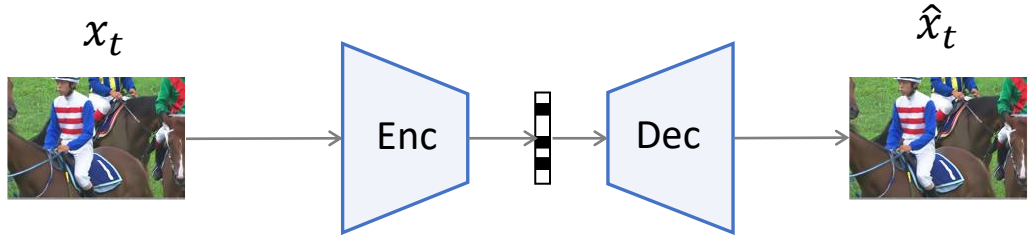
Deep Contextual Video Compression

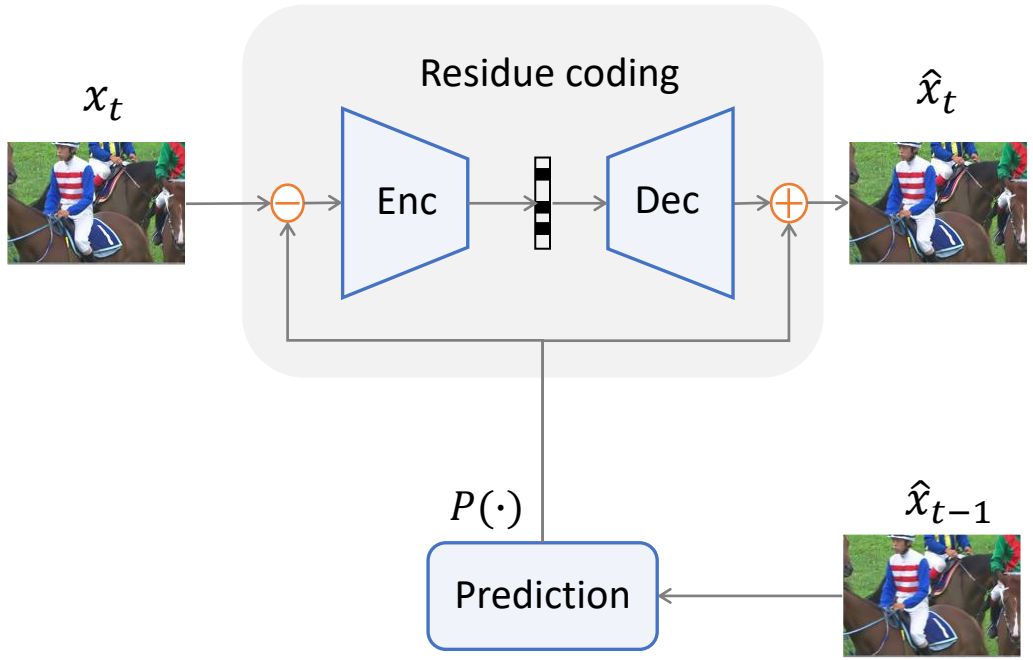
Jiahao Li, Bin Li, and Yan Lu
Microsoft Research Asia

