Beyond Pixel Loss: Video-INRs Prefer Perceptual Optimization

Junqi Shi, Wuyang Cong, Ming Lu, Bowei Xu, Zhan Ma
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
Corresponding author: minglu@nju.edu.cn

Abstract

Implicit neural representations (INRs) have recently emerged as a powerful paradigm for video modeling, representing videos as continuous functions parameterized by network weights rather than storing raw pixels or latent codes. However, most existing video-INR methods still rely on pixel-wise supervision (MSE or ℓ_1), which—through the lens of variational inference—implicitly assumes Gaussian or Laplacian reconstruction noise. We show that such assumptions are statistically misaligned with per-video characteristics*, where reconstruction errors are highly structured and temporally correlated in real-world videos. We argue that INRs, by their sequence-specific nature, are inherently better suited to* perceptual *rather than pixel alignment. To validate this perspective, we propose* ***POVI*** *(Perceptually Optimized Video Implicit representation), a perceptually aligned learning framework that shifts INR supervision into multi-level visual feature domains. POVI integrates two complementary perceptual objectives: Multi-Vision Feature Similarity (MVFS) for spatial fidelity and Vision Subject Similarity (VSS) for temporal coherence. Even with a lightweight INR backbone using simple cascaded upsampling, POVI achieves superior perceptual quality compared to state-of-the-art VAE- and diffusion-based codecs, while maintaining real-time decoding at ∼125 FPS on 1080p videos. Our findings reveal that perceptual optimization is not merely a heuristic improvement, but a principled objective shift essential for advancing video-INR representation and reconstruction.*

1. Introduction

Implicit neural representations (INRs) [12, 30, 62, 65, 86] have recently emerged as a powerful paradigm for modeling visual data as continuous functions parameterized by network weights, rather than storing discrete samples. Unlike rasterized latent representations used in variational autoencoders (VAEs) [5, 16, 26, 53] and diffusion models [44, 60], INRs directly map spatiotemporal coordinates to signal values, enabling continuous reconstruction, differentiable processing, and consistent train–test behavior. In video compression, this functional parameterization replaces raw frames with compact network weights, offering a lightweight and interpretable alternative to VAEs. Methods such as NeRV [12] and its extensions [13, 36, 83] show that even simple feed-forward architectures can overfit individual video sequences, achieving high-quality reconstructions and supporting tasks like denoising, inpainting, and interpolation [45, 81].

Despite architectural progress, video-INRs remain constrained by pixel-level optimization. Most existing works still adopt mean squared error (MSE) or ℓ_1 loss [13, 20, 31, 64, 71, 83, 84], sometimes supplemented with SSIM or MS-SSIM [36, 45, 81]. These metrics are largely heuristic, optimized for PSNR, and seldom grounded in statistical principles. For instance, Zhao et al. [83] observed that ℓ_1 tends to preserve texture under small motion while MSE handles large motion better—yet such behavior lacks theoretical justification, and the notion of "large motion" is ill-defined. Meanwhile, many efforts have focused on architectural design [23, 67, 88], sometimes at the expense of efficiency: compact INRs with only 1M parameters can decode slower than VAEs with over 20M parameters [28], and their compression rate–distortion performance still trails advanced VAE codecs [29, 77].

We argue that the bottleneck lies not only in model design, but in the objective itself. From a variational inference viewpoint, pixel-wise losses correspond to fixed likelihood assumptions: MSE implies Gaussian noise, and ℓ_1 implies Laplacian noise [24, 34]. While acceptable in VAEs, where dataset-level statistics are amortized, these assumptions misalign with single-video INRs, where errors are structured, temporally dependent, and sequence-specific. Enforcing such distributions constrains representational flexibility and hinders perceptual fidelity.

This motivates a paradigm shift—**beyond pixel loss**. Instead of measuring reconstruction in raw RGB space, we advocate optimizing INRs in perceptual feature domains, where error statistics are more stable and semantically meaningful. Pretrained vision models provide a natural basis for such perceptual supervision, as they implic-

itly encode hierarchical structures and temporal semantics. Building on this insight, we propose **POVI** (Perceptually Optimized Video Implicit representation), a perceptually aligned learning framework for video-INRs. POVI introduces two complementary objectives: *Multi-Vision Feature Similarity* (MVFS), which aggregates features from multiple pretrained backbones to enhance intra-frame fidelity, and *Vision Subject Similarity* (VSS), which promotes inter-frame temporal coherence by aligning subject-level representations. Even with a lightweight INR backbone using simple cascaded upsampling, POVI surpasses state-of-the-art VAE- and diffusion-based codecs [48, 54, 74] in perceptual quality, while maintaining real-time decoding at ∼125 FPS on 1080p videos.

Our main contributions are as follows:

- We present a new perspective on video-INR optimization, showing that conventional pixel-level losses implicitly impose Gaussian or Laplace error models that are inconsistent with single-video statistics due to strong temporal dependencies and structured content.
- We propose a perceptual supervision framework leveraging pretrained vision models to relax restrictive distributional assumptions and align optimization with perceptual semantics. This includes *Multi-Vision Feature Similarity* (MVFS) for intra-frame fidelity and *Vision Subject Similarity* (VSS) for inter-frame consistency.
- We demonstrate that even simple INRs trained under this perceptual paradigm achieve superior visual quality and temporal consistency, outperforming sophisticated VAE- and diffusion-based codecs while retaining real-time decoding efficiency.

2. Preliminary

Implicit Neural Representations. Implicit neural representations (INRs) [15, 62, 72] model signals as continuous functions that map coordinates to values, instead of storing discrete samples. This functional formulation enables resolution-agnostic reconstruction, consistent train–test behavior, and differentiable signal manipulation. In the context of compression, INRs encode raw data into compact network parameters, which can then be quantized and entropy coded [37].

Originally popularized in 3D scene modeling [50], INRs have been extended to various modalities—images [20, 72], videos [12, 67], audio [38, 65], and hyperspectral data [14, 57]—enabling diverse applications such as super-resolution [3, 78], denoising [59, 73], and inpainting [13]. This generality establishes INRs as a versatile signal representation framework across space, time, and spectrum.

Video-INRs. Video INRs extend this paradigm into the temporal domain by learning a function $\mathcal{F} : [0,1]^{d_{in}} \rightarrow \mathbb{R}^{d_{out}}$ that maps temporal coordinates t to corresponding frames V_t. In practice, a neural decoder $\mathcal{D}$ with cascaded upsampling and nonlinearities, conditioned on temporal embeddings $\mathcal{E}$, reconstructs frames as $\hat{V}_t = \mathcal{D}(\mathcal{E}(t))$.

