLC-video distorts the chimney

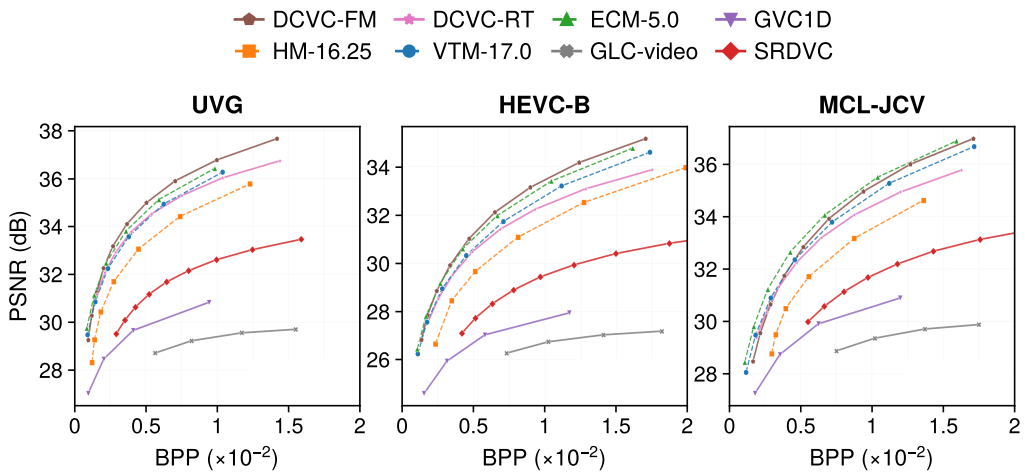


Figure 8: PSNR-based RD curves on UVG (Mercat et al., 2020), MCL-JCV (Wang et al., 2016), and HEVC-B (Bossen, 2013).

shape. In the second row, DCVC-RT over-smooths the hair details, while GLC-video distorts the text. In the third row, DCVC-RT completely blurs the horse hair, while GLC-video loses the facial details. In the fourth row, DCVC-RT severely blurs the dog's leg, while GLC-video changes the shape of the pebbles, causing semantic distortion. In the fifth row, DCVC-RT blurs the horse leg and grass details, while GLC-video causes severe red fading in the boxed region and distorts the corner of the curtain. In the sixth row, DCVC-RT loses the tree texture, while GLC-video loses semantic structure on the top of the boat. In the seventh row, DCVC-RT heavily blurs the floor texture, while GLC-video changes the shape of the sock pattern. Overall, these visual comparisons are consistent with SRDVC preserving fine structures and semantic details more reliably at lower bitrate. We will additionally provide an anonymized supplementary comparison video demo to reviewers to make the temporal differences easier to inspect.

C Analysis Experiments

C.1 Frame-Group RD under Uninterrupted Reference Propagation

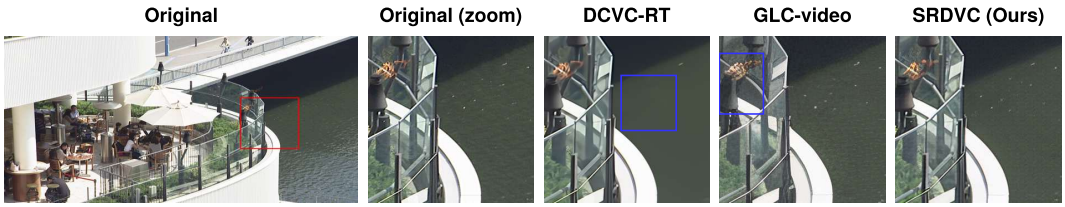
We examine frame-group RD under uninterrupted reference propagation on 48-frame UVG sequences. The first frame is encoded as an I-frame and all subsequent frames as P-frames without intra refresh. We divide the sequence into G1 (frames 1–16), G2 (17–32), and G3 (33–48), use G1 as the reference group, and exclude the I-frame from the BD-Rate calculation.

Figure 10 shows different group-wise trends. GLC-video's BD-Rate relative to G1 changes from approximately -5% in G2 to approximately $+1\%$ in G3, whereas SRDVC reaches approximately -29% and -37%, respectively. These are favorable finite-horizon observations for SRDVC under this protocol. Because the groups contain different temporal segments, their content difficulty and interaction with the codec also affect the comparison; the experiment does not isolate accumulated reference perturbations from those factors.

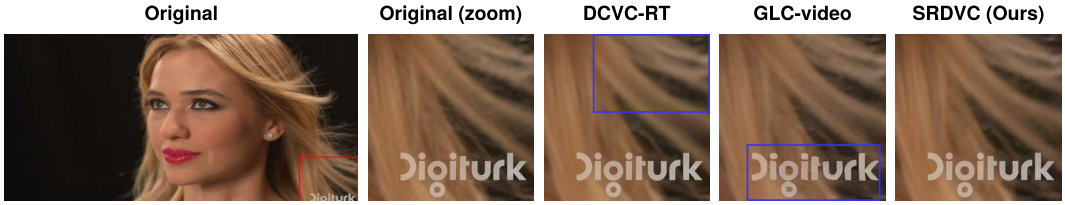
C.2 Sensitivity to Input Noise

We compare clean and noisy input runs on UVG. In the noisy run, Gaussian noise with standard deviation 0.005 is added to input frames 17–32. LPIPS BD-Rate is computed relative to each model's clean run, which is normalized to zero.

Table 6 reports a penalty of 32.35% for SRDVC and 70.62% for GLC-video. SRDVC is therefore less sensitive under this input-noise protocol. Since the perturbation changes the current-frame coding problem as well as information entering the reference path, the comparison does not isolate attenuation by the recurrent update alone.



