DiT-IC: Aligned Diffusion Transformer for Efficient Image Compression

Supplementary Material

1. Method Details

Model Architecture. The overall architecture is illustrated in Fig. 1. Our entropy model follows the classical hyperprior framework and further incorporates the autoregressive context model introduced in StableCodec [24]. Different from StableCodec, we replace the original context modules with a lightweight *DepthConvBlock* [13], which significantly reduces computational complexity while preserving effective spatial–channel context modeling capability. Given the quantized latent representation $\hat{\mathbf{z}}$, the autoregressive module predicts the Gaussian distribution parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ via a 4-step autoregressive procedure. These parameters are then fed into an arithmetic coder to convert quantized symbols into a bitstream during encoding, or to reconstruct symbols from the bitstream during decoding.

Resolution Generalization. DiT-IC adopts a Diffusion Transformer without positional encoding (NoPE) [14], avoiding positional extrapolation issues commonly encountered in standard Transformers. By removing positional embeddings, the model does not bind representations to fixed spatial indices, thereby improving length and resolution generalization. Although trained on small patches, the model generalizes reliably to higher resolutions at inference time without architectural modification.

Self-Distillation Alignment. The key idea is to collapse multi-step diffusion supervision into a self-aligned single-step objective without introducing an external teacher model. We adopt alignment-style objectives to approximate diffusion behavior under a one-step formulation, and term this strategy *Self-Distillation Alignment* to distinguish it from conventional teacher–student distillation methods. This formulation preserves diffusion-style supervision while avoiding additional model overhead.

Variance–Timestep Mapping. As shown in Fig. 4, the predicted variance exhibits a strong correlation with compressed noise (with cosine similarity up to 0.94). From a variational inference perspective, higher latent variance corresponds to higher conditional entropy and greater reconstruction uncertainty, which manifests as stronger noise components. This observation motivates a monotonic variance-to-timestep mapping strategy: larger variance is mapped to a larger diffusion timestep, implying stronger denoising. Consequently, entropy modeling and timestep prediction are naturally aligned. Empirically, blocking gradients from the $\mathcal{F} : \sigma \rightarrow t$ branch results in negligible bitrate change, indicating that joint optimization introduces minimal conflict between compression and diffusion objectives.

Distortion–Perception Trade-off. Under a fixed information rate, distortion and perceptual quality cannot be simultaneously optimized, as established by the rate–distortion–perception trade-off principle [3, 4, 19]. DiT-IC adheres to this information-theoretic constraint, which explains why perceptual optimization may lead to reduced PSNR. The trade-off is controlled by the weighting parameter λ in Eq. (10). In practice, sweeping λ produces smooth distortion–perception operating curves, allowing flexible control over reconstruction fidelity and perceptual realism.