Recent architectural advances improve efficiency and reconstruction fidelity through techniques such as patch-wise modeling [4, 49], temporal residual embeddings [83], optical-flow compensation [39], hierarchical grids [36], and adaptive backbones [67]. However, these improvements often come with significant decoding cost, approaching that of large VAE-based codecs. In contrast, the role of *optimization objectives*—how INRs are trained—has received comparatively little attention, despite being a fundamental determinant of their representational behavior. We posit that loss design, alongside architecture, plays a central role in unlocking the true potential of video-INRs.

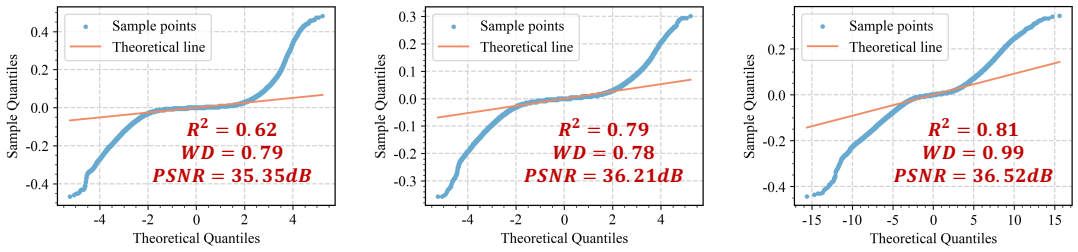
Optimization Objectives for Signal Reconstruction. The training objective defines the implicit assumptions governing signal reconstruction. Most INR-based methods still rely on pixel-level losses such as MSE or ℓ_1, which directly penalize discrepancies in RGB space. While simple and numerically stable, these objectives implicitly correspond to Gaussian or Laplacian error models (see Sec. 3.1), assumptions that are often violated in real-world videos exhibiting structured, temporally correlated residuals. Such mismatches can bias optimization, limiting both representational flexibility and perceptual fidelity.

A large body of research in generative modeling has shown that reducing pixel distortion does not necessarily improve perceptual quality [10, 22]. To address this, perceptual losses were introduced, comparing images in learned feature spaces or via adversarial discriminators. Representative metrics such as LPIPS [79], DISTS [19], and GAN-based objectives [18, 25] emphasize semantic and structural similarity over raw pixel accuracy.

In contrast, perceptual optimization within the INR framework remains largely underexplored. Ballé et al. [6] introduced a perceptual objective for image INRs via Wasserstein Distortion (WD) [52], achieving improved visual quality over conventional pixel-based training. However, its extension to video INRs—where temporal consistency, structured motion, and perceptual coherence are inherently intertwined—has not been systematically studied. This gap motivates our work: we show that video-INRs, by nature of their sequence-specific and continuous representation, are *especially well-suited* to perceptual supervision. Even lightweight architectures, when optimized in feature space, can achieve state-of-the-art perceptual quality and real-time decoding efficiency.

3. Method

We revisit the training of video-INRs through the lens of variational inference, revealing that pixel-level supervision is fundamentally misaligned with the single-video optimization setting (Sec. 3.1). Building on this insight,



(a) L1 - Gaussian **(b) L2 - Gaussian** **(c) L1+SSIM - Laplace**

Figure 1. Q–Q (Quantile–Quantile) plots of reconstruction errors trained with different pixel-level loss functions on a sample from YouHQ [87]. Each plot compares empirical error distributions with their theoretical counterparts (Gaussian or Laplace). $R^2 \uparrow$ measures the linear alignment with the reference distribution ($R^2 = 1$ indicates perfect fit, though tail deviations may be underestimated). $WD \downarrow$ denotes the Wasserstein Distance [52], where larger values indicate stronger distributional mismatch. Perfect distributional alignment would result in points lying along the diagonal reference line.

we argue that INRs naturally favor perceptual optimization (Sec. 3.2), and propose two complementary feature-domain objectives that enhance both intra-frame fidelity and inter-frame coherence.

3.1. Revisiting Pixel-wise Losses: A Variational Perspective

Video-INRs under a Variational Framework. Training a video-INR can be formulated as a rate–distortion optimization problem, where the *rate* measures model complexity (the number of bits required to encode network parameters), and the *distortion* quantifies reconstruction fidelity. From a variational inference view, this corresponds to approximating the true posterior $p_{\tilde{\boldsymbol{w}}|\boldsymbol{x}}(\tilde{\boldsymbol{w}}|\boldsymbol{x})$ with a variational density $q(\tilde{\boldsymbol{w}}|\boldsymbol{x})$ by minimizing the expected KL divergence over the data distribution $p_{\boldsymbol{x}}$ [5, 37, 58]:

$$\mathbb{E}_{\boldsymbol{x}} D_{KL}[q||p_{\tilde{\boldsymbol{w}}|\boldsymbol{x}}] = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}\mathbb{E}_{\tilde{\boldsymbol{w}}\sim q}\Big[\underbrace{-\log p_{\boldsymbol{x}|\tilde{\boldsymbol{w}}}(\boldsymbol{x}|\tilde{\boldsymbol{w}})}_{\text{Distortion } \mathcal{L}_D} \quad (1)$$

$$\underbrace{-\log p_{\tilde{\boldsymbol{w}}}(\tilde{\boldsymbol{w}})}_{\text{Rate } \mathcal{L}_R} + \log p_{\boldsymbol{x}}(\boldsymbol{x})\Big], \quad (2)$$

where $\log p_{\boldsymbol{x}}(\boldsymbol{x})$ is constant, reducing the objective to the canonical rate–distortion trade-off.

While prior works explore different priors for $p(\tilde{\boldsymbol{w}})$—e.g., uniform [82], Gaussian [37, 81], or Laplacian [31, 40]—our focus lies on the distortion likelihood $p(\boldsymbol{x}|\tilde{\boldsymbol{w}})$, as it directly defines the supervision space for INR learning.

Distributional Assumptions in Pixel-wise Losses. Pixel-wise losses inherently impose specific probabilistic assumptions on the reconstruction error distribution. Minimizing an ℓ_p loss is equivalent to performing maximum likelihood estimation (MLE) under a generalized Gaussian distribution (GGD) [21] with shape parameter $\beta = p$:

$$p(e) = \frac{p}{2\alpha\Gamma(1/p)} \exp\left(-\left|\frac{e}{\alpha}\right|^p\right), \quad (3)$$

where $e = \boldsymbol{x} - \tilde{\boldsymbol{x}}$ denotes the reconstruction error, α is a scale parameter controlling the spread of the distribution, and $\Gamma(\cdot)$ is the Gamma function ensuring proper normalization.

