Machine Oriented Compression

Dan Jacobellis

TEXAS
The University of Texas at Austin

UT-SysML
Systems & Machine Learning

Why do we need compression?

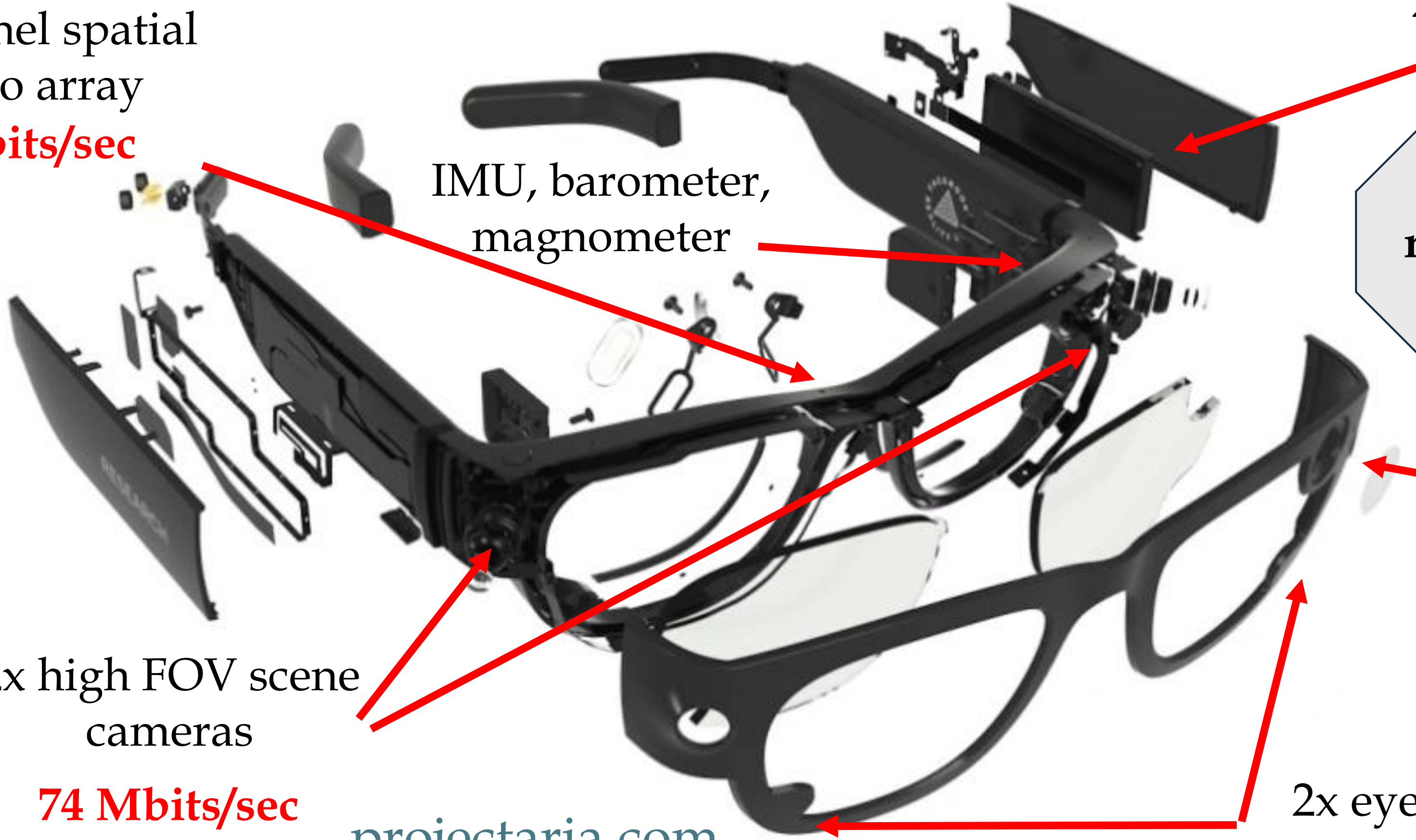
Modern sensor hardware provides incredible power efficiency

7-channel spatial audio array
6 Mbits/sec

IMU, barometer, magnetometer

2.5 Watt-hour battery

30 Hours of recording on a single charge



8MP Front facing camera
200 Mbits/sec

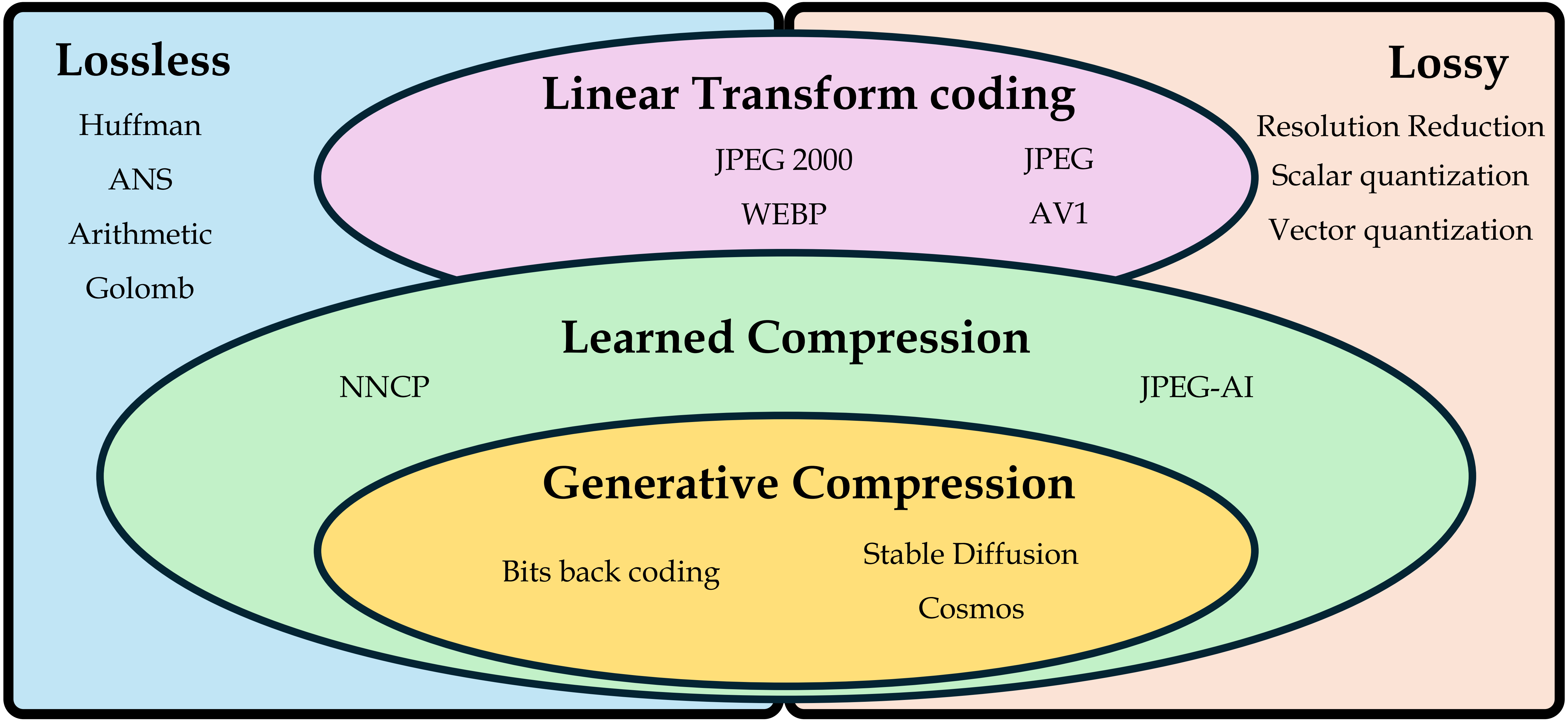
2x high FOV scene cameras
74 Mbits/sec

projectaria.com

2x eye tracking cameras
36 Mbits/sec

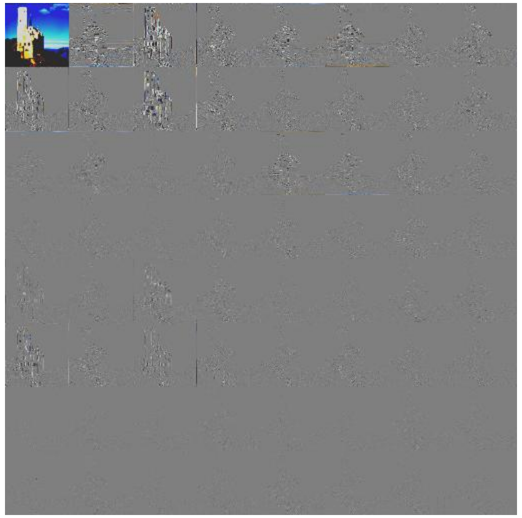
High resolution signals are too large to process on device or transmit to the cloud

Types of compression systems



Linear transform coding

- Most of the bits that move across the internet today use linear transform coding (e.g. JPEG, AV1)
- These codecs use energy compacting transforms (e.g. DCT) to create a sparse representation
- Bits are allocated to different components using models of human perception
- Exploit sparsity via entropy coding (RLE, huffman, etc)

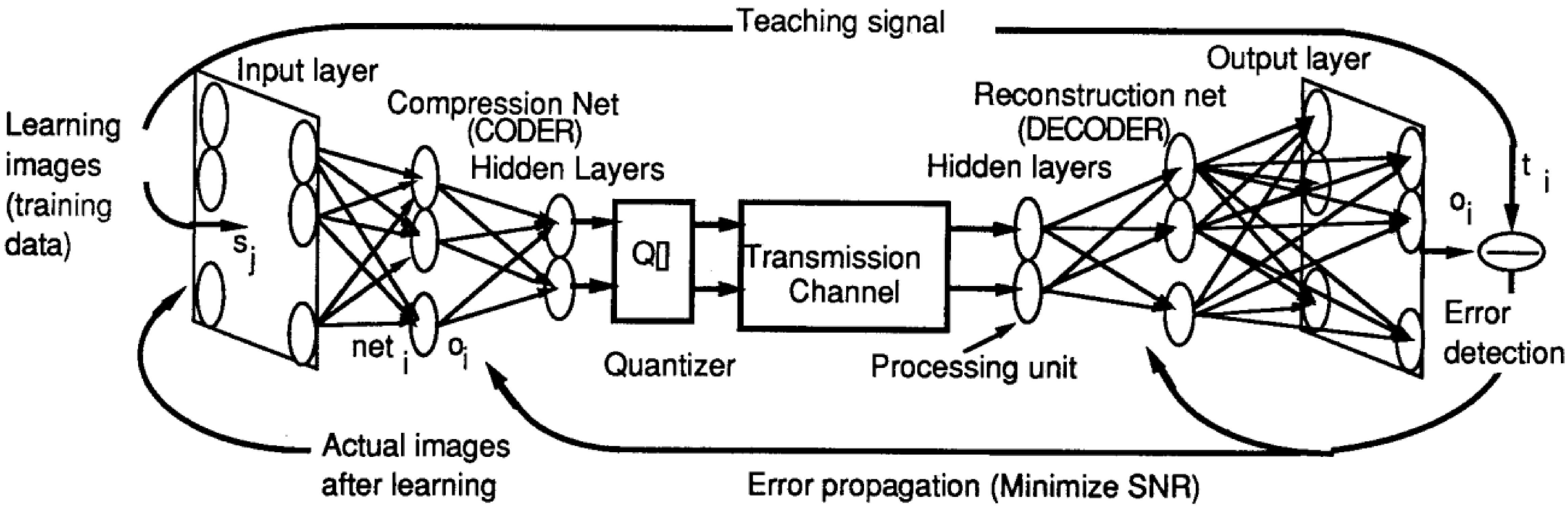


$$\begin{bmatrix} 52 & 55 & 61 & 66 & 70 & 61 & 64 & 73 \\ 63 & 59 & 55 & 90 & 109 & 85 & 69 & 72 \\ 62 & 59 & 68 & 113 & 144 & 104 & 66 & 73 \\ 63 & 58 & 71 & 122 & 154 & 106 & 70 & 69 \\ 67 & 61 & 68 & 104 & 126 & 88 & 68 & 70 \\ 79 & 65 & 60 & 70 & 77 & 68 & 58 & 75 \\ 85 & 71 & 64 & 59 & 55 & 61 & 65 & 83 \\ 87 & 79 & 69 & 68 & 65 & 76 & 78 & 94 \end{bmatrix}$$

$$\begin{bmatrix} -26 & -3 & -6 & 2 & 2 & -1 & 0 & 0 \\ 0 & -2 & -4 & 1 & 1 & 0 & 0 & 0 \\ -3 & 1 & 5 & -1 & -1 & 0 & 0 & 0 \\ -3 & 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

0	-1	1	-3	2	-4	-2	5	-6	-26
1	000	001	010 1	011 0	010 00	010 01	011 10	011 110	011 111

Learned compression using neural networks



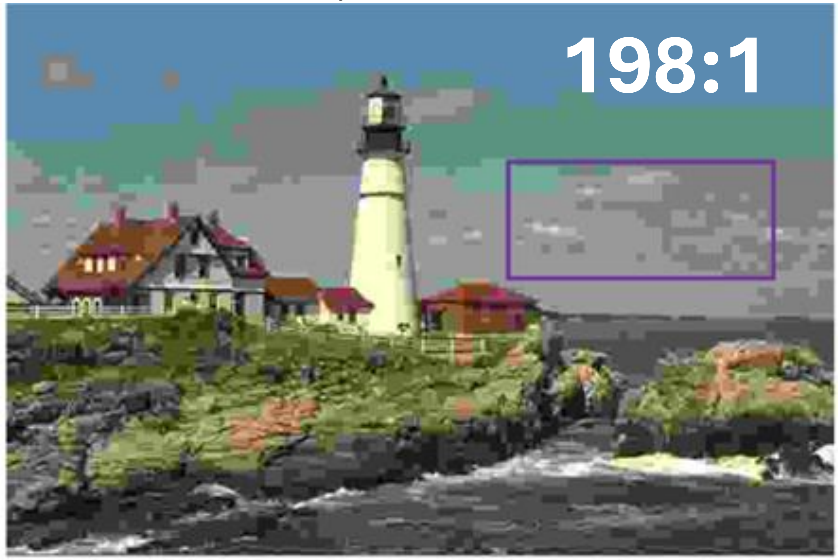
Learn the transform and quantizer from representative data

High compression efficiency **Poor computational efficiency**

Sonehara, et al. "Image data compression using a neural network model.“ *International 1989 Joint Conference on Neural Networks*. IEEE, 1989.

Compression efficiency vs computational efficiency

JPEG

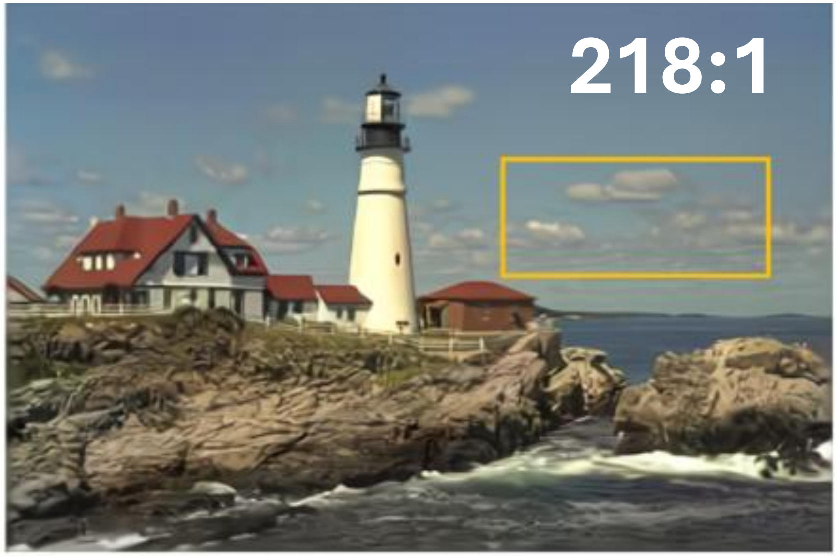


64 parameters

<500 MACs/pixel

Original

Neural Network



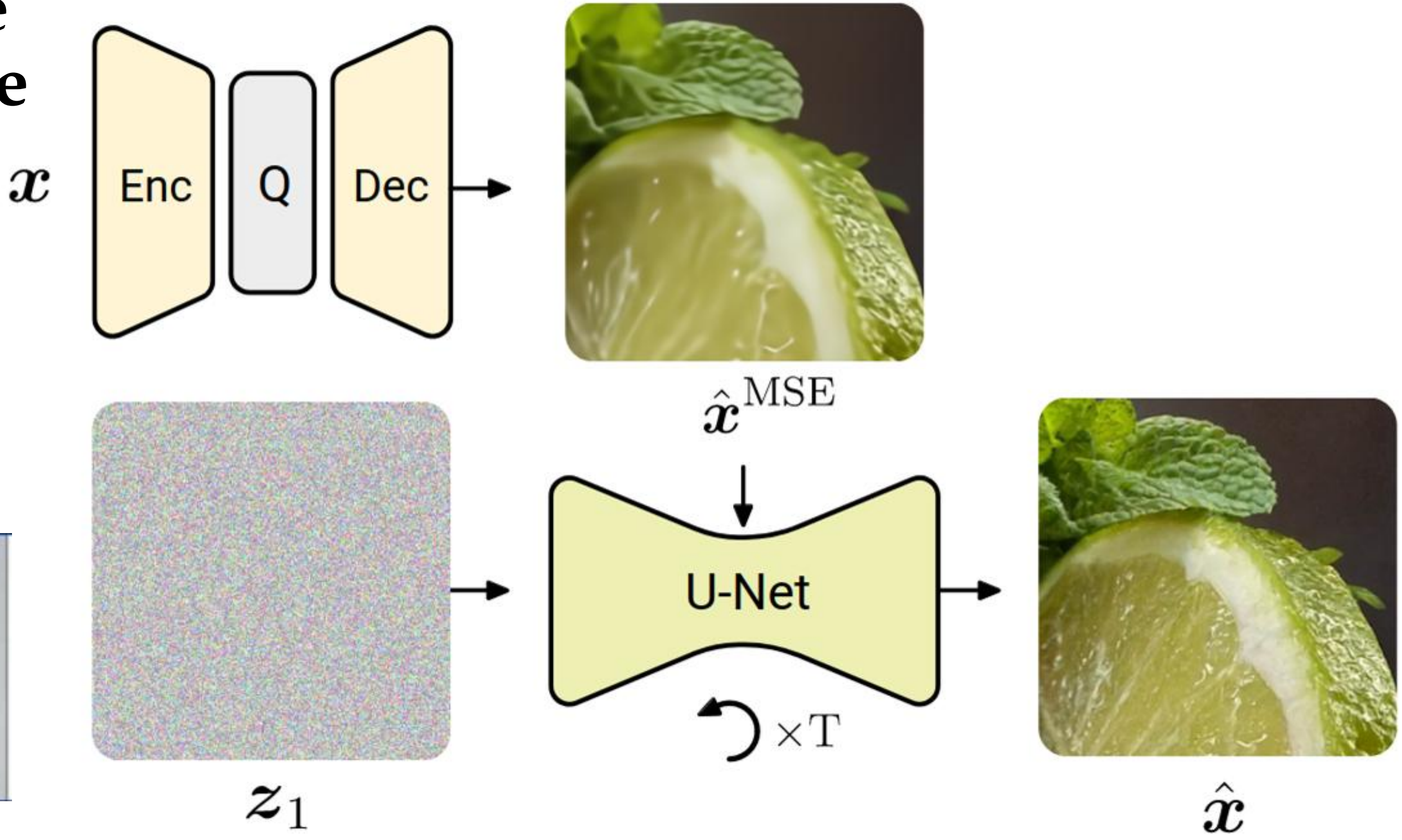
Millions of Parameters

>100k MACs/pixel

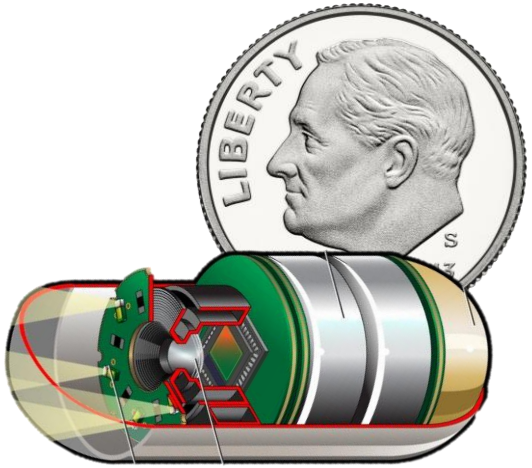
Generative compression

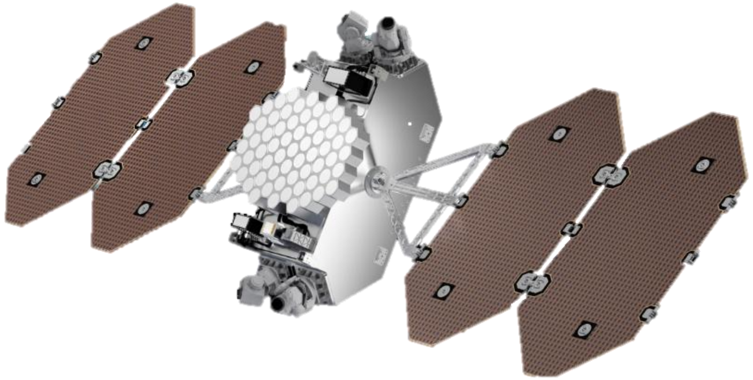
- Autoencoder will struggle to preserve **details**, **texture** and **high frequencies**

- Use a **generative model** to resynthesize the details



Who is the perceiver?





