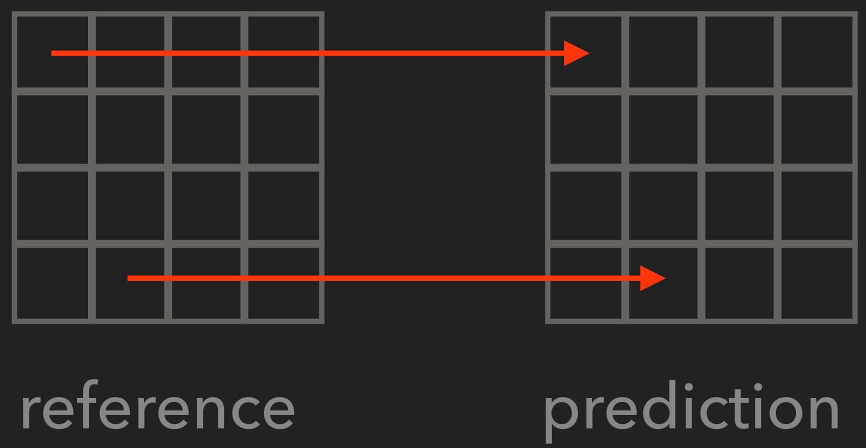
Jointly Optimized Transform Domain Temporal Prediction (TDTP) and Sub-pixel Interpolation

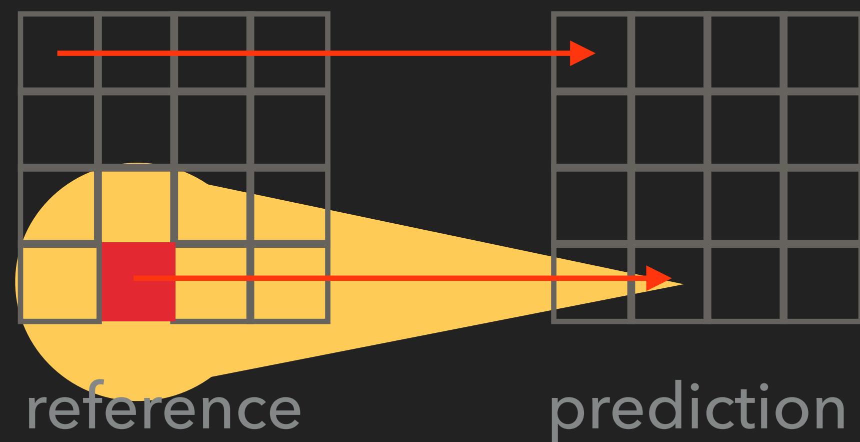
Shunyao Li, Tejaswi Nanjundaswamy, Kenneth Rose
University of California, Santa Barbara

MOTIVATION



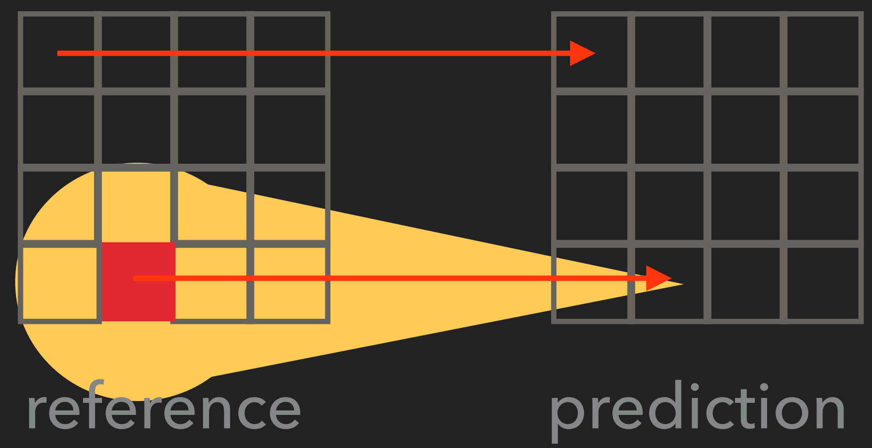
- Conventional temporal prediction: pixel-to-pixel

MOTIVATION



- Conventional temporal prediction: pixel-to-pixel
  - which ignores the spatial correlation -> suboptimal

MOTIVATION



- Conventional temporal prediction: pixel-to-pixel
  - which ignores the spatial correlation -> suboptimal
- Usually, people account for this in very complex ways:
  - Multi-tap filtering, 3D subband coding, etc.

TDTP

- A different perspective:
  - Spatial correlation is de-correlated in DCT domain
  - Optimal one-to-one prediction!

Transform Domain Temporal Prediction (TDTP)[1]

[1]J. Han et al. 2010, "Transform-domain temporal prediction in video coding: exploiting correlation variation across coefficients"

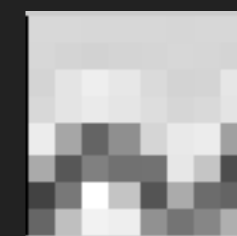
TEMPORAL CORRELATION

Reference block

Original block

Pixel domain

$\rho \approx 1$



TEMPORAL CORRELATION

Pixel domain

Reference block

$$\rho \approx 1$$

Original block

DCT domain

Reference block:

1497	-2	-33	-4	-21	81	14	0
229	-10	64	52	1	-70	-26	2
8	47	-70	-146	39	-15	1	5
-136	-38	18	130	-35	69	20	-4
78	-2	39	-17	10	-54	-30	8
43	17	-46	-82	-6	-20	19	4
-25	1	15	37	-10	35	-12	-5
-6	2	4	6	2	-17	5	1

Original block:

1505	1	-44	-10	-47	41	29	-15
230	-11	62	50	51	-40	-34	19
-41	38	-53	-136	-9	-8	14	-15
-110	-39	24	143	-32	44	19	5
80	1	26	-3	46	-33	-50	8
0	23	-44	-82	-30	4	42	-10
1	-8	21	29	4	10	-10	7
-1	-2	-3	3	8	-12	-7	-2

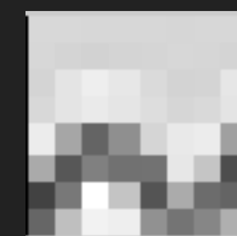
TEMPORAL CORRELATION

Reference block

Original block

Pixel domain

$\rho \approx 1$



At low frequency, $\rho \approx 1$

DCT domain

1497	-2	-33	-4	-21	81	14	0
229	-10	64	52	1	-70	-26	2
8	47	-70	-146	39	-15	1	5
-136	-38	18	130	-35	69	20	-4
78	-2	39	-17	10	-54	-30	8
43	17	-46	-82	-6	-20	19	4
-25	1	15	37	-10	35	-12	-5
-6	2	4	6	2	-17	5	1

1505	1	-44	-10	-47	41	29	-15
230	-11	62	50	51	-40	-34	19
-41	38	-53	-136	-9	-8	14	-15
-110	-39	24	143	-32	44	19	5
80	1	26	-3	46	-33	-50	8
0	23	-44	-82	-30	4	42	-10
1	-8	21	29	4	10	-10	7
-1	-2	-3	3	8	-12	-7	-2

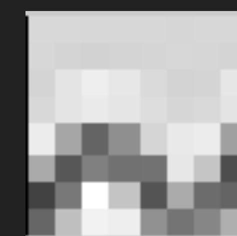
TEMPORAL CORRELATION

Pixel domain

Reference block

$\rho \approx 1$

Original block



At low frequency, $\rho \approx 1$

DCT domain

Reference block:

1497	-2	-33	-4	-21	81	14	0
229	-10	64	52	1	-70	-26	2
8	47	-70	-146	39	-15	1	5
-136	-38	18	130	-35	69	20	-4
78	-2	39	-17	10	-54	-30	8
43	17	-46	-82	-6	-20	19	4
-25	1	15	37	-10	35	-12	-5
-6	2	4	6	2	-17	5	1

Original block:

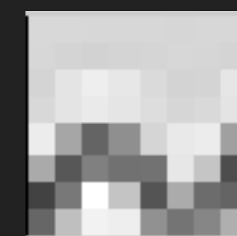
1505	1	-44	-10	-47	41	29	-15
230	-11	62	50	51	-40	-34	19
-41	38	-53	-136	-9	-8	14	-15
-110	-39	24	143	-32	44	19	5
80	1	26	-3	46	-33	-50	8
0	23	-44	-82	-30	4	42	-10
1	-8	21	29	4	10	-10	7
-1	-2	-3	3	8	-12	-7	-2

TEMPORAL CORRELATION

Reference block | Original block

Pixel domain

$\rho \approx 1$



At low frequency, $\rho \approx 1$

DCT domain

Reference block:

1497	-2	-33	-4	-21	81	14	0
229	-10	64	52	1	-70	-26	2
8	47	-70	-146	39	-15	1	5
-136	-38	18	130	-35	69	20	-4
78	-2	39	-17	10	-54	-30	8
43	17	-46	-82	-6	-20	19	4
-25	1	15	37	-10	35	-12	-5
-6	2	4	6	2	-17	5	1

Original block:

1505	1	-44	-10	-47	41	29	-15
230	-11	62	50	51	-40	-34	19
-41	38	-53	-136	-9	-8	14	-15
-110	-39	24	143	-32	44	19	5
80	1	26	-3	46	-33	-50	8
0	23	-44	-82	-30	4	42	-10
1	-8	21	29	4	10	-10	7
-1	-2	-3	3	8	-12	-7	-2

At high frequency, $\rho < 1$

TEMPORAL CORRELATION

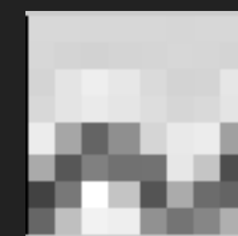
Dominated by low frequency part

Reference block

Original block

Pixel domain

$\rho \approx 1$



At low frequency, $\rho \approx 1$

DCT domain

Reference block:

1497	-2	-33	-4	-21	81	14	0
229	-10	64	52	1	-70	-26	2
8	47	-70	-146	39	-15	1	5
-136	-38	18	130	-35	69	20	-4
78	-2	39	-17	10	-54	-30	8
43	17	-46	-82	-6	-20	19	4
-25	1	15	37	-10	35	-12	-5
-6	2	4	6	2	-17	5	1

Original block:

1505	1	-44	-10	-47	41	29	-15
230	-11	62	50	51	-40	-34	19
-41	38	-53	-136	-9	-8	14	-15
-110	-39	24	143	-32	44	19	5
80	1	26	-3	46	-33	-50	8
0	23	-44	-82	-30	4	42	-10
1	-8	21	29	4	10	-10	7
-1	-2	-3	3	8	-12	-7	-2

At high frequency, $\rho < 1$

TEMPORAL CORRELATION

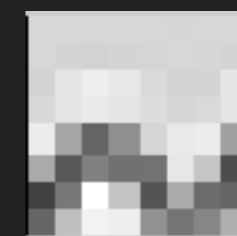
Dominated by low frequency part

Pixel domain

Reference block

$\rho \approx 1$

Original block



DCT domain

At low frequency, $\rho \approx 1$

Reference block (DCT):

1497	-2	-33	-4	-21	81	14	0
229	-10	64	52	1	-70	-26	2
8	47	-70	-146	39	-15	1	5
-136	-38	18	130	-35	69	20	-4
78	-2	39	-17	10	-54	-30	8
43	17	-46	-82	-6	-20	19	4
-25	1	15	37	-10	35	-12	-5
-6	2	4	6	2	-17	5	1