Common special cases include: (a) $p = 2$, Gaussian (ℓ_2, MSE); (b) $p = 1$, Laplacian (ℓ_1); (c) $p < 1$, yielding heavy-tailed distributions that emphasize outliers; and (d) $p > 2$, producing sharp-peaked distributions that strongly penalize small deviations. For instance, MSE minimization corresponds to Gaussian MLE:

$$\min \tfrac{1}{2\sigma^2}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2^2 \quad \Leftrightarrow \quad \max \log \mathcal{N}(\boldsymbol{x}|\tilde{\boldsymbol{x}}, \sigma^2), \quad (4)$$

while ℓ_1 minimization corresponds to Laplacian MLE:

$$\min \tfrac{1}{b}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_1 \quad \Leftrightarrow \quad \max \log \text{Laplace}(\boldsymbol{x}|\tilde{\boldsymbol{x}}, b). \quad (5)$$

Hence, choosing a pixel-wise loss effectively fixes a parametric form of reconstruction noise. However, for video-INRs—trained per sequence rather than across datasets—these assumptions break down. Reconstruction errors are structured, temporally dependent, and video-specific: high-motion sequences produce heavy tails, while static scenes yield more concentrated errors. A single global noise model cannot faithfully capture these dynamics. By contrast, VAE-based codecs benefit from population-level amortization across datasets, making Gaussian or Laplacian assumptions more reasonable and training more stable. We provide more analysis in supplementary materials.

Empirical Evidence. Fig. 1 presents Q–Q plots of reconstruction errors. Within a narrow central range ($[-0.1, 0.1]$), empirical errors moderately fit Gaussian/Laplacian distributions. However, extreme errors exhibit pronounced heavy tails, with Wasserstein Distance (WD) > 0.7, highlighting systematic mismatch.

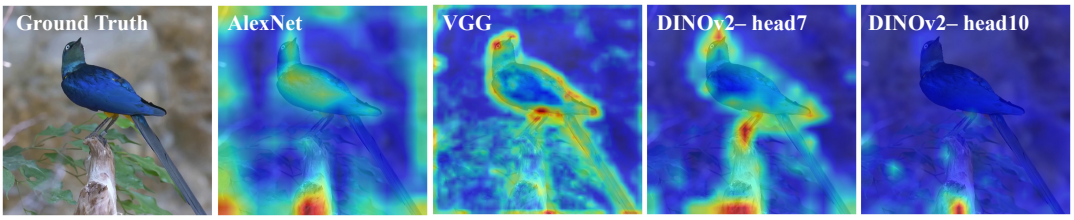


Figure 2. Illustrative heatmaps showing the feature sensitivity of different pretrained vision models on a sample sequence from YouHQ [87]. From left to right: AlexNet [35], VGG [61] and DINOv2 [51].

Interestingly, we observe that pixel-wise losses suffer from poor distributional matching, which directly leads to lower reconstruction fidelity (e.g., PSNR drops from 36.21 dB to 35.35 dB as the fit deteriorates). This mismatch highlights the inherent limitation of MSE/ℓ_1 losses as statistically inconsistent error models. Prior studies [32, 36, 81] similarly report that even simple alternatives such as ℓ_1+SSIM can outperform pure MSE—even under MSE-based evaluation. Fig. 1(c) also echoes this trend.

Remark 1. *Pixel-wise supervision enforces rigid, dataset-agnostic error assumptions (e.g., Gaussian or Laplacian) that are systematically violated in single-video INRs. Empirical evidence reveals that reconstruction errors are video-dependent, heavy-tailed, and temporally correlated. These findings motivate the search for alternative representational spaces, where error statistics are more stable and better aligned with the optimization objective.*

3.2. Turn to Perceptual Optimization

Pretrained Vision Models as Transformation. The preceding analysis reveals a key challenge for video-specific INRs: pixel-space errors exhibit video-dependent, heavy-tailed statistics that violate Gaussian or Laplacian assumptions. Hence, an effective learning objective should operate in a space where reconstruction errors are statistically stable. Designing such a transformation analytically is intractable due to strong temporal correlations and the lack of dataset-level averaging. As a practical solution, we turn to pretrained vision models as empirical transformation functions, naturally motivating a perceptual optimization paradigm.

Perceptual objectives measure discrepancies in deep feature spaces rather than raw RGB pixels, capturing texture, structure, and semantics that better correlate with human visual perception. Representative examples include LPIPS [79], which measures feature-space Euclidean distance, and DISTS [19], which combines structural similarity with learned features.

From a probabilistic viewpoint, these objectives correspond to implicit likelihood assumptions in feature space—where pretrained representations, learned over large-scale natural image distributions, provide more stable priors than any single video can offer. Thus, perceptual optimization inherently relaxes the overly restrictive distributional constraints of pixel-level supervision, making it particularly suitable for sequence-specific INR training.

Multi-Vision Feature Similarity (MVFS). On the other hand, the pixel-to-feature mapping induced by pretrained vision models is highly nonlinear and analytically intractable. Consequently, assuming Gaussian errors in feature space does not directly translate to a corresponding distributional constraint in pixel space; such equivalence would only hold for a linear transformation. Nevertheless, even as a loose or indirect constraint, relying on a single pretrained model may bias the optimization toward specific patterns captured by that model (Fig. 2). To mitigate this potential bias and enhance generalization, we employ multiple pretrained vision models and aggregate their feature-based losses, thereby reducing the suboptimality introduced by any single inductive bias or distributional assumption. Supplementary materials provide a more detailed comparison of visual models.

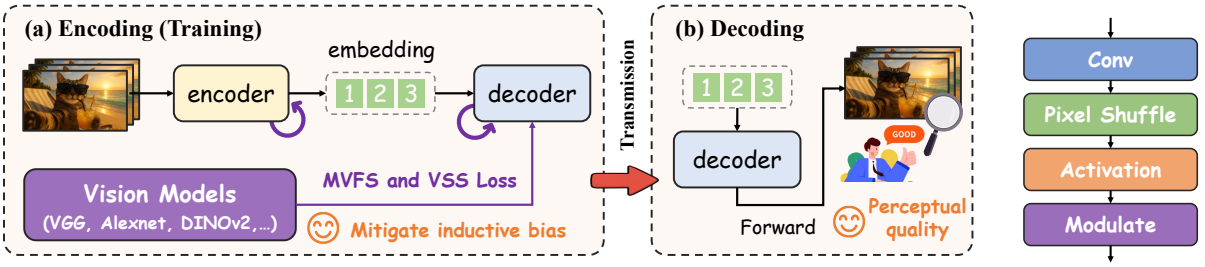
Formally, let $\phi_m^l(\cdot)$ denote the activation of layer l of the m-th pretrained vision model. Given reconstructed frames $\tilde{\boldsymbol{x}}$ and ground-truth frames $\boldsymbol{x}$, the **MVFS** loss aggregates multi-level feature distances:

$$\mathcal{L}_{\text{MVFS}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_{m=1}^{M} w_m \sum_{l \in \mathcal{A}_m} d\Big(\phi_m^l(\boldsymbol{x}), \phi_m^l(\tilde{\boldsymbol{x}})\Big), \quad (6)$$

where $d(\cdot,\cdot)$ denotes LPIPS- or DISTS-style distances, $\mathcal{A}_m$ is the layer set for model m, w_m is a model-specific weight, and M is the total number of vision models. Empirically, this multi-model fusion improves both spatial fidelity and perceptual stability across diverse content types.