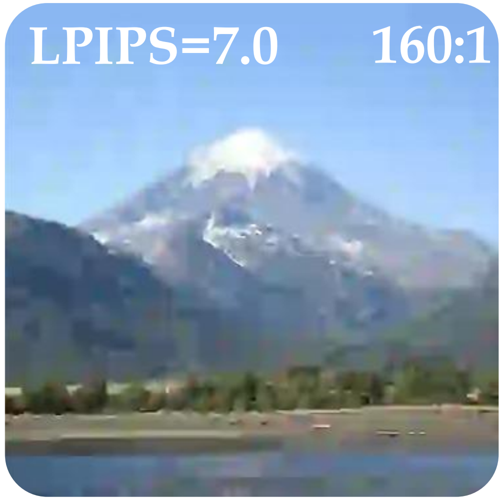
We need **machine-oriented** compression systems

Perceptual quality

Legacy transform codec

Human
<

Modern transform codec



Human
<

MSE Autoencoder

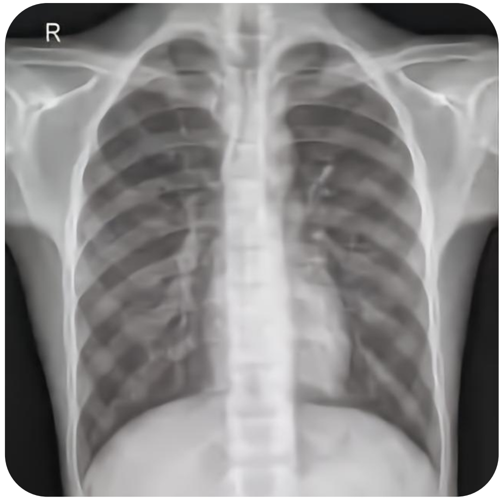
Human
<

Generative Autoencoder

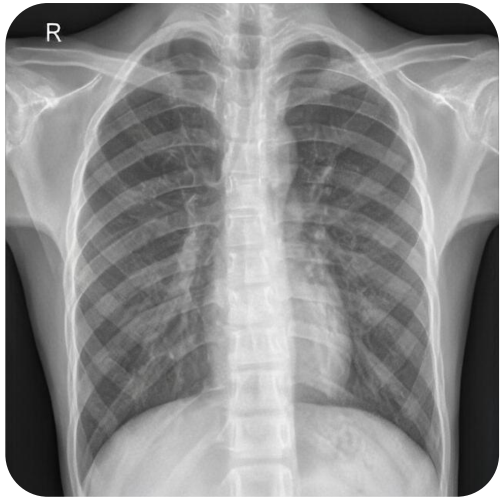
Human perceptual quality is well modeled (e.g. LPIPS)
What about machine perception for specific applications?

?

?



?

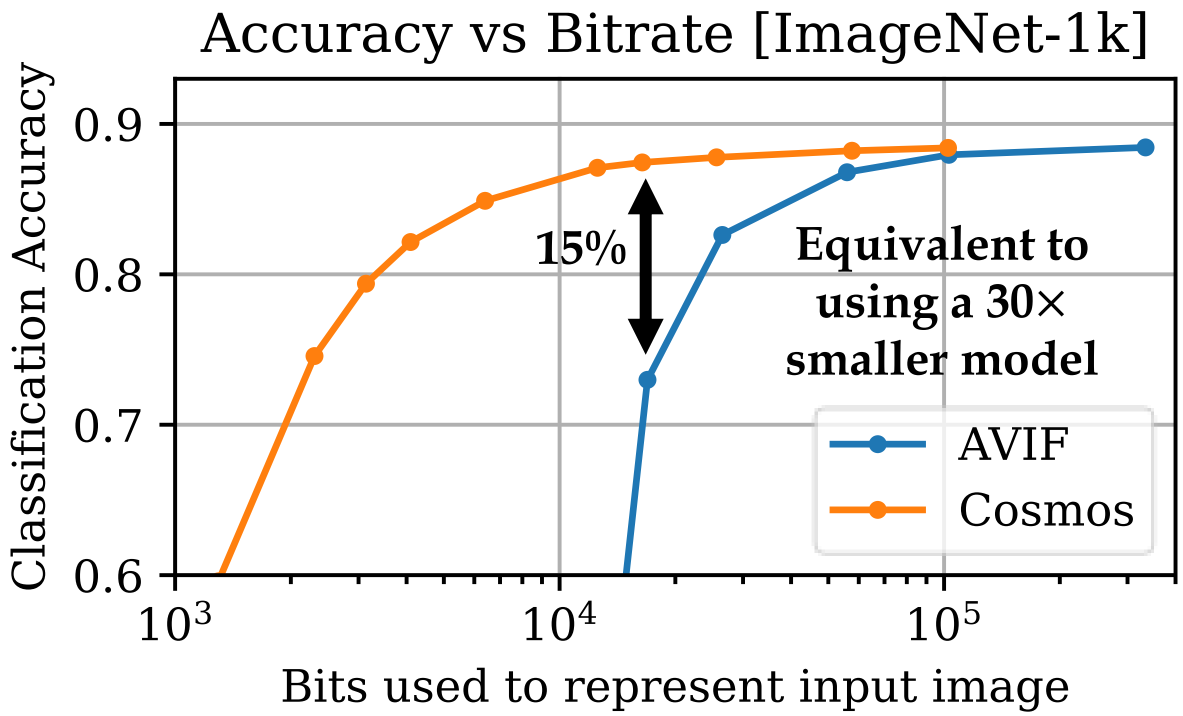


Machine perceptual quality

Original (0.1 MP)	AVIF	Cosmos
		
2.5×10^6 bits	3.1×10^4 bits	2.5×10^4 bits

Does the high human perceptual quality of generative codecs translate to high machine perceptual quality?

Yes; generative codecs often provide **better downstream performance** than conventional methods at **lower rates**



Does lossy compression always hurt accuracy?

How could lossy compression *increase* performance?

Datasets used to pre-train foundation models use legacy JPEG and MPEG compression at default settings

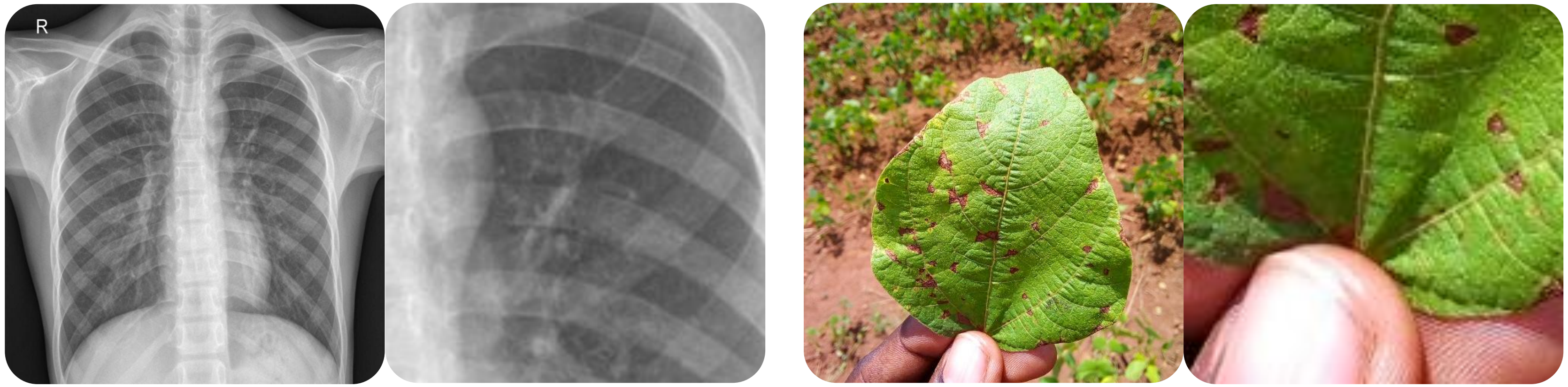
High quality, lossless samples are **out of distribution!**

Legacy transform coding

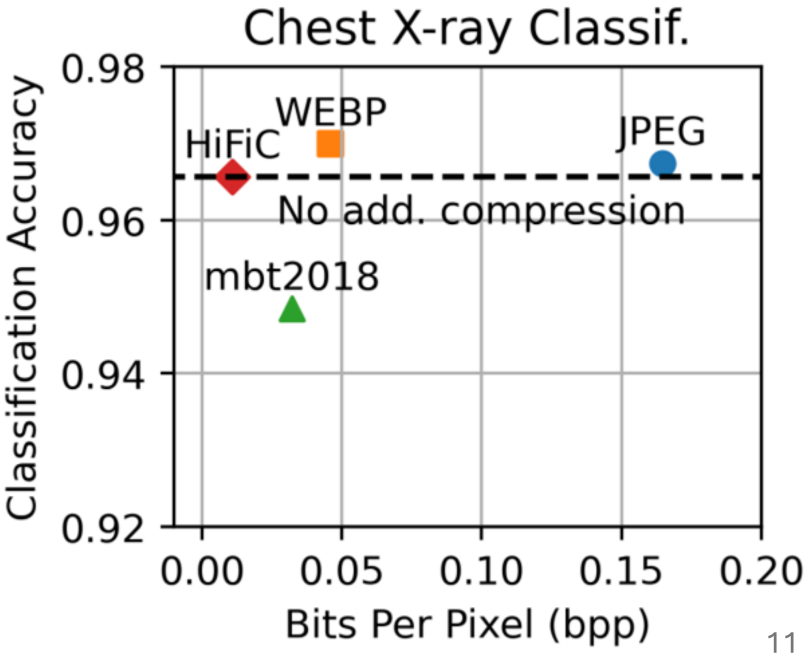
Modern transform coding

MSE-optimized autoencoder

Generative compression

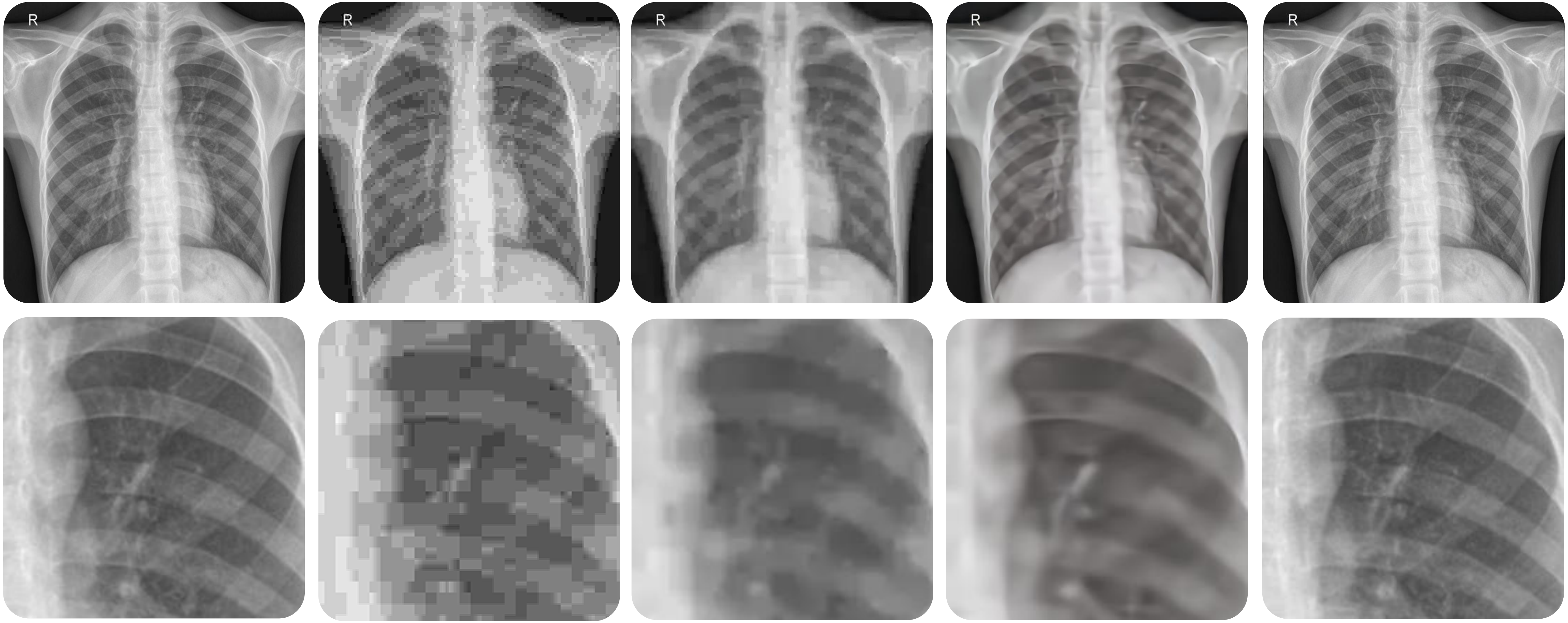


Pristine, high resolution inputs



Denoising effect of lossy compression

Original	Legacy transform coding (JPEG)	Modern transform coding (WEBP)	MSE-optimized Autoencoder	Generative model (HiFiC)



How much power is available for sensing?

Mobile, remote, and wearable sensors produce constant streams of **high resolution** signals

Sensor efficiency is increasing, while ML models get more expensive

Solution: divide computation between sensor and cloud

10^1 10^2 10^4 10^5

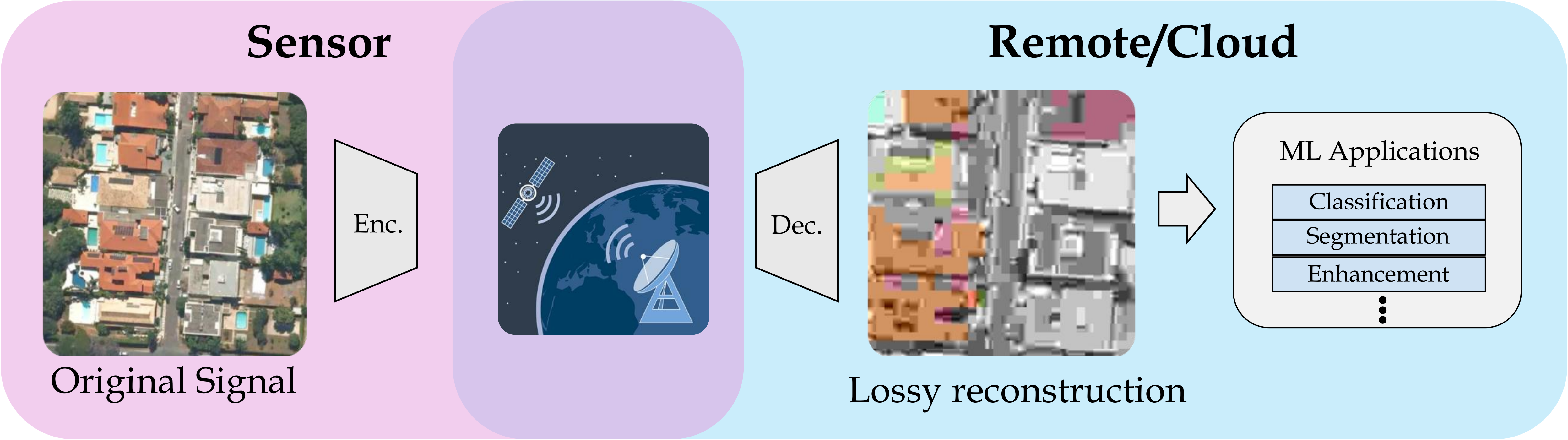
Battey capacity (Watt-hours)

Compression for mobile and remote sensing

Mobile, remote, and wearable sensors produce constant streams of **high resolution** signals

Sensor efficiency is increasing, while ML models get more expensive

Solution: divide computation between sensor and cloud



Demands high compression ratio

Degrades accuracy

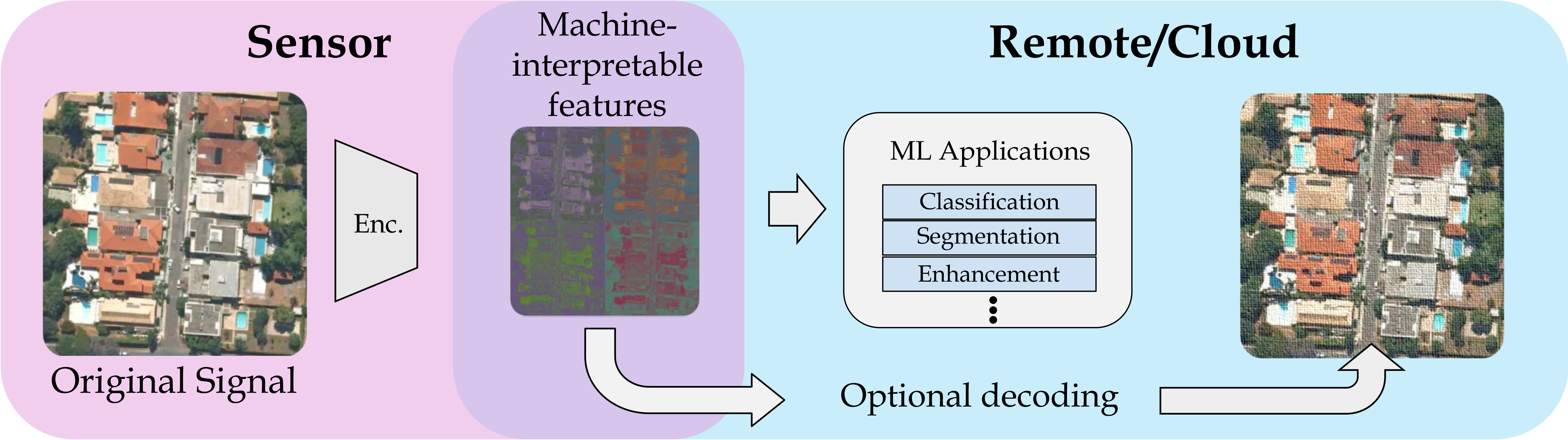
Adds decoding overhead

Machine-oriented compression

Mobile, remote, and wearable sensors produce constant streams of **high resolution** signals

Sensor efficiency is increasing, while ML models get more expensive

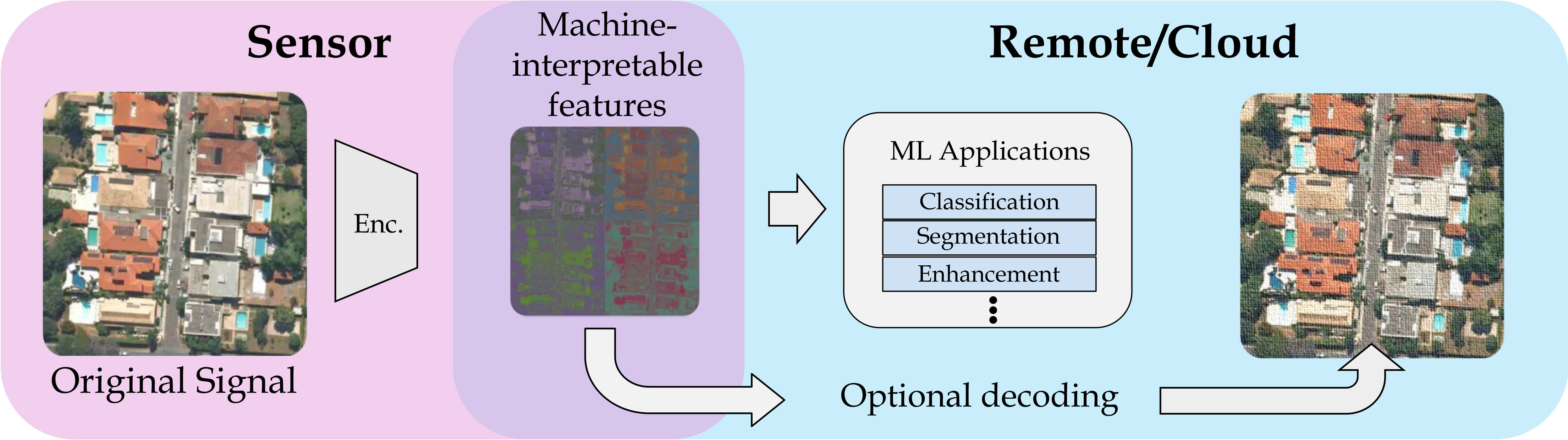
Solution: divide computation between sensor and cloud



Less bandwidth **Enhanced accuracy** **More efficient ML**

Machine-oriented compression

What are ideal characteristics of the compression system?