Original block (DCT):

1505	1	-44	-10	-47	41	29	-15
230	-11	62	50	51	-40	-34	19
-41	38	-53	-136	-9	-8	14	-15
-110	-39	24	143	-32	44	19	5
80	1	26	-3	46	-33	-50	8
0	23	-44	-82	-30	4	42	-10
1	-8	21	29	4	10	-10	7
-1	-2	-3	3	8	-12	-7	-2

At high frequency, $\rho < 1$

- TDTP: Better exploit the temporal correlation

TDTP

- For each DCT coefficient, its prediction is:

$$\tilde{x}_n = \rho \hat{x}_{n-1}$$

$$\rho = \frac{E(x_n \hat{x}_{n-1})}{E(\hat{x}_{n-1}^2)} \longrightarrow \text{Correlation between source and reference}$$

- TDTP: scale reference with temporal correlation for each DCT coefficient

CHALLENGE: SUB-PIXEL INTERPOLATION

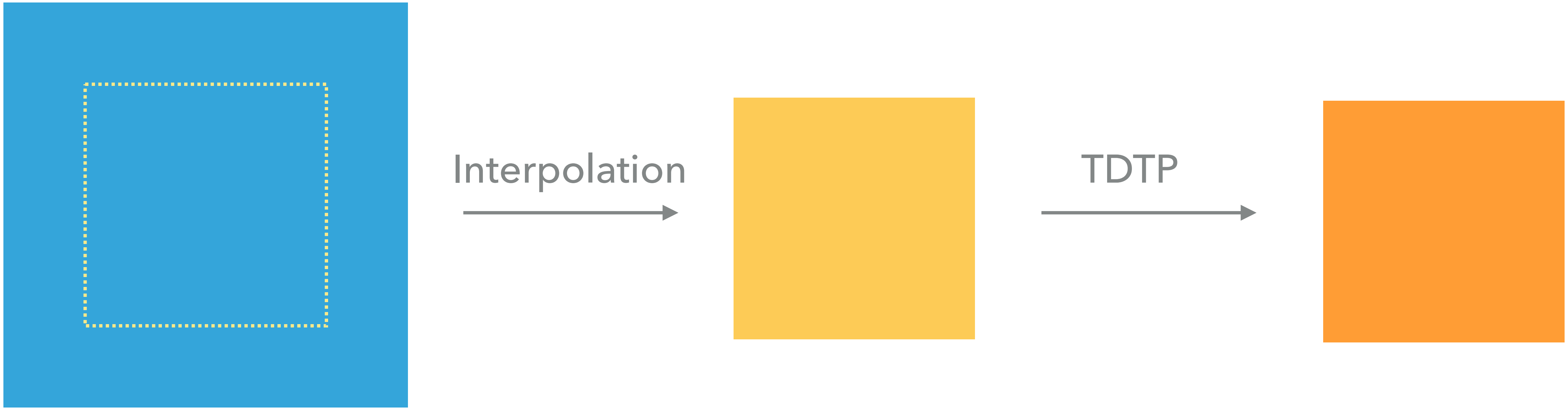
$$\tilde{x}_n = \rho \hat{x}_{n-1}$$

0.999	0.998	0.997	…				
0.996	0.978	…					
0.983	…						
…							
							…
						…	0.748
					…	0.700	0.512
				…	0.640	0.470	0.339

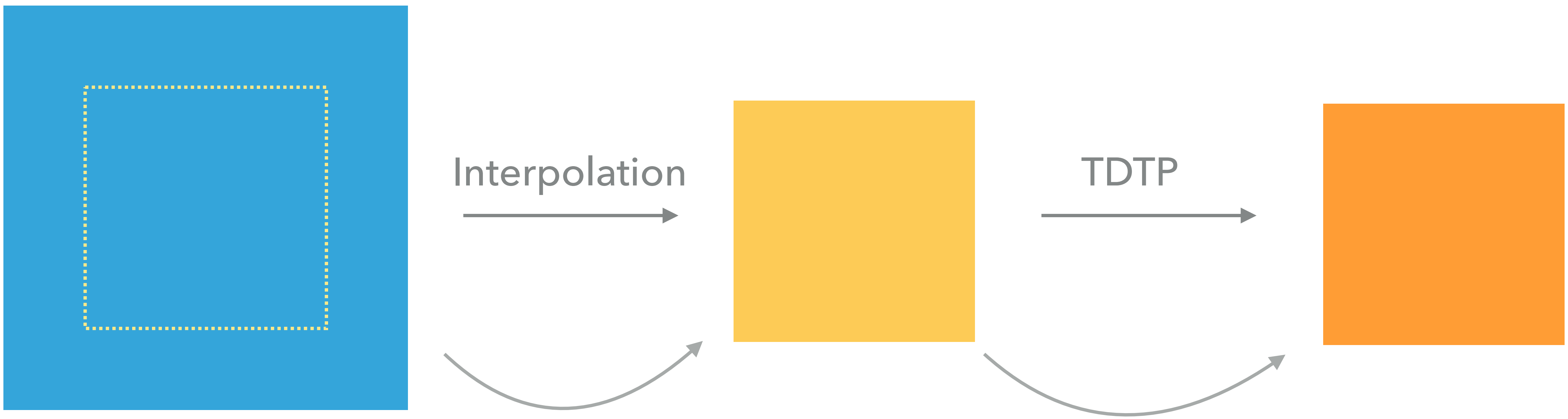
Example ρ values in 8x8 blocks

- High-freq are scaled down more than low-freq
- Similar to the interpolation filters’ low-pass frequency response
- The gain drops significantly!