2. Captions Generated by a VLM

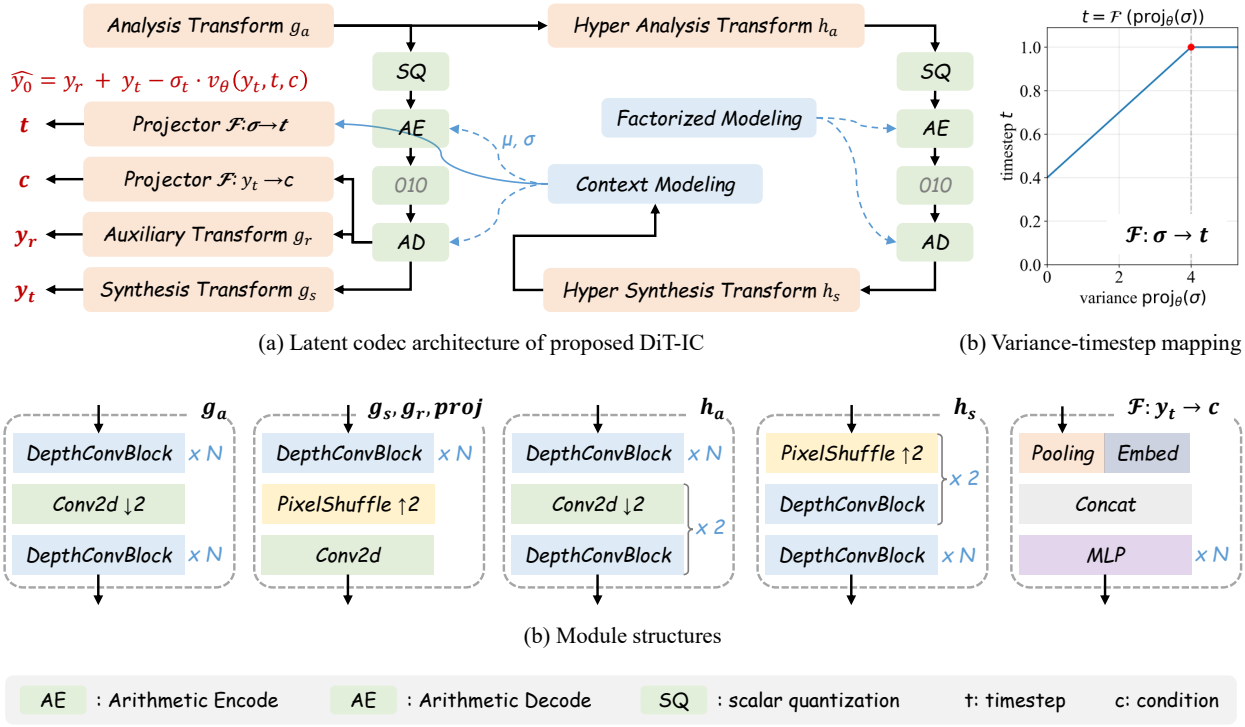
To avoid manual annotation and ensure scalable supervision, we employ a Vision–Language Model (VLM) to automatically generate semantic captions for training. Specifically, we adopt InternVL [6, 22], which is consistent with the captioning pipeline used in the original text-to-image DiT-SANA [23] pretraining. Representative caption examples produced by the VLM are shown in Fig. 2.

3. More Implementation Details

Our DiT-IC model is trained on two NVIDIA RTX Pro 6000 GPUs using PyTorch 2.8.0 and CUDA 12.8. For fair comparison, we reproduce several open-source baselines within the same environment to obtain detailed results. Due to differences in software versions and numerical kernels, minor deviations from the originally reported numbers may occur.

The training consists of two stages. In Stage 2, we initially disable the adversarial loss by setting $\lambda_{\text{adv}} = 0$, and enable it only after 30% of iterations to stabilize optimization. We also gradually anneal the contrastive co-alignment loss that aligns latent embeddings with text embeddings, controlled by a temperature parameter τ. This loss is used only during the initial 30% of Stage 2 to provide early semantic guidance while avoiding unstable or noisy text-driven updates in later iterations.

After the two-stage training, the model typically reaches a stable convergence point. At this stage, the Self-Distillation Alignment module becomes less essential, and jointly finetuning the entire model—including the encoder—could potentially yield further improvements. Although encoder finetuning is not included in this work, exploring this unified training strategy remains a promising direction for future research.



(a) Latent codec architecture of proposed DiT-IC

(b) Variance-timestep mapping

(b) Module structures

Figure 1. Overall architecture of our model. The entropy model is based on the hyperprior framework and an autoregressive context model similar to StableCodec [24], but replaces heavy components with lightweight *DepthConvBlocks* [13].

Table 1. Quantitative perceptual comparison between DiT-IC and StableCodec ($\lambda = 2.0$) at similar bitrates (∼0.03–0.04 bpp). We report a suite of perceptual metrics including FID, KID, NIQE, CLIPIQA and MUSIQ. Lower is better for FID/KID/NIQE, and higher is better for the remaining metrics. DiT-IC consistently outperforms StableCodec across most datasets and metrics, with a slight exception on KID for CLIC2020. The best results are highlighted in **red**.