Vision Subject Similarity (VSS). While perceptual supervision improves spatial realism, it may inadvertently introduce temporal inconsistency—manifesting as flickering or identity drift across frames. To address this, inspired by temporal consistency assessment techniques [27], we introduce Vision Subject Similarity that enforces temporal coherence at the subject level.

We employ DINOv2 [51] to extract features that capture



(a) Overview of the proposed POVI framework. (b) Upsampling layer.

Figure 3. **Left:** Overview of the proposed POVI framework. The encoding is equivalent to training the network, where the video signal is parameterized as a neural function. During decoding, only a forward pass through the trained decoder is required for reconstruction. **Right:** Upsampling layer. Detailed network specifications are provided in supplementary materials.

subject identity. Unlike standard classification networks, which are trained to collapse intra-class variations, DINOv2 is self-supervised and produces representations that are both semantically meaningful and sensitive to subtle identity differences [69]. Let f_t denote the DINOv2 feature of frame t; we enforce cross-frame consistency within a local temporal window $\mathring{U}(t, \delta)$:

$$\mathcal{L}_{\text{VSS}} = \sum_{i \in \mathring{U}(t,\delta)} \text{dist}(f_i, f_t), \tag{7}$$

where $\text{dist}(\cdot, \cdot)$ is a mariginal cosine feature distance. This loss constrains global identity coherence, complementing MVFS's frame-wise perceptual fidelity.

DINOv2 is utilized in both MVFS and VSS. In our implementation, the feature f of each reconstructed frame is computed only once per epoch and cached in a buffer for reuse. Therefore, the additional computational overhead introduced by VSS is negligible.

Overall Pipeline. The full framework is illustrated in Fig. 3a. Encoding corresponds to training the INR decoder $\mathcal{D}$ to fit a specific video, with frames parameterized as $\hat{V}_t = \mathcal{D}(\mathcal{E}(t))$. During decoding, reconstruction requires only a single forward pass. Our decoder adopts a lightweight cascaded upsampling design (Fig. 3b) to ensure high efficiency. For transmission, only the embeddings and decoder weights are quantized and entropy-coded. The overall objective integrates pixel-level and perceptual losses:

$$\begin{aligned}\mathcal{L}(\phi) = \lambda_1 \mathcal{L}_{\ell_1} + \lambda_2 \mathcal{L}_{\text{SSIM}} \\ + \lambda_3 \mathcal{L}_{\text{MVFS}} + \lambda_4 \mathcal{L}_{\text{VSS}} + \lambda_5 \mathcal{L}_{\text{GAN}},\end{aligned} \tag{8}$$

where, following successful practices in image and video reconstruction [63, 75, 81], ℓ_1 and SSIM [70] losses are included as regularization terms to stabilize training and preserve low-level fidelity, while MVFS and VSS drive perceptual alignment across space and time.

Despite its architectural simplicity—without motion prediction, patch-wise modeling, or hierarchical grids—our approach achieves superior perceptual quality and temporal stability compared to state-of-the-art VAE- and diffusion-based codecs. This confirms our central claim: *perceptual optimization is a principled and efficient learning objective for advancing video-INR representations.*

4. Experiment

Compared Methods. We compare our approach with representative perceptual-optimized video compression methods from three paradigms: (1) *GAN-based*: PLVC [74], which enhances perceptual quality through adversarial learning; (2) *VAE-based*: GLC [54], which introduces perceptual coding in the generative latent space and achieves state-of-the-art perceptual fidelity, and DVC-P [80], an early exploration of perceptual representation; (3) *Diffusion-based*: DiffVC [48], a recent state-of-the-art approach leveraging inter-frame diffusion priors for perceptually faithful reconstructions.

For comparison with pixel-optimized codecs, we include strong neural baselines such as DCVC [41], DHVC 2.0 [47], DCVC-FM [43], DCVC-RT [28], NeuroQuant [58], HNeRV [13], and NVRC [37], as well as conventional codecs H.265/HEVC [66] and H.266/VVC [11] for reference.

Test Datasets. We evaluate on two benchmarks. UVG[1] contains seven 1920 × 1080 videos at 120 FPS, 300–600 frames each, and serves as a standard compression benchmark. YouHQ [87] comprises ∼37K high-definition (1080 × 1920) YouTube clips covering diverse scenes—streets, landscapes, animals, faces, night scenes. For ablations, we select 10 sequences and center-crop to 960 × 960 for efficiency. UVG provides standard evaluation, while YouHQ offers a diverse, lightweight testbed.

Training and Implementation. We train models using the Adam optimizer [33] with a batch size of 1 and an ini-

[1]Beauty, Bosphorus, HoneyBee, Jockey, ReadySetGo, ShakeNDry, YachtRide

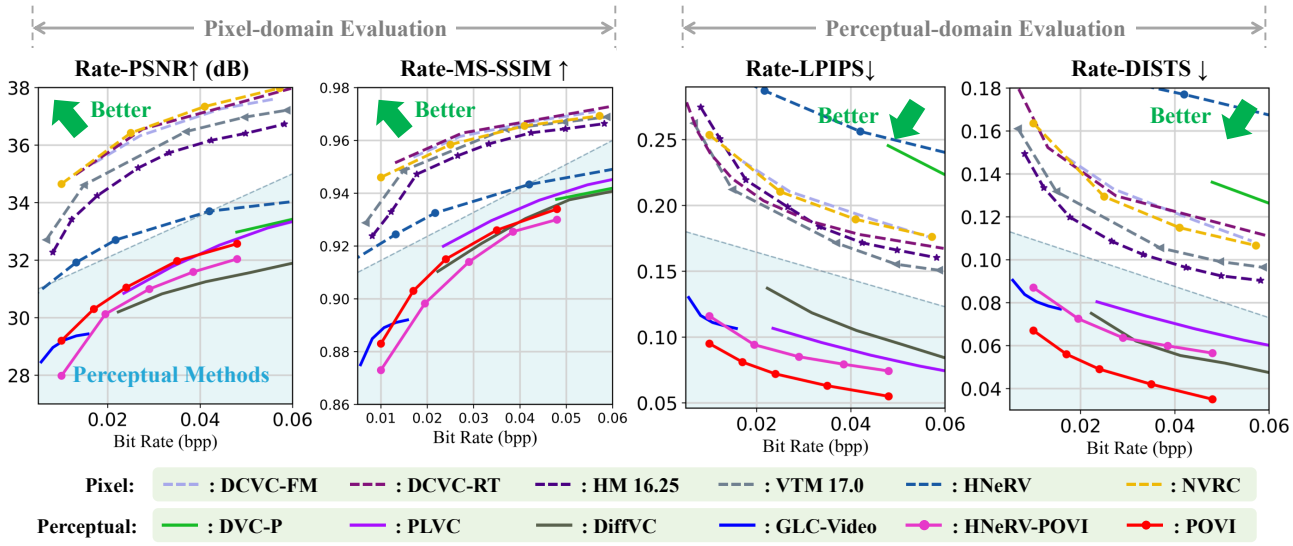


Figure 4. Rate–distortion performance on the UVG dataset. Solid lines represent perceptual-optimized methods, while dashed lines indicate pixel-optimized codecs. Our method achieves the highest perceptual quality across bitrates. The proposed POVI framework is further validated on other video-INR architectures such as HNeRV.