INTERPOLATION FILTER VS TDTP



INTERPOLATION FILTER VS TDTP



Interpolation filter maps the pixels as well as its neighbor pixels into a subspace

TDTP de-correlates spatial correlation in the subspace

EXTENDED BLOCK TDTP (EB-TDTP)

EXTENDED BLOCK TDTP (EB-TDTP)

$$\tilde{\mathbf{Y}} = \mathbf{F}_1 \mathbf{D}'_{B_2} (\mathbf{D}_{B_2} \mathbf{X} \mathbf{D}'_{B_2}) \circ \mathbf{P}_{B_2}) \mathbf{D}_{B_2} \mathbf{F}_2$$

DCT

EB-TDTP

Back to pixel domain interpolation

EXTENDED BLOCK TDTP (EB-TDTP)

$$min||\mathbf{Y} - \tilde{\mathbf{Y}}||^2$$

$$\tilde{\mathbf{Y}} = \mathbf{F}_1 \mathbf{D}'_{B_2} (\mathbf{D}_{B_2} \mathbf{X} \mathbf{D}'_{B_2}) \circ \mathbf{P}_{B_2}) \mathbf{D}_{B_2} \mathbf{F}_2$$

DCT

EB-TDTP

Back to pixel domain interpolation

JOINT OPTIMIZATION

- Design $\{\mathbf{P}_{B_2}, \mathbf{F}_1, \mathbf{F}_2\}$ to minimize the MSE
- Use ***an iterative approach*** to optimize one of them while fixing the others
  - Fixing $\{\mathbf{F}_1, \mathbf{F}_2\}$, optimize $\mathbf{P}_{B_2}$ → optimize EB-TDTP
  - Fixing $\{\mathbf{P}_{B_2}, \mathbf{F}_2\}$, optimize $\mathbf{F}_1$ → optimize interpolation filter
  - Fixing $\{\mathbf{P}_{B_2}, \mathbf{F}_1\}$, optimize $\mathbf{F}_2$ → optimize interpolation filter

$$min||\mathbf{Y} - \tilde{\mathbf{Y}}||^2$$

$$\tilde{\mathbf{Y}} = \mathbf{F}_1 \mathbf{D}'_{B_2} (\mathbf{D}_{B_2} \mathbf{X} \mathbf{D}'_{B_2}) \circ \mathbf{P}_{B_2}) \mathbf{D}_{B_2} \mathbf{F}_2$$

JOINT OPTIMIZATION

$$J = ||\mathbf{Ax} - \mathbf{b}||^2$$

$$\mathbf{x}_{opt} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

- Fixing $\{\mathbf{F}_1, \mathbf{F}_2\}$, optimize $\mathbf{P}_{B_2}$ → optimize EB-TDTP
- Fixing $\{\mathbf{P}_{B_2}, \mathbf{F}_2\}$, optimize $\mathbf{F}_1$ → optimize interpolation filter
- Fixing $\{\mathbf{P}_{B_2}, \mathbf{F}_1\}$, optimize $\mathbf{F}_2$ → optimize interpolation filter

$$min||\mathbf{Y} - \tilde{\mathbf{Y}}||^2$$

$$\tilde{\mathbf{Y}} = \mathbf{F}_1\mathbf{D}'_{B_2}(\mathbf{D}_{B_2}\mathbf{X}\mathbf{D}'_{B_2}) \circ \mathbf{P}_{B_2})\mathbf{D}_{B_2}\mathbf{F}_2$$

RE-CAP

- TDTP ***de-correlates spatial correlation*** and ***exploits real temporal correlation across frequencies***
- TDTP ***interferes*** with interpolation filter
- ***Joint design*** by an iterative approach

NON-SEPARABLE FILTERS

- Separable filters cannot perfectly capture the spatial correlation

NON-SEPARABLE FILTERS

- Separable filters cannot perfectly capture the spatial correlation
- Alternative: non-separable filters (at the same complexity)
  2D 4x4 non-separable filters = two 1D 8-tap separable filters
- A similar iterative optimization approach to design $\{\mathbf{P}_{B_2}, \mathbf{F}\}$

$$min||\mathbf{Y} - \tilde{\mathbf{Y}}||^2$$

$$\tilde{\mathbf{Y}} = (\mathbf{D}'_{B_2}((\mathbf{D}_{B_2}\mathbf{X}\mathbf{D}'_{B_2}) \circ \mathbf{P}_{B_2})\mathbf{D}_{B_2}) * \mathbf{F}$$