Datasets	Kodak		CLIC 2020		DIV2K		Average		Avg. Difference	
Methods	DiT-IC	StableCodec	DiT-IC	StableCodec	DiT-IC	StableCodec	DiT-IC	StableCodec	$\lvert\Delta\rvert \uparrow$	$\lvert\Delta\rvert(\%) \uparrow$
FID ↓	-	-	**3.750**	3.940	**8.650**	10.350	**6.200**	7.145	0.945	13.23%
KID ↓	-	-	0.00083	**0.00066**	**0.00060**	0.00080	**0.00072**	0.00073	0.00002	2.06%
NIQE ↓	**3.099**	3.557	**3.833**	4.459	**3.270**	3.603	**3.400**	3.873	0.473	12.21%
CLIPIQA ↑	**0.735**	0.716	**0.582**	0.531	**0.626**	0.570	**0.648**	0.606	0.042	6.92%
MUSIQ ↑	**74.494**	73.177	**60.606**	58.663	**65.818**	63.822	**66.972**	65.221	1.752	2.69%

4. Complexity

Training Complexity. Multi-stage training is commonly adopted in diffusion-based codecs (e.g., StableCodec, OneDC, and ResULIC), often involving external teacher inference or multi-step diffusion supervision. In contrast, our training pipeline is strictly sequential and does not require additional teacher models or iterative diffusion sampling during optimization. In practice, the model converges within approximately 3 days on two NVIDIA A100 GPUs.

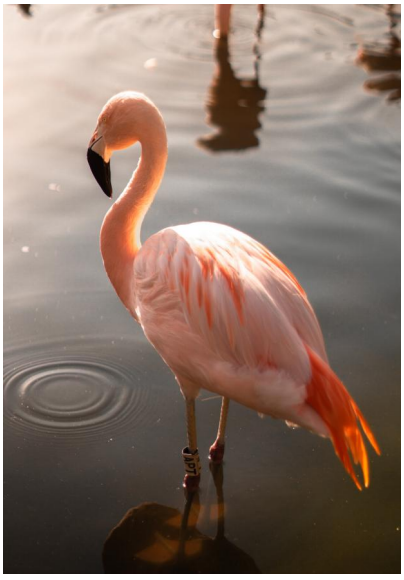
Memory Usage. The reported 16GB memory footprint corresponds to full-frame 2K decoding without tiling. When using 1024×1024 tiled decoding, peak memory consumption decreases to below **7GB** without any observable quality degradation. Employing smaller tiles can further reduce memory usage if necessary. Moreover, applying INT8 quantization lowers memory consumption to approximately **4GB**, making deployment feasible on consumer-grade GPUs.

5. Quantitative Evaluation

Rate-Distortion Curves. In Fig. 3, we present full rate–distortion curves on Kodak [8], CLIC 2020 [20], and DIV2K [1] as a supplement to Fig. 11 of main paper. As discussed in main paper, pixel-level metrics such as PSNR

Caption: "An urban cityscape with towering skyscrapers of varying heights and textures, including glass facades and concrete structures, lining both sides of a busy street. The buildings cast long shadows, creating a dynamic interplay of light and dark. In the foreground, two bright yellow taxis with black numbers and rooftop signs dominate the scene. The taxi on the left displays the number '4K72' on its rooftop sign, while the taxi on the right shows '7V53.' One taxi is sleek and rectangular, while the other is larger and more angular. The street is filled with other vehicles, including a silver SUV, adding to the congestion. The road stretches into the distance, flanked by buildings that narrow the view towards a bright, slightly cloudy sky. A McDonald's sign is visible on the right side, capturing the essence of a bustling, vibrant city."

Caption: "A single vibrant yellow rose with a gradient of orange hues stands gracefully in a dark glass bottle labeled 'Montanaro Vermouth di Torino Rosso,' which is placed on a wooden table. The rose's green leaves are prominently visible, adding a touch of freshness. In the background, a window with white frames reveals a view of a building with a red door and a weathered roof. On the windowsill, there is a clear glass vase holding delicate white dried flowers, alongside a small glass jar. The scene is softly lit by natural light, creating a warm and serene atmosphere, enhanced by a white radiator below the window."



Caption: "A vibrant pink flamingo stands gracefully in shallow water, its long, slender legs partially submerged, creating delicate ripples that spread outward. The flamingo's plumage is a soft gradient of pink and orange, with darker accents near the tips of its feathers. Its neck curves elegantly, leading to a distinctive black-tipped beak. The water reflects the flamingo's form, adding depth to the scene. The sunlight bathes the flamingo in a warm glow, highlighting the texture and color variations in its feathers, capturing a serene and tranquil moment in nature."

Figure 2. Illustrative VLM-generated captions used for semantic conditioning.