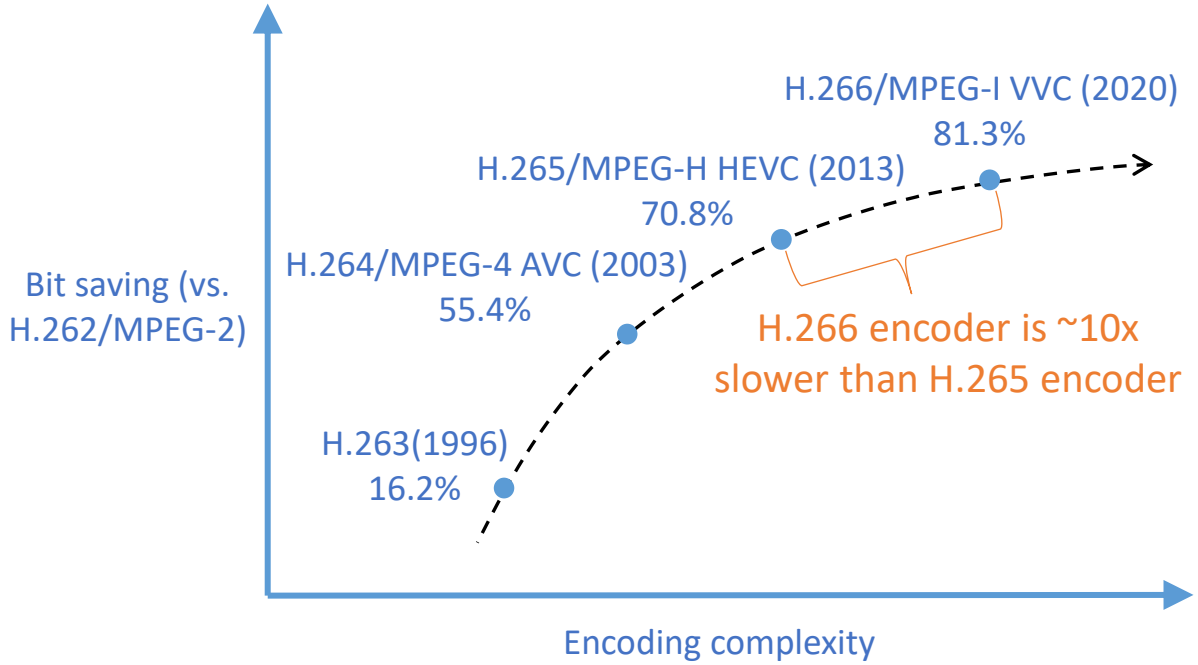
Image compression



Video compression via residue coding

$$H(x_t) \Longrightarrow H(x_t - P(\hat{x}_{t-1})) \ , where \ \ P(\cdot) = \{\text{Motion, block partition, interpolation...}\}$$

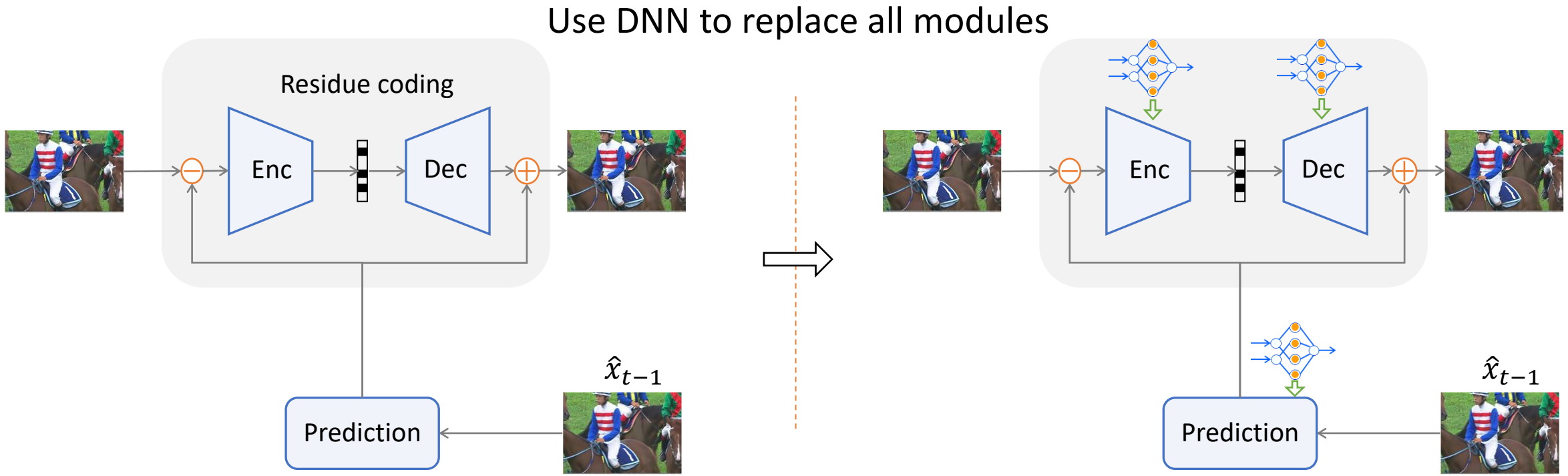




- All Standards in recent 30 years adopt this framework
- Continuously refine $P(\cdot)$: bit saving becomes marginal, but the cost increase is non-trivial