Method	BPP ↓	Relative BPP	LPIPS ↓	DISTS ↓
DCVC-RT	0.0048	1.26×	0.2934	0.2042
GLC-video	0.0050	1.32×	0.1513	0.1797
SRDVC (Ours)	**0.0038**	**1.00×**	**0.0738**	**0.1224**



Method	BPP ↓	Relative BPP	LPIPS ↓	DISTS ↓
DCVC-RT	0.0082	1.09×	0.3730	0.2311
GLC-video	0.0094	1.25×	0.0966	0.1279
SRDVC (Ours)	**0.0075**	**1.00×**	**0.0615**	**0.1069**



Method	BPP ↓	Relative BPP	LPIPS ↓	DISTS ↓
DCVC-RT	0.0159	1.18×	0.2666	0.2117
GLC-video	0.0192	1.42×	0.1014	0.1164
SRDVC (Ours)	**0.0135**	**1.00×**	**0.0583**	**0.0791**

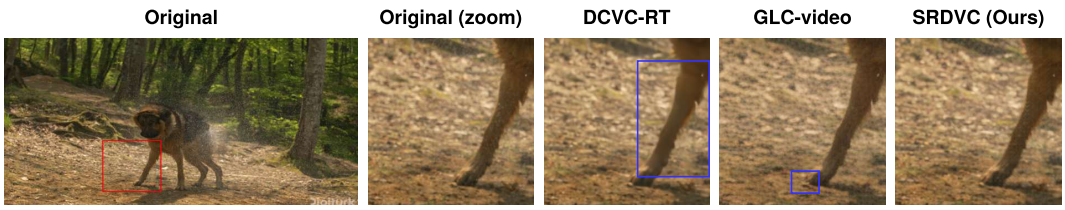
Figure 9: Visual comparisons among DCVC-RT (Jia et al., 2025), GLC-video (Qi et al., 2025), and SRDVC. Each row reports BPP, LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2022); lower LPIPS and DISTS indicate better perceptual quality.

Table 6: Sensitivity to input noise on UVG. LPIPS BD-Rate changes are relative to each model's own clean run, normalized to 0%. Noise is added to input frames 17–32 under the protocol described above.

Method	Clean	Noisy
GLC-video	0.0%	+70.62%
SRDVC (Ours)	0.0%	**+32.35%**

C.3 Open-Loop Reference Prediction

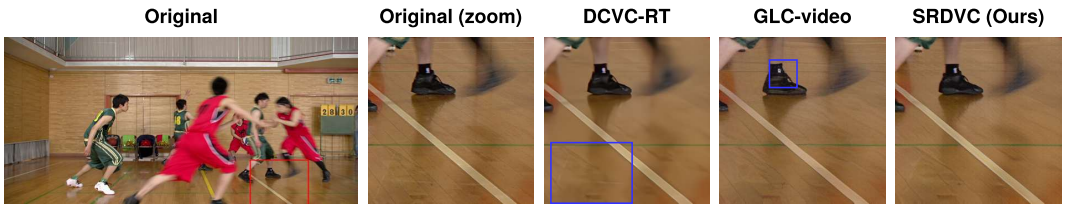
We examine open-loop reconstruction on UVG. The first 16 frames are coded normally. After frame 16, residual correction is disabled and the latent variables are replaced by the entropy-model means. We report LPIPS over frames 16–32, with frame 16 serving as the last normally coded anchor.



Method	BPP ↓	Relative BPP	LPIPS ↓	DISTS ↓
DCVC-RT	0.0222	1.30×	0.3300	0.2448
GLC-video	0.0183	1.07×	0.1504	0.1287
SRDVC (Ours)	**0.0171**	**1.00×**	**0.0810**	**0.0826**



Method	BPP ↓	Relative BPP	LPIPS ↓	DISTS ↓
DCVC-RT	0.0155	1.19×	0.1149	0.1320
GLC-video	0.0143	1.10×	0.0992	0.1328
SRDVC (Ours)	**0.0130**	**1.00×**	**0.0352**	**0.0592**



Method	BPP ↓	Relative BPP	LPIPS ↓	DISTS ↓
DCVC-RT	0.0309	1.39×	0.3464	0.2448
GLC-video	0.0235	1.05×	0.0994	0.1518
SRDVC (Ours)	**0.0223**	**1.00×**	**0.0810**	**0.1324**

Figure 9: Visual comparisons (continued). The remaining three examples use the same methods and metrics.

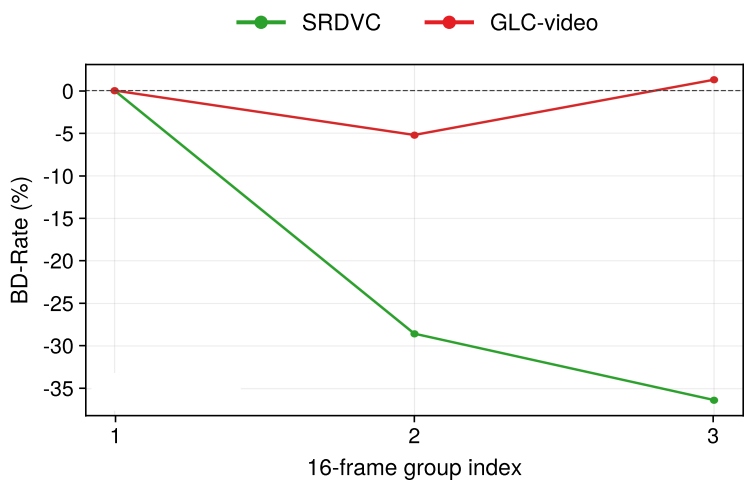


Figure 10: LPIPS BD-Rate of 16-frame groups relative to the first group on UVG, with the I-frame excluded and no subsequent intra refresh. SRDVC has lower relative BD-Rate in the later groups, while GLC-video's result rises from G2 to G3. The comparison is between different temporal segments of each sequence.

Both models degrade during the open-loop continuation. SRDVC has lower average LPIPS over the reported interval (0.17 versus 0.24 for GLC-video) and a smaller increase from frame 16 to frame 32 (0.17 versus 0.21). These results characterize the complete prediction and reconstruction paths under the tested continuation. The models enter the open-loop phase at different reconstruction qualities, so the comparison does not isolate the transition rule from its initial condition and the remaining codec components.