DCT

EB-TDTP

Back to pixel domain interpolation

non-separable wiener filter

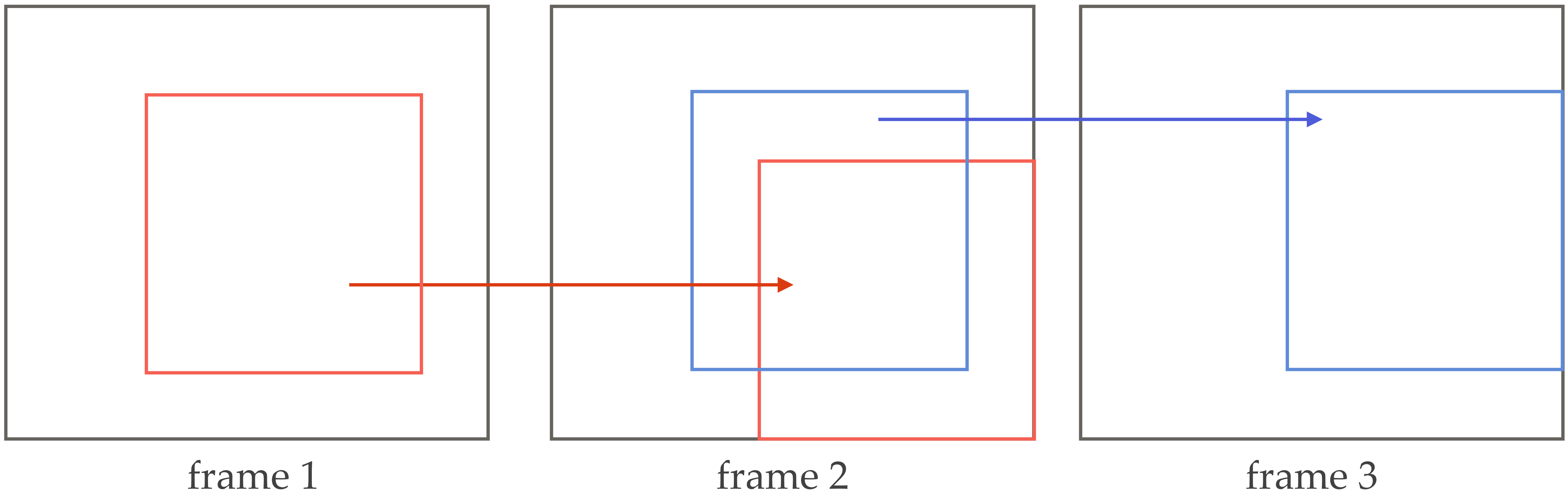
INSTABILITY PROBLEM IN TRAINING

INSTABILITY PROBLEM IN TRAINING

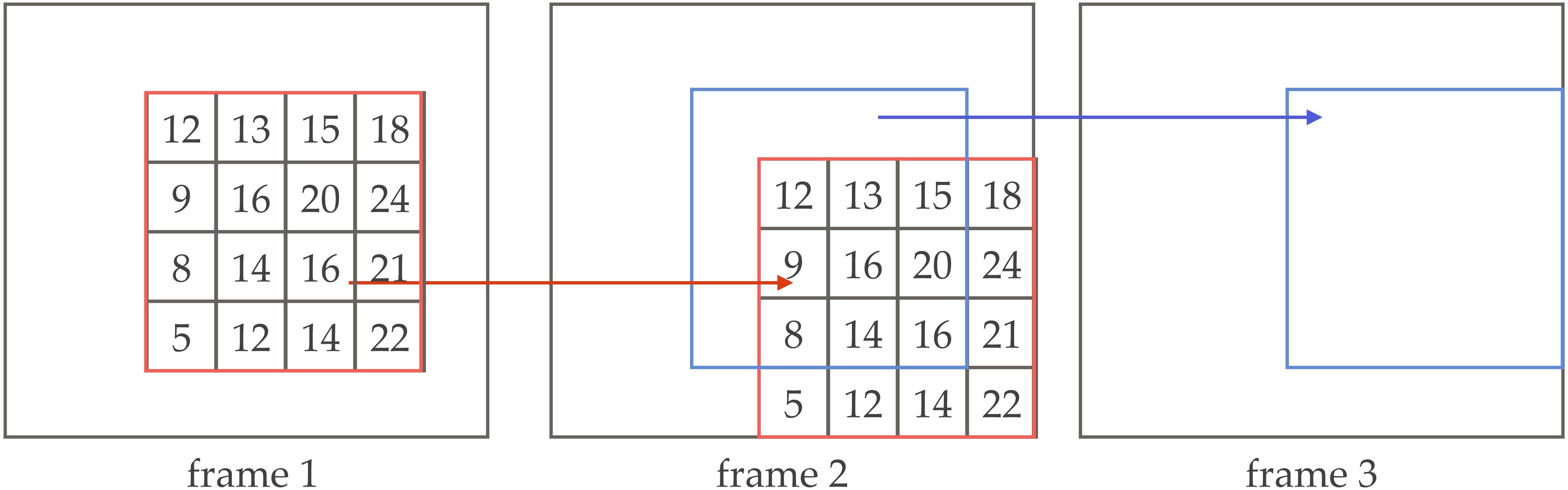
Whatever statistics we designed for will be **changed**
when we apply the new predictor on it

Because in a closed-loop system
each frame is **referencing from a different reconstruction** now.