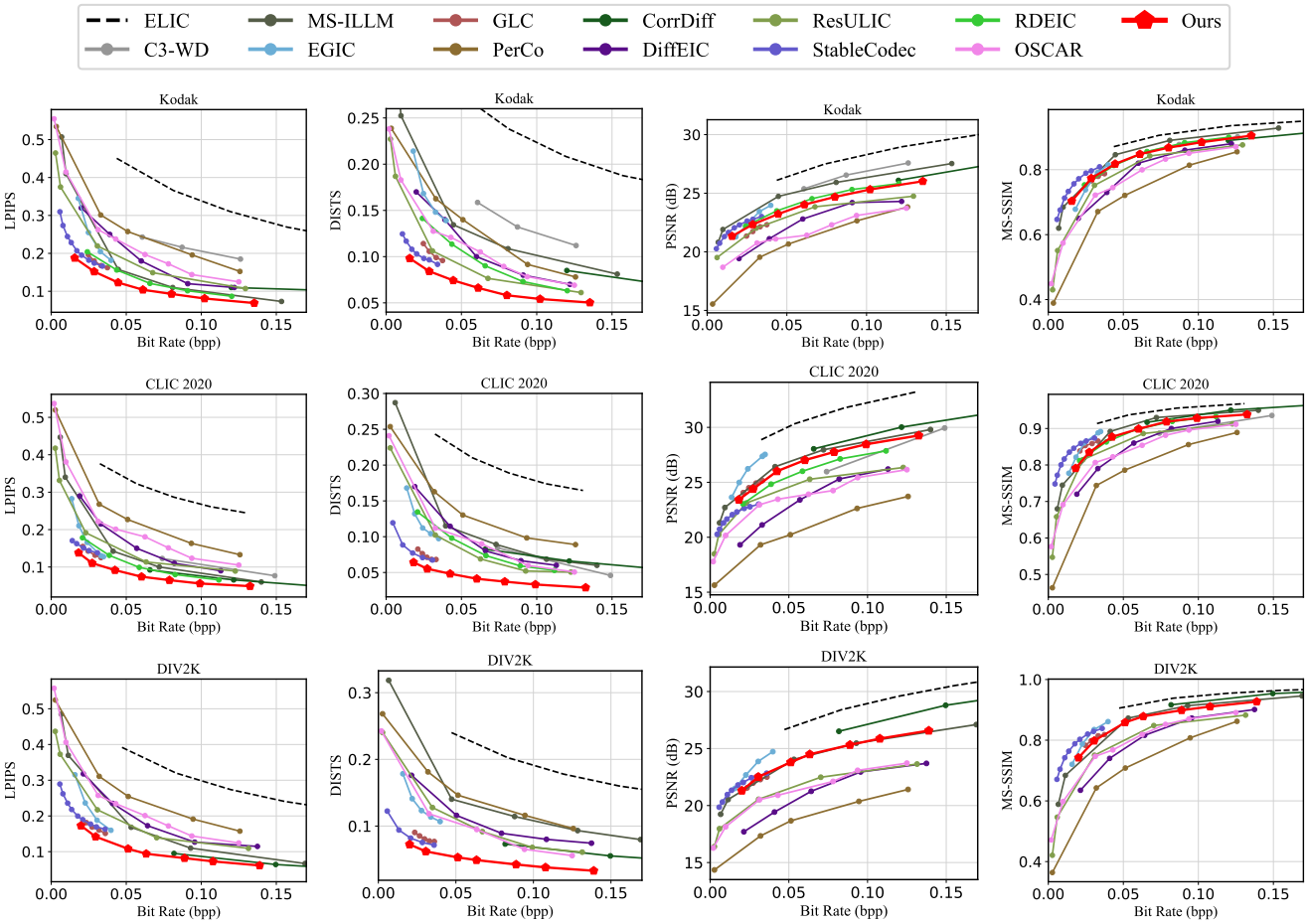


Figure 3. Detailed Rate-distortion-perception curve comparisons of different methods on the Kodak, CLIC2020 and DIV2K dataset.

and MS-SSIM exhibit notable limitations [5, 7, 24]. These metrics primarily emphasize pixel fidelity rather than semantic consistency or perceptual realism, making them less suitable for evaluating compression performance in the ultra–low bitrate regime.