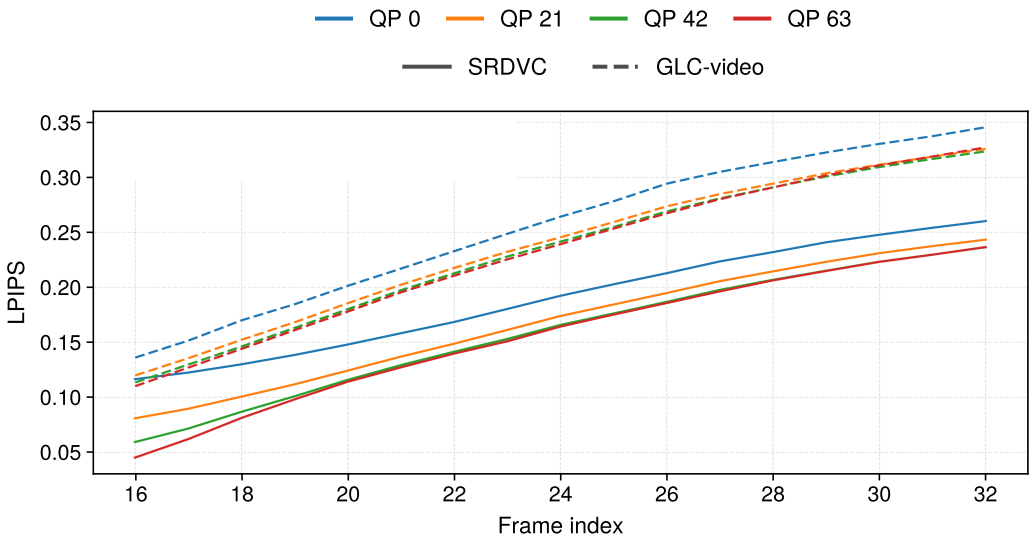


Figure 11: Open-loop reconstruction LPIPS on UVG. Frame 16 is the last normally coded anchor; subsequent frames are reconstructed with residual correction disabled and latents replaced by entropy-model means. Solid and dashed curves show SRDVC and GLC-video under the listed QPs.

C.4 2AFC Subjective Preference Study

To complement the objective metrics and diagnostic analyses, we conduct a two-alternative forced choice (2AFC) subjective study. We compare SRDVC against DCVC-RT (Jia et al., 2025) and GLC-video (Qi et al., 2025) on 18 representative clips drawn from UVG (Mercat et al., 2020) and HEVC-B (Bossen, 2013). For each clip, the compared methods are matched using the nearest available operating points with average BPP mismatch below 3%. Each participant therefore evaluates 36 pairwise trials in total (18 clips $\times$ 2 baselines).

We recruit 60 volunteer participants who are not authors of the paper. In each trial, two compressed videos of the same source clip are displayed in randomized left-right order at the same spatial resolution, and the participant is asked to choose the version they would prefer to watch overall. The instruction shown to the participants is: *"You will watch two compressed versions of the same video clip. Please select the one you prefer overall in terms of perceptual quality. Base your decision on overall viewing quality, including naturalness, visible artifacts, and temporal stability. If the difference is small, please still choose the one you prefer slightly more."* Participants may replay each pair once before making a decision. Since this is a 2AFC setup, no tie option is provided.

Participant protections. The study followed the applicable ethical requirements for human-participant research. Participants provided informed consent, took part voluntarily, and could discontinue participation. Participant information and responses were handled to protect privacy.

Table 7: 2AFC subjective preference results. Each comparison contains 1080 binary judgments (18 clips $\times$ 60 participants). Higher preference rate is better.

Comparison	Preferred	Not preferred	Preference rate
SRDVC vs. DCVC-RT	783	297	72.5%
SRDVC vs. GLC-video	744	336	68.9%

Table 7 shows a majority preference for SRDVC against both baselines, with a higher preference rate against DCVC-RT. The LPIPS/DISTS advantage is thus accompanied by viewer preference at matched bitrate.

Each participant contributes choices on multiple clips, and each clip is viewed by multiple participants. The 1080 judgments per comparison therefore have a repeated-measures structure. We report aggregate preference rates descriptively, without an independent-trial significance claim.

We further examine whether the preference gap is concentrated on clips with stronger temporal challenges. Splitting the 18 clips into a higher-motion subset (8 clips) and a lower-motion subset (10 clips), SRDVC is preferred over DCVC-RT in 368 of 480 judgments (76.7%) on the higher-motion subset and 415 of 600 judgments (69.2%) on the lower-motion subset. Against GLC-video, the corresponding preference rates are 72.7% (349/480) and 65.8% (395/600), respectively. The gap is larger where temporal prediction and propagation are more demanding, although this split is descriptive and the clip subsets differ in content in other respects as well.

C.5 Training-Subset BPP and the Effect of MASP

We examine how the H.265 coding-cost distribution of the fine-tuning data relates to perceptual RD performance. Because MASP is applied only in the 32-frame fine-tuning stage, we study that dataset. We sort the training sequences by average H.265 P-frame BPP and construct four equal-sized subsets (2000 sequences each) with mean BPP 0.05, 0.16, 0.31, and 0.92, respectively. Starting from the same pretrained weights, we fine-tune one model on each subset and evaluate LPIPS (Zhang et al., 2018) BD-Rate (Bjøntegaard, 2001) on UVG (Mercat et al., 2020), using the 0.05-BPP subset as the reference so that its result is normalized to 0%. We also fine-tune a model on a random 2000-sample subset, whose mean BPP is 0.35. Fig. 12 shows a non-monotonic trend: the lowest-BPP subset is not best, and the strongest performance is obtained by the moderate-BPP subsets at 0.16 and 0.31, which achieve -34.76% and -39.98%, respectively. The high-BPP subset at 0.92 is also worse, yielding -30.95%. Thus, the moderate-BPP subsets provide better fine-tuning data than either extreme in this experiment.