INSTABILITY PROBLEM IN TRAINING

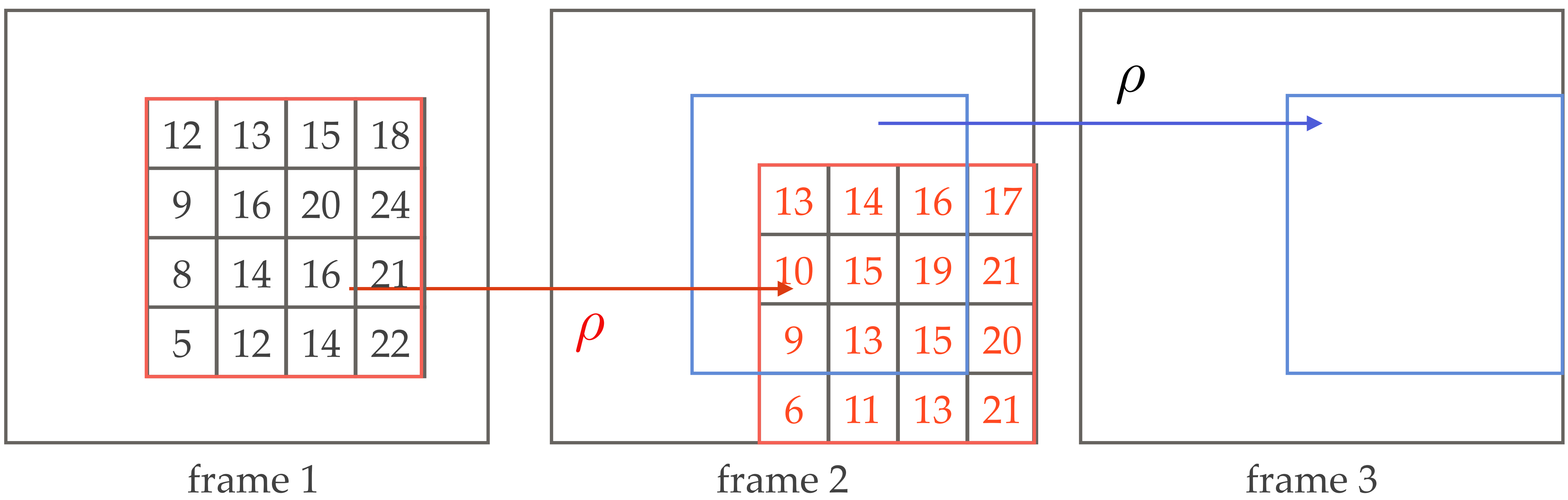


INSTABILITY PROBLEM IN TRAINING

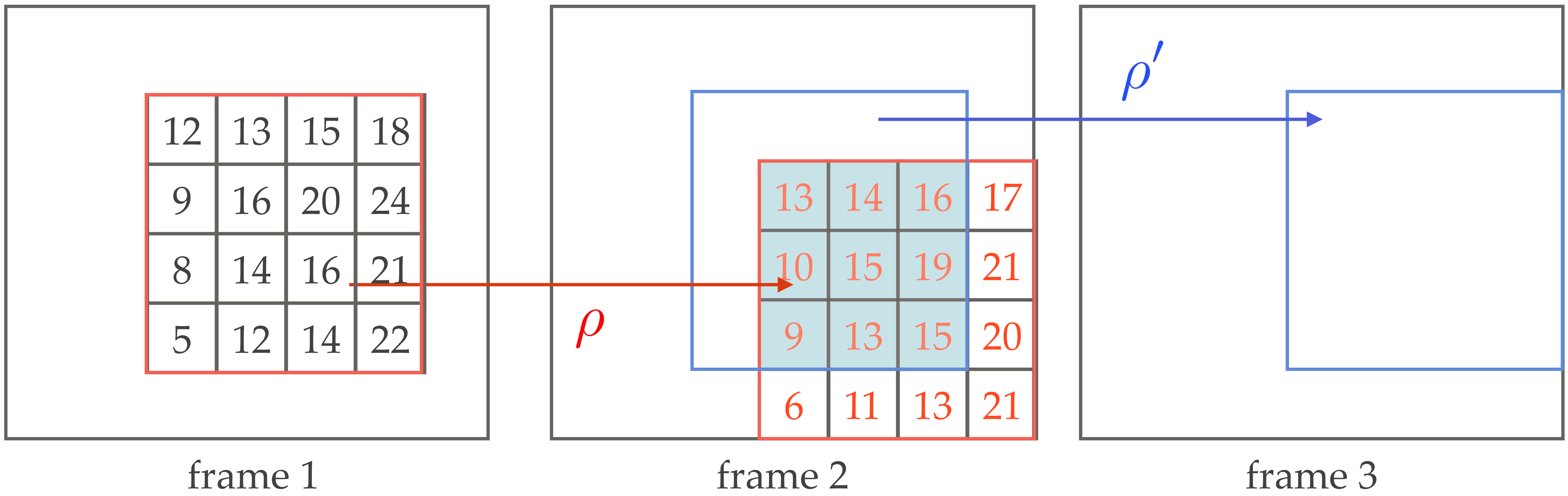


Get some ρ from the reference blocks and original blocks

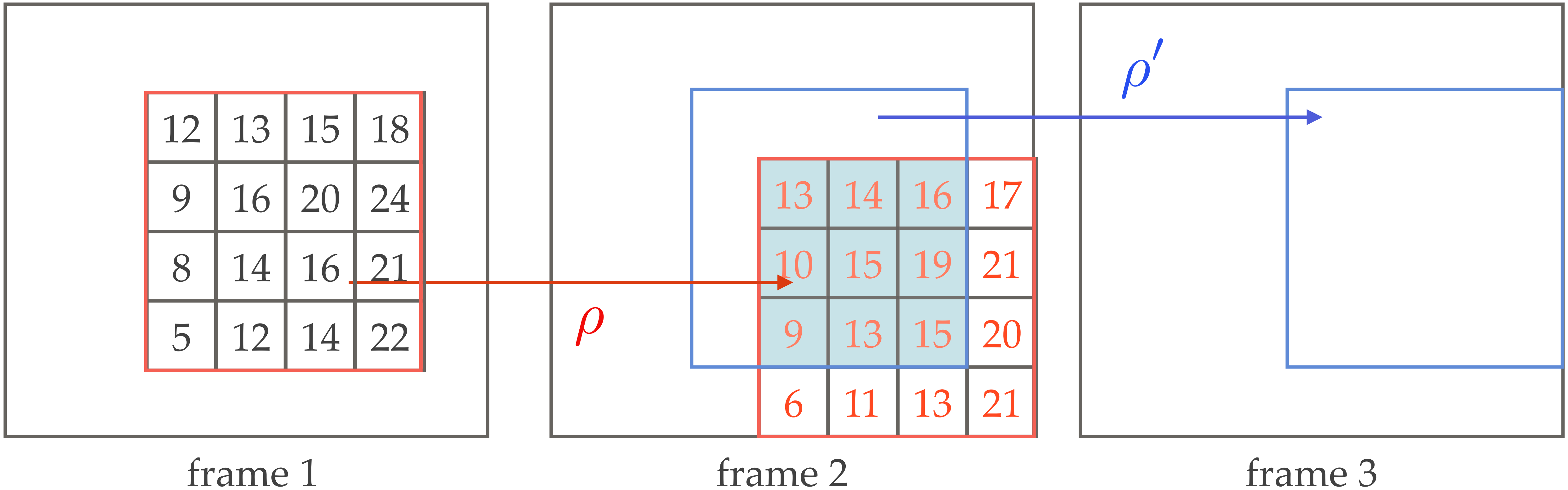
INSTABILITY PROBLEM IN TRAINING



INSTABILITY PROBLEM IN TRAINING



INSTABILITY PROBLEM IN TRAINING

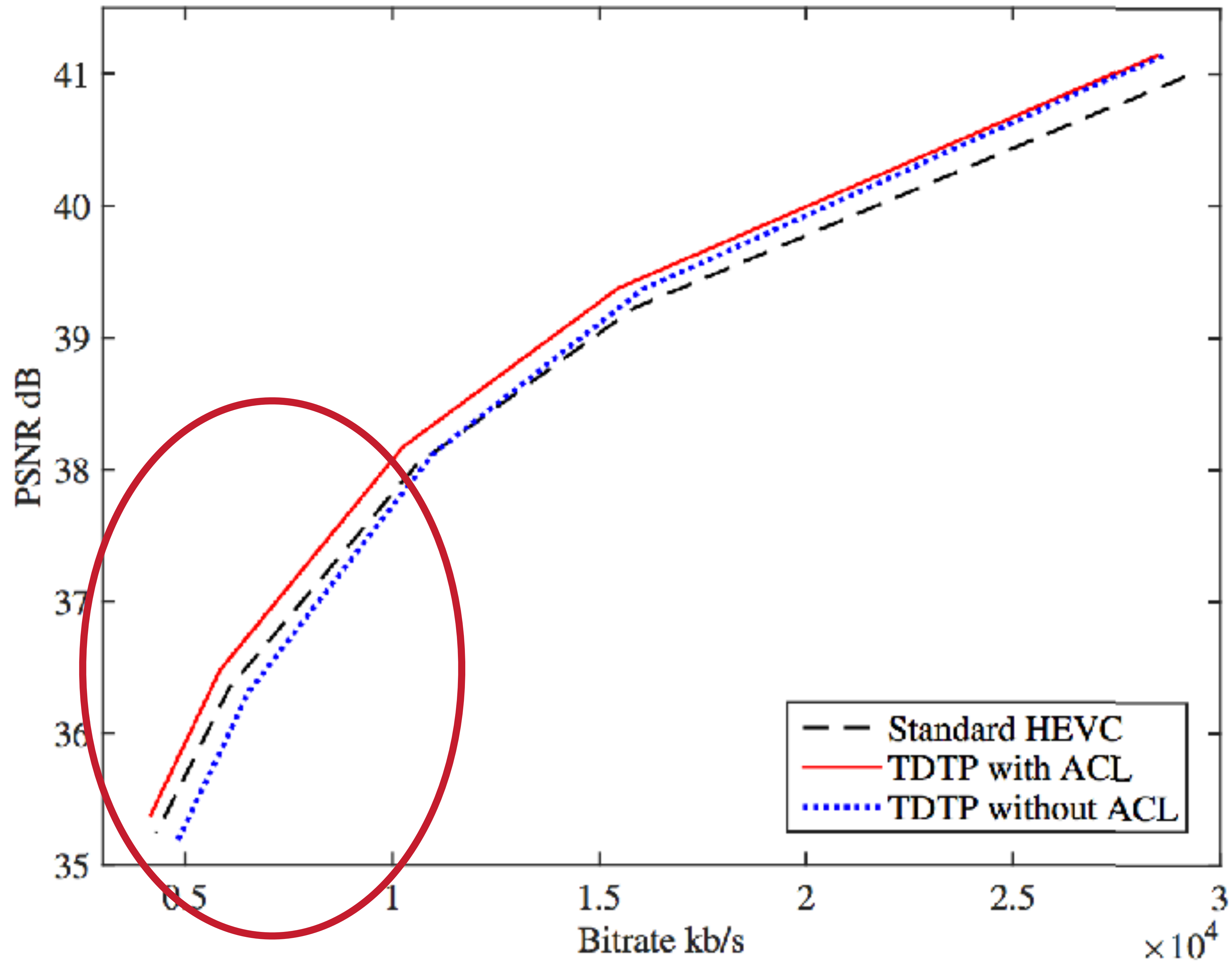


- The change in reconstruction will keep propagating to the following frames… and change the statistics completely in the end!

SOLUTION — ASYMPTOTIC CLOSED-LOOP (ACL) DESIGN

[1] H. Khalil, K. Rose, and S. L. Regunathan, "The asymptotic closed-loop approach to predictive vector quantizer design with application in video coding," TIP 2001

[2] S. Li, T. Nanjundaswamy, Y. Chen, and K. Rose, "Asymptotic Closed-loop Design for Transform Domain Temporal Prediction", ICIP 2015

