HyperCool: Reducing Encoding Cost in Overfitted Codecs with Hypernetworks

Anonymous submission

Abstract

Overfitted image codecs like Cool-chic achieve strong compression by tailoring lightweight models to individual images, but encoding is slow and costly. Non-Overfitted (N-O) Cool-chic accelerates encoding with a learned inference model, trading compression performance for speed. We introduce HyperCool, a hypernetwork that generates content-adaptive parameters for a Cool-chic decoder in a single forward pass, avoiding per-image fine-tuning. Our method reduces bitrate by 4.9% over N-O Cool-chic with minimal overhead and provides a strong initialization for further optimization, reducing steps to approach fully overfitted performance. With fine-tuning, HEVC-level compression is achieved at 60.4% of the cost of fully overfitted Cool-chic. This approach offers a practical way to accelerate overfitted image codecs under tight compute budgets.

Introduction

Learned image compression methods can outperform traditional codecs in rate-distortion (RD) performance, particularly at low bitrates (Liu, Sun, and Katto 2023; Jiang et al. 2025). These methods train neural networks end-to-end to optimize RD metrics, but often impose substantial computational demands. To address the decoding cost, Cool-chic (Ladune et al. 2023) and the C3 framework (Kim et al. 2024) introduce a novel approach: instead of relying on large, fixed, pre-trained models, they overfit lightweight neural networks to individual images and transmit the network parameters as the compressed representation. This per-image overfitting yields competitive compression with minimal decompression cost, offering a compelling alternative to autoencoder and diffusion-based schemes, which remain compute-intensive, particularly at decode time.

Despite its fast decoding, Cool-chic suffers from slow encoding, requiring iterative optimization of both weights and latents from scratch per image. To address this, Blard et al. propose Non-Overfitted (N-O) Cool-chic (Blard et al. 2024), which replaces per-image optimization with an analysis transform and a universal decoder that produces latents directly, without iterative rate-distortion optimization. This yields a substantial encoding speed-up and maintains Cool-chic's low decoding complexity. However, it also degrades compression efficiency, incurring a 56.5% rate increase on the CLIC2020 dataset.

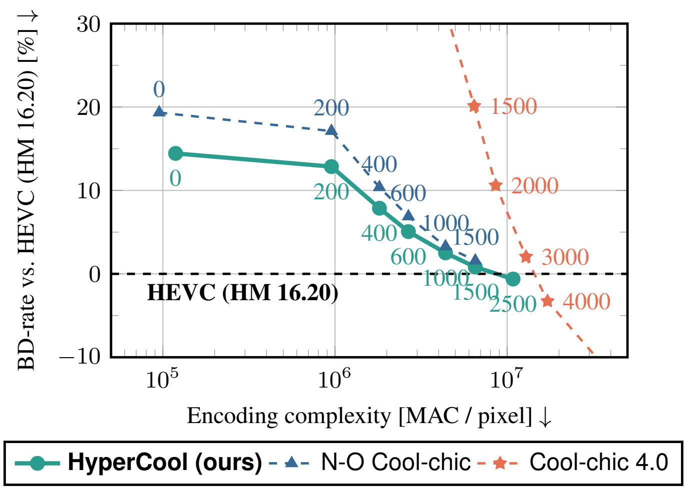


Figure 1: BD-rate against encoding complexity when fine-tuning from different initializations on the CLIC2020 dataset. Numbers next to data points indicate optimization steps.

This work aims to recover the compression efficiency lost in N-O Cool-chic while retaining its fast encoding and low decoding cost. We introduce *HyperCool*, a new variant of Cool-chic that restores image-dependent information in the decoder by employing a hypernetwork to predict decoder weights conditioned on the input image. HyperCool improves compression performance over N-O Cool-chic while retaining its fast encoding and maintaining the same low decoding cost. On the CLIC2020 dataset, it achieves a 4.9% BD-rate reduction compared to N-O Cool-chic, narrowing the gap to fully overfitted methods.