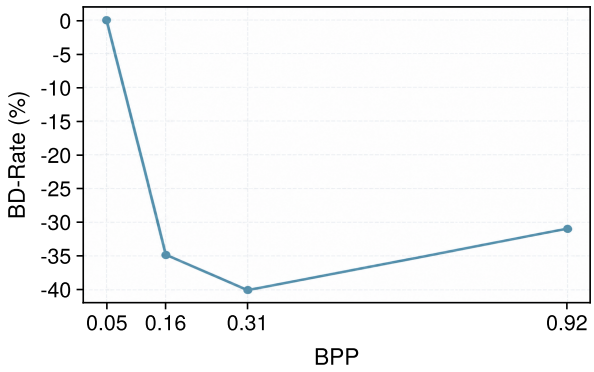


Figure 12: UVG (Mercat et al., 2020) LPIPS (Zhang et al., 2018) BD-Rate for models fine-tuned on equal-sized subsets grouped by mean H.265 P-frame BPP. Lower is better.

The random subset at mean BPP 0.35 achieves -24.68%, underperforming the nearby 0.31-BPP subset despite their similar means. The selected and random subsets differ in content composition as well as mean BPP. Their comparison supports evaluating data selection through downstream RD performance, while leaving the contributions of texture, motion, and other content factors unresolved.

These subset results motivate filtering the high-coding-cost tail while retaining a range of BPP values. The direct test of MASP is the A/E comparison in Table 2: removing selection increases LPIPS/DISTS BD-Rate by 5.9%/6.3% under the same protocol on UVG.

Table 8 reports the sensitivity of the MASP threshold on UVG (Mercat et al., 2020). The BD-Rate values are reported relative to the $\tau = 0.4$ setting, which is normalized to 0%.

Table 8: Sensitivity of the MASP threshold τ on UVG (Mercat et al., 2020). The BD-Rate values are reported relative to the $\tau = 0.4$ setting, which is normalized to 0%. Lower is better.

τ	BD-Rate (LPIPS)	BD-Rate (DISTS)
0.3	$+3.2\%$	$+2.8\%$
0.4	$\mathbf{0.0}\%$	$\mathbf{0.0}\%$
0.5	$+1.7\%$	$+1.4\%$

Table 8 shows that $\tau = 0.4$ performs best among the three tested thresholds. Reducing τ to 0.3 excludes more sequences and increases LPIPS/DISTS BD-Rate by 3.2%/2.8%. Increasing τ to 0.5 admits additional higher-BPP sequences and increases BD-Rate by 1.7%/1.4%. Within the three tested thresholds and the reported UVG evaluation, $\tau = 0.4$ has the lowest measured BD-Rate. This sensitivity study does not establish a dataset-independent optimum.

C.6 Mapping the Abstract Reference Variable to Implemented States

The variable M_t in Sec. 3.1–3.3 is intentionally model-agnostic: it denotes the propagated reference representation that supports current-frame coding, without committing to a single parameterization shared by all codecs. In SRDVC, this propagated reference is implemented by a recurrent reference path rather than by a single tensor. Specifically, S_t is the compact recurrent state that governs reference evolution, $\hat{O}_t = f_O(S_t, O_{t-1})$ is the predicted reference used by DMA for coding frame t, and O_t is the reconstructed observation stored in the recurrent buffer for the next step. In contrast, for GLC-video (Qi et al., 2025), the corresponding propagated reference is the reference latent together with the reference features derived from it by the reference module. Thus the implemented recurrent reference path contains both State and Observation buffers, while $\hat{O}_t$ is the prediction derived for current coding. The State-only local probes analyze a component of this path with specified inputs fixed. They should not be read as measurements of the full joint-buffer Jacobian or as verification of closed-loop boundedness.

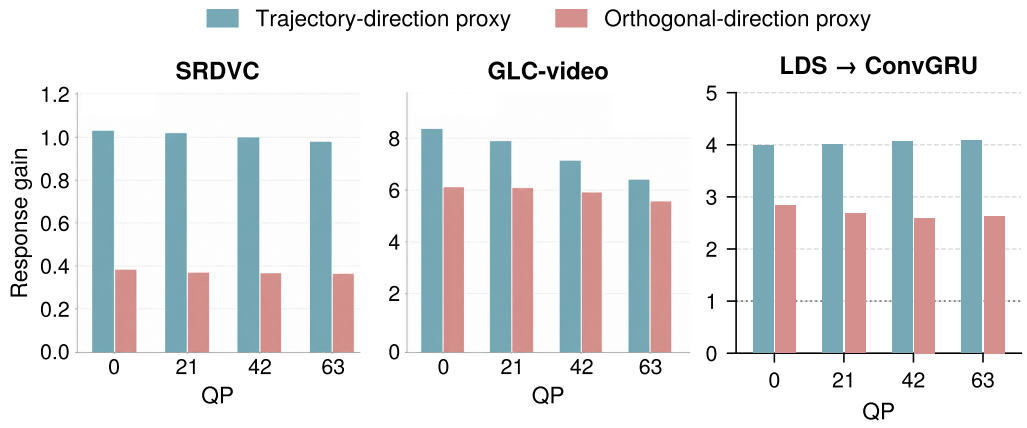


Figure 13: Cross-model one-step directional gains. Axes use model-specific scales, and unity denotes unchanged magnitude. Gains are normalized within each model's own representation and characterize sampled response patterns; absolute gains are not treated as a common coordinate-invariant stability metric across models.

C.7 Local Temporal-Subspace Response Analysis

We probe the learned reference update using temporal directions estimated from observed State evolution. The consecutive-State difference is the $K = 1, r = 1$ special case of the local subspace construction below. Figure 13 reports the complementary cross-model comparison; the main-text Fig. 6 compares the temporal-subspace responses of SRDVC State and GLC-video reference latent. Table 9 gives the full SRDVC sensitivity results and the matched GLC-video measurements.