tial learning rate of 1.5×10^{-4}, scheduled by warm-up cosine annealing. Unless specified otherwise, training is conducted for 150K iterations on UVG and 15K on YouHQ. Quantization-aware training with a 7-bit precision and the straight-through estimator (STE) [7, 76] is applied. Downsampling uses Conv2d with stride patterns $(5, 4, 2, 2, 2)$ for 960/1920-pixel inputs and $(5, 3, 3, 2, 2)$ for 1080-pixel inputs; other resolutions are adjusted adaptively. The loss coefficients in Eq. 8 are set to $\lambda_1 \in [1, 3]$, $\lambda_2 \in [0.3, 1]$, $\lambda_3 = 1.0$, $\lambda_4 = 0.1$, and $\lambda_5 = 0.1$. All experiments run on PyTorch with NVIDIA RTX A6000 and 4090 GPUs.

Evaluation Protocol. For fairness, we report official results from published papers when source code is unavailable or outdated (e.g., PLVC). For methods with well-maintained implementations, such as HNeRV, we reproduce results under our unified environment to ensure consistency.

4.1. Main Results

Frame-level Evaluation. We evaluate reconstruction quality using four widely adopted metrics: PSNR and MS-SSIM for pixel-level fidelity, and LPIPS and DISTS for perceptual similarity. As shown in Fig. 4, our method consistently achieves superior perceptual scores (LPIPS, DISTS) while maintaining competitive distortion performance against other perceptual-optimized baselines. Table 1 further reports the corresponding BD-metrics [9], summarizing average performance differences across bitrates. Slight interpolation discrepancies may arise due to variations in reported bitrates across prior works, but the overall trends remain consistent.

Notably, our approach achieves state-of-the-art perceptual quality while maintaining high efficiency, decoding 1080p videos at ∼75 FPS (FP32) and ∼125 FPS (FP16) on UVG sequences under serial decoding—well above real-time playback requirements. A key advantage of INR-based codecs is their fully parallel decoding, where all frames can, in principle, be reconstructed in a single forward pass,

Table 1. Frame-level BD-metrics [9] on the UVG dataset using HM as the anchor. FPS indicates decoding efficiency (FP16 inference) measured on an NVIDIA RTX 4090. Symbols ↑ / ↓ denote that higher/lower values are better. HNeRV-POVI refers to applying the proposed POVI optimization to the HNeRV architecture. We adopt the HiNeRV [36] backbone from NeuroQuant [58]. The best results under perceptual optimization are highlighted in **bold**, and our method is additionally marked with a gray background.

Methods	Type	PSNR↑	MS-SSIM↑	LPIPS↓	DISTS↓	FPS↑
HandCraft						
HM 16.25 [1]	H.265	0.	0.	0.	0.	∼40
VTM 17.0 [2]	H.266	0.68	0.006	-0.017	-0.056	∼25
Pixel Optimization						
DHVC 2.0 [47]	HVAE	0.50	0.005	/	/	<8
DCVC-FM [43]	VAE	1.06	0.006	0.015	0.024	<5
DCVC-RT [28]	VAE	1.18	**0.007**	**-0.007**	0.022	∼105
HNeRV [13]	INR	-2.00	-0.016	0.07	0.069	**∼165**
NVRC [37]	INR	**1.25**	0.006	0.007	**0.020**	∼15
NeuroQuant [58]	INR	0.43	0.001	0.025	0.031	∼60
Perceptual Optimization						
PLVC [74]	GAN	-3.80	**-0.026**	-0.088	-0.029	/
GLC [54]	VAE	-3.92	-0.043	-0.123	-0.056	< 5
DVC-P [80]	VAE	**-3.37**	-0.026	0.071	0.040	/
DiffVC [48]	Diff.	-4.80	-0.034	-0.069	-0.041	<0.1
HNeRV-POVI	INR	-4.4±0.2	-0.05±0.01	-0.11±0.01	-0.044±0.01	**165**±5
Ours-POVI	INR	-3.8±0.3	-0.04±0.01	**-0.13**±0.01	**-0.064**±0.01	**125**±5

Table 2. Sequence-level evaluation on UVG using VBench [27].

Methods	Architecture	Perceptual Opt.	Bpp↓	Subject Consistency↑	Background Consistency↑	Motion Smoothness↑	Average Score↑
DCVC-FM [43]	VAE		0.051	84.5%	**90.6%**	**99.2%**	91.4%
DCVC [41]	VAE		0.025	**85.5%**	90.0%	99.1%	**91.5%**
DVC-P [80]	VAE	✓	0.048	83.7%	90.1%	98.9%	90.9%
HNeRV [13]	INR		0.010	80.1%	88.3%	98.5%	89.0%
NVRC [37]	INR		0.010	82.8%	88.5%	98.9%	90.1%
NeuroQuant [58]	INR		0.010	82.3%	88.2%	98.7%	89.7%
Ours	INR	✓	0.010	**84.5**±1.5%	**89.4**±1.0%	**99.0**±0.5%	**91.0**±1.8%
Empirical Min	/	/	/	14.62%	26.15%	70.60%	/
Empirical Max	/	/	/	100%	100%	99.75%	/

Table 3. Encoding complexity, measured as the number of training steps per second achieved on 1080p inputs. All results are evaluated on an NVIDIA RTX 4090 GPU with FP16 precision.

Methods / bpp	0.01	0.03	0.05	Avg.
HNeRV [13]	41.15	30.21	25.00	32.12
NeuroQuant [58]	19.40	10.40	6.00	11.93
NVRC [37]	6.40	3.60	2.20	4.07
(V0) ℓ_1 + SSIM (baseline)	30.61	21.15	14.00	21.92
(V1) w/ DISTS [19]	9.46	8.47	7.07	8.33
(V2) w/ MVFS	7.50	6.88	6.00	6.79
(V3) w/ MVFS + VSS	7.31	6.70	5.85	6.62
(V4) w/ MVFS + VSS + GAN	5.35	4.93	4.33	4.87

avoiding the inter-frame dependencies that limit conventional VAE-based methods.