Semantic Study. We further evaluate semantic fidelity using the OCRBench v2 evaluation pipeline [9]. This protocol measures high-level semantic consistency by applying a unified OCR-based recognition framework to reconstructed images and comparing semantic accuracy against ground truth. Unlike pixel-level metrics, this evaluation directly assesses whether compressed reconstructions preserve semantically meaningful content. As shown in Fig. 4 (right), DiT-IC maintains strong semantic consistency, indicating that the perceptual enhancement does not compromise high-level semantic integrity.

User Study. We conduct a large-scale user study with 61 participants to evaluate perceptual realism. Each participant is presented with randomized pairwise comparisons among ResULIC, PerCo, StableCodec, OSCAR, and DiT-IC at matched bitrates, and is asked to select the visually more realistic reconstruction. The aggregated preference scores are $8.2\%, 1.0\%, 27.5\%, 6.5\%, \mathbf{56.8}\%$ for ResULIC, PerCo, StableCodec, OSCAR, and DiT-IC, respectively. DiT-IC receives the highest preference by a substantial margin, demonstrating its clear advantage in perceptual realism under controlled bitrate settings.

Perceptual Evaluation. To provide a comprehensive perceptual assessment beyond pixel-level measures, we additionally report several widely used perceptual metrics, including FID [12], KID [2], NIQE [18], CLIPIQA [21] and MUSIQ [16]. FID and KID measure the distributional discrepancy between reconstructed and reference images in the feature space of pretrained classifiers, serving as holistic indicators of realism. NIQE is a no-reference metric that evaluates natural scene statistics, reflecting perceived image naturalness. CLIPIQA leverages CLIP embeddings to assess semantic fidelity, while MUSIQ is modern deep IQA models designed to capture high-level perceptual quality across diverse content and resolutions.

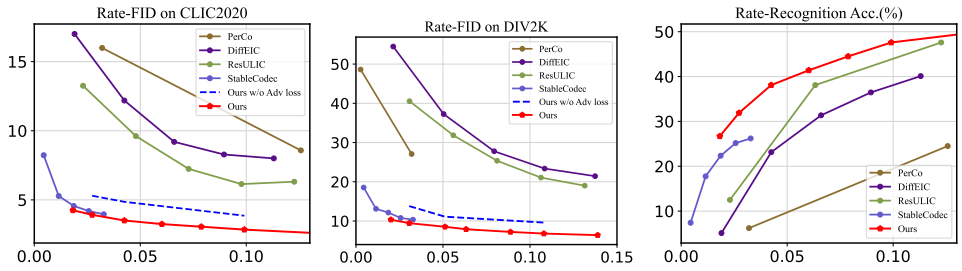


Figure 4. DiT-IC achieves superior FID and Semantic accuracy.