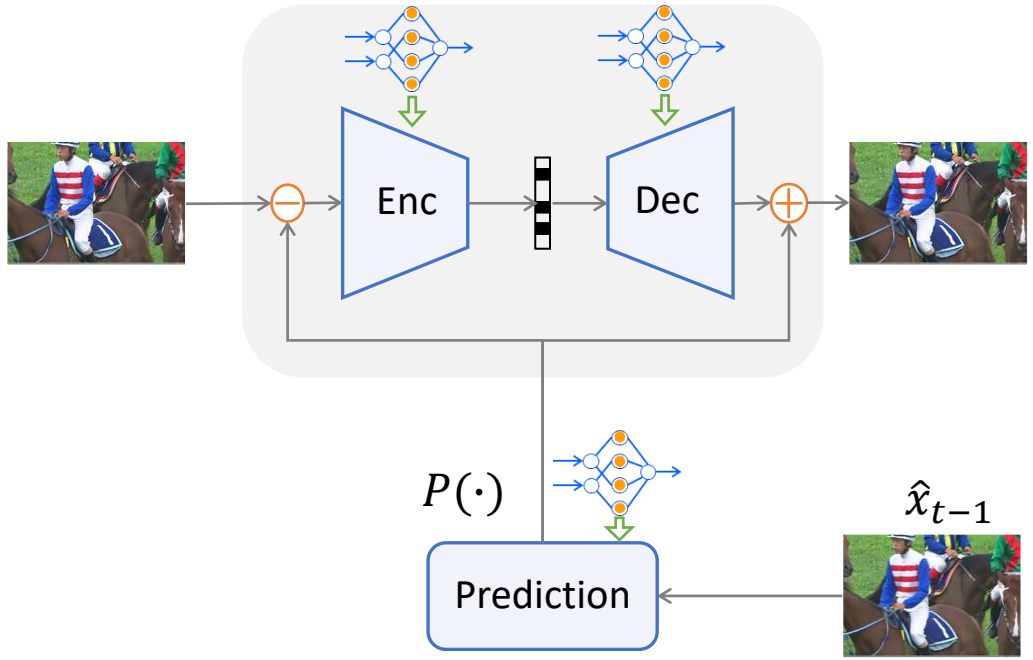
Deep video compression

- Most existing solutions follow the residue coding-based framework



*Lu et al. "An end-to-end learning framework for video compression." IEEE TPAMI (2020)

Limitation: fixed temporal redundancy removal



Step 1: use **subtraction** to remove temporal redundancy

$$\underbrace{H(x_t - P(\hat{x}_{t-1}))}$$

Step 2: remove spatial redundancy in residue

Two-stage learning

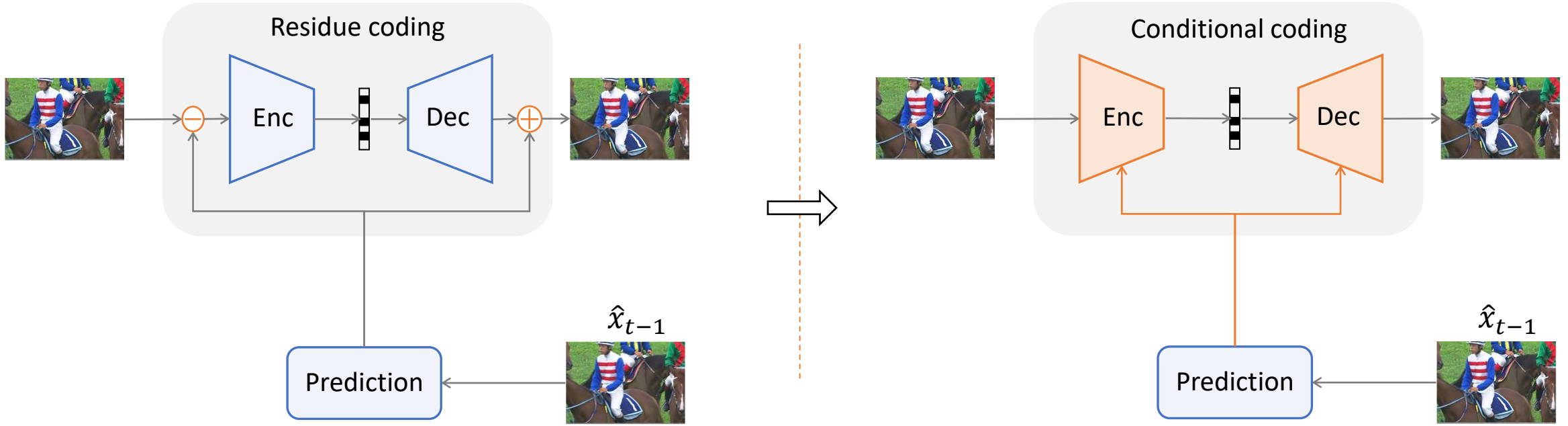
Failure case of fixed temporal redundancy removal

It should be easy to compress x_t given $P(\hat{x}_{t-1})$, but subtraction leads to residue with large energy

Problem: how to better utilize temporal correlation?