In addition to providing fast and adaptive compression, HyperCool supports optional fine-tuning of the predicted decoder on a single image, effectively using it as a warm start for full Cool-chic overfitting. This hybrid strategy reaches HEVC-level compression while requiring only 60.4% of the original Cool-chic encoding cost and preserving its decoding efficiency. We also provide a detailed analysis of the trade-offs between hypernetwork inference, optional per-image fine-tuning, and the resulting rate–distortion performance.

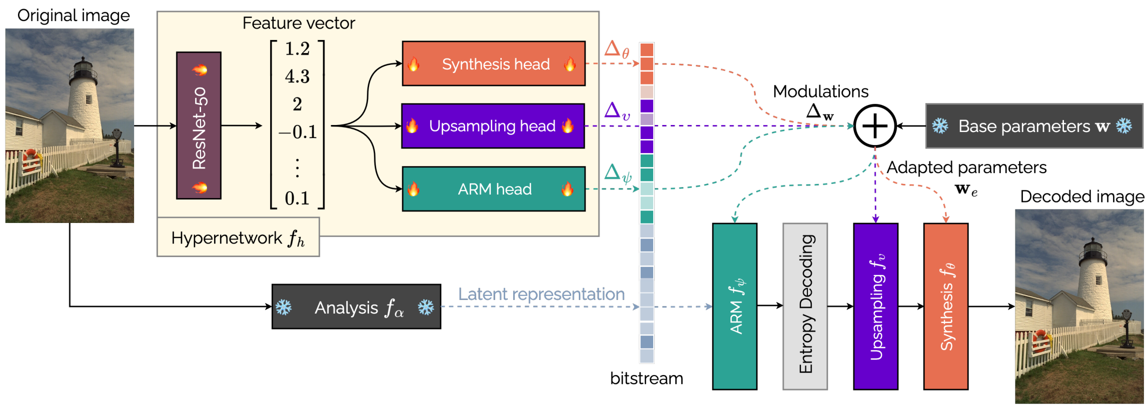


Figure 2: Architecture of the proposed HyperCool. The hypernetwork takes an input image and produces weight modulations for the synthesis, upsampling, and ARM composing a Cool-chic decoder. Only the weight modulations are transmitted.

Related Work and Background

Learned Image Compression

Autoencoder-based learned image codecs (Balle, Laparra, and Simoncelli 2017; Ballé et al. 2018) work by an encoder mapping the image $\mathbf{x}$ to latents $\mathbf{y}$, which are quantized to $\hat{\mathbf{y}}$ and entropy-coded. A decoder reconstructs the image from $\hat{\mathbf{y}}$. These models are trained end-to-end with a rate-distortion loss that balances reconstruction quality and bitrate:

$$\mathcal{L} = R(\hat{\mathbf{y}}) + \lambda D(\mathbf{x}, \hat{\mathbf{x}}) \quad (1)$$

where D is a distortion metric (e.g., MSE), R estimates the bitrate, and λ controls the trade-off.

Overfitted Codecs

Overfitted codecs train a dedicated model per image. COIN (Dupont et al. 2021) encodes each image as a fully connected network mapping coordinates to RGB values. COIN++ (Dupont et al. 2022) introduces a meta-learned base network shared across images and small per-image modulations, which are quantized and entropy-coded.

Cool-chic (Ladune et al. 2023) extends these by: (1) Representing images with hierarchical latent grids $\hat{\mathbf{y}} = \hat{y}_1, \dots, \hat{y}_N$. (2) Using a small synthesis network f_θ to reconstruct images from upsampled latents. (3) Compressing latents with an image-specific autoregressive entropy model f_ψ conditioned on causal context. Cool-chic overfits $\{\hat{\mathbf{y}}, \theta, \psi\}$ per image by minimizing a rate-distortion loss:

$$\mathcal{L} = \mathbb{E}_\mathbf{x}\left[\lambda D(\mathbf{x}, f_\theta(\hat{\mathbf{y}})) - \log p_\psi(\hat{\mathbf{y}})\right], \quad (2)$$

where p_ψ is modeled autoregressively:

$$p_\psi(\hat{\mathbf{y}}) = \prod_{i,j,k} p_\psi(\hat{y}_{ijk} | \mathbf{c}_{ijk}). \quad (3)$$

Cool-chic offers strong compression with a lightweight decoder but, high encoding cost.