Cross-model one-dimensional protocol. Let B_t denote the variable being perturbed: the compact State S_t for SRDVC, the replacement hidden state for the LDS-to-ConvGRU ablation (B), and the reference latent for GLC-video (Qi et al., 2025). The trajectory-direction proxy and sampled orthogonal-direction proxy are

$$u_t = B_t - B_{t-1}, \qquad \Delta_t^{\parallel} = u_t, \qquad \Delta_t^{\perp} = z_t - \frac{\langle z_t, u_t \rangle}{\|u_t\|_2^2} u_t, \tag{8}$$

where z_t is a random vector. Holding other update inputs fixed, we measure the one-step response of the update Φ as

$$G(\Delta_t) = \frac{\|\Phi(B_t + \epsilon\Delta_t) - \Phi(B_t)\|_2}{\epsilon\|\Delta_t\|_2}. \tag{9}$$

The cross-model probe uses $\epsilon = 0.005$. Gains are dimensionless and normalized within each model's latent space. At QPs $0, 21, 42, 63$, ConvGRU gives trajectory-direction gains of 3.984, 4.001, 4.067, and 4.082, and orthogonal gains of 2.837, 2.680, 2.595, and 2.632. For example, a response of 398.38% corresponds to $G = 3.984$. ConvGRU amplifies both directions, with a smaller orthogonal response. SRDVC retains trajectory-direction gains around 1.0 and orthogonal gains around 0.36–0.38; GLC-video amplifies both directions. The cross-model comparison concerns these directional response patterns within each parameterization.

Multi-dimensional temporal-subspace protocol. For SRDVC, flatten each State difference into a column vector and form

$$d_{t-i} = S_{t-i} - S_{t-i-1}, \quad i = 0, \ldots, K-1, \qquad D_t = [d_t, d_{t-1}, \ldots, d_{t-K+1}]. \tag{10}$$

The leading left singular vectors of $D_t = U\Sigma V^\top$ define

$$\mathcal{T}_t = \mathrm{span}(U_{1:r}), \qquad P_t = U_{1:r}U_{1:r}^\top. \tag{11}$$

We sample paired perturbation directions $v_{\mathrm{in}} \in \mathcal{T}_t$ and $v_{\mathrm{out}} \in \mathcal{T}_t^{\perp}$. The corresponding directions can be represented by $P_t z_{\mathrm{in}}$ and $(I - P_t)z_{\mathrm{out}}$ for nonzero projected samples. We measure $G(v)$

Table 9: Local temporal-subspace sensitivity for SRDVC and matched GLC-video responses. All entries use equal weights for QPs 0, 21, 42, and 63. A averages paired out/in gain ratios. Bold marks the default main-text setting in each model.

Memory update	K	r	G_{in}	G_{out}	Mean A
SRDVC (State)					
$S_{16} \to S_{17}$	8	2	0.813	0.329	0.460
$S_{24} \to S_{25}$	1	1	1.147	0.329	0.288
$S_{24} \to S_{25}$	4	2	0.842	0.329	0.445
$S_{24} \to S_{25}$	8	1	1.265	0.329	0.259
$\mathbf{S_{24} \to S_{25}}$	**8**	**2**	**0.852**	**0.329**	**0.442**
$S_{24} \to S_{25}$	8	4	0.653	0.329	0.556
$S_{24} \to S_{25}$	16	2	0.852	0.329	0.448
GLC-video (reference latent)					
$B_{24} \to B_{25}$	1	1	11.963	11.328	0.975
$\mathbf{B_{24} \to B_{25}}$	**8**	**2**	**11.043**	**11.249**	**1.081**
$B_{24} \to B_{25}$	16	2	10.478	11.237	1.102

with Eq. (9), using $B_t = S_t$, to obtain paired in-subspace and out-of-subspace gains. The selectivity statistic averages the ratio within each pair before aggregation:

$$A = \frac{1}{4} \sum_{q \in \{0,21,42,63\}} \mathbb{E}_{\text{pairs}|q} \left[\frac{G(v_{\text{out}})}{G(v_{\text{in}})} \right]. \tag{12}$$

The reported G_{in} and G_{out} are also averaged equally over these four QPs. Thus A is the mean pairwise ratio, not the ratio of the two reported mean gains. The $K = 1, r = 1$ setting uses the same one-dimensional temporal-proxy definition as the cross-model test, evaluated here at the specified recurrent position.

For GLC-video, we apply the same construction to its reference latent B_t in place of S_t, using the same videos, QPs, perturbation normalization, and aggregation at frame $24 \to 25$.

Sensitivity and recurrent position. For SRDVC, Table 9 shows stronger transmission within the estimated temporal subspace throughout the tested settings. At $S_{24} \to S_{25}$, the out-of-subspace gain remains at 0.3286–0.3287 as K and r change, while A stays between 0.259 and 0.556. Increasing the default rank from 2 to 4 reduces the mean in-subspace response from 0.852 to 0.653, yet the out-of-subspace response remains 0.329. Repeating $K = 8, r = 2$ at $S_{16} \to S_{17}$ gives gains of 0.813/0.329 and $A = 0.460$, close to the default update's 0.852/0.329 and $A = 0.442$. The anisotropic signature holds for both one-dimensional and multi-dimensional temporal proxies and at both tested update positions. The subspace is estimated from observed memory evolution and is not an identified tangent space of the data manifold. These are sampled local responses, not worst-case operator norms. Gains are normalized within each representation; cross-model magnitudes are not assumed to be invariant to the choice of latent coordinates.

C.8 Directional Responses of Structural Ablations

We extend the directional analysis to the recurrent State of the full model, denoted A(S), and to the Observation representations of the four structural variants in Table 2, denoted A(O), B(O), C(O), and D(O). State is a recurrent component in SRDVC and the direct object of the local design target. Observation-space probes provide a corresponding comparison across architectural variants.

Sequences and quality settings. The experiment uses the first 40 frames of each of the seven UVG videos: Beauty, Bosphorus, HoneyBee, Jockey, ReadySteadyGo, ShakeNDry, and YachtRide, at their original 1920×1080 resolution. Base QPs are $0, 21, 42, 63$, with each codec's original frame-level QP schedule retained. The protocol tests independent injections after frames 16 and 24, followed by normal propagation over frames 17–32 and 25–40, respectively. The Observation-space ratios in Table 2 and the gains in Table 10 use the frame-24 injection.