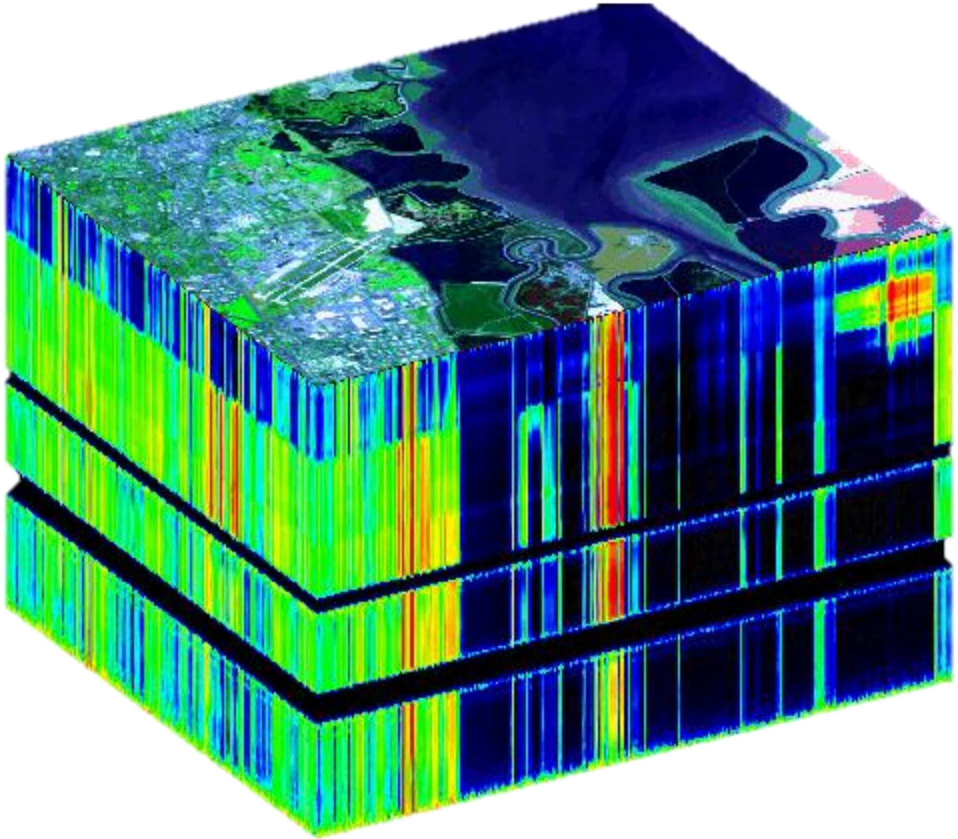
Less bandwidth **Enhanced accuracy** **More efficient ML**

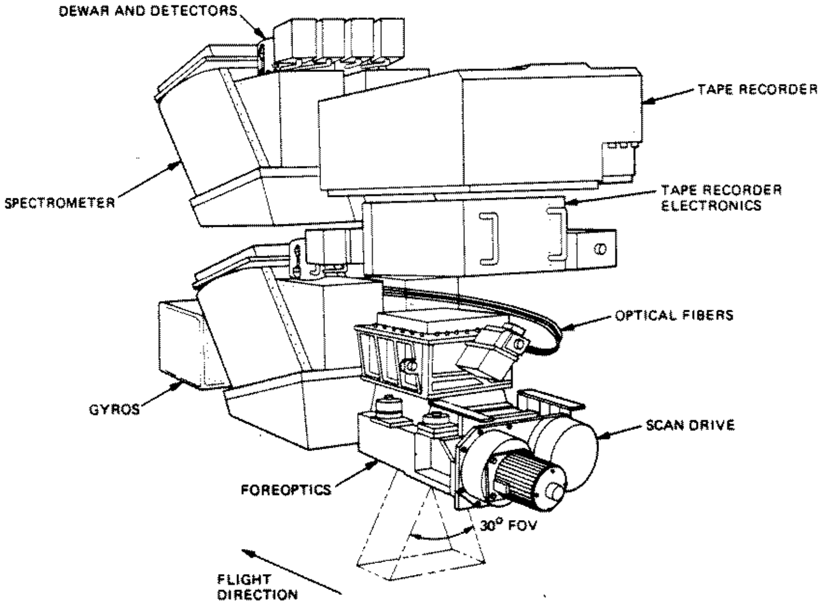
Machine-oriented compression

What are ideal characteristics of the compression system?

- **Support many modalities**

- **Hyperspectral**



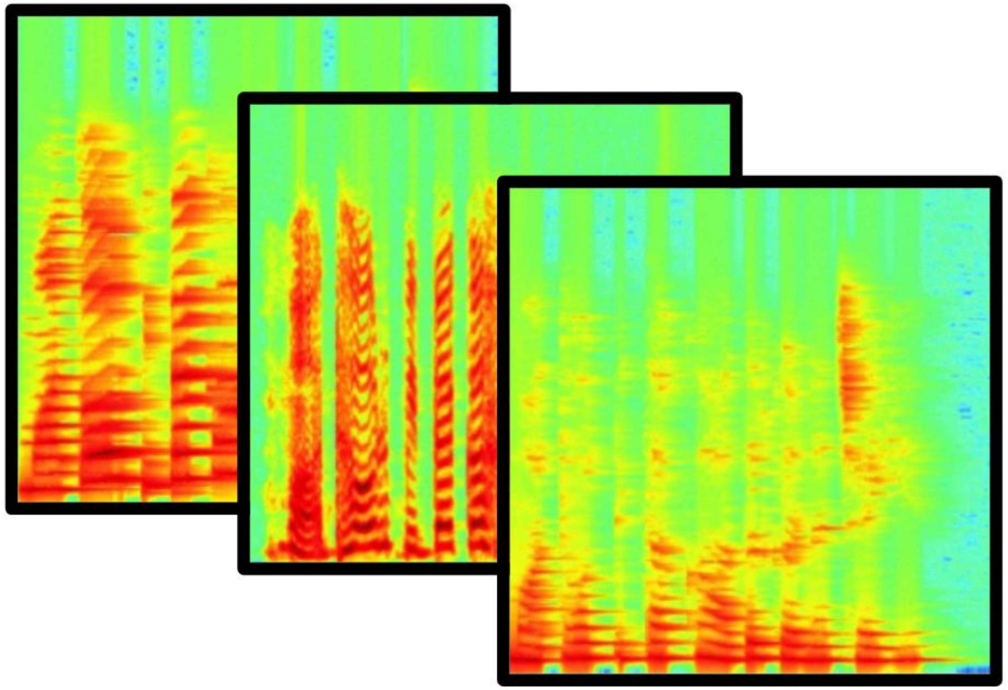


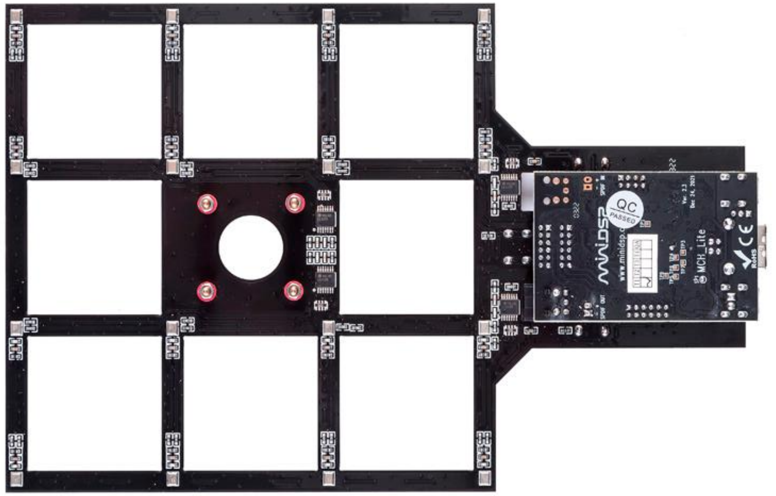
Machine-oriented compression

What are ideal characteristics of the compression system?

- **Support many modalities**

- Hyperspectral
- **Spatial Audio**



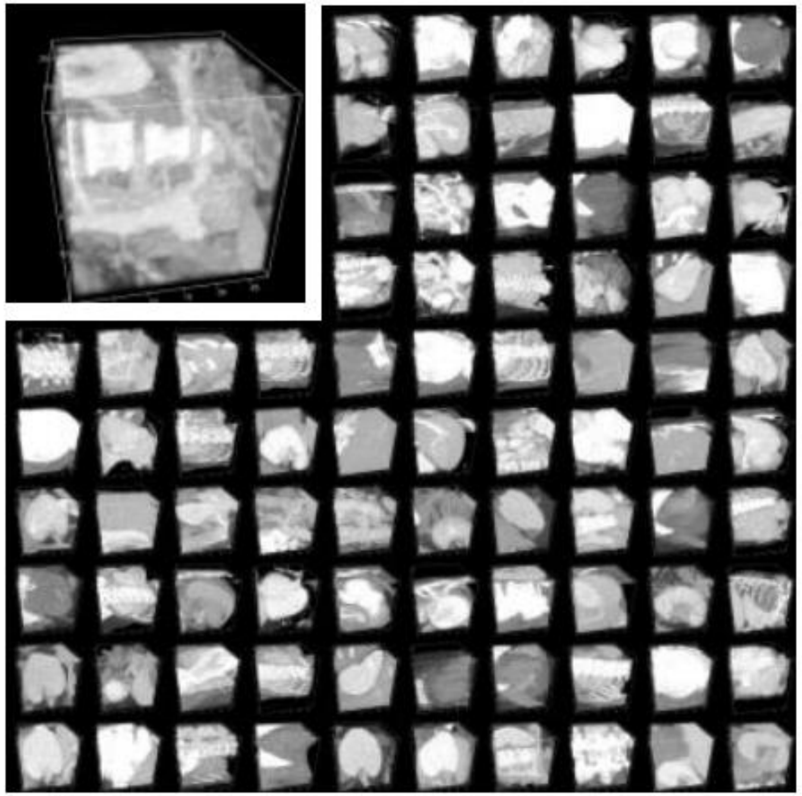


Machine-oriented compression

What are ideal characteristics of the compression system?

- **Support many modalities**

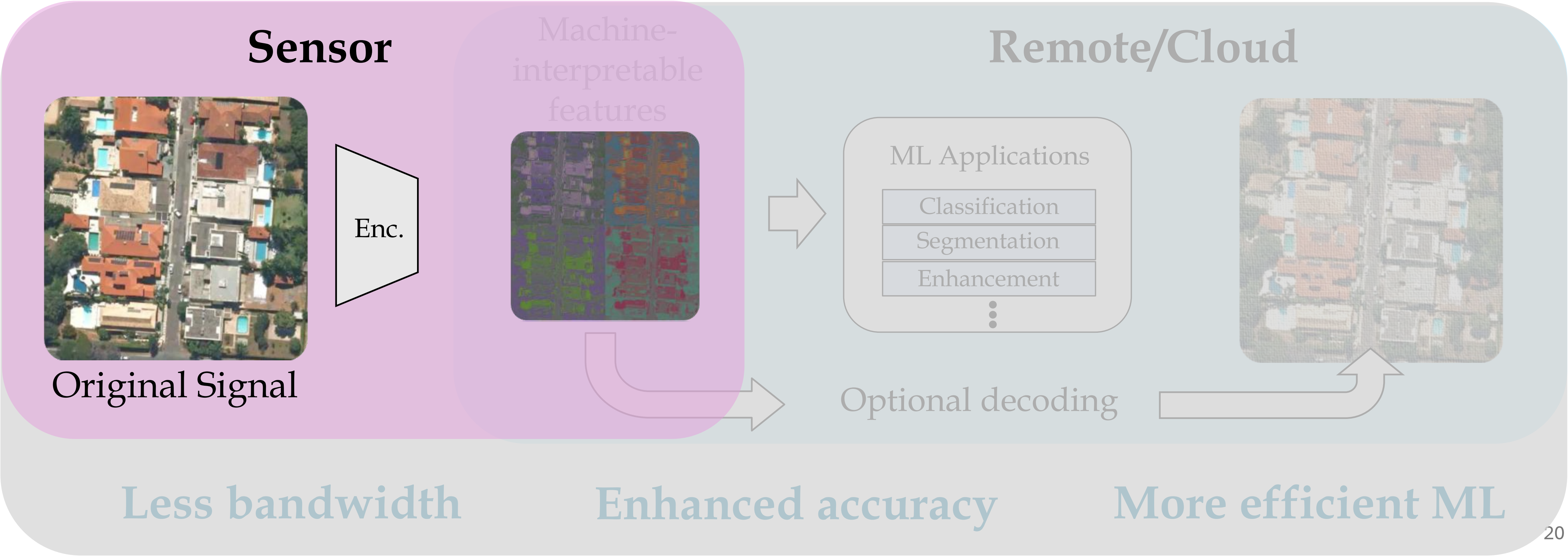
- Hyperspectral
- Spatial Audio
- **3D volumes, medical images**



Machine-oriented compression

What are ideal characteristics of the compression system?

- Support many modalities
- **Allow efficient encoding**



Machine-oriented compression

What are ideal characteristics of the compression system?

- Support many modalities
- Allow efficient encoding
- **Preserve details**

Generative models synthesize details
For recognition, we must **preserve** details

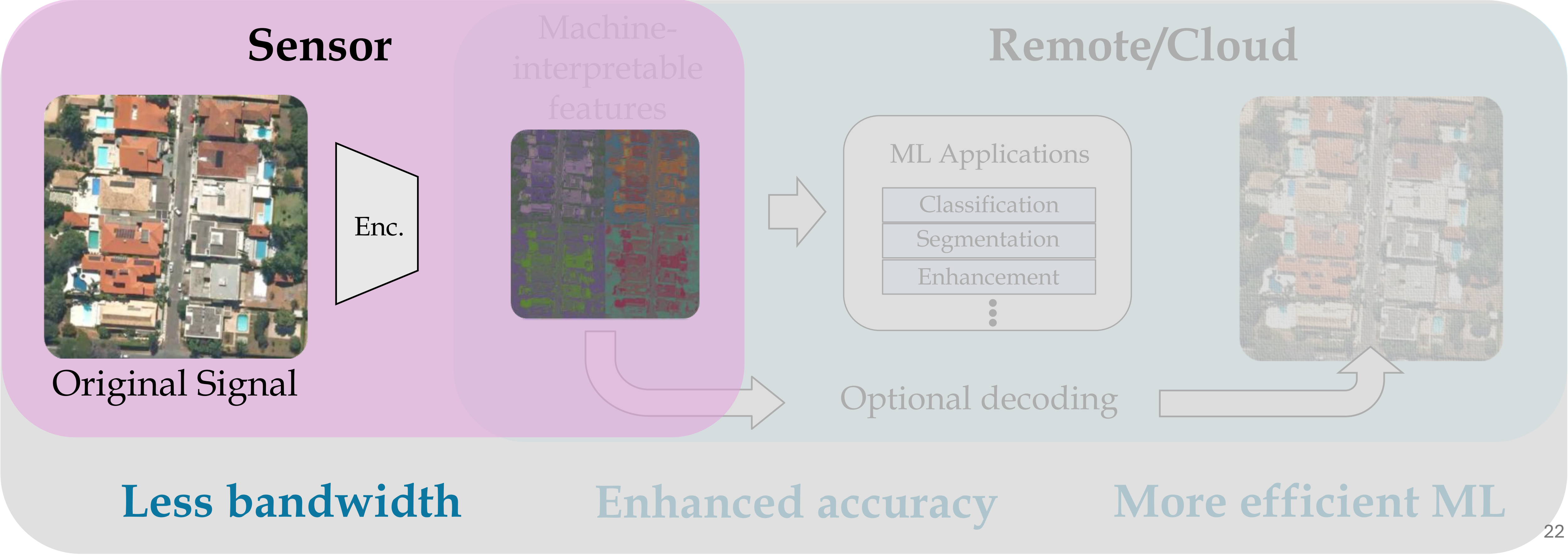
Original

Stable Diff. VAE

Machine-oriented compression

What are ideal characteristics of the compression system?

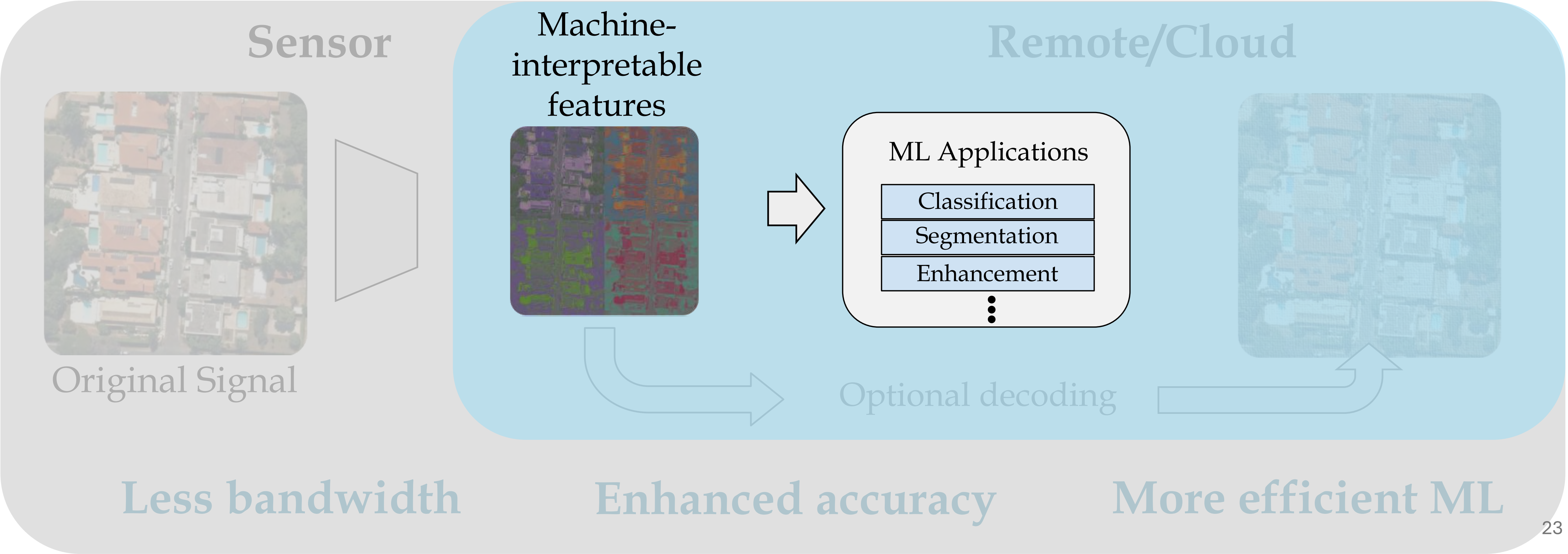
- Support many modalities
- Allow efficient encoding
- Preserve details
- **Achieve high compression rate**



Machine-oriented compression

What are ideal characteristics of the compression system?