Sequence-level Evaluation. Frame-level metrics are widely used but fail to capture temporal dynamics in videos. VAE-based codecs often show frame-wise quality fluctuations due to inter-frame dependencies (e.g., hierarchical quality control [42]) [17], which are hidden under conventional evaluations. This issue is amplified under perceptual optimization, where viewers are sensitive to temporal inconsistencies.

Recent studies [27, 85] have demonstrated that commonly used video-level metrics—such as Inception Score (IS) [56], Fréchet Video Distance (FVD) [68], and CLIP-SIM [55]—often correlate poorly with subjective perception. To address this, we adopt the recently proposed VBench [27], which offers a more reliable and fine-grained evaluation of perceptual video quality at the sequence level.

On UVG (Table 2), our method achieves strong temporal consistency at a lower bitrate compared with VAE- and INR-based baselines. We recommend combining frame- and sequence-level metrics for a more comprehensive assessment. Additional definitions and results are provided in the supplementary materials.

4.2. Deep Dive

Encoding Complexity. As shown in Table 3, our lightweight variant is only marginally slower than HNeRV under the same pixel-wise training regime (V0). Introducing perceptual supervision increases complexity mainly due to the additional vision model—a cost inherent to any perceptual optimization rather than specific to our framework (V0 vs. V1). Compared with DISTS, the extra overhead brought by POVI remains modest (V1 vs. V2), as DINOv2 operates on downsampled inputs and VSS reuses cached features (Sec. 3.2), contributing negligible computational cost (V2 vs. V3).

Although INR-based approaches generally incur higher encoding latency—making real-time streaming challenging [8]—they are well-suited for video-on-demand (VOD) and large-scale storage applications, where decoding throughput dominates overall system efficiency [46].

Loss Ablation. Table 4 reports results on YouHQ [87]. Among pixel-wise losses, ℓ_1+SSIM yields the best PSNR/MS-SSIM but still suffers from high LPIPS/DISTS, showing limited perceptual fidelity. Adding perceptual losses (LPIPS or DISTS) significantly reduces these distances, supporting our key finding that *INRs inherently benefit from perceptual optimization*. A single VFS (VGG) introduces trade-offs across perceptual metrics, while our MVFS aggregation alleviates the bias of any single feature extractor. With VSS enforcing temporal consistency, our full model achieves the best overall performance, reaching LPIPS 0.019 and DISTS 0.0085.

Perceptual Visualization. Qualitative comparisons with HNeRV (INR), DCVC-FM (VAE), and VVC (traditional codec) are shown in Fig. 5. Our method produces sharper reconstructions and preserves richer textures even under extreme compression, whereas competing approaches often suffer from oversmoothing or perceptual artifacts. More visualizations are provided in the supplementary materials.

Convergence. As illustrated in Fig. 6, video-INRs trained under the POVI framework exhibit a stable convergence pattern. Notably, good perceptual quality emerges

Table 4. Loss ablation on subset from YouHQ [87].

	Pixel Optimization			Perceptual Optimization				
	MSE	ℓ_1	ℓ_1+SSIM	w/ LPIPS [79]	w/ DISTS [19]	w/ VFS-VGG	w/ MVFS	w/ MVFS &VSS
PSNR	36.21	35.34	**36.51**	34.45	34.32	34.41	34.61	34.75
MS-SSIM	0.9742	0.9773	**0.9824**	0.9752	0.9758	0.9756	0.9774	0.9782
LPIPS	0.1630	0.1702	0.1496	0.0241	0.0298	0.0284	0.0225	**0.0190**
DISTS	0.1362	0.1414	0.1235	0.0185	0.0130	0.0152	0.0103	**0.0085**

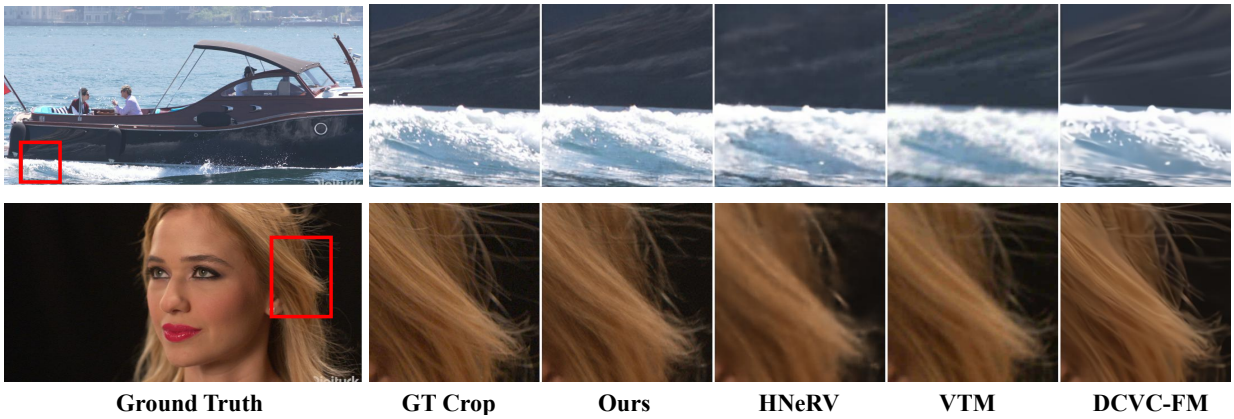


Figure 5. Visual comparison on a UVG sequence at 0.014 bpp ($\sim$ 1700$\times$ compression).

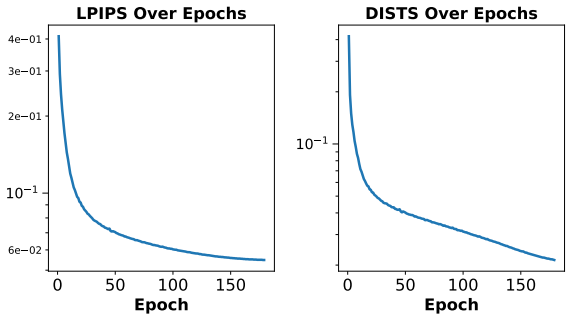


Figure 6. Loss convergence behavior.

within the first 50 epochs, offering a favorable trade-off between performance and encoding cost. Extending training further continues to yield measurable improvements though with increased runtime. This behavior demonstrates that POVI enables efficient optimization while remaining scalable to longer training schedules.