Our conditional coding-based solution

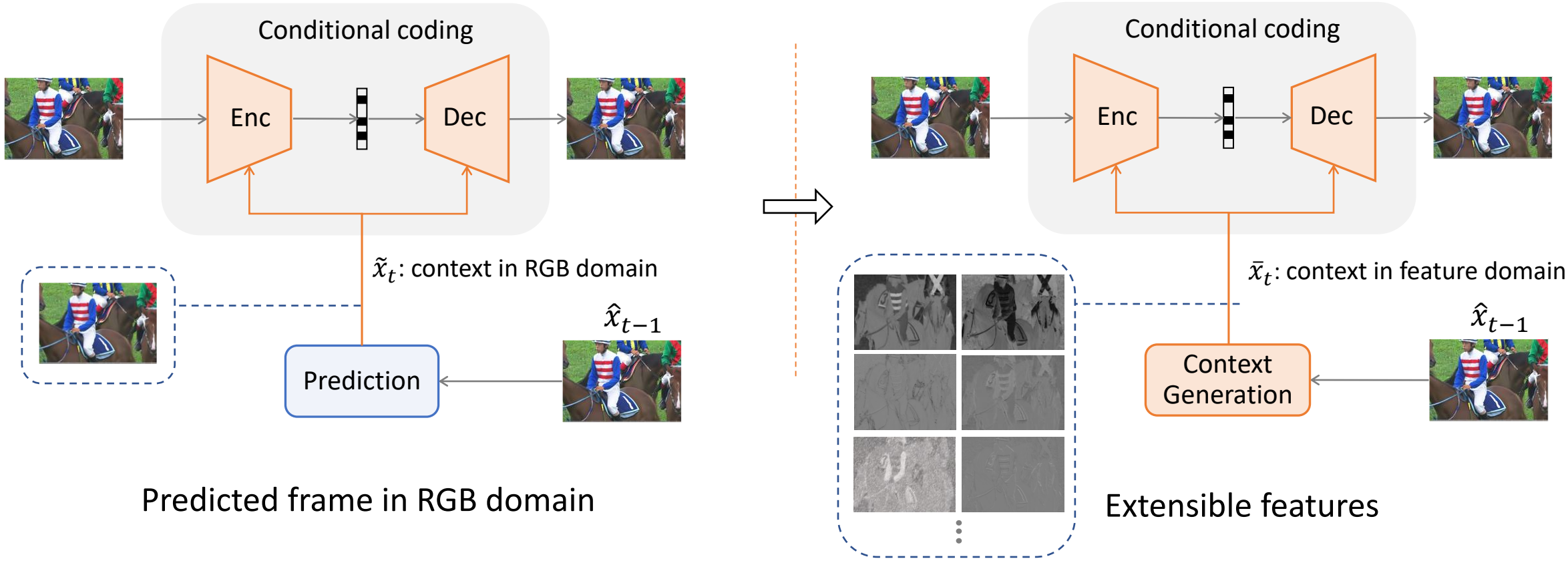
- From fixed subtraction to adaptive learning manner



$$H(x_t - P(\hat{x}_{t-1})) \geq H(x_t \mid P(\hat{x}_{t-1}))$$

Deep contextual video compression (DCVC)