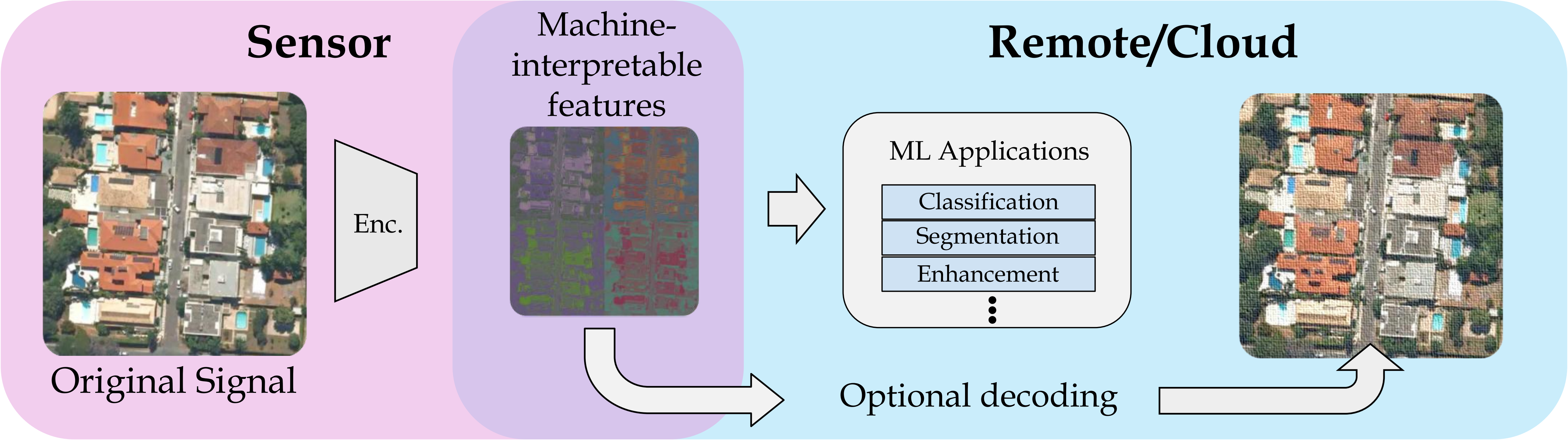
- Support many modalities
- Allow efficient encoding
- Preserve details
- Achieve high compression rate
- **Accelerate downstream ML models**



Machine-oriented compression

What are ideal characteristics of the compression system?

- Support many modalities
- Allow efficient encoding
- Preserve details
- Achieve high compression rate
- Accelerate downstream ML models



Less bandwidth **Enhanced accuracy** **More efficient ML**

Comparison of existing codec designs

Resample

WEBP

DGML (Cheng2020)

Stable Diff. VAE

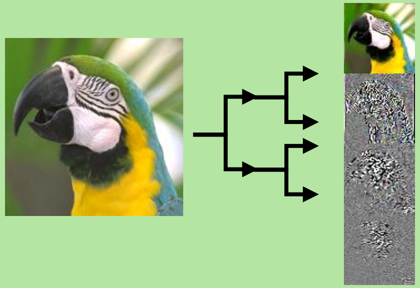
	RR	LTC	E2ELC	GenAE	Goal
Allow efficient encoding	✅	✅	❌	❌	✅
Accelerate downstream ML	✅	❌	❌	✅	✅
Achieve high compression rate	❌	✅	✅	❌	✅
Preserve details	❌	✅	✅	❌	✅
Support many modalities	✅	❌	✅	❌	✅

Proposed design

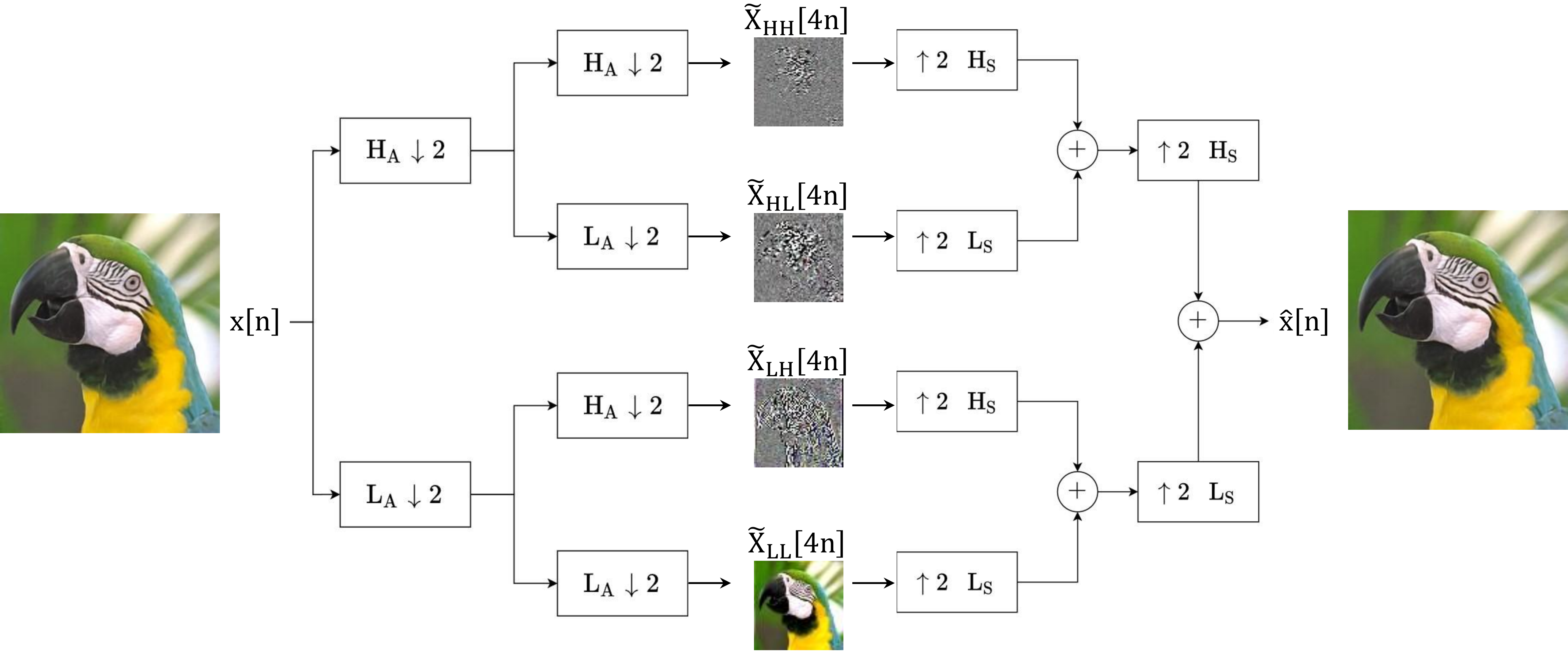
Encoding efficiency

Inspired by linear transform coding

Forgo expensive DNN-based analysis transform; leverage efficient, separable transform for energy compaction instead (wavelet packet decomposition)



Wavelet packet transform



WPT exchanges spatial resolution with channels

No information loss

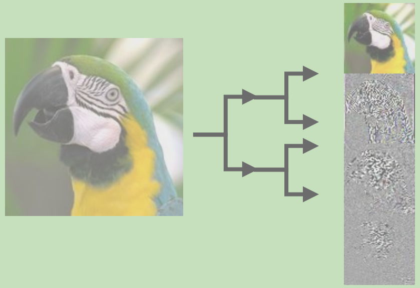
Energy compaction

Proposed design

Encoding efficiency

Inspired by linear transform coding

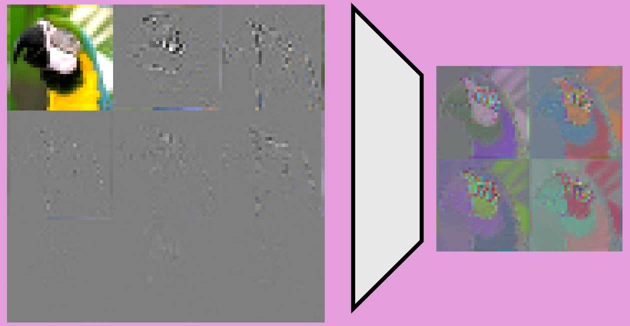
Forgo expensive DNN-based analysis transform; leverage efficient, separable transform for energy compaction instead (wavelet packet decomposition)



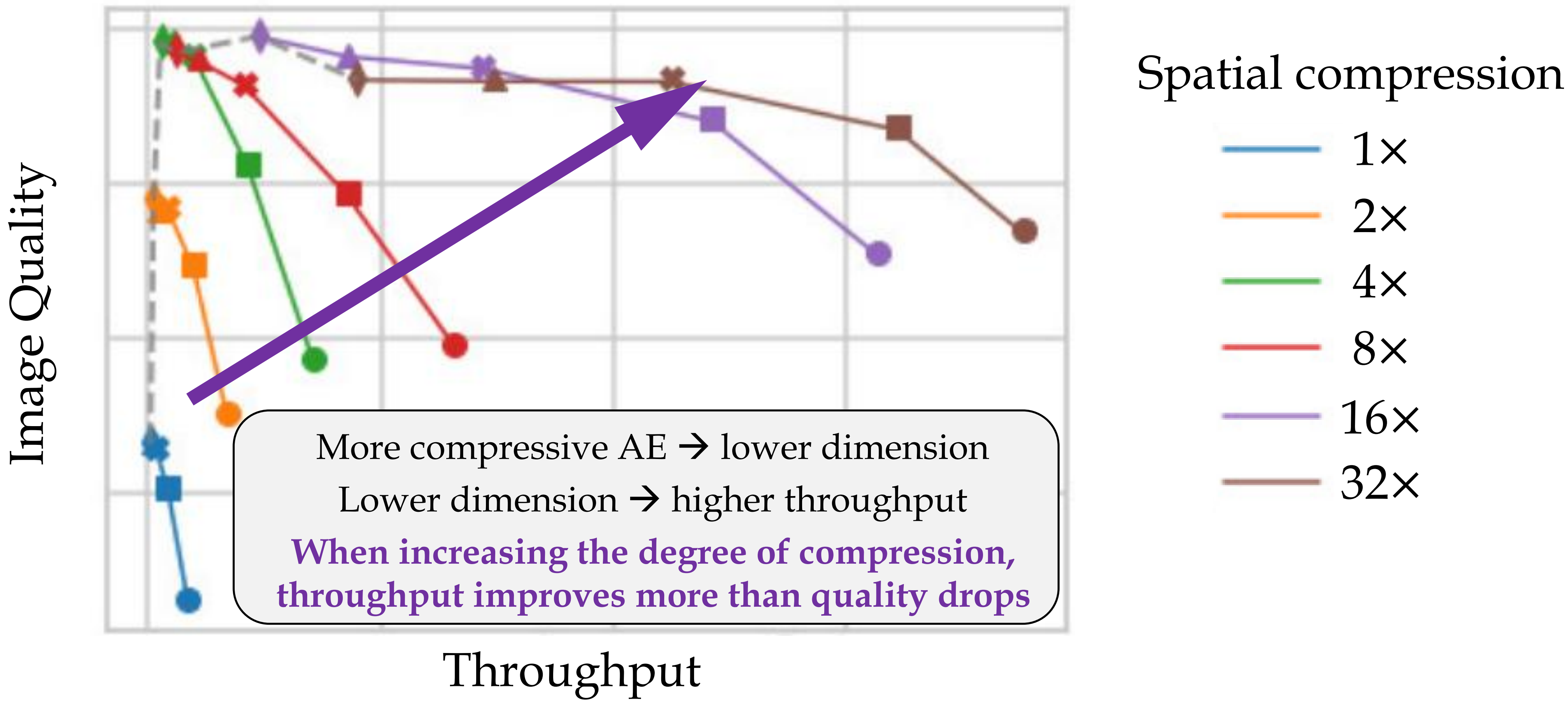
Dimension reduction

Inspired by generative autoencoders

Don’t rely exclusively on sparsity; use channel bottleneck to provide guaranteed, uniform dimensionality reduction to accelerate downstream models

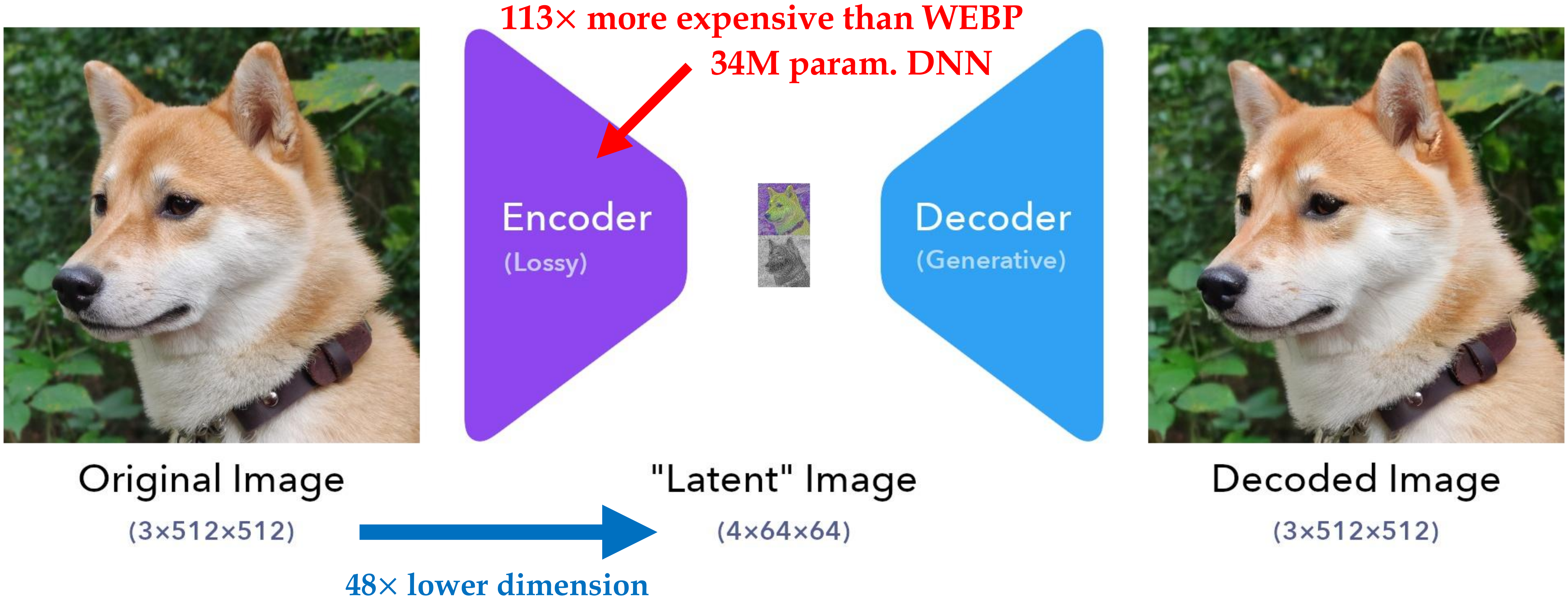


Autoencoder for dimension reduction



"High-Resolution Image Synthesis with Latent Diffusion Models"
(aka "Stable Diffusion") Rombach et al. 2021

Autoencoder for dimension reduction



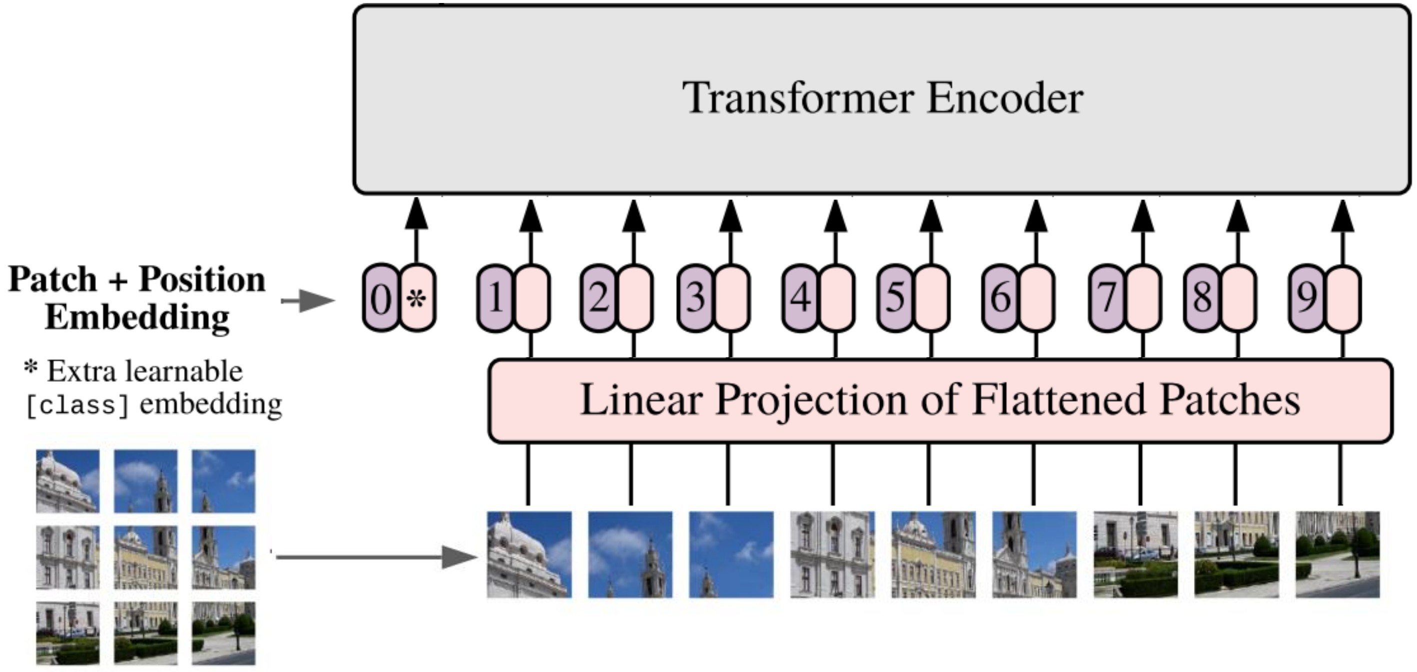
“High-Resolution Image Synthesis with Latent Diffusion Models”
(aka “Stable Diffusion”) Rombach et al. 2021

Does the encoder need to be so expensive?