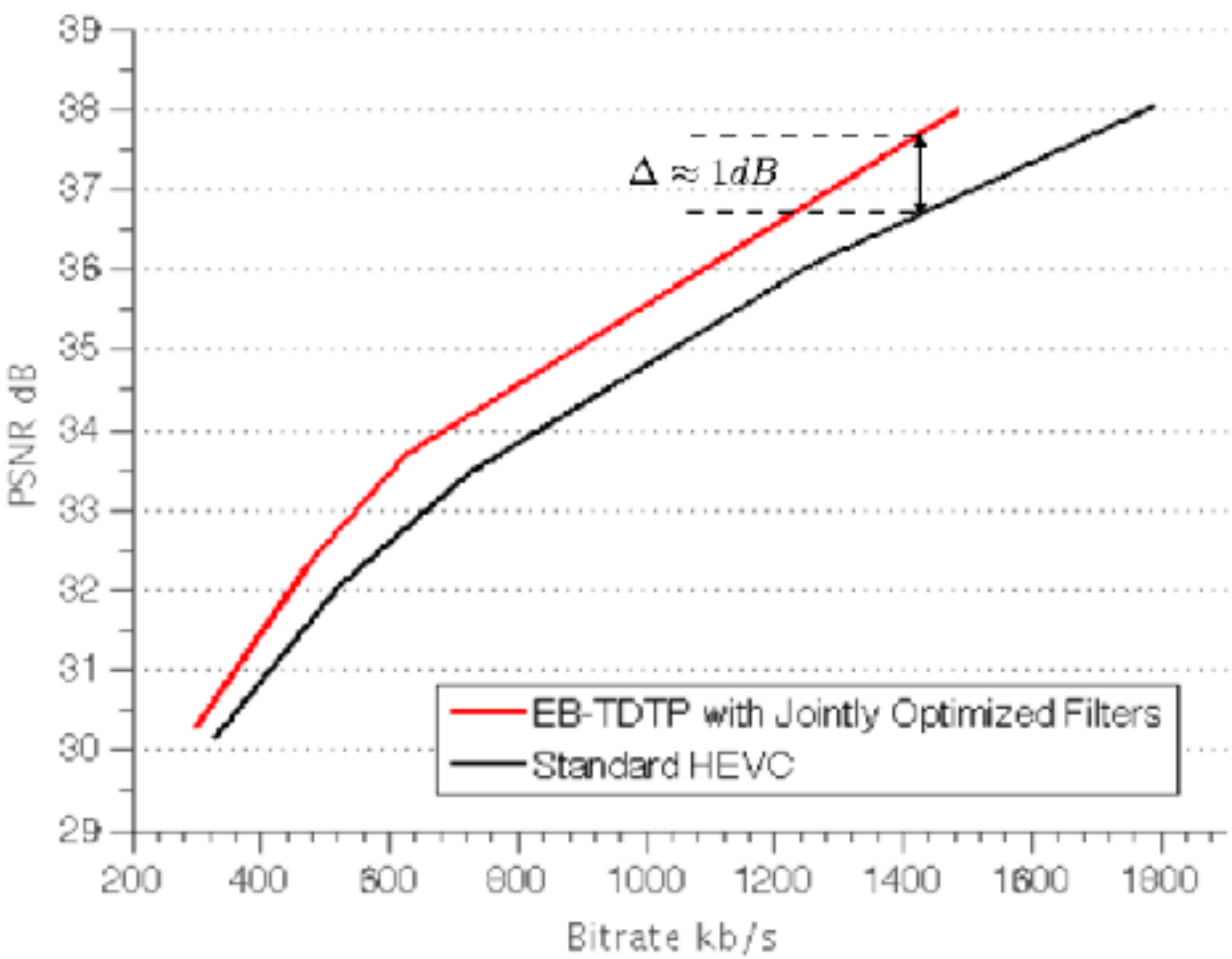


RESULTS

- HEVC baseline: lowdelay P; using previous frame as reference frame; fixing CU/TU size to be 8x8; SAO disabled
- ***Experiment 1***: design the EB-TDTP and interpolation for each sequence, aiming for ***offline encoding application***

RESULTS

- HEVC baseline: lowdelay P; using previous frame as reference frame; fixing CU/TU size to be 8x8; SAO disabled
- ***Experiment 1***: design the EB-TDTP and interpolation for each sequence, aiming for ***offline encoding application***



RD curve for sequence *BQSquare*

	TDTP	EB-TDTP	JointOpt
coastguard (CIF)	8.61	10.03	10.63
bridge-far (CIF)	9.20	10.58	11.35
mobile(CIF)	4.60	7.44	8.58
highway (CIF)	3.94	6.10	7.73
stefan (CIF)	3.67	3.90	4.67
BQSquare (240p)	0.74	1.90	14.44
BlowingBubbles (240p)	1.06	1.01	1.20
BQMall (480p)	2.04	1.83	3.43
PartyScene (480p)	1.21	1.41	5.72
Keiba (480p)	4.04	4.52	4.48
vidyo1 (720p)	1.53	2.51	2.51
BQTerrace (1080p)	12.78	15.03	20.14
ParkScene (1080p)	2.53	2.57	2.57
Kimono (1080p)	7.21	6.91	8.16
AVERAGE	**4.51**	**5.41**	**7.54**

RESULTS

- ***Experiment 2***: provide 8 modes of the trained parameters for encoder to choose for each sequence (with an overhead of 3 bits/sequence)

RESULTS

- ***Experiment 2***: provide 8 modes of the trained parameters for encoder to choose for each sequence (with an overhead of 3 bits/sequence)

- For simplicity, we use the 8 most distinct sets of predictors from the training set -> huge potential for proper mode design and adaptivity exploration

	JointOpt
container (CIF)	9.16
bridge-close (CIF)	6.26
bus (CIF)	3.77
tempete (CIF)	3.67
waterfall (CIF)	1.81
flower (CIF)	0.21
city (CIF)	0.92
FourPeople (720p)	6.27
vidyo3 (720p)	4.18
vidyo4 (720p)	3.59
BasketballDrive (1080p)	4.71
Cactus (1080p)	3.21
Tennis (1080p)	1.09
AVERAGE	**3.76**

SUMMARY

Paper #2324: Jointly Optimized Transform Domain Temporal Prediction (TDTP) and Sub-pixel Interpolation

- Transform domain temporal prediction (TDTP) ***disentangles the spatial and temporal correlation***, and exploits the ***true temporal correlation at each frequency***
- ***TDTP interferes with interpolation filter***
- Extended blocks TDTP (EB-TDTP) accounts for the spatial correlations ***outside the block***
- We ***jointly design*** the EB-TDTP and (separable and non-separable) sub-pixel interpolation filters in ***an iterative approach*** **(main contribution of this paper)**
- We use the asymptotic closed-loop (ACL) approach to ***avoid the instability problem*** due to quantization error propagation
- Future research includes proper mode design and adaptivity exploration for real-time encoding applications