Table 10: Directional gains in State and Observation space. The full model's sampled State response attenuates orthogonal perturbations. Observation-space variants are compared through their paired directional ratios.

Model / space	$G_{\parallel}$	$G_{\perp}$	$A_{\mathrm{sel}} \downarrow$
A(S)	0.886	0.361	0.467
A(O)	6.661	1.875	**0.324**
C(O)	8.209	3.839	0.502
B(O)	13.127	8.688	0.703
D(O)	14.351	10.816	0.811

Directional perturbations. Let B_t denote the memory being perturbed: S_t for A(S), or O_t for the Observation-space variants. Using raw memory coordinates, we estimate the temporal subspace from the leading $r = 2$ left singular vectors of the most recent $K = 8$ adjacent memory differences, following Appendix C.7. Each draw samples a unit direction $v_{\parallel}$ in this subspace and a unit direction $v_{\perp}$ in its orthogonal complement. The injection length and perturbed memories are

$$\eta_t = 0.005 \, \|B_t - B_{t-1}\|_2, \qquad \widetilde{B}_t^{s,v} = B_t + s\eta_t v, \quad s \in \{-1, +1\}, \quad \|v\|_2 = 1. \tag{13}$$

The norm covers the entire memory tensor. Parallel and orthogonal perturbations have the same length for a given condition; each variant is normalized by its own preceding memory difference. Thus 0.5% is relative to the temporal change, rather than the memory norm or an elementwise offset.

Four direction draws, indexed $d = 0, 1, 2, 3$, are used for each video, QP, and injection position. Both signs are run for each parallel and orthogonal direction, giving 16 perturbation branches per condition. These are perturbation-direction seeds, not repeated training runs. For zero-based video index i in the sequence order above, base QP q, injection frame t, and draw d, the seed is

$$\mathrm{seed} = 20260913 + 10000i + 100q + 100000t + 10000000d.$$

For Beauty at QP 0 after frame 24, the four seeds are 22660913, 32660913, 42660913, and 52660913.

One-step gains and paired aggregation. Each perturbed run is compared with its clean counterpart at the next frame. If $B_{t+1}^{s,v}$ and B_{t+1}^{clean} are the corresponding next-frame memories, the gain is

$$G_s(v) = \frac{\|B_{t+1}^{s,v} - B_{t+1}^{\mathrm{clean}}\|_2}{\eta_t}. \tag{14}$$

These are one-step gains; the following 16 frames constitute the propagation window, not an averaging window for G. For each video, QP, direction draw, and sign, we pair the parallel and orthogonal gains. The directional selectivity statistic is

$$A_{\mathrm{sel}} = \mathbb{E}_{\mathrm{pairs}} \left[\frac{G_{\perp}}{G_{\parallel}} \right]. \tag{15}$$

Each ratio is computed within a paired directional probe before averaging. It is therefore distinct from the ratio of the two reported mean gains. Lower A_{sel} indicates a smaller orthogonal response relative to the trajectory-aligned response.

A(S) has a larger sampled in-subspace response than orthogonal response under this protocol, with mean gains of 0.886 and 0.361. Both Observation-space gains exceed one, with A_{sel} quantifying their directional balance within each representation. Absolute gains depend on the representation and its parameterization, so the State and Observation results are interpreted separately. A(S) differs slightly from Table 9 (0.852/0.329) because this protocol scales perturbations by the preceding memory difference, uses both signs, and compares next frames under normal propagation.

The associated single-impulse LPIPS BD-Rate changes relative to each model's clean run remain close to zero and show no consistent directional ordering. Thus the one-step gain differences are not accompanied by a consistent directional coding-cost effect at the tested perturbation scale. We use this experiment to characterize the learned response, not to establish that directional selectivity causes the RD differences between architectures.

D Additional Theoretical Derivations

D.1 Local Perturbation Recursions and Conditional Bounds

We analyze perturbations of the recurrent State along a reference trajectory. Let ϵ_t denote the State perturbation, J_t the local State Jacobian with other update inputs fixed, and δ_t the contribution from changes in those inputs and newly introduced errors. The local model is

$$\epsilon_t \approx J_t \epsilon_{t-1} + \delta_t. \tag{16}$$

For zero-mean perturbations, define $\sigma^2_{\epsilon,t} = \mathbb{E}\|\epsilon_t\|_2^2$, the total variance. If J_t approximately preserves perturbation energy, the carried-over and incoming perturbations are uncorrelated, and $\mathbb{E}\|\delta_t\|_2^2 = \sigma^2_\delta$ is constant, then

$$\sigma^2_{\epsilon,t} \approx \sigma^2_{\epsilon,t-1} + \sigma^2_\delta \approx \sigma^2_{\epsilon,0} + t\sigma^2_\delta. \tag{17}$$

This derives the preservation-oriented accumulation model in Eq. (1).

Orthogonal propagation with bounded total input. Let $P_t^\perp$ be the orthogonal projector onto the idealized local orthogonal subspace at step t, with $\epsilon_t^\perp = P_t^\perp \epsilon_t$ and $P_t^\parallel = I - P_t^\perp$. Projecting the local update separates orthogonal carry-over from the total input to that component:

$$\begin{aligned} \epsilon_t^\perp &= B_t \epsilon_{t-1}^\perp + \delta_t^\perp, \qquad B_t = P_t^\perp J_t P_{t-1}^\perp, \\ \delta_t^\perp &= P_t^\perp J_t P_{t-1}^\parallel \epsilon_{t-1} + P_t^\perp \delta_t + \rho_t^\perp. \end{aligned} \tag{18}$$

Here $\rho_t^\perp$ collects the projected linearization remainder. Thus $\delta_t^\perp$ includes cross-direction coupling, changes in other update inputs, new errors, and the remainder. Its boundedness is a condition on their combined effect, including feedback through those inputs.