Synthesizing details is hard

Discarding details is easy

→Use a simple encoder (e.g. linear projection)



"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (aka "ViT") Beyer et al. 2021

ViT-B/16

Patch size	3×16×16
Sequence Len	196
Embedding Dim	768
Compression	1:1
Accuracy	86.1

ViT-B/32

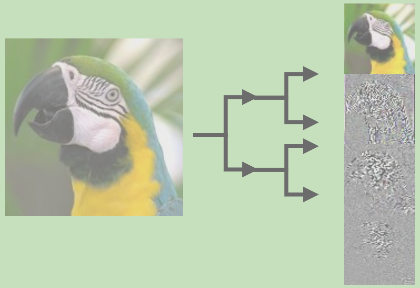
Patch size	3×32×32
Sequence Len.	49
Embedding Dim	768
Compression	4:1
Accuracy	83.3

Proposed design

Encoding efficiency

Inspired by linear transform coding

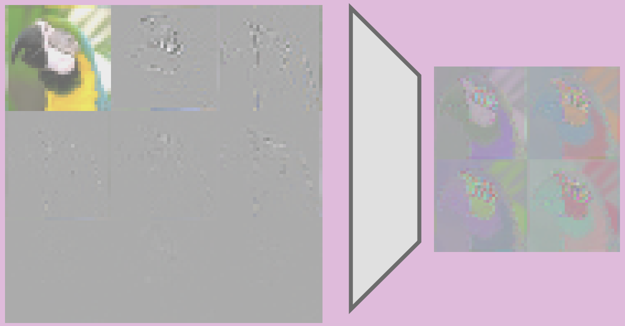
Forgo expensive DNN-based analysis transform; leverage efficient, separable transform for energy compaction instead (wavelet packet decomposition)



Dimension reduction

Inspired by generative AEs

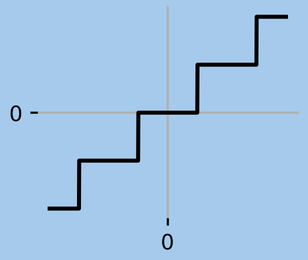
Don't rely exclusively on sparsity; use channel bottleneck to provide guaranteed, uniform dimensionality reduction to accelerate downstream models



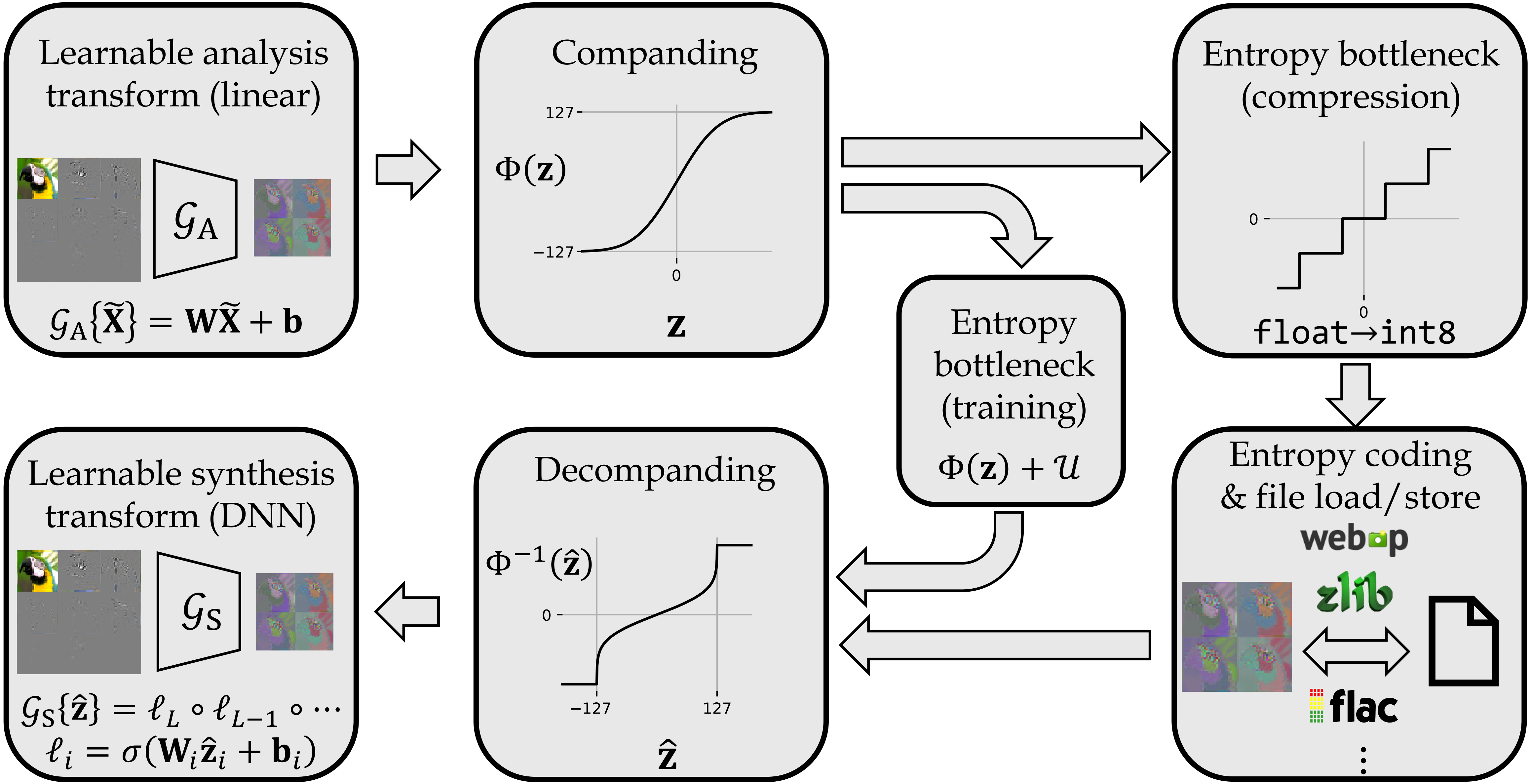
Compression ratio

Inspired by E2E learned compression

Guarantee resilience to quantization via additive noise during training. Leverage existing lossless codecs as a compression multiplier.



E2E learned compression: quantization and entropy coding

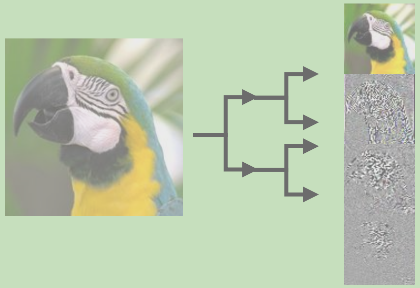


Proposed design

Encoding efficiency

Inspired by linear transform coding

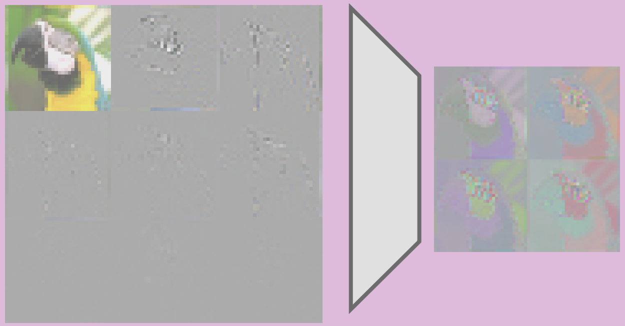
Forgo e
analysi
efficien
energy compaction instead
(wavelet packet decomposition)



Dimension reduction

Inspired by generative AEs

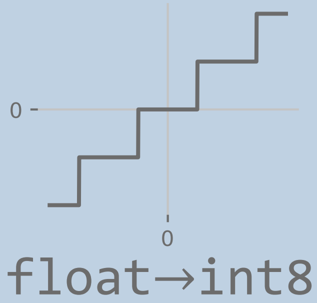
dimensionality reduction to
accelerate downstream models



Compression ratio

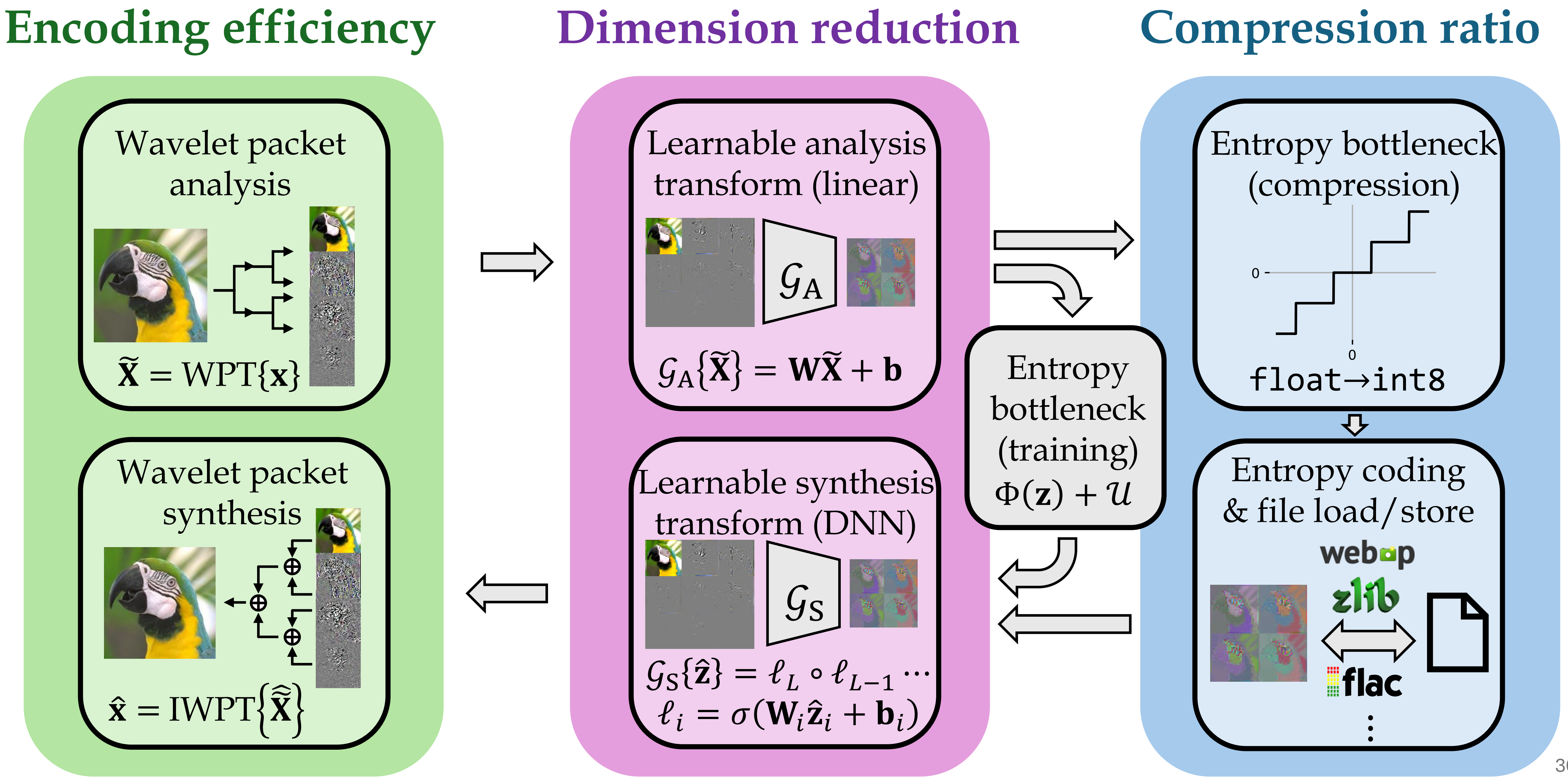
Inspired by E2E learned compression

e noise
ge
existing lossless codecs as a
compression multiplier.



WaLLoC: Wavelet Learned Lossy Compression

WaLLoC workflow



How to avoid the pitfalls of generative autoencoders?

Resample | WEBP | DGML (Cheng2020) | Stable Diff. VAE

	RR	LTC	E2ELC	GenAE	Goal
Allow efficient encoding	✅	✅	❌	❌	✅
Accelerate downstream ML	✅	❌	❌	✅	✅
Achieve high compression rate	❌	✅	✅	❌	✅
Preserve details	❌	✅	✅	❌	✅
Support many modalities	✅	❌	✅	❌	✅

Loss function

$$\mathcal{L}(x,\hat{x}) = \cancel{\underbrace{\mathrm{MSE}(\mathrm{LPF}\{x\}, \mathrm{LPF}\{\hat{x}\})}_{\substack{\text{Pooled MSE} \\ \text{(does not penalize high frequencies)}}}} + \cancel{\underbrace{\mathcal{L}_{\mathrm{LPIPS}}(x,\hat{x})}_{\substack{\text{Lerned perceptual} \\ \text{patch similarity}}}} + \cancel{\underbrace{\mathcal{L}_{\mathrm{GAN}}(x,\hat{x})}_{\substack{\text{Adversarial loss using} \\ \text{VGG16 discriminator}}}}$$

Only preserves low frequency details

Requires pre-trained models specific to RGB images

$$\mathcal{L}(x,\hat{x}) = \mathrm{MSE}(x,\hat{x})$$

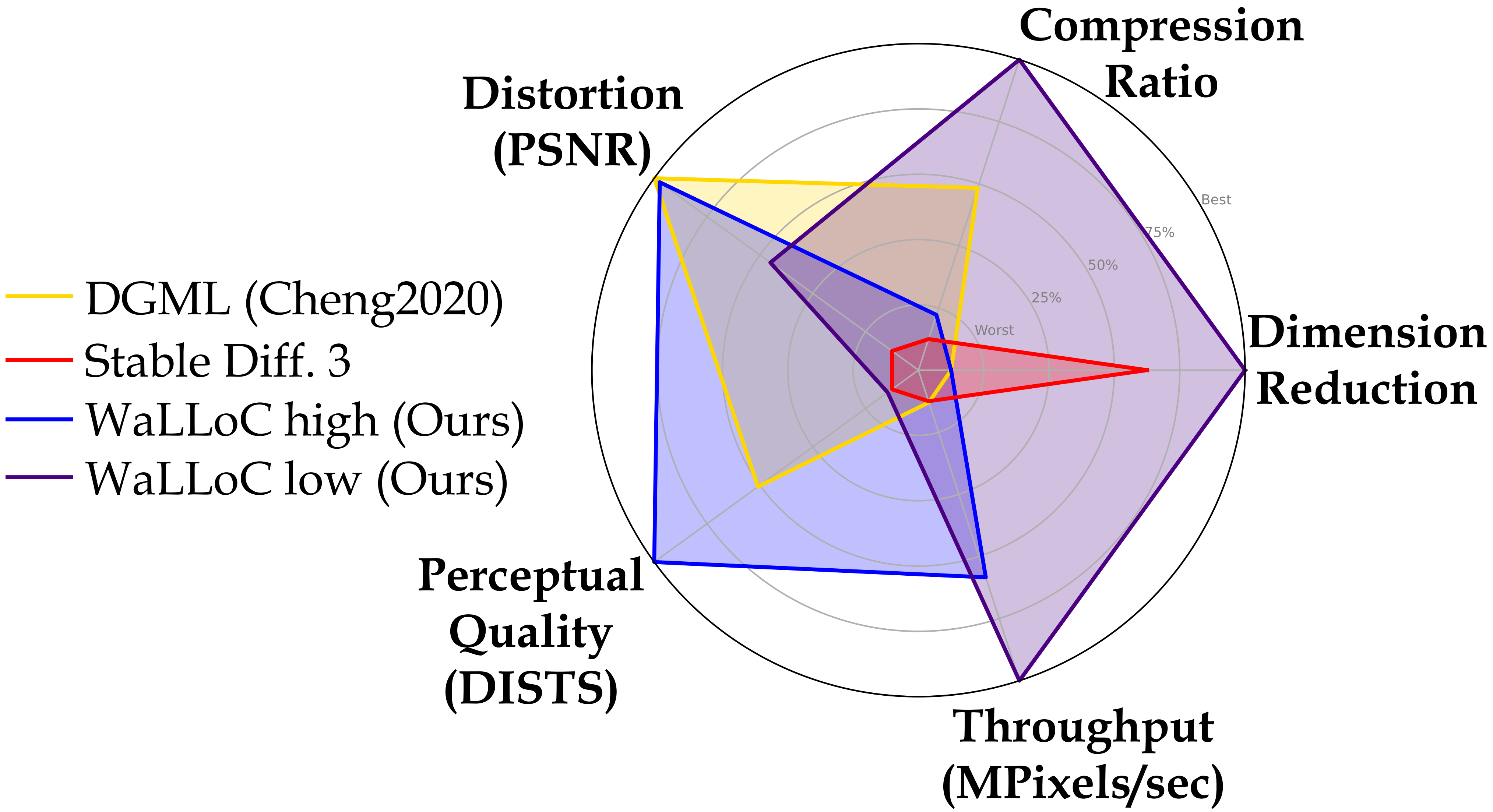
$$\hat{x} = \mathrm{decode}(\mathrm{encode}(x) + \mathcal{U})$$

Better preservation of high frequency details

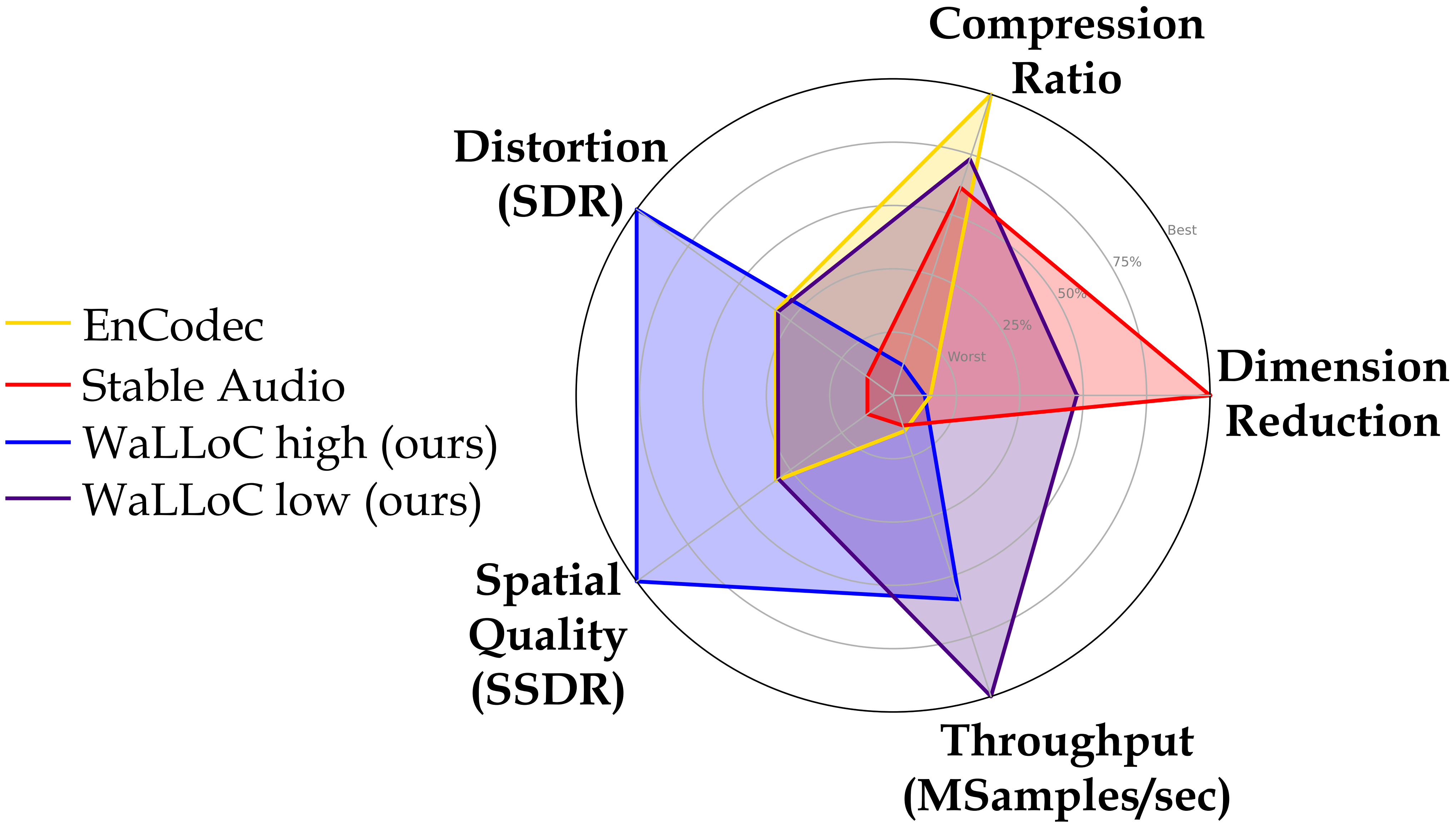
Supports a wide range of modalities

“Training VQGAN and VAE, with detailed explanation”
S. Ryu, 2024. github.com/cloneofsimo/vqgan-training

Comparison of autoencoder designs (RGB image)



Comparison of autoencoder designs (stereo audio)



How does it perform on downstream applications?