Subsequent works improved Cool-chic via refined architecture, quantization, and training strategies (Kim et al. 2024; Leguay et al. 2023; Philippe et al. 2024). The Cool-chic implementation (Orange OpenSource 2025) integrates these improvements and serves as our starting point.

Reducing Encoding Complexity

Non-Overfitted (N-O) Cool-chic (Blard et al. 2024) speeds up encoding by removing per-image optimization and learning: (1) An analysis transform f_α that maps images to latents in a single forward pass. (2) A universal upsampling, synthesis network, and entropy model. The model is trained end-to-end by minimizing:

$$\min_{\alpha,\theta,\psi} \mathbb{E}_\mathbf{x}\left[\lambda D(\mathbf{x}, f_\theta(\mathrm{Ups}(f_\alpha(\mathbf{x})))) - \log p_\psi(f_\alpha(\mathbf{x}))\right]. \quad (4)$$

N-O Cool-chic enables fast encoding but loses some compression efficiency relative to fully optimized Cool-chic.

Metalearning methods like MLIIC (Zhang et al. 2025) use meta-learned initializations to speed up adaptation, but the code is unreleased and the results unverified.

Method

We propose a hypernetwork-based method that merges N-O Cool-chic's efficiency with the adaptability of overfitted decoders, reducing encoding time while boosting compression. Figure 2 illustrates the encoding and decoding process.

Starting from a pretrained N-O Cool-chic base model with decoder parameters $\mathbf{w}$ and an analysis transform f_α mapping images to latent grids $\hat{\mathbf{y}}$, we train a hypernetwork f_h to produce image-conditioned *modulation parameters* $\Delta_\mathbf{w}$:

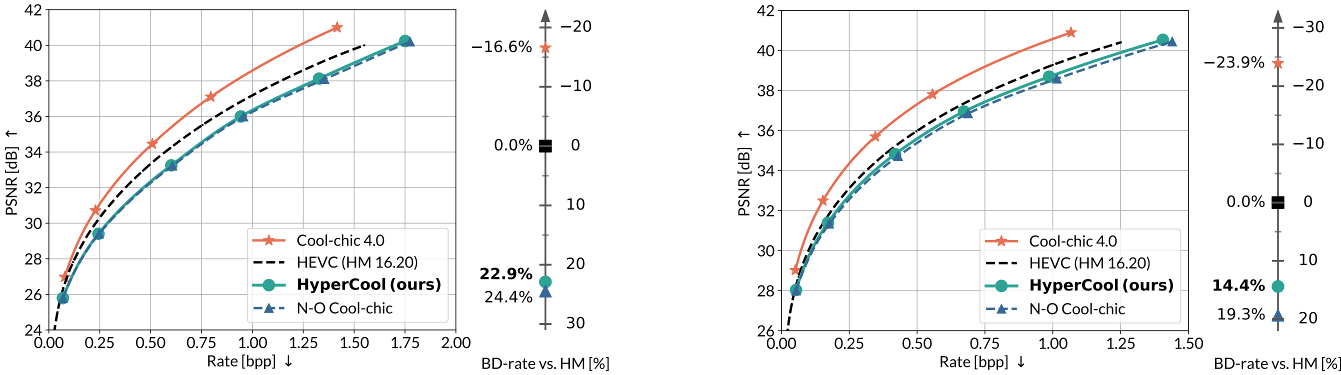
$$\Delta_\mathbf{w} = f_h(\mathbf{x}). \quad (5)$$

As shown in Figure 2, the hypernetwork f_h has two components: a pretrained `ResNet-50` backbone, followed by separate MLP heads generating modulations for the upsampling, synthesis, and autoregressive entropy modules.