- Extensible features as condition rather than RGB prediction

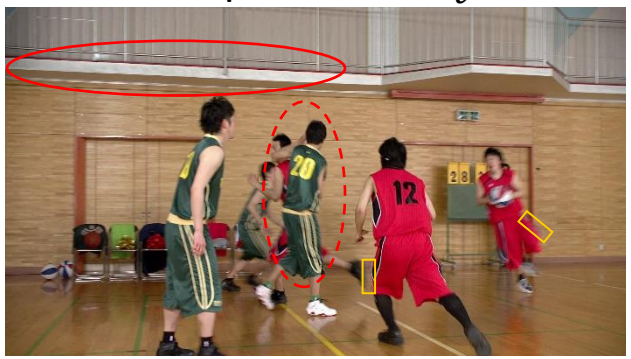


DCVC: better video quality

- Smaller reconstruction error for high frequency region and object boundary

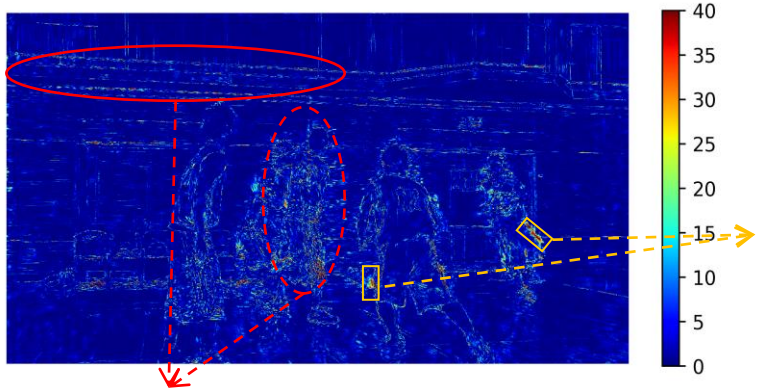
Previous decoded frame $\hat{x}_{t-1}$

Input frame x_t



High frequency in x_t