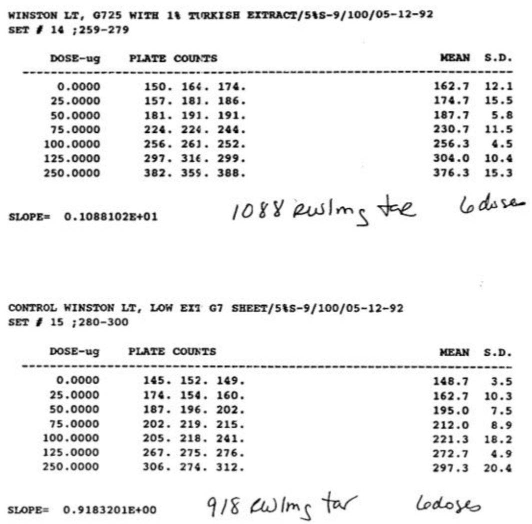
Image Classification

Cat

Document Understanding

Q: What is the date mentioned in the second table?

A: [“05-12-92”]



WINSTON LT, G725 WITH 1% TURKISH EXTRACT/5%S-9/100/05-12-92
SET # 14 ;259-279

DOSE-ug	PLATE COUNTS	MEAN	S.D.
0.0000	150. 164. 174.	162.7	12.1
25.0000	157. 181. 186.	174.7	15.5
50.0000	181. 191. 191.	187.7	5.8
75.0000	224. 224. 244.	230.7	11.5
100.0000	256. 261. 252.	256.3	4.5
125.0000	297. 316. 299.	304.0	10.4
250.0000	382. 355. 388.	376.3	15.3

SLOPE= 0.1088102E+01 1088 revlmg tar 6dose

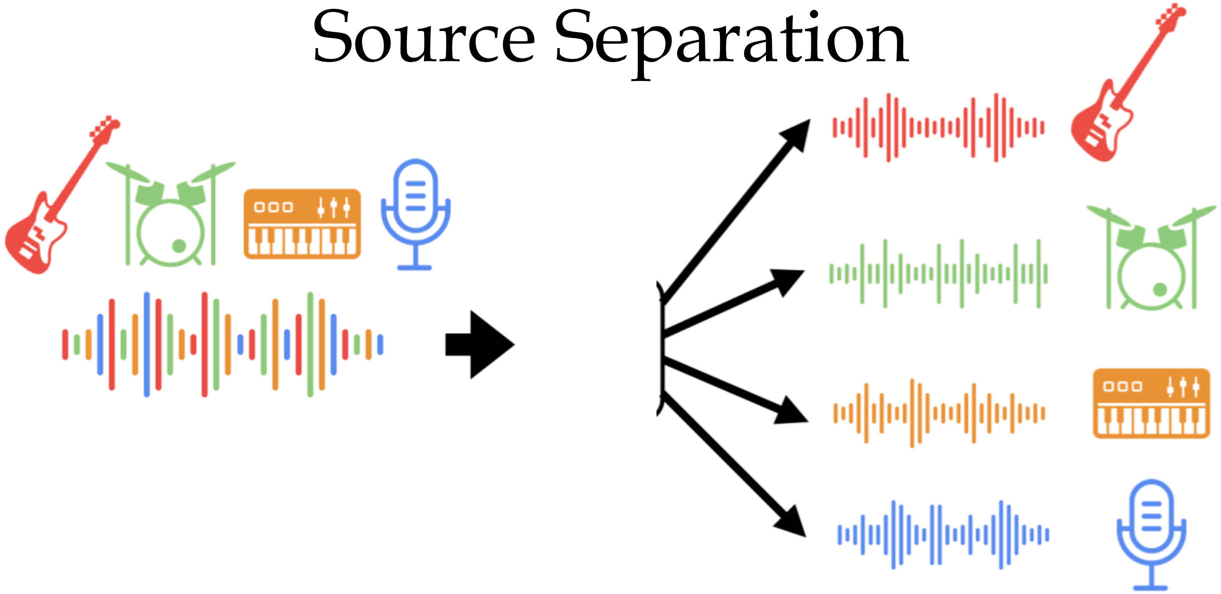
CONTROL WINSTON LT, LOW EXT G7 SHEET/5%S-9/100/05-12-92
SET # 15 ;280-300

DOSE-ug	PLATE COUNTS	MEAN	S.D.
0.0000	145. 152. 149.	148.7	3.5
25.0000	174. 154. 160.	162.7	10.3
50.0000	187. 196. 202.	195.0	7.5
75.0000	202. 219. 215.	212.0	8.9
100.0000	205. 218. 241.	221.3	18.2
125.0000	267. 275. 276.	272.7	4.9
250.0000	306. 274. 312.	297.3	20.4

SLOPE= 0.9183201E+00 918 revlmg tar 6doses

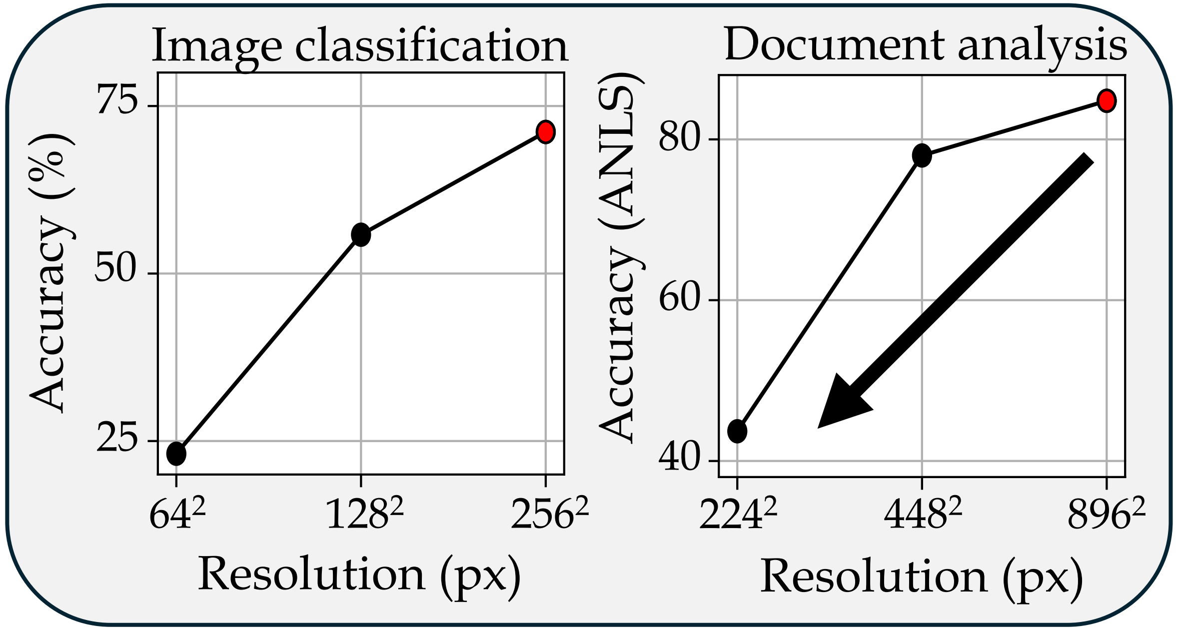
Colorization

Source Separation



Comparison vs. resolution reduction

Reduced resolution & reduced computation



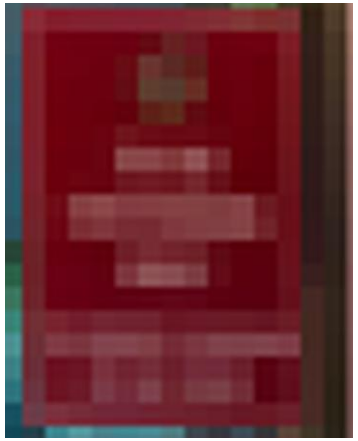
4× lower latency

21GB→8GB GPU Mem

85% → 44% Accuracy

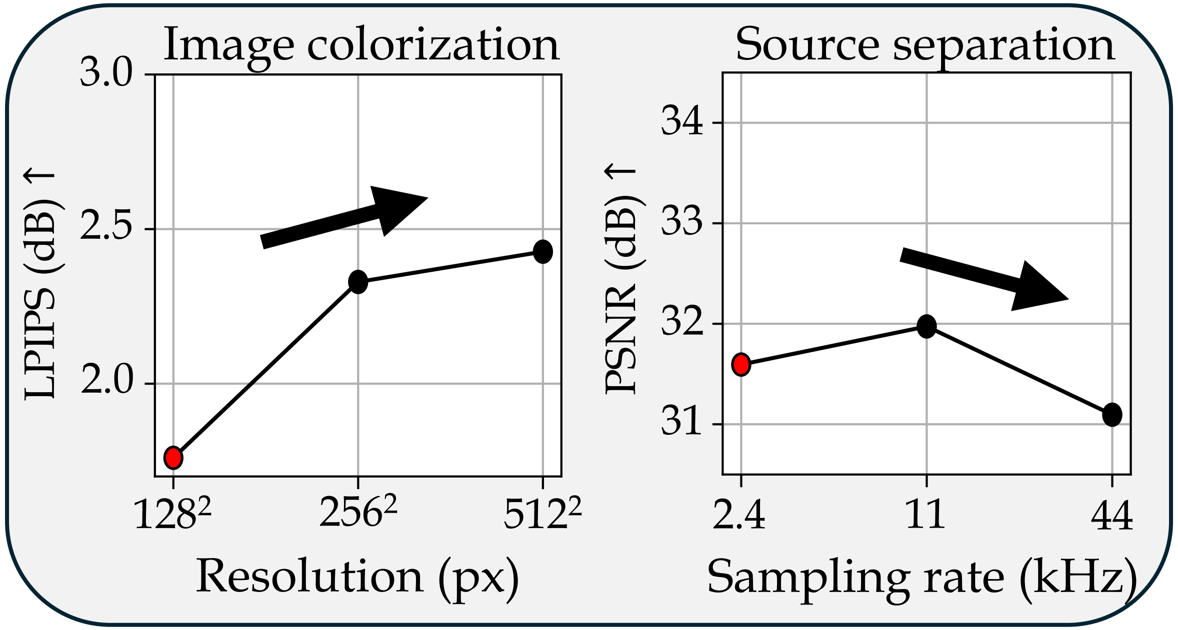
Baseline

Resample



Comparison vs. resolution reduction

Increased resolution & fixed computation



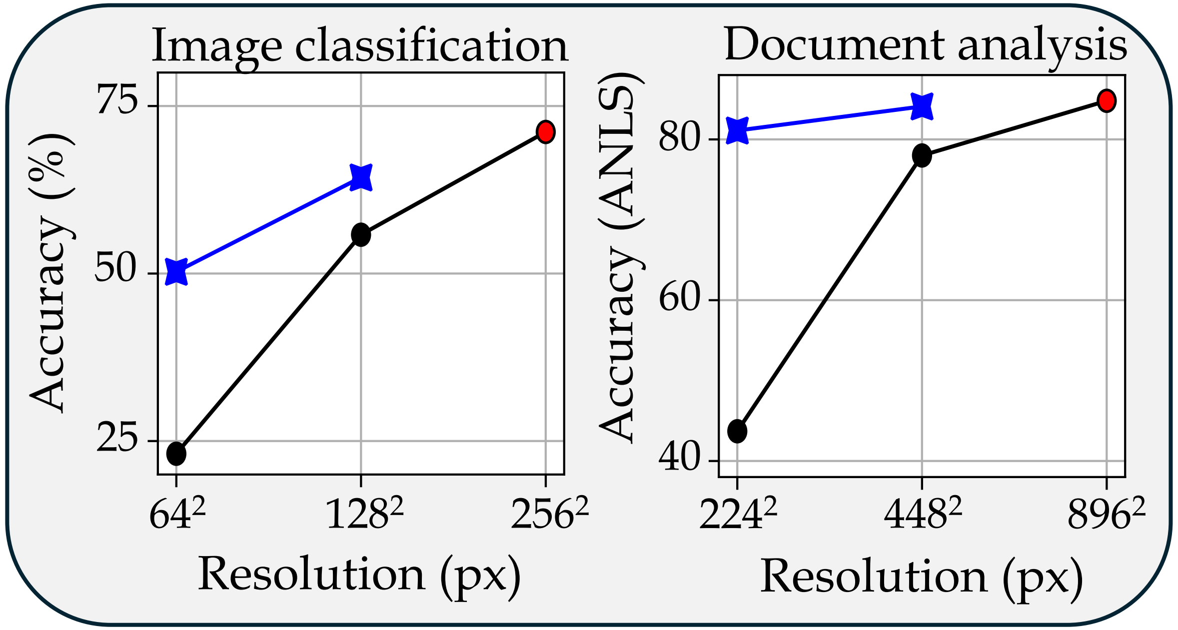
Diminishing return of larger patches / filters

Baseline

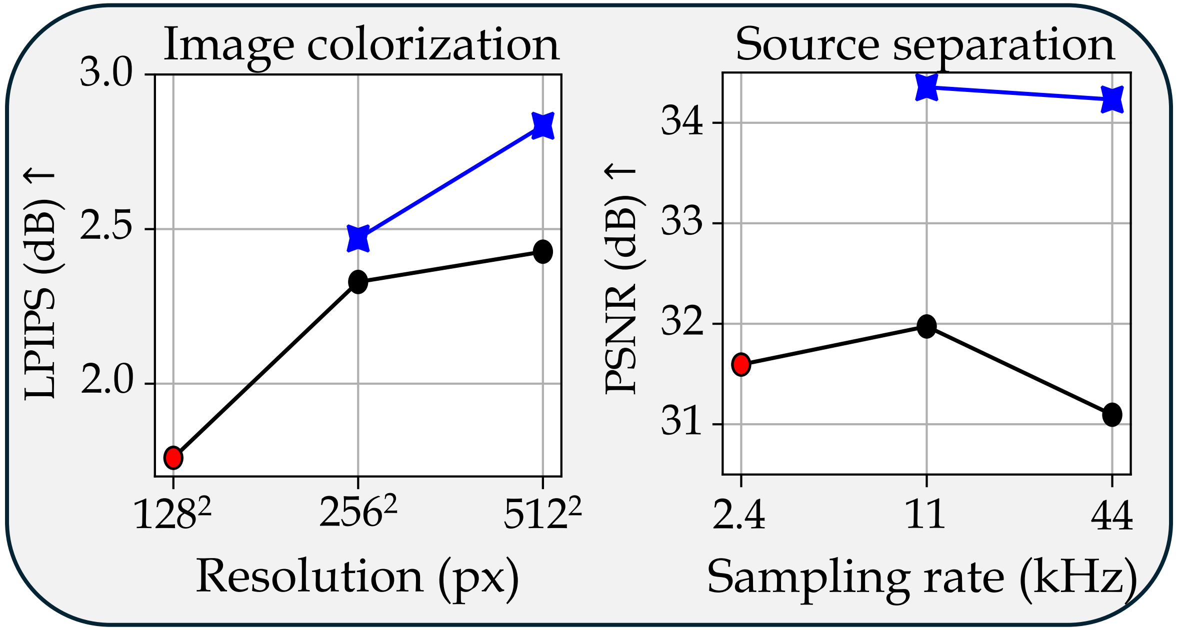
Larger patches

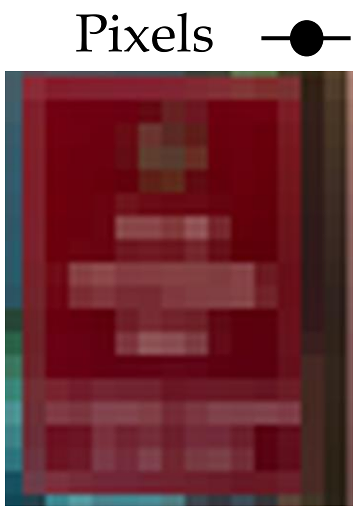
Comparison vs. resolution reduction

Reduced resolution & reduced computation



Increased resolution & fixed computation





Visual Comparison

786 KB

6 KB

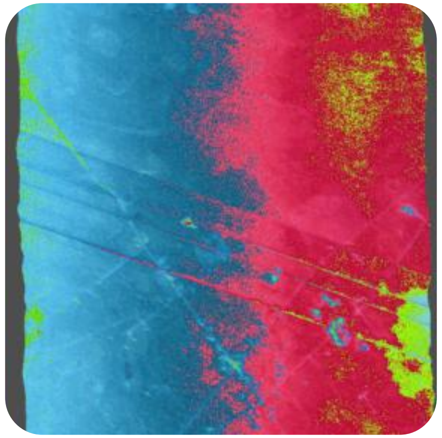
Visual Comparison

786.43 KB **5.5 KB**

Areas for improvement

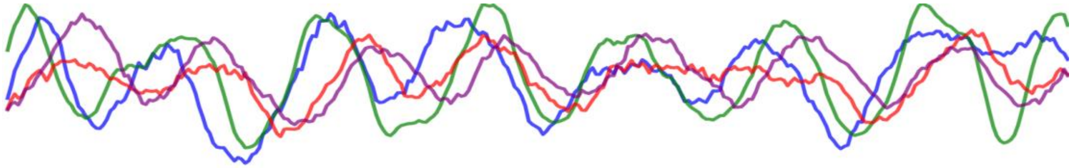
- Can we make it competitive in terms of the rate-distortion-complexity trade-off?
- How can we support a wider range of specialized signals types and modalities?
- Can we decouple the “generative” part of the decoding process to make it optional?