The modulation $\Delta_\mathbf{w}$ is transmitted alongside the latent representation $\hat{\mathbf{y}}$. Modulations are encoded like the Cool-chic neural network parameters: quantized, then entropy-coded using Exp-Golomb coding. To decode the image, the image-adapted parameters $\mathbf{w}_e$ are constructed by adding the base decoder parameters $\mathbf{w}$ and the modulation $\Delta\mathbf{w}$:

$$\mathbf{w}_e = \mathbf{w} + \Delta_\mathbf{w}. \quad (6)$$

These image-adapted parameters are then used to compute the decoded image from the latent representation.



(a) Kodak dataset (b) CLIC2020 pro validation dataset

Figure 3: HyperCool rate-distortion performance. Results are averaged across the whole test dataset.

At inference, the hypernetwork predicts modulation parameters. These modulations adapt the decoder to the image, improving compression. However, transmitting the modulations introduces a small rate overhead. At the encoder side, it is verified if the modulations improve performance. If not, they are discarded. This ensures our method never underperforms the N-O Cool-chic and often improves upon it.

Results

Training

HyperCool is trained on 500,000 images from the OpenImages dataset (Kuznetsova et al. 2020), using random 256×256 patches. The hypernetwork is learned on top of pre-trained N-O Cool-chic models[1] using our method. One optimization step consists of the encoding and decoding described in Section and depicted in Fig. 2.

Only the hypernetwork parameters h are trained *i.e.*, the backbone and the different MLP heads. All the N-O Cool-chic parameters remain fixed, including the base decoder parameters $\mathbf{w}$ and analysis transform f_α. Since the latent is not optimized, latent quantization remains non-differentiable, simplifying training. The training loss is the standard rate-distortion objective, defined in Equation (1). Note that during training, the rate term only accounts for the latent representation's bitrate (via the adapted ARM), excluding the modulation parameters' rate.

Compression and Encoding Complexity Trade-Off

We evaluated our methods on the Kodak (Eastman Kodak Company 1999) and CLIC2020 professional validation (CLIC Challenge Organizers 2020) datasets. Kodak contains 24 images at 768×512 resolution, while CLIC2020 includes 41 images ranging from 512×384 to 2048×1370.

Figure 3 shows the rate-distortion performance of HyperCool compared to the N-O Cool-chic baseline and the original overfitted Cool-chic 4.0. Our method improves compression over N-O Cool-chic on both datasets. Gains are more pronounced at higher bitrates and on larger images, such as those in CLIC2020.

[1]We thank **Théophile Blard** for training these models.

Table 1: Encoding complexity and BD-rate against HEVC of the proposed HyperCool compared to N-O Cool-chic.

Method	Complexity [kMAC / pix] ↓			BD-rate [%] ↓	
	Analysis	Hypernet	Total	Kodak	CLIC20
N-O Cool-chic	99	/	99	24.4	19.3
HyperCool	99	24	123	22.9	14.4
Cool-chic fast	/	/	64,000	-11.8	-16.9
Cool-chic slow	/	/	450,000	-16.6	-23.9

Table 1 compares the BD-rates of the proposed HyperCool against HEVC, along with encoding complexity. It shows that HyperCool improves compression over N-O Cool-chic, with only a slight increase in encoding cost. We also compare HyperCool's encoding complexity to standard Cool-chic using the *fast* and *slow* presets from the official open-source implementation (Orange OpenSource 2025). HyperCool is 500 to 3000 times cheaper to encode than fully overfitted Cool-chic, though at the cost of reduced compression performance.

Modulations Rate Overhead and Usage

Adapting decoder parameters to the image using modulation parameters $\Delta_{\mathbf{w}}$ requires transmitting them, adding rate overhead. Therefore, modulations are only used if the compression improvement outweighs their signaling cost. This is determined at the encoder via a simple test, which disables modulations when counterproductive.

Figure 4 shows the proportion of images using modulations under different rate constraints and datasets. At higher rates, nearly all images use the hypernetwork modulations, as more bits are available for parameter signaling. However, under stricter rate constraints, many images do not use modulations *e.g.*, only 20 % of the images at the lowest rate on CLIC2020. This behavior explains the improved performance of HyperCool on CLIC2020, where larger images permit greater use of modulations due to higher bit budgets.