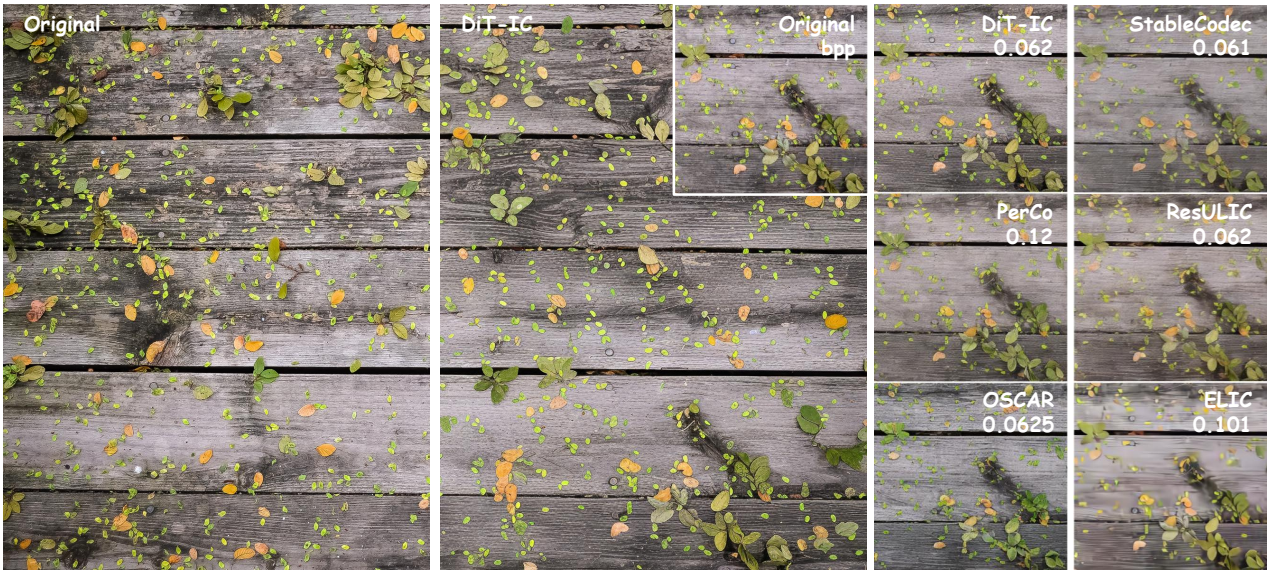


Figure 5. Visual examples and comparisons.

As shown in Table 1, we compare our DiT-IC with the state-of-the-art StableCodec ($\lambda = 2.0$) at similar bitrates (approximately 0.03–0.04 bpp). Due to differences in implementation environments, our reproduced results exhibit minor deviations from the originally reported values. We neglect the FID and KID results on Kodak as it is too small for calculating. DiT-IC achieves consistent improvements across most perceptual metrics on Kodak, DIV2K, and CLIC2020. The only exception is KID on CLIC2020, where StableCodec shows a slight advantage, but DiT-IC maintains overall superior perceptual performance across datasets and metrics..

Fig. 4 further presents FID as a function of bitrate. DiT-IC consistently outperforms prior codecs across operating points. The performance margin on CLIC is smaller than on DIV2K, likely because StableCodec is trained on CLIC, resulting in better dataset alignment. Similar trends are observed for KID. In addition, incorporating adversarial training further enhances perceptual realism, as evidenced by the comparison with *Ours w/o Adv loss* in Fig. 4.

6. Visualization

We provide additional qualitative results and comparisons on high-quality images from DIV2K [1] and CLIC 2020 [20]. We compare our DiT-IC with representative compression models, including StableCodec [24], ELIC [11], PerCo [17], OSCAR [10], and ResULIC[15]. As shown, DiT-IC delivers superior semantic consistency and textural realism while operating at lower bitrates than competing methods.

Figure 6. Visual examples and comparisons.

Figure 7. Visual examples and comparisons.

Figure 8. Visual examples and comparisons.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 2, 5

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 4

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 1

[4] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019. 1

[5] Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2023. 4

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1

[7] 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, 2020. 4

[8] R. Franzen. Kodak lossless true color image suite. http://r0k.us/graphics/kodak/, 1993. Accessed: 2025-11-06. 2

[9] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024. 4

[10] Jinpei Guo, Yifei Ji, Zheng Chen, Kai Liu, Min Liu, Wang Rao, Wenbo Li, Yong Guo, and Yulun Zhang. Oscar: One-step diffusion codec across multiple bit-rates. *arXiv preprint arXiv:2505.16091*, 2025. 5

[11] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5718–5727, 2022. 5

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[13] 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 Computer Vision and Pattern Recognition Conference*, pages 12543–12552, 2025. 1, 2

[14] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36: 24892–24928, 2023. 1

[15] Anle Ke, Xu Zhang, Tong Chen, Ming Lu, Chao Zhou, Jiawen Gu, and Zhan Ma. Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. In *Forty-second International Conference on Machine Learning*, 2025. 5

[16] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 4

[17] Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha Hauke, Daniel Mueller-Gritschneder, and Björn Schuller. Perco (SD): Open perceptual compression. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024. 5

[18] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 4

[19] Xueyan Niu, Deniz Gündüz, Bo Bai, and Wei Han. Conditional rate-distortion-perception trade-off. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 1068–1073. IEEE, 2023. 1

[20] George Toderici, Lucas Theis, Nick Johnston, Eirikur Agustsson, Fabian Mentzer, Johannes Ballé, Wenzhe Shi, and Radu Timofte. Clic 2020: Challenge on learned image compression. *Retrieved March*, 29:2021, 2020. 2, 5

[21] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 4

[22] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1

[23] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 1

[24] Tianyu Zhang, Xin Luo, Li Li, and Dong Liu. Stablecodec: Taming one-step diffusion for extreme image compression. *arXiv preprint arXiv:2506.21977*, 2025. 1, 2, 4, 5