Medical Images and 3D volumes



Hyperspectral & HDR

Video

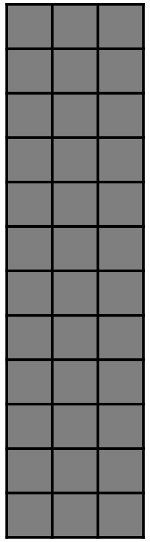


Multi-channel & spatial audio

Even a single linear projection can be expensive

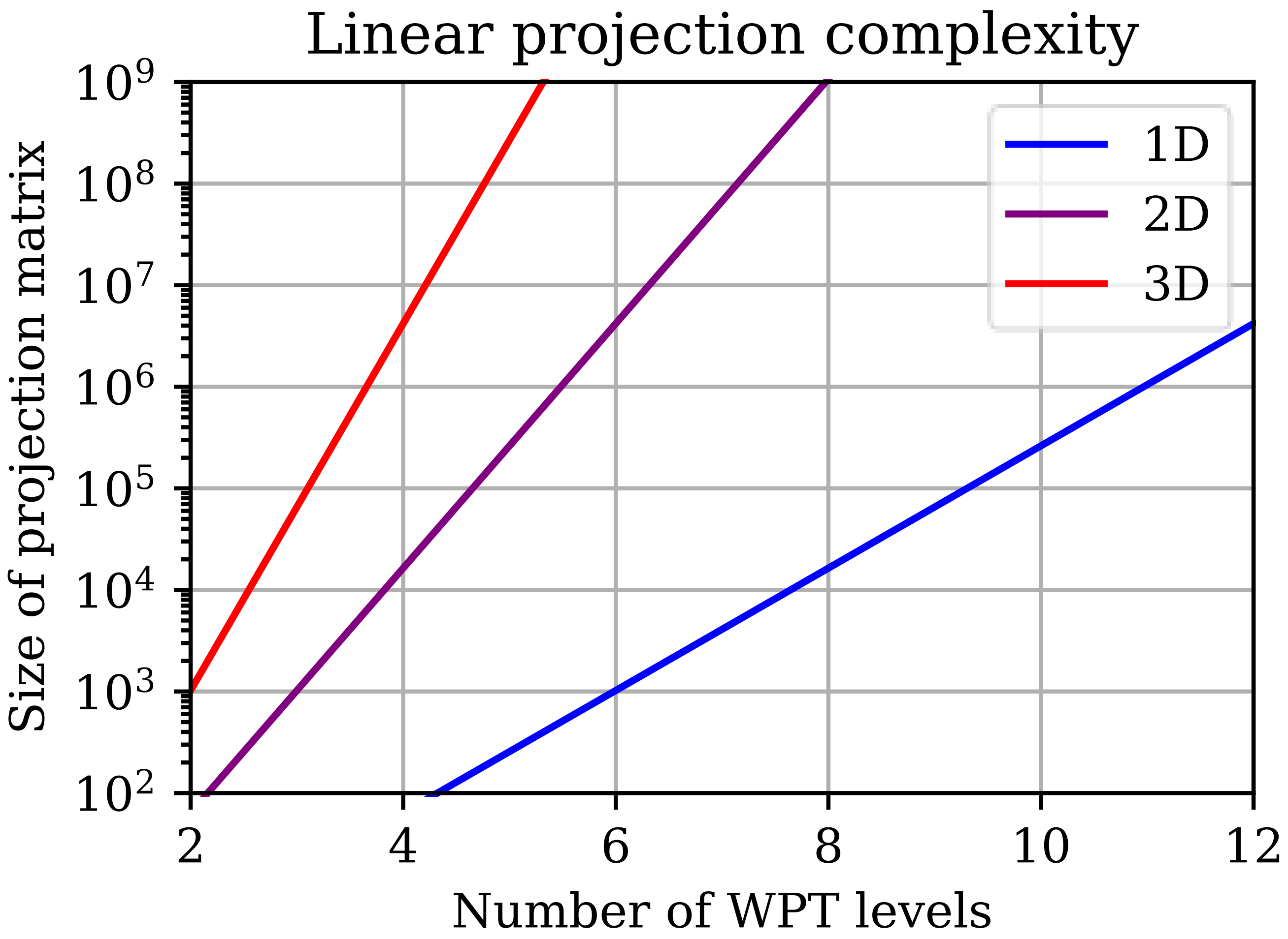
Dense 1×1 conv. layer

$N_{\text{in}} = (2^J)^D$
input dimension



$N_{\text{out}} = \frac{N_{\text{in}}}{\text{R}}$
Latent dimension

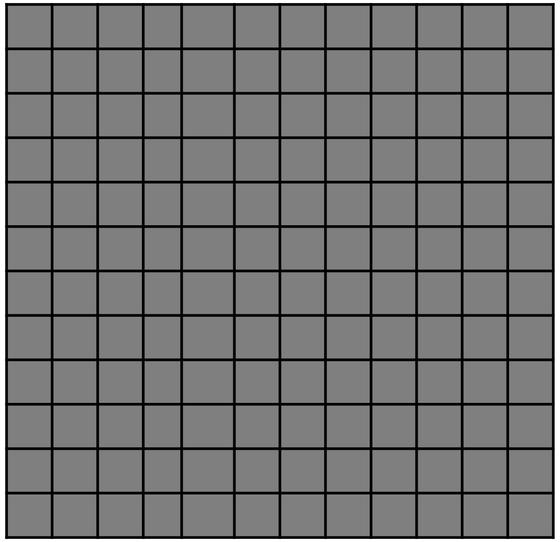
$\text{MACs} = 2^{2JD}/\text{R}$



Lightweight, FFT-inspired structured operations

Dense 1×1 conv. layer

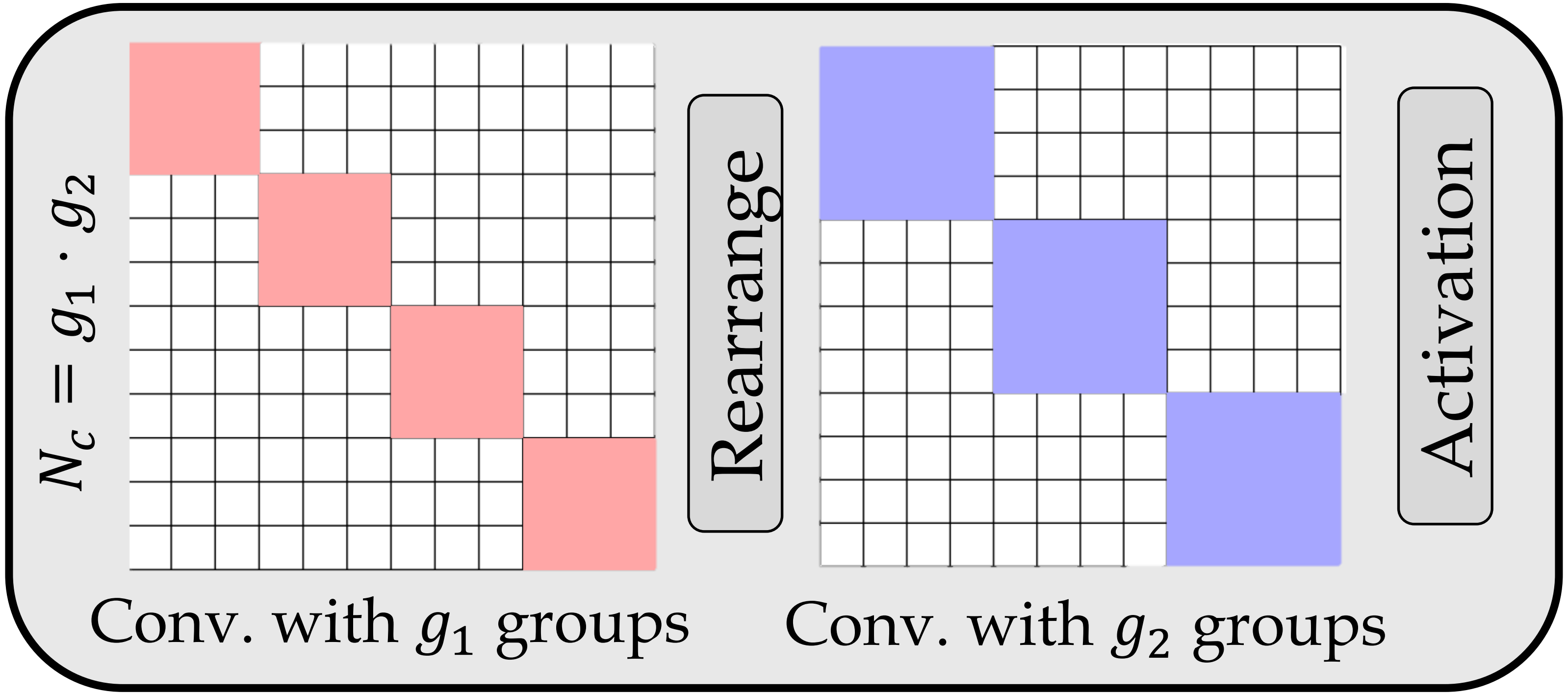
N_c = Input dimension



N_c = Output dimension

$$\text{MACs} = N_C^2$$

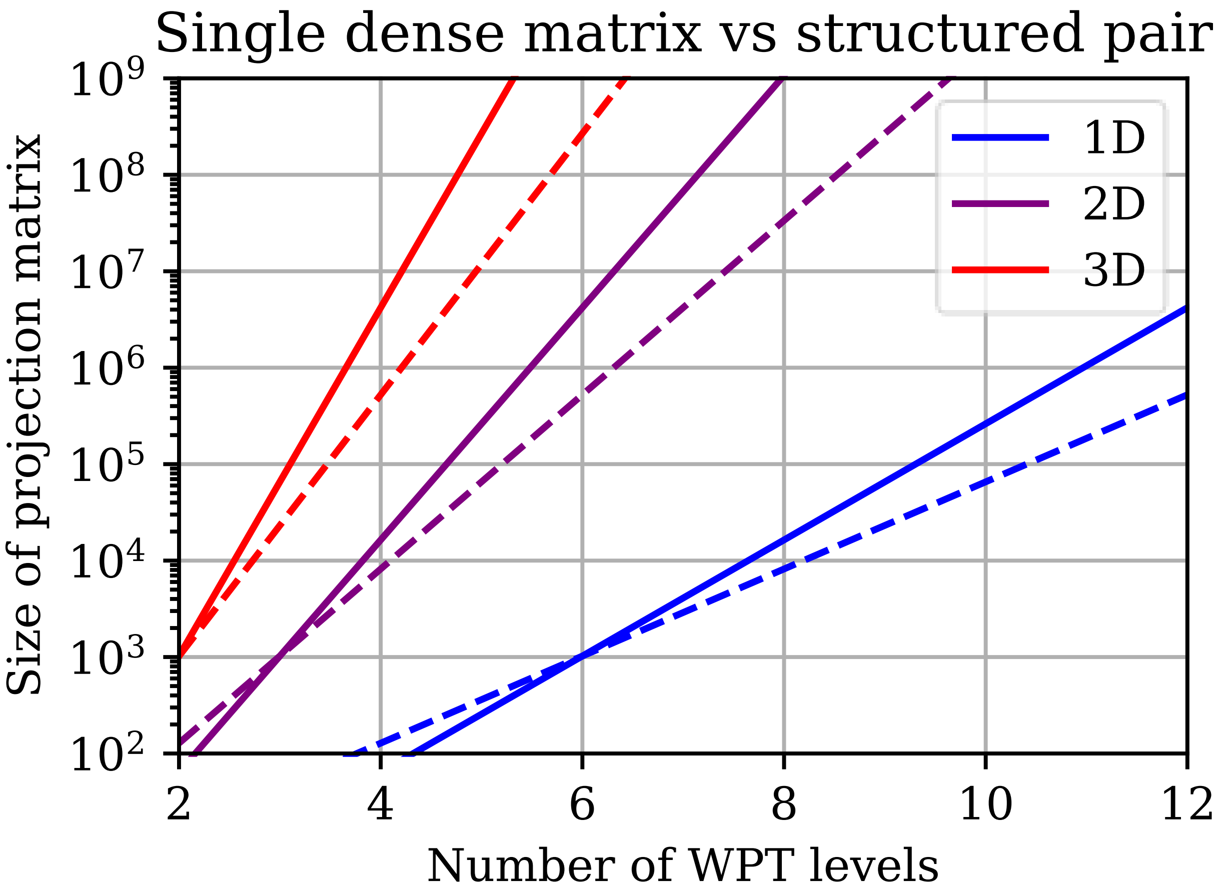
Structured 1×1 conv. layer

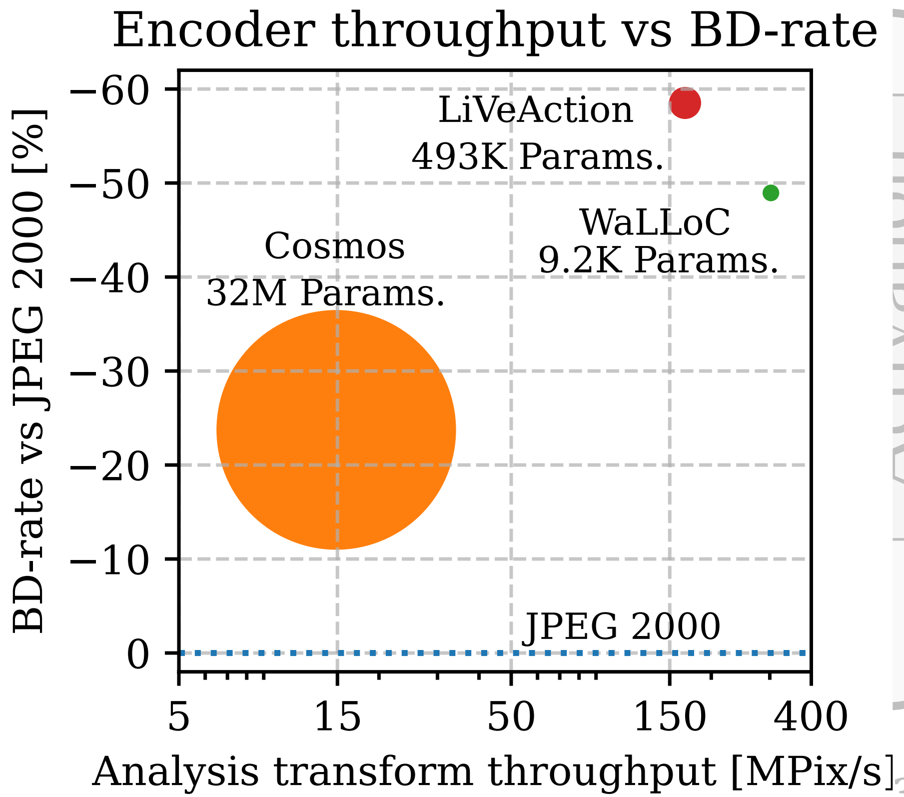


$$g_1 \approx g_2 \approx \sqrt{C_{\text{input}} \cdot 2^{J \cdot D}}$$

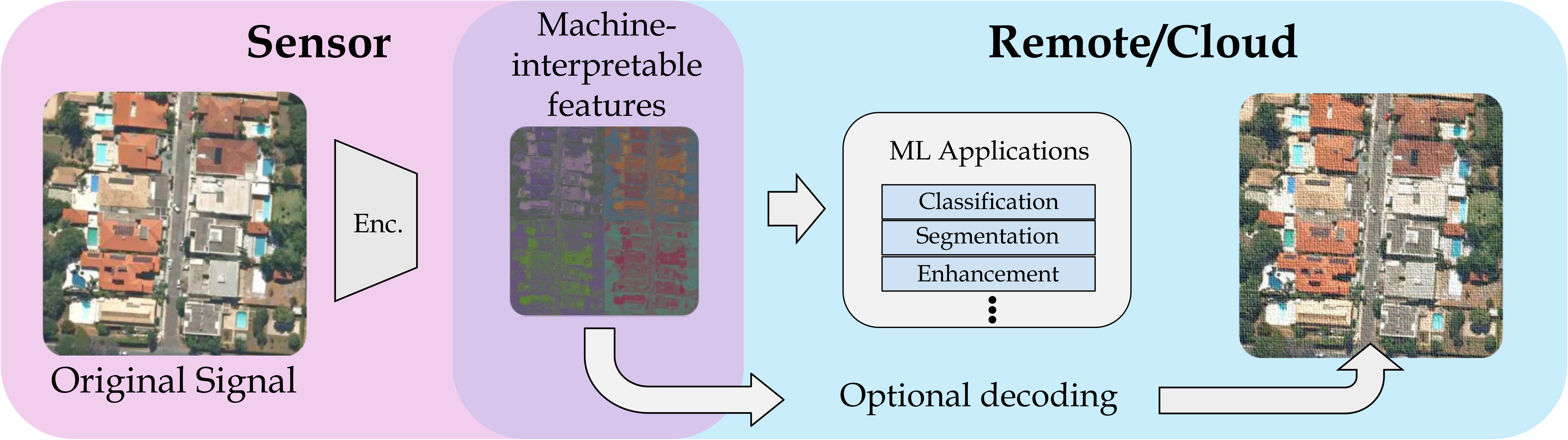
$$N_C \log N_C \leq \text{MACs} \leq N_C^{3/2}$$

Lightweight, FFT-inspired structured operations





Asymmetric Design

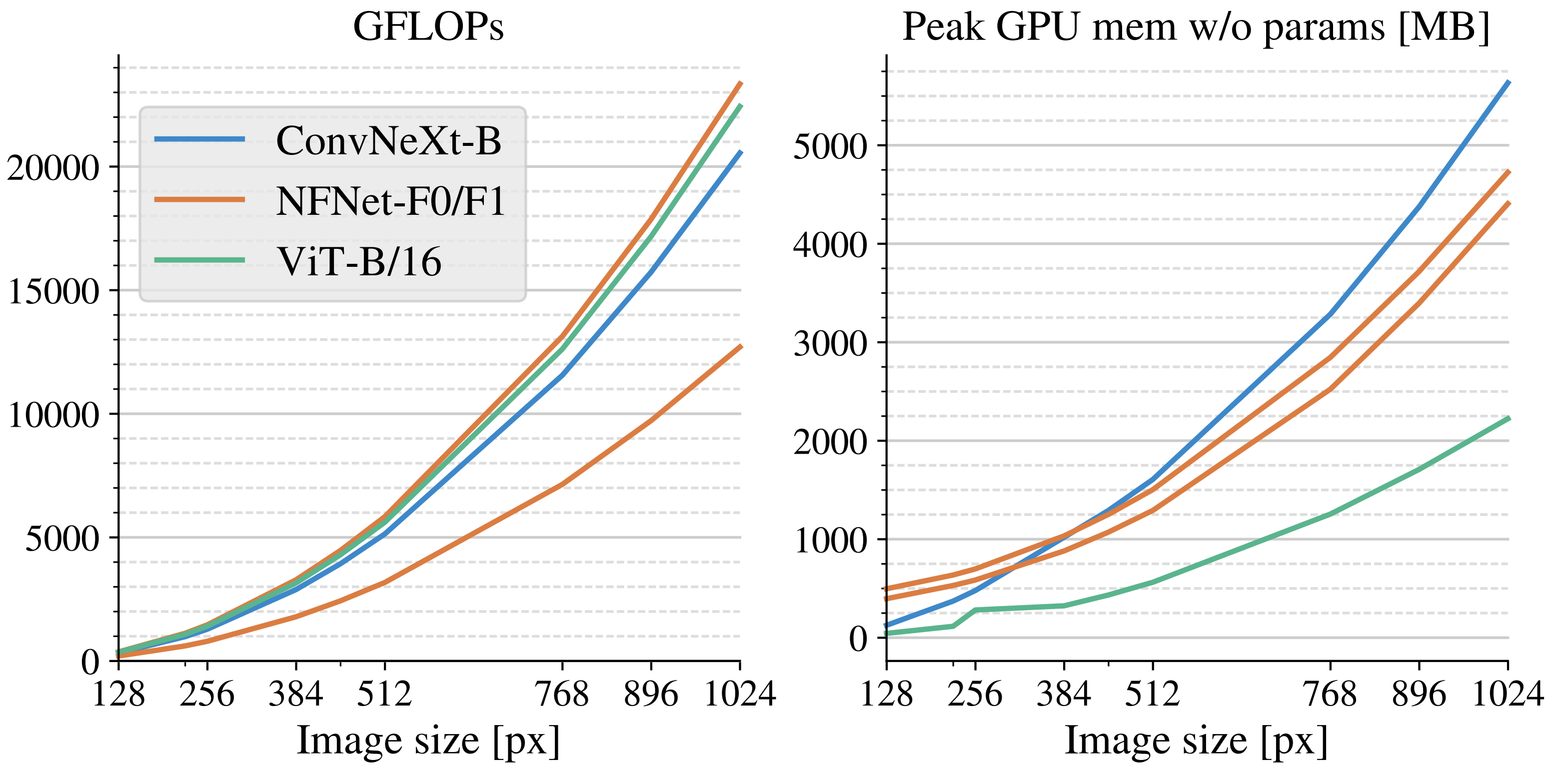


Encoding efficiency is very important

Decoder can run on cloud AI supercomputers (or throw it away completely)

Improving the synthesis transform

- High complexity synthesis transform is tolerable at runtime
- Training should still be possible in reasonable number of GPU hours
- Support modalities with high spatial and/or temporal resolution



Beyer, Lucas. "On the speed of ViTs and CNNs." (2024).

ND-generalized ViT decoder with linear attention

- Global receptive field

→Exploit non-local redundancies

- No excessive compute requirements

→Train at high resolution on single GPU

- No position encoding or batch norm

→ Works for any modality

Cai, Han, et al. "Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction." *Proceedings of the IEEE/CVF international conference on computer vision*. 2023.

LiVeAction Overview

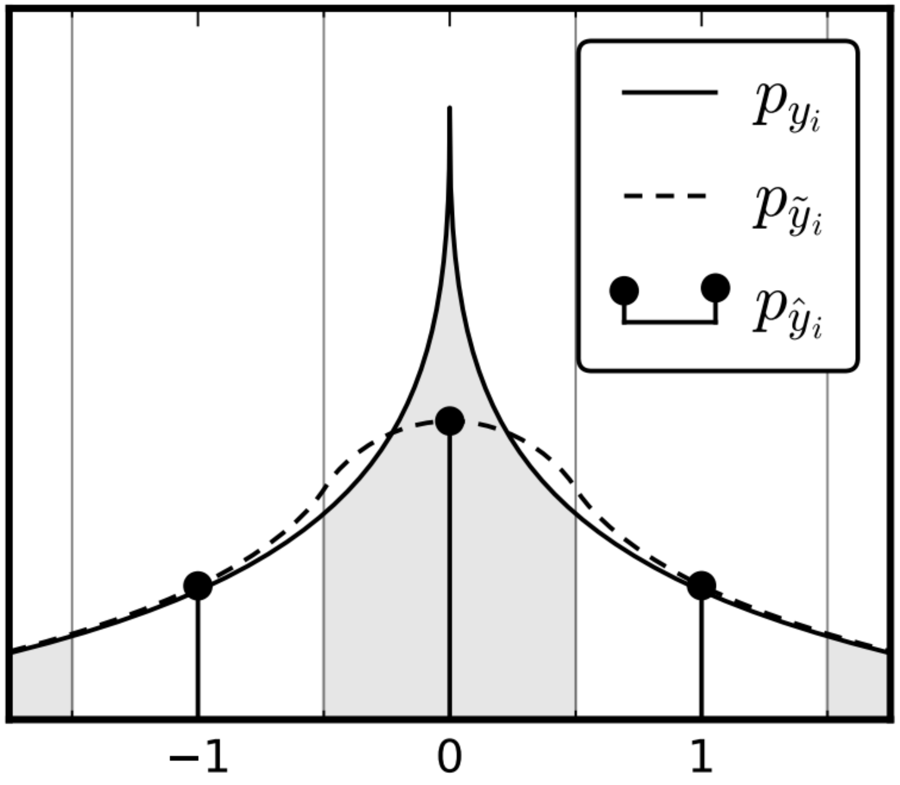
- Can we make it competitive in terms of the rate-distortion-complexity trade-off?
- How can we support a wider range of specialized signals types and modalities?
- Can we decouple the “generative” part of the decoding process to make it optional?