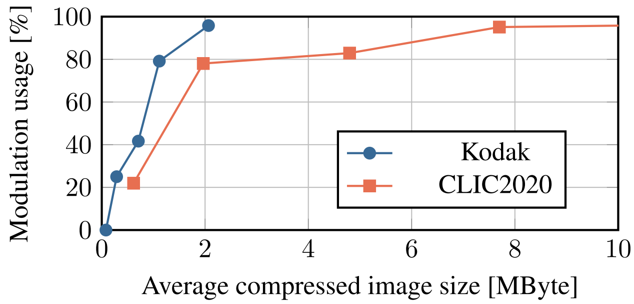


Figure 4: Usage of the modulation parameters $\Delta_{\mathbf{w}}$ across different bitrates.

Table 2: Change in compared to N-O Cool-chic when using different modulations. Averaged across rates.

Modulations ARM	Ups	Syn	Rate [bpp] ↓ Modulation	Latent	Total	PSNR [dB] ↑
✓	✓	✓	+0.008	-0.019	-0.011	+0.071
✓			+0.003	-0.019	-0.016	0
	✓	✓	+0.005	0	+0.005	+0.071

Figure 5 illustrates that modulation parameters $\Delta_{\mathbf{w}}$ are more compact than the full parameters $\mathbf{w}_e$. We confirm this by comparing the standard deviations of $\Delta\mathbf{w}$ and $\mathbf{w}_e$, computed from a Laplace distribution fitted to the parameters. Modulations show lower variance, indicating better compressibility with Exp-Golomb coding.

Hypernetwork Ablation Experiments

To assess the contribution of each hypernetwork module, we start from the full HyperCool model and selectively disable modulations for different components. Table 2 summarizes the average change in bitrate and PSNR compared to the base N-O Cool-chic across rate points. Using only ARM modulations reduces the latent bitrate without improving PSNR. In contrast, applying only upsampling and synthesis modulations improves PSNR but increases the total bitrate. Combining all modulations yields a bitrate reduction of 0.011 bpp and a PSNR increase of 0.071 dB. Together, they improve both compression rate and image quality, with only a slight increase in modulation bitrate.

HyperCool as an Overfitting Initialization

Standard Cool-chic encodes an image through the overfitting of the latent representation and decoder parameters, starting from a random initialization. Both N-O Cool-chic and the proposed HyperCool provide a strong initial guess for the latent and decoder parameters, improving initialization for subsequent overfitting.

Figure 1 compares Cool-chic encoding using three different initializations: random, N-O Cool-chic, and HyperCool. Across all encoding complexities, HyperCool initialization consistently outperforms N-O Cool-chic, highlighting the hypernetwork's effectiveness. Moreover, HyperCool enables reaching HEVC-level compression 40% faster than random initialization. However, standard Cool-chic with random initialization achieves better asymptotic performance, suggesting HyperCool may converge to a local minimum.

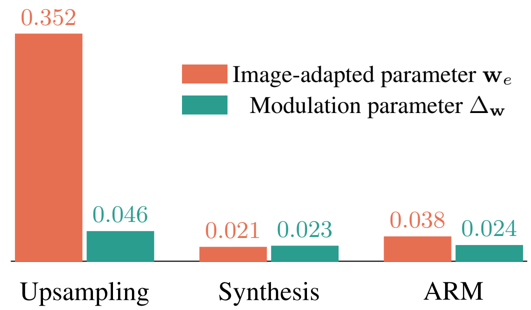


Figure 5: Comparison of the standard deviation of image-adapted and modulation parameters on CLIC2020.

Limitations and Future Directions

Although our results are positive, there are notable limitations that must be further investigated. The performance advantage of our hypernetwork is most pronounced at medium to high bitrates. At low bitrates, the quantization process often favors excluding the hypernetwork's output to save on the additional rate, leading to performance nearly identical to the underlying N-O Cool-chic model. Additionally, our approach depends on the quality of the pre-trained N-O Cool-chic base model, as the hypernetwork only generates modulation parameters for it.