Assume along the trajectory that $\|B_t\|_2 \le \alpha < 1$, that $(\mathbb{E}\|\delta_t^\perp\|_2^2)^{1/2} \le d$, and that the initial perturbation has a finite second moment. Define $m_t = (\mathbb{E}\|\epsilon_t^\perp\|_2^2)^{1/2}$. The triangle inequality in L^2 gives

$$m_t \le \alpha m_{t-1} + d. \tag{19}$$

Iterating establishes the bound

$$m_t \le \alpha^t m_0 + d\frac{1-\alpha^t}{1-\alpha}, \qquad \limsup_{t\to\infty} m_t \le \frac{d}{1-\alpha}. \tag{20}$$

This bound holds for correlated inputs as well as independent ones. Uniform contraction controls orthogonal carry-over, and bounded total input controls the perturbations transferred into that component. The directional gains in Fig. 6 characterize sampled local responses; the bound assumes a uniform gain along the trajectory.

Constant-gain variance model. For the special case $\epsilon_t^\perp = \alpha\epsilon_{t-1}^\perp + \delta_t^\perp$ in orthonormal local coordinates, assume zero means, uncorrelated carried-over and incoming terms, and constant input variance $\sigma^2_{\delta^\perp}$. Then

$$\begin{aligned} \sigma^2_{\epsilon^\perp,t} &= \alpha^2 \sigma^2_{\epsilon^\perp,t-1} + \sigma^2_{\delta^\perp} \\ &= \alpha^{2t}\sigma^2_{\epsilon^\perp,0} + \frac{1-\alpha^{2t}}{1-\alpha^2}\sigma^2_{\delta^\perp}. \end{aligned} \tag{21}$$

Taking the limit yields

$$\lim_{t\to\infty} \sigma^2_{\epsilon^\perp,t} = \frac{\sigma^2_{\delta^\perp}}{1-\alpha^2}. \tag{22}$$

This expression quantifies the balance between attenuation and continuing perturbation input.

D.2 Illustrative Rate-Penalty Analysis under an Additive-Noise Model

This section provides an illustrative scalar residual model rather than a quantitative prediction of the codec's measured perceptual RD performance. Gaussianity, independence of the clean residual and perturbation, a sufficiently fine fixed quantizer, and the stated residual-sensitivity assumptions are modeling conditions, not properties verified by the directional probes.

We connect State perturbation energy to residual coding cost through local linearization, an additive Gaussian model, and high-rate quantization (Cover & Thomas, 2006; Gray & Neuhoff, 1998). Rates below are per scalar residual coordinate, using natural logarithms and units of nats; division by $\ln 2$ converts them to bits. Let $n_t \sim \mathcal{N}(0, \sigma_N^2)$ denote the clean prediction residual with fixed $\sigma_N^2 > 0$, and let $\zeta_t \sim \mathcal{N}(0, \sigma_{\zeta,t}^2)$ denote the residual-domain perturbation induced by the accumulated orthogonal component of the reference perturbation. Assuming that n_t and ζ_t are independent, the noisy residual becomes

$$r_t = n_t + \zeta_t \sim \mathcal{N}(0, \sigma_N^2 + \sigma_{\zeta,t}^2). \tag{23}$$

Under a high-rate approximation (Gray & Neuhoff, 1998) with the same sufficiently small quantization step Δ, the clean and noisy residual rates can be compared through Gaussian differential entropies (Cover & Thomas, 2006), with the common quantization term shown explicitly:

$$R_t^{\text{clean}} \approx \frac{1}{2}\ln(2\pi e\sigma_N^2) - \ln\Delta, \qquad R_t^{\text{noisy}} \approx \frac{1}{2}\ln\big(2\pi e(\sigma_N^2 + \sigma_{\zeta,t}^2)\big) - \ln\Delta. \tag{24}$$

Therefore, the additional rate caused by this residual perturbation is approximately

$$\Delta R_t = R_t^{\text{noisy}} - R_t^{\text{clean}} \approx \frac{1}{2}\ln\left(1 + \frac{\sigma_{\zeta,t}^2}{\sigma_N^2}\right). \tag{25}$$

For small noise, where $x_t = \sigma_{\zeta,t}^2/\sigma_N^2 \ll 1$, we have

$$\Delta R_t \approx \frac{1}{2}\ln(1 + x_t) \approx \frac{x_t}{2} = \frac{\sigma_{\zeta,t}^2}{2\sigma_N^2}. \tag{26}$$

Thus, in the small-noise regime, the additional rate is approximately proportional to the residual-domain perturbation variance.

To connect this quantity with the recurrent reference perturbation model in Sec. 3.2–3.3, let $\epsilon_t^\perp$ denote the orthogonal component of the reference-path perturbation and adopt a local linearization $\zeta_t \approx A_t\epsilon_t^\perp$. If the average sensitivity varies slowly so that the perturbation energy satisfies $\mathbb{E}\|\zeta_t\|_2^2 \approx c\,\mathbb{E}\|\epsilon_t^\perp\|_2^2$ for some approximately constant $c > 0$, then, at the variance level,

$$\sigma_{\zeta,t}^2 \approx c\,\sigma_{\epsilon^\perp,t}^2. \tag{27}$$

That is, the residual-domain perturbation inherits the same growth trend as the orthogonal reference perturbation up to a scaling factor.

Assuming zero-mean innovations and negligible cross-correlation between carried-over noise and newly injected perturbations, the preservation-oriented model applied to this component follows Eq. (1), which yields

$$\sigma_{\zeta,t}^2 \approx c\big(\sigma_{\epsilon^\perp,0}^2 + t\sigma_{\delta^\perp}^2\big). \tag{28}$$

Substituting this into the high-rate expression gives

$$\Delta R_t \approx \frac{1}{2}\ln\left(1 + \frac{c\big(\sigma_{\epsilon^\perp,0}^2 + t\sigma_{\delta^\perp}^2\big)}{\sigma_N^2}\right). \tag{29}$$

Therefore, the rate penalty monotonically increases with recurrent steps. In the small-noise regime, it further admits the approximation

$$\Delta R_t \approx \frac{c\big(\sigma_{\epsilon^\perp,0}^2 + t\sigma_{\delta^\perp}^2\big)}{2\sigma_N^2}, \tag{30}$$

which explains the approximately linear growth at the early stage.