- FFT-like structured matrix operations in encoder
- ND-generalized vision transformer decoder with linear attention
- Simplified rate penalty

Rate objective

$$\min_{\mathcal{E},\mathcal{D}} \underbrace{\|x - \mathcal{D}((\mathcal{E}(x))\|^2}_{\text{MSE Distortion}} + \underbrace{H(\mathcal{E}(x))}_{\text{latent rate}}$$

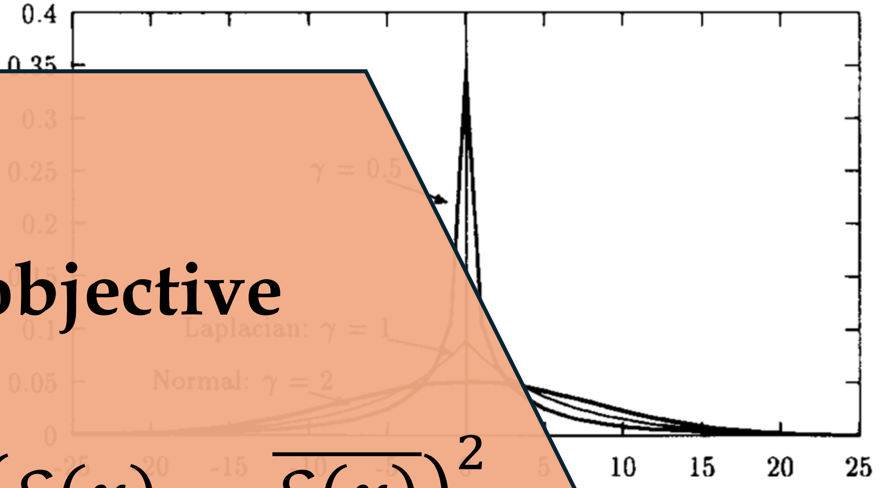
- Standard approach to optimize $H(\mathcal{E}(x))$ involves fitting a continuous proxy distribution p_{y_i}
- Requires an auxiliary optimizer with additional hyperparameters and separate learning rate

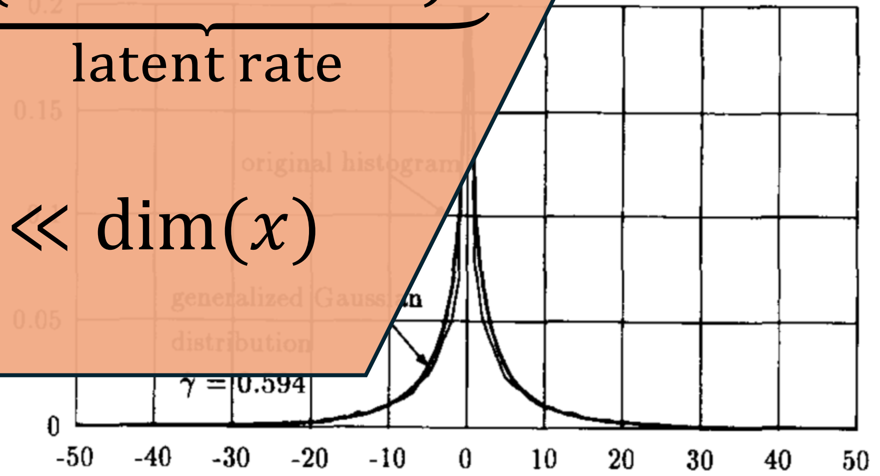


Ballé, Jona, Valero Laparra, and Eero P. Simoncelli. "End-to-end Optimized Image Compression." *International Conference on Learning Representations*. 2017.

Simplified rate penalty

- Intensity of sub-band filter outputs follow generalized gaussian distribution (GGD) for natural signals
- Empirically, outputs of DNN analysis transform also follow GGD
- For exponential family, minimizing variance is equivalent to minimizing rate





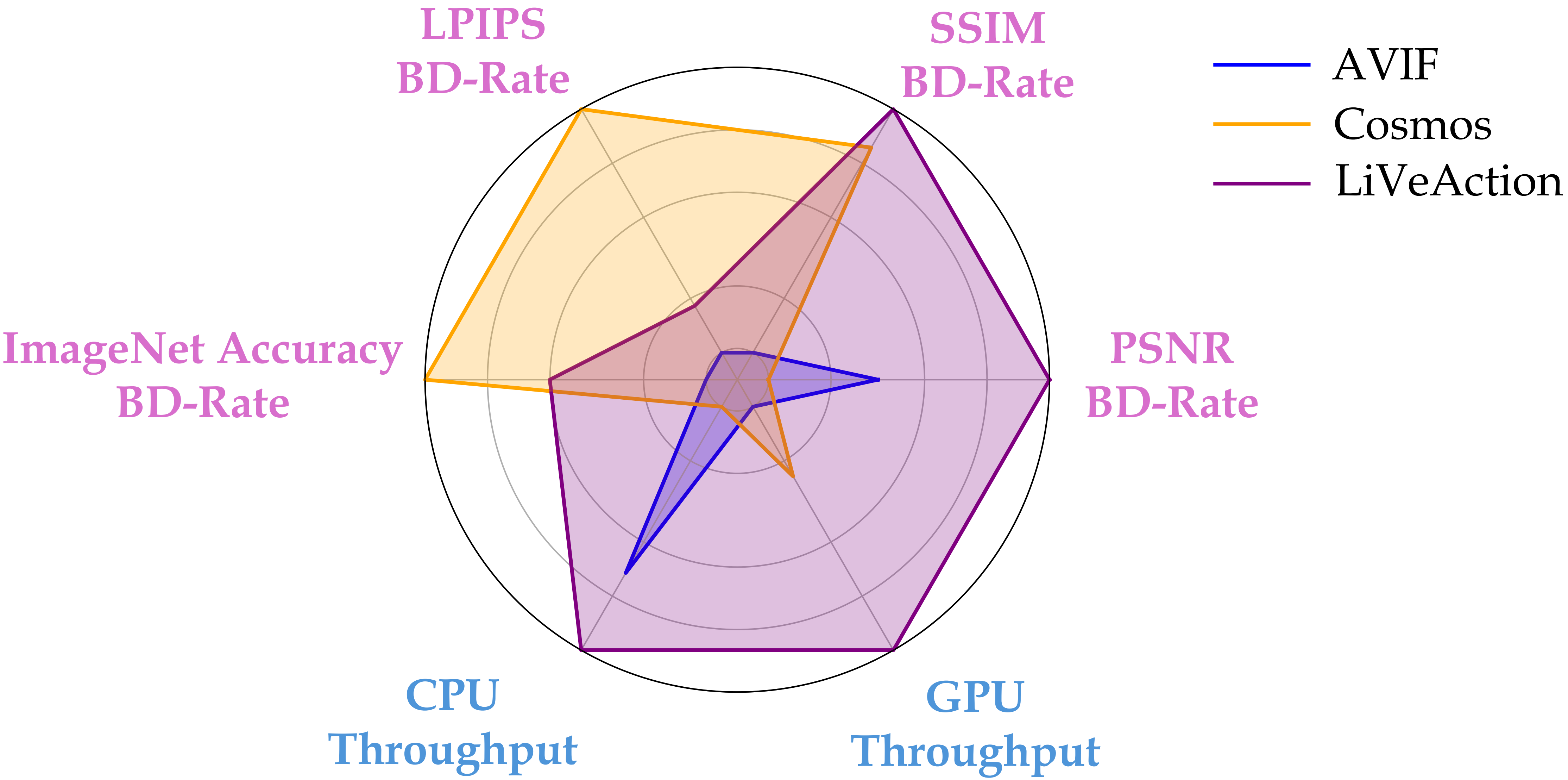
Simplied training objective

$$\min_{\mathcal{E},\mathcal{D}} \underbrace{\|x - \mathcal{D}((\mathcal{E}(x)) \|^2}_{\text{MSE Distortion}} + \underbrace{\sum\left(\mathcal{E}(x)_i - \overline{\mathcal{E}(x)}\right)^2}_{\text{latent rate}}$$

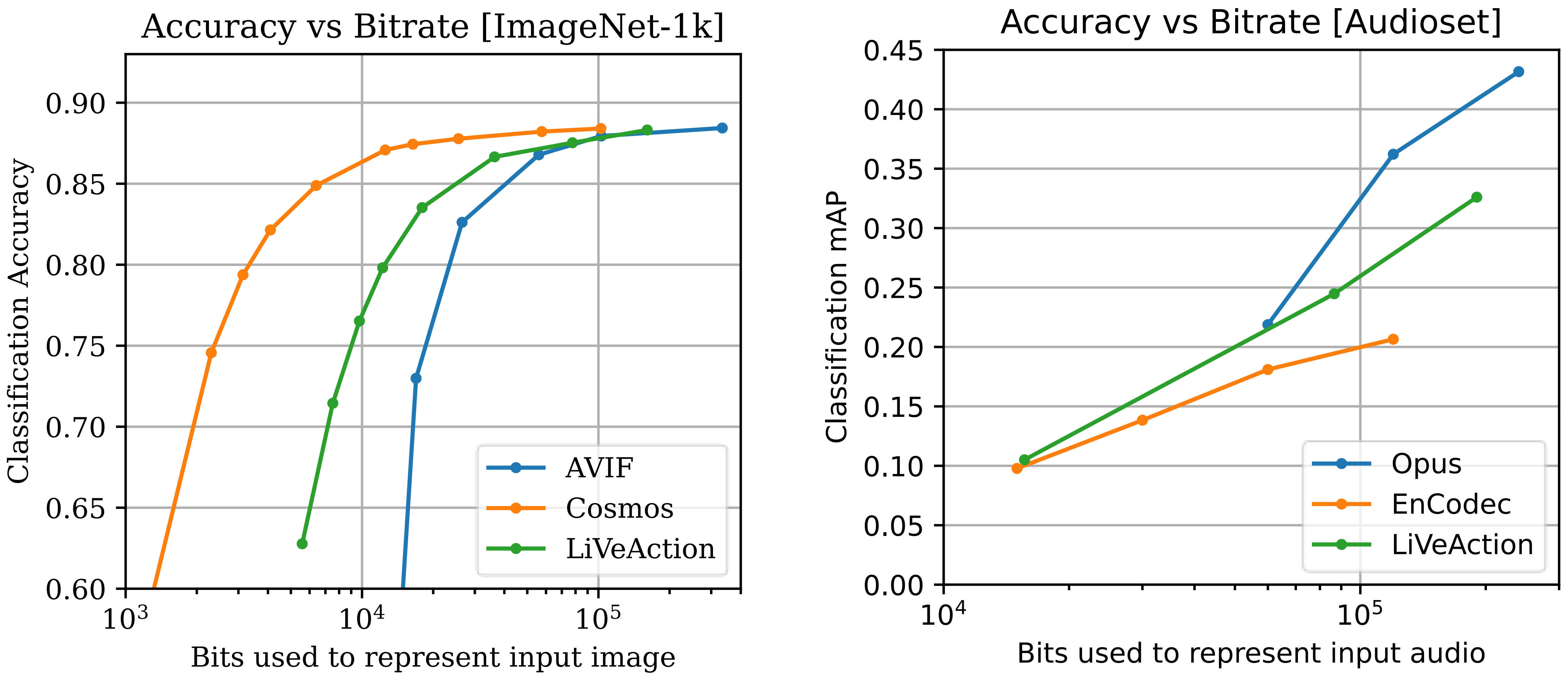
Subject to $\dim\big(\mathcal{E}(x)\big) \ll \dim(x)$

Sharifi, Karnran, and Alberto Leon-Garcia. "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video." *IEEE Transactions on Circuits and Systems for Video Technology* 5.1 (1995)

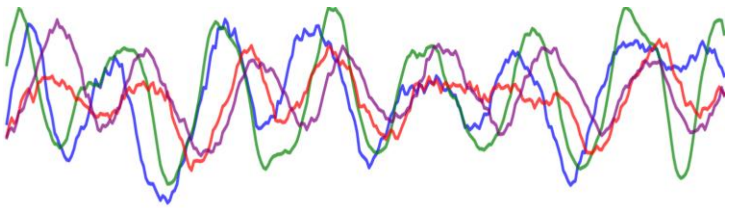
Compression efficiency and computational efficiency

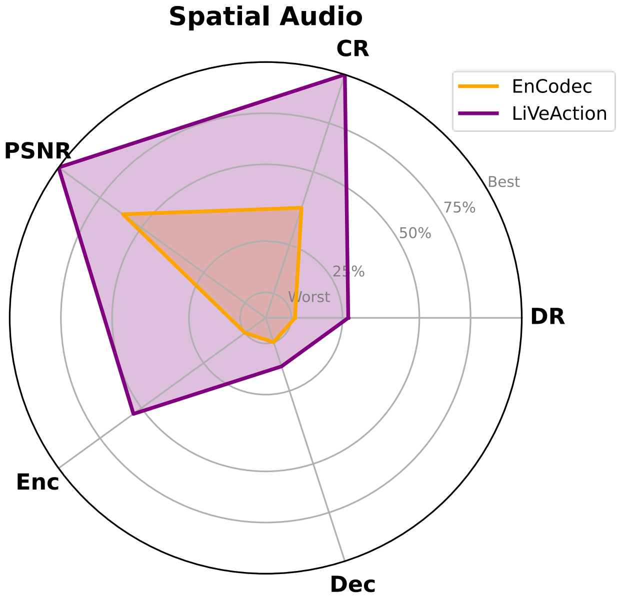


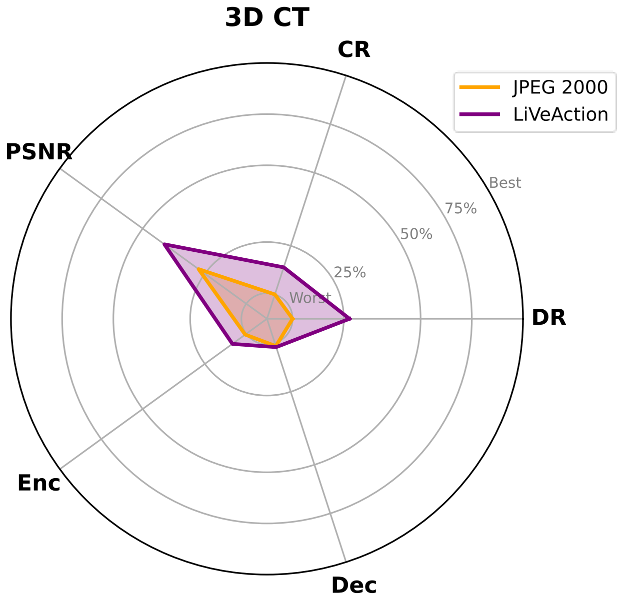
Machine Perception

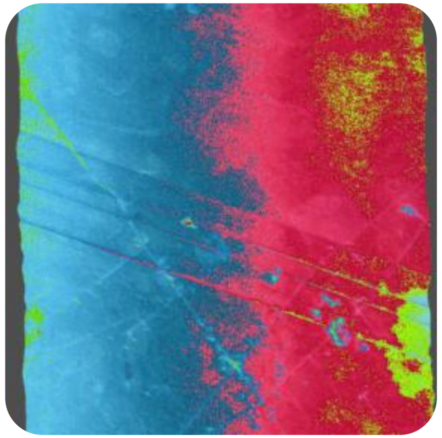


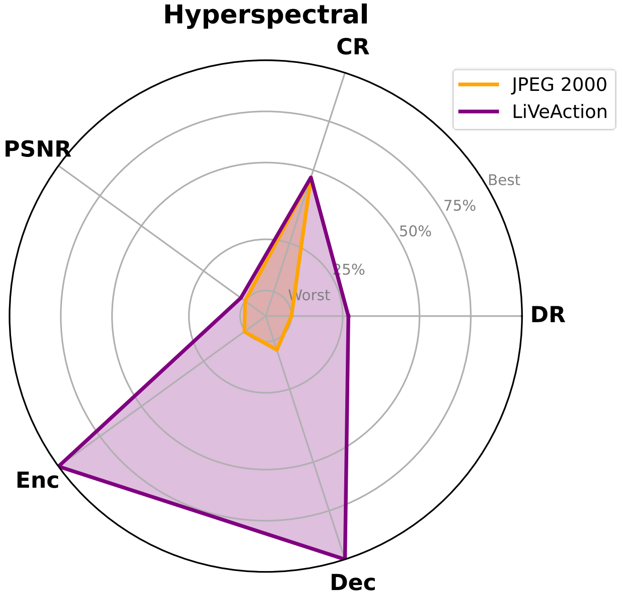
Other modalities

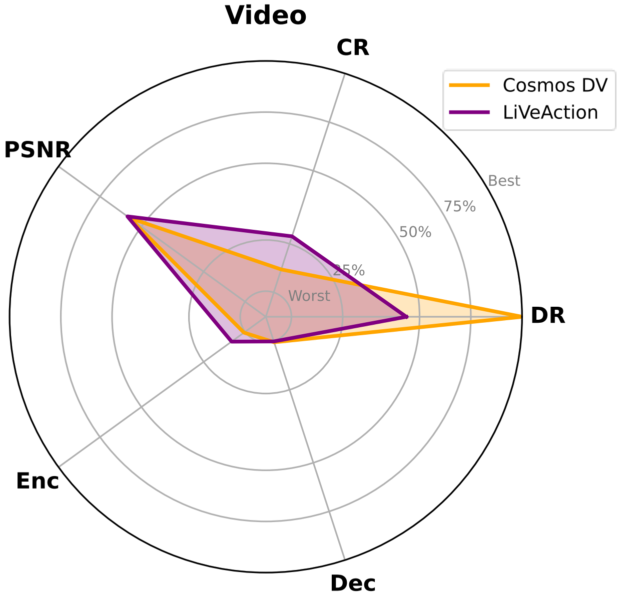












Generative Enhancement

(a) Original

(b) Cosmos

(c) LiVeAction

(d) LiVeAction + FLUX

Publications

[1] D. Jacobellis, D. Cummings, and N.J. Yadwadkar. "Machine Perceptual Quality: Evaluating the Impact of Severe Lossy Compression on Audio and Image Models.“ *Data Compression Conference*. IEEE, 2024.

[2] D. Jacobellis and N.J. Yadwadkar."Learned Compression for Compressed Learning.“ *Data Compression Conference*. IEEE, 2025.

[3] D. Jacobellis and N.J. Yadwadkar. "LiVeAction: a Lightweight, Versatile, and Asymmetric Neural Codec Design for Real-time Operation.“ Under Review.

[4] D. Jacobellis, M. Ulhaq, F. Racapé, H. Choi, and N.J. Yadwadkar.“ Dedelayed: Deleting remote inference delay via on-device correction.“ Under Review.

Software releases

Installation → `pip install walloc`

Audio → Pre-trained codec

Images → Pre-trained codec

Training (1D) → Tutorial

Training (2D) → Tutorial

More details available:

https://ut-sysml.org/walloc/

Contact: danjacobellis@utexas.edu

Software releases

Installation → `pip install livecodec`

Dozen+ pre-trained codecs available on Hugging Face
https://hf.co/danjacobellis/liveaction

Training → Tutorial

More details available:
https://ut-sysml.org/liveaction/

Contact: danjacobellis@utexas.edu