Reduction of reconstruction error



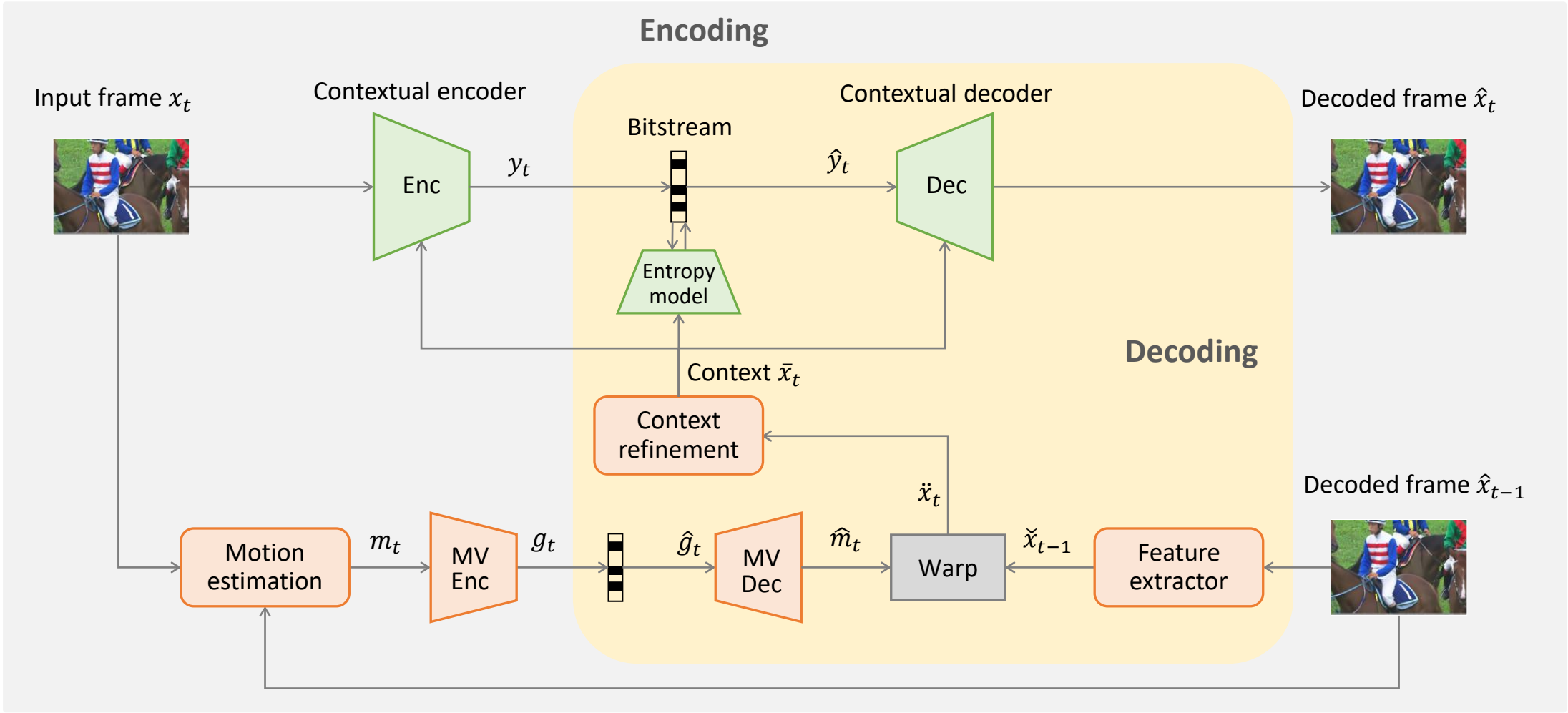
Object boundary

DCVC enables the adaptivity of intra coding and inter coding

High frequency region in foreground/ background

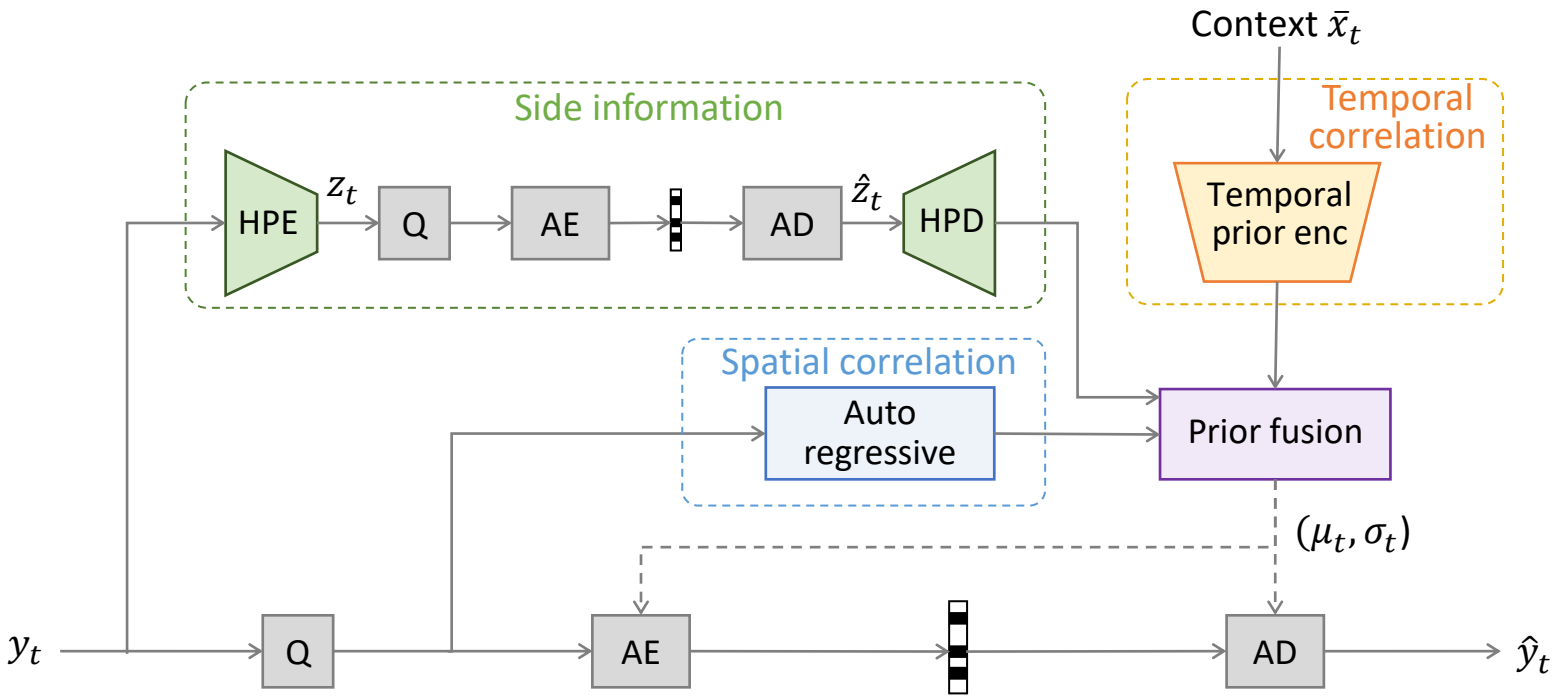
high dimension context carries rich information to help reconstruct the high frequency contents