Under the same approximation, a contracting orthogonal update follows Eq. (21), which gives

$$\sigma_{\zeta,t}^2 \approx c\left(\alpha^{2t}\sigma_{\epsilon^\perp,0}^2 + \frac{1-\alpha^{2t}}{1-\alpha^2}\sigma_{\delta^\perp}^2\right). \tag{31}$$

The corresponding rate penalty is

$$\Delta R_t \approx \frac{1}{2} \ln \left(1 + \frac{c \left(\alpha^{2t} \sigma^2_{\epsilon^\perp,0} + \frac{1-\alpha^{2t}}{1-\alpha^2} \sigma^2_{\delta^\perp} \right)}{\sigma^2_N} \right). \tag{32}$$

Taking the limit gives

$$\lim_{t\to\infty} \Delta R_t \approx \frac{1}{2} \ln \left(1 + \frac{c\, \sigma^2_{\delta^\perp}}{(1-\alpha^2)\sigma^2_N} \right). \tag{33}$$

More generally, adopting the local residual model $\zeta_t = A_t \epsilon_t^\perp$ with uniformly bounded residual sensitivity $\|A_t\|_2 \le L$, the local residual model gives $\sigma^2_{\zeta,t} \le L^2 m_t^2$. Equation (20) therefore yields, within the Gaussian/high-rate approximation,

$$\limsup_{t\to\infty} \Delta R_t \lesssim \frac{1}{2} \ln \left(1 + \frac{L^2 d^2}{(1-\alpha)^2 \sigma^2_N} \right). \tag{34}$$

Selective contraction bounds the additional coding cost attributable to the orthogonal State perturbation under bounded total input and residual sensitivity.

Historical information and conditional entropy. For a deterministic transformation g of the existing reference, the Data Processing Inequality (Cover & Thomas, 2006) gives

$$I(g(M_{t-1}); X_{1:t-1}) \le I(M_{t-1}; X_{1:t-1}). \tag{35}$$

This statement concerns deterministic reprocessing of an existing reference; the complete codec also receives new information from the current bitstream when updating that reference. Dependence on past observations does not rank predictive usefulness or specify which components to retain. The entropy trends below instead follow from the additive-noise model.

For a scalar reference coordinate under the additive-noise model, consider

$$M_t = \bar{M}_t(X_{1:t-1}) + \epsilon_t, \tag{36}$$

where M_t here denotes that coordinate and $\bar{M}_t(X_{1:t-1})$ denotes the predictable component determined by the past observations available to the codec. If ϵ_t is zero-mean Gaussian and independent of $X_{1:t-1}$, its conditional differential entropy is (Cover & Thomas, 2006)

$$h(M_t \mid X_{1:t-1}) = h(\epsilon_t) = \frac{1}{2} \ln(2\pi e \sigma^2_{\epsilon,t}). \tag{37}$$

For a fixed high-rate quantizer (Gray & Neuhoff, 1998) with step size Δ, the corresponding discrete conditional entropy satisfies

$$H_\Delta(M_t \mid X_{1:t-1}) = h(M_t \mid X_{1:t-1}) - \ln \Delta + o(1), \tag{38}$$

so the monotone trend in differential entropy carries over to the discrete entropy up to an additive quantization-dependent constant.

Applying the same scalar additive assumptions to the preceding reference conditioned on $X_{1:t-1}$, the preservation-oriented variance model gives the entropy increment

$$\Delta h_t = \frac{1}{2} \ln \frac{\sigma^2_{\epsilon,t}}{\sigma^2_{\epsilon,t-1}} \ge 0. \tag{39}$$

Under these additive Gaussian, uncorrelated-input, and fixed high-rate quantization assumptions, the preservation-oriented variance growth yields

$$H_\Delta(M_t \mid X_{1:t-1}) \gtrsim H_\Delta(M_{t-1} \mid X_{1:t-1}), \tag{40}$$

where H_Δ denotes discrete conditional entropy at quantization step Δ. In the contraction model, the orthogonal variance approaches its fixed point, so its entropy increment approaches zero and can be negative when the initial variance exceeds that fixed point. These entropy trends follow from the perturbation model, independently of DPI.

E LIMITATIONS AND BROADER IMPACTS

E.1 LIMITATIONS

The perturbation analysis is local and conditional. The RMS bound requires uniform orthogonal contraction and bounded total input, including cross-direction coupling and feedback through other inputs. The coding-cost interpretation additionally assumes the stated residual model and Gaussian/high-rate approximation. These conditions are not verified as a closed-loop property of the complete codec.

The directional diagnostics use subspaces estimated from recent memory differences. They do not identify data-manifold tangent spaces, estimate worst-case operator norms, or isolate selectivity as the cause of the RD gains. In particular, the single-impulse BD-Rate experiment shows no consistent directional ordering at the tested scale. The finite-horizon propagation studies describe their respective protocols rather than establish unrestricted long-term stability.

Evaluation covers UVG, MCL-JCV, and HEVC-B under the reported low-delay settings, with A100 measurements for locally timed codecs. The study does not report variance across repeated training runs or a broader cross-dataset generalization evaluation. Subjective preferences summarize the tested participants, clips, and baselines. Code and weights are planned for release after acceptance.

E.2 BROADER IMPACTS

This work studies perceptual video compression, which can have both positive and negative societal impacts. On the positive side, stronger compression can reduce storage, bandwidth, and serving costs for video communication, streaming, and archival systems, and may also reduce the compute and energy required for large-scale video delivery. On the negative side, more efficient video compression can also improve the scalability of large video collection and distribution systems, including surveillance or other privacy-sensitive deployments. In addition, perceptual optimization may alter fine image details in ways that are acceptable for human viewing but less desirable for downstream forensic, restoration, or scientific-analysis use cases. We therefore view SRDVC primarily as a systems and compression contribution whose deployment should still consider application-specific requirements on faithfulness, privacy, and downstream use.