4.3. Discussion and Future Work

To maintain fast decoding, we adopt a lightweight feed-forward architecture without complicated modules, which naturally limits the achievable rate–distortion tradeoff. Future work could explore hierarchical grids [36], context modeling [37, 81], or learned quantization [58] to further improve efficiency. Perceptual optimization, by nature, increases training complexity due to vision-model feature supervision. This is a common trait of perceptual methods. Developing lightweight perceptual surrogates is an important direction for more efficient optimization.

5. Conclusion

We revisited the optimization of video-INRs through the lens of variational inference and highlighted a key limitation of conventional pixel-wise losses: their simplistic error models are statistically misaligned with per-video reconstruction, where errors are structured and temporally correlated. To address this, we propose **POVI**, which shifts supervision from pixels to feature spaces using multiple pre-trained vision models. POVI combines MVFS for spatial fidelity and VSS for temporal coherence, relaxing restrictive pixel-domain assumptions and aligning optimization with perceptual semantics. Experiments show that even with a lightweight cascaded-upsampling INR, POVI substantially improves perceptual quality over state-of-the-art VAE- and diffusion-based methods, while enabling real-time decoding at $\sim$125 FPS on 1080p videos. These results demonstrate that perceptual optimization is not a heuristic but a principled objective shift, advancing video-INR representation beyond pixel loss.

Acknowledgments

This work was supported in part by Natural Science Foundation of China (Grant No. 62401251, 62431011) and Natural Science Foundation of Jiangsu Province (Grant No. BK20241226, BK20243038). The authors would like to express their sincere gratitude to the Interdisciplinary Research Center for Future Intelligent Chips (Chip-X) and Yachen Foundation for their invaluable support.

References

[1] HM-16.25: HEVC Test Model Reference Software. https://vcgit.hhi.fraunhofer.de/jvet/HM/. 6

[2] VTM-17.0: VVC Test Model Reference Software. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM. 6

[3] Mary Aiyetigbo, Wanqi Yuan, Feng Luo, and Nianyi Li. Implicit neural representation for video and image super-resolution. *arXiv preprint arXiv:2503.04665*, 2025. 2

[4] Yunpeng Bai, Chao Dong, Cairong Wang, and Chun Yuan. Ps-nerv: Patch-wise stylized neural representations for videos. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 41–45. IEEE, 2023. 2

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. 1, 3

[6] Jona Ballé, Luca Versari, Emilien Dupont, Hyunjik Kim, and Matthias Bauer. Good, cheap, and fast: Overfitted image compression with wasserstein distortion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23259–23268, 2025. 2

[7] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 6

[8] Abdelhak Bentaleb, May Lim, Mehmet N Akcay, Ali C Begen, Sarra Hammoudi, and Roger Zimmermann. Toward one-second latency: Evolution of live media streaming. *IEEE Communications Surveys & Tutorials*, 2025. 7

[9] Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *ITU-T SG16, Doc. VCEG-M33*, 2001. 6

[10] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019. 2

[11] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 5

[12] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems*, 34:21557–21568, 2021. 1, 2

[13] Hao Chen, Matthew Gwilliam, Ser-Nam Lim, and Abhinav Shrivastava. Hnerv: A hybrid neural representation for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10270–10279, 2023. 1, 2, 5, 6, 7

[14] Huan Chen, Wangcai Zhao, Tingfa Xu, Guokai Shi, Shiyun Zhou, Peifu Liu, and Jianan Li. Spectral-wise implicit neural representation for hyperspectral image reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3714–3727, 2023. 2

[15] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 2

[16] Wuyang Cong, Junqi Shi, Ming Lu, Xu Zhang, and Zhan Ma. Taming hierarchical image coding optimization: A spectral regularization perspective. In *The Fourteenth International Conference on Learning Representations*, 2026. 1

[17] Wuyang Cong, Junqi Shi, Lizhong Wang, Weijing Shi, Ming Lu, Hao Chen, and Zhan Ma. Reinforced rate control for neural video compression via inter-frame rate-distortion awareness. *arXiv preprint arXiv:2601.19293*, 2026. 7

[18] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 2

[19] 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. 2, 4, 7, 8

[20] Emilien Dupont, Adam Goliński, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. Coin: Compression with implicit neural representations. *arXiv preprint arXiv:2103.03123*, 2021. 1, 2

[21] Alex Dytso, Ronit Bustin, H Vincent Poor, and Shlomo Shamai. Analytical properties of generalized gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(1):6, 2018. 3

[22] Dror Freirich, Tomer Michaeli, and Ron Meir. A theory of the distortion-perception tradeoff in wasserstein space. *Advances in Neural Information Processing Systems*, 34: 25661–25672, 2021. 2

[23] Ge Gao, Ho Man Kwan, Fan Zhang, and David Bull. Pnvc: Towards practical inr-based video compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3068–3076, 2025. 1

[24] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016. 1

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[26] Amirhossein Habibian, Ties van Rozendaal, Jakub M Tomczak, and Taco S Cohen. Video compression with rate-distortion autoencoders. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7033–7042, 2019. 1

[27] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 4, 7

[28] 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, 5, 6

[29] Wei Jiang, Junru Li, Kai Zhang, and Li Zhang. Ecvc: Exploiting non-local correlations in multiple frames for contextual video compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7331–7341, 2025. 1

[30] Alper Kayabasi, Anil Kumar Vadathya, Guha Balakrishnan, and Vishwanath Saragadam. Bias for action: Video implicit neural representations with bias modulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27999–28008, 2025. 1

[31] Hyunjik Kim, Matthias Bauer, Lucas Theis, Jonathan Richard Schwarz, and Emilien Dupont. C3: High-performance and low-complexity neural compression from a single image or video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9347–9358, 2024. 1, 3

[32] Jina Kim, Jihoo Lee, and Je-Won Kang. Snerv: Spectra-preserving neural representation for video. In *European Conference on Computer Vision*, pages 332–348. Springer, 2024. 4

[33] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 4

[36] Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. Hinerv: Video compression with hierarchical encoding-based neural representation. *Advances in Neural Information Processing Systems*, 36:72692–72704, 2023. 1, 2, 4, 6, 8

[37] Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. Nvrc: Neural video representation compression. *Advances in Neural Information Processing Systems*, 37:132440–132462, 2024. 2, 3, 5, 6, 7, 8

[38] Luca A Lanzendörfer and Roger Wattenhofer. Siamese siren: Audio compression with implicit neural representations. *arXiv preprint arXiv:2306.12957*, 2023. 2

[39] Joo Chan Lee, Daniel Rho, Jong Hwan Ko, and Eunbyung Park. Ffnerv: Flow-guided frame-wise neural representations for videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7859–7870, 2023. 2

[40] Thomas Leguay, Théo Ladune, Pierrick Philippe, and Olivier Déforges. Cool-chic video: Learned video coding with 800 parameters. In *2024 Data Compression Conference (DCC)*, pages 23–32. IEEE, 2024. 3

[41] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34:18114–18125, 2021. 5, 7