Detailed framework



Context guided entropy model

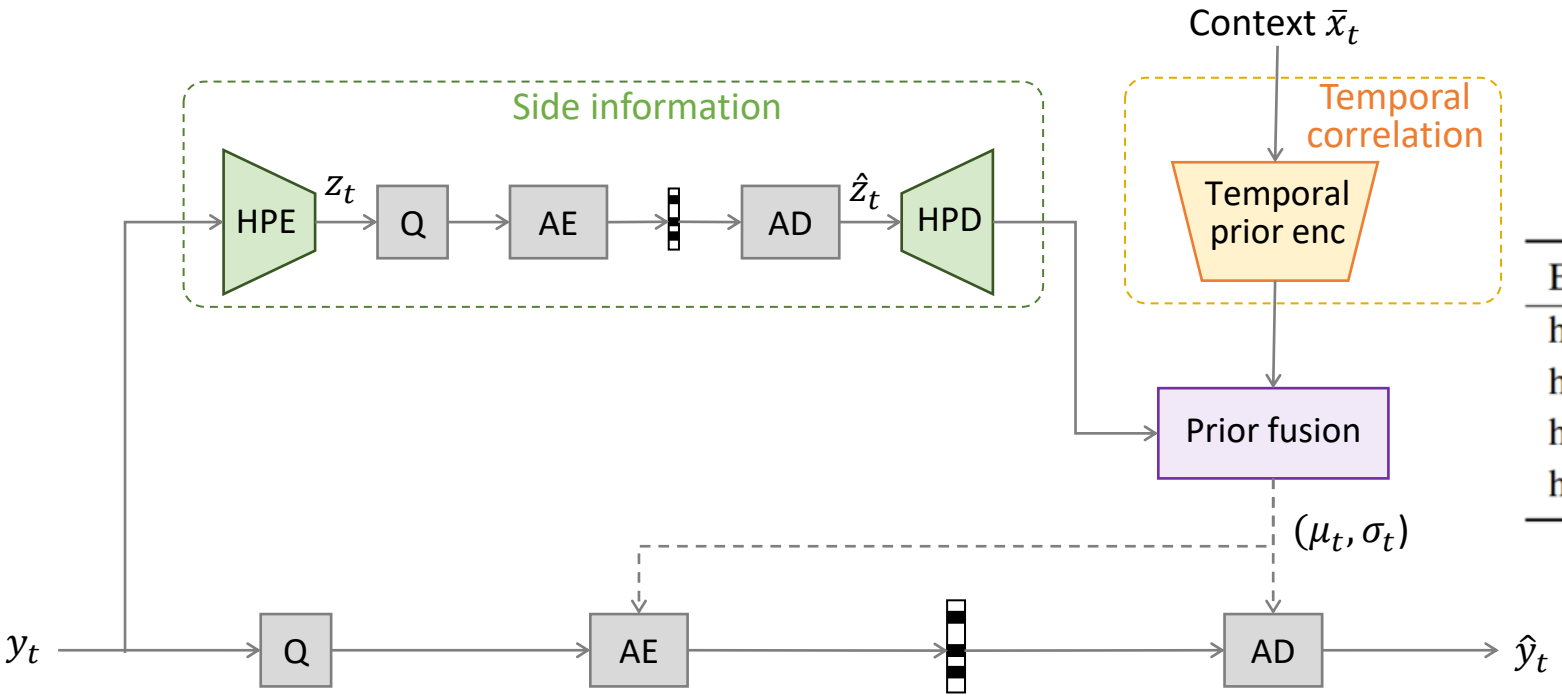
- More accurate entropy model



- AE/AD: arithmetic encoder/decoder
- Q: quantization

Context guided entropy model

- Also support fast encoding/decoding mode by removing auto regressive model

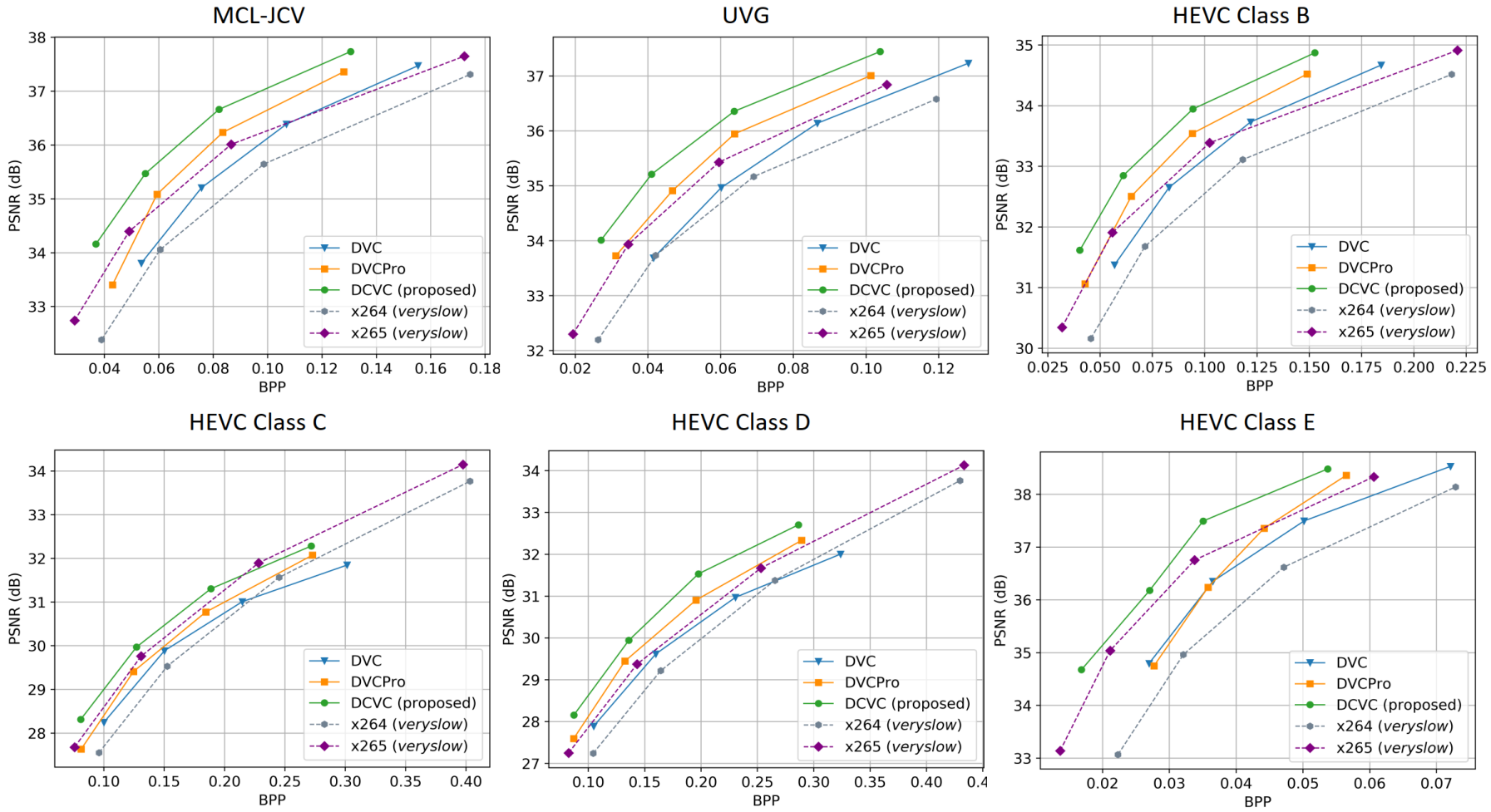


Entropy model	Bitrate increase
hyper prior + spatial prior + temporal prior	0.0%
hyper prior + temporal prior	3.8%
hyper prior + spatial prior	4.6%
hyper prior	60.9%

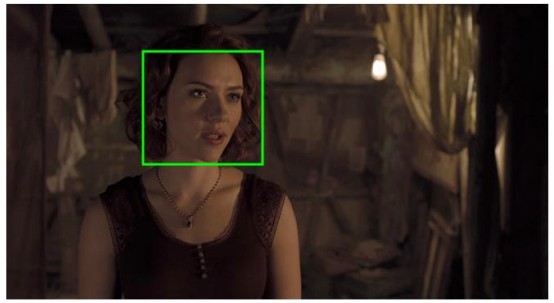
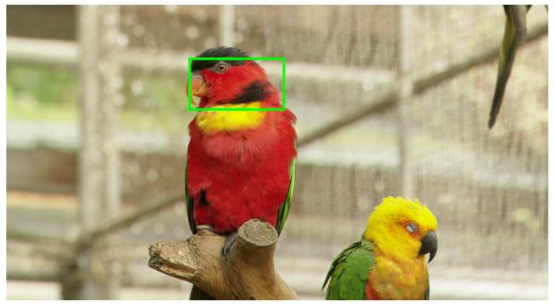
- AE/AD: arithmetic encoder/decoder
- Q: quantization

Quantitative results

- Improvement on datasets with various resolutions and contents



Qualitative comparison

	Original	DVCPro	Our DCVC
			
	BPP/PSNR	0.012/38.9	0.011/43.0
			
	BPP/PSNR	0.032/25.6	0.031/33.7
			
	BPP/PSNR	0.020/34.0	0.019/36.2

Summary

- Design a conditional coding framework for deep video compression
  - Enables the adaptivity of intra coding and inter coding
- Feature domain condition rather than pixel domain condition
  - Richer information to help reconstruct the high frequency contents
- Context guided entropy model
  - More accurate probability estimation
- Extensible framework, where the condition can be flexibly designed
  - Great potential

Thank You