Future work could explore several directions. Alternative hypernetwork architectures may yield further improvements. It would be valuable to compare HyperCool with other meta-learning strategies. For example, COIN++ (Dupont et al. 2022) and MLIIC (Zhang et al. 2025) apply MAML (Finn, Abbeel, and Levine 2017) to learn a base network for task-wise adaptation. A hybrid method combining MAML-based adaptable bases with our hypernetwork modulation could better parametrize the base model, improving BD-rate while keeping computational cost unchanged.

Conclusion

This work addresses the main drawback of overfitted codecs: slow encoding that requires per-image optimization. We introduce a novel hypernetwork that builds upon the Non-Overfitted Cool-chic framework to generate image-adaptive parameters in a single forward pass. HyperCool improves compression efficiency without per-image optimization, providing a step toward practical overfitted codecs.

Our method achieves a 4.9% BD-rate reduction over the N-O Cool-chic baseline with minimal computational overhead. Additionally, the hypernetwork output provides a strong initialization for full Cool-chic decoder optimization, reducing the number of fine-tuning steps by 40%. This makes our approach a practical way to accelerate overfitted codecs and broaden their range of applications.

References

Balle, J.; Laparra, V.; and Simoncelli, E. P. 2017. End-to-end Optimized Image Compression. In *International Conference on Learning Representations (ICLR) 2017*.

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*.

Blard, T.; Ladune, T.; Philippe, P.; Clare, G.; Jiang, X.; and Déforges, O. 2024. Overfitted Image Coding at Reduced Complexity. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, volume 1, 927–931. Lyon, France: IEEE. ISBN 978-94-6459-361-7.

CLIC Challenge Organizers. 2020. Workshop and Challenge on Learned Image Compression (CLIC) 2020. Accessed: 2025-10-16.

Dupont, E.; Goliński, A.; Alizadeh, M.; Teh, Y. W.; and Doucet, A. 2021. COIN: COmpression with Implicit Neural representations. ArXiv:2103.03123 [eess].

Dupont, E.; Loya, H.; Alizadeh, M.; Golinski, A.; Teh, Y. W.; and Doucet, A. 2022. COIN++: Neural Compression Across Modalities. *Transactions on Machine Learning Research*.

Eastman Kodak Company. 1999. Kodak Lossless True Color Image Suite. Accessed: 2025-10-16.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135. PMLR. ISSN: 2640-3498.

Jiang, W.; Yang, J.; Zhai, Y.; Gao, F.; and Wang, R. 2025. MLIC++: Linear Complexity Multi-Reference Entropy Modeling for Learned Image Compression. *ACM Trans. Multimedia Comput. Commun. Appl.* Just Accepted.

Kim, H.; Bauer, M.; Theis, L.; Schwarz, J. R.; and Dupont, E. 2024. C3: High-Performance and Low-Complexity Neural Compression from a Single Image or Video. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9347–9358. Seattle, WA, USA: IEEE. ISBN 9798350353006.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; Duerig, T.; and Ferrari, V. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Ladune, T.; Philippe, P.; Henry, F.; Clare, G.; and Leguay, T. 2023. COOL-CHIC: Coordinate-based Low Complexity Hierarchical Image Codec. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13469–13476. Paris, France: IEEE. ISBN 9798350307184.

Leguay, T.; Ladune, T.; Philippe, P.; Clare, G.; and Henry, F. 2023. Low-complexity Overfitted Neural Image Codec. ArXiv:2307.12706 [eess].

Liu, J.; Sun, H.; and Katto, J. 2023. Learned Image Compression with Mixed Transformer-CNN Architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1–10.

Orange OpenSource. 2025. Cool-Chic. https://github.com/Orange-OpenSource/Cool-Chic. Accessed: 2025-07-07.

Philippe, P.; Ladune, T.; Clare, G.; Henry, F.; Blard, T.; and Leguay, T. 2024. Upsampling Improvement for Overfitted Neural Coding. ArXiv:2411.19249 [eess].

Zhang, Z.; Li, H.; Jiang, W.; Zhai, Y.; and Wang, R. 2025. MLIIC: Meta-Learned Implicit Image Codec with 15× Faster Encoding Speed and Higher Performance. In *2025 Data Compression Conference (DCC)*, 103–112. ISSN: 2375-0359.