[42] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1503–1511, 2022. 7

[43] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26099–26108, 2024. 5, 6, 7

[44] Xingchen Li, Junzhe Zhang, Junqi Shi, Ming Lu, and Zhan Ma. Yoda: Yet another one-step diffusion-based video compressor. *arXiv preprint arXiv:2601.01141*, 2026. 1

[45] Zizhang Li, Mengmeng Wang, Huaijin Pi, Kechun Xu, Jianbiao Mei, and Yong Liu. E-nerv: Expedite neural video representation with disentangled spatial-temporal context. In *European Conference on Computer Vision*, pages 267–284. Springer, 2022. 1

[46] Mufan Liu, Le Yang, Yiling Xu, Ye-Kui Wang, and Jenq-Neng Hwang. Evan: Evolutional video streaming adaptation via neural representation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 7

[47] Ming Lu, Zhihao Duan, Wuyang Cong, Dandan Ding, Fengqing Zhu, and Zhan Ma. High-efficiency neural video compression via hierarchical predictive learning. *arXiv preprint arXiv:2410.02598*, 2024. 5, 6

[48] Wenzhuo Ma and Zhenzhong Chen. Diffusion-based perceptual neural video compression with temporal diffusion information reuse. *arXiv preprint arXiv:2501.13528*, 2025. 2, 5, 6

[49] Shishira R Maiya, Sharath Girish, Max Ehrlich, Hanyu Wang, Kwot Sin Lee, Patrick Poirson, Pengxiang Wu, Chen Wang, and Abhinav Shrivastava. Nirvana: Neural implicit representations of videos with adaptive networks and autoregressive patch-wise modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14378–14387, 2023. 2

[50] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[52] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, 2019. 2, 3

[53] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and cap-

tions. *Advances in neural information processing systems*, 29, 2016. 1

[54] Linfeng Qi, Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image and video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2, 5, 6

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7

[56] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 7

[57] Junqi Shi, Mingyi Jiang, Ming Lu, Tong Chen, Xun Cao, and Zhan Ma. Hiner: Neural representation for hyperspectral image. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9837–9846, 2024. 2

[58] Junqi Shi, Zhujia Chen, Hanfei Li, Qi Zhao, Ming Lu, Tong Chen, and Zhan Ma. On quantizing neural representation for variable-rate video coding. In *The Thirteenth International Conference on Learning Representations*, 2025. 3, 5, 6, 7, 8

[59] Junqi Shi, Qirui Zhang, Ming Lu, and Zhan Ma. Compression as restoration: A unified implicit approach to self-supervised hyperspectral image representation. *IEEE Journal of Selected Topics in Signal Processing*, 2025. 2

[60] Junqi Shi, Ming Lu, Xingchen Li, Anle Ke, Ruiqi Zhang, and Zhan Ma. Dit-ic: Aligned diffusion transformer for efficient image compression, 2026. 1

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[62] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 1, 2

[63] Jake Snell, Karl Ridgeway, Renjie Liao, Brett D Roads, Michael C Mozer, and Richard S Zemel. Learning to generate images with perceptual similarity metrics. In *2017 IEEE international conference on image processing (ICIP)*, pages 4277–4281. IEEE, 2017. 5

[64] Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari. Implicit neural representations for image compression. In *European Conference on Computer Vision*, pages 74–91. Springer, 2022. 1

[65] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022. 1, 2

[66] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 5

[67] Lv Tang, Jun Zhu, Xinfeng Zhang, Li Zhang, Siwei Ma, and Qingming Huang. Canerv: Content adaptive neural representation for video compression. *arXiv preprint arXiv:2502.06181*, 2025. 1, 2

[68] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[69] Ani Vanyan, Alvard Barseghyan, Hakob Tamazyan, Vahan Huroyan, Hrant Khachatrian, and Martin Danelljan. Analyzing local representations of self-supervised vision transformers. *arXiv preprint arXiv:2401.00463*, 2023. 5

[70] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[71] Chang Wu, Guancheng Quan, Gang He, Xin-Quan Lai, Yunsong Li, Wenxin Yu, Xianmeng Lin, and Cheng Yang. Qsnerv: Real-time quality-scalable decoding with neural representation for videos. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2584–2592, 2024. 1

[72] Shaowen Xie, Hao Zhu, Zhen Liu, Qi Zhang, You Zhou, Xun Cao, and Zhan Ma. Diner: Disorder-invariant implicit neural representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6143–6152, 2023. 2

[73] Wentian Xu and Jianbo Jiao. Revisiting implicit neural representations in low-level vision. *arXiv preprint arXiv:2304.10250*, 2023. 2

[74] Ren Yang, Radu Timofte, and Luc Van Gool. Perceptual learned video compression with recurrent conditional gan. In *IJCAI*, pages 1537–1544, 2022. 2, 5, 6

[75] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025. 5

[76] Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*, 2019. 6

[77] Chun Zhang, Heming Sun, and Jiro Katto. Flavc: Learned video compression with feature level attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28019–28028, 2025. 1

[78] Kaiwei Zhang, Dandan Zhu, Xiongkuo Min, and Guangtao Zhai. Implicit neural representation learning for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2022. 2

[79] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 4, 8

[80] Saiping Zhang, Marta Mrak, Luis Herranz, Marc Górriz Blanch, Shuai Wan, and Fuzheng Yang. Dvc-p: Deep video compression with perceptual optimizations. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021. 5, 6, 7

[81] Xinjie Zhang, Ren Yang, Dailan He, Xingtong Ge, Tongda Xu, Yan Wang, Hongwei Qin, and Jun Zhang. Boosting neural representations for videos with a conditional decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2556–2566, 2024. 1, 3, 4, 5, 8

[82] Yunfan Zhang, Ties Van Rozendaal, Johann Brehmer, Markus Nagel, and Taco Cohen. Implicit neural video compression. *arXiv preprint arXiv:2112.11312*, 2021. 3

[83] Qi Zhao, M Salman Asif, and Zhan Ma. Dnerv: Modeling inherent dynamics via difference neural representation for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2031–2040, 2023. 1, 2

[84] Qi Zhao, M Salman Asif, and Zhan Ma. Pnerv: Enhancing spatial consistency via pyramidal neural representation for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19103–19112, 2024. 1

[85] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 7

[86] Xingguang Zhong, Yue Pan, Cyrill Stachniss, and Jens Behley. 3d lidar mapping in dynamic environments using a 4d implicit neural representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15417–15427, 2024. 1

[87] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2545, 2024. 3, 4, 5, 7, 8

[88] Jun Zhu, Xinfeng Zhang, Lv Tang, and JunHao Jiang. Msnerv: Neural video representation with multi-scale feature fusion. *arXiv preprint arXiv:2506.15276*, 2025. 1