Beyond Pixel Loss: Video-INRs Prefer Perceptual Optimization

Supplementary Material

1. Rate–Distortion–Perception Trade-off and Why Perception.

The trade-off between rate, distortion, and perceptual quality is a fundamental principle in modern compression theory. Under a fixed entropy budget, distortion and perceptual fidelity cannot be simultaneously optimized, as formally analyzed in prior studies [3, 10]. This theoretical limitation explains a common phenomenon in practice: methods optimized for perceptual quality often exhibit reduced performance in distortion-oriented metrics such as PSNR.

Empirical observations further support this insight. In equal-time experiments, where different methods share the same optimization budget, pixel-oriented implicit neural representations (INRs) tend to encounter a perceptual quality bottleneck. In contrast, the proposed POVI framework continues to yield noticeable perceptual improvements under the same computational budget, suggesting that perceptual optimization can more effectively utilize limited optimization resources.

From a practical perspective, the evaluated bitrate regime aligns with realistic network conditions, where bandwidth remains constrained and perceptual differences between reconstructions are visually significant. As shown in the main paper and supplementary material, perceptual optimization produces clearer structures and more natural textures in reconstructed videos under such conditions.

Beyond visual appearance, perceptual optimization can also benefit downstream vision tasks. Prior work has shown that perceptually faithful reconstructions often preserve semantic structures important for machine perception [25]. Consistently, POVI improves the accuracy of video classification and captioning tasks by approximately 6%, indicating that perceptual quality can translate into improved task-level performance.

Finally, emphasizing perceptual quality is also motivated by the characteristics of the human visual system, which is perception-driven rather than pixel-accurate. While MSE-optimized reconstructions may converge visually at very high bitrates, perceptual optimization effectively lowers the bitrate threshold required to achieve perceptually faithful results, making it particularly valuable in bandwidth-limited scenarios.

2. INR in Variational Viewpoint

Variational inference provides a unifying theoretical framework for probabilistic representation learning [2, 14]. In this section, we reinterpret implicit neural representations (INRs) through this lens and reveal why the common practice of training INRs with pixel-wise losses is fundamentally sub-optimal.

2.1. From VAE to INR

In variational autoencoders (VAEs), the generative process constructs a stochastic mapping from latent variable $\boldsymbol{z}$ to observation $\boldsymbol{x}$, i.e., $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$, while inference approximates the posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$.

In contrast, INRs parameterize the signal as a deterministic function of network weights $\boldsymbol{w}$ and coordinates $\boldsymbol{t}$:

$$\boldsymbol{x} = \mathcal{F}(\boldsymbol{w}, \boldsymbol{t}), \tag{1}$$

where $\boldsymbol{t}$ denotes spatial or temporal coordinates. Thus, while VAEs adopt a latent-centric representation, INRs realize a function-centric representation: the role of the latent variable is played by the neural weights $\boldsymbol{w}$, which uniquely encode the signal.

This correspondence enables us to view INR training as a special case of variational inference over functions: $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{w})$, while inference corresponds to approximating the posterior $p(\boldsymbol{w}|\boldsymbol{x})$. The distribution of $\boldsymbol{w}$ is implicitly shaped by optimization dynamics and the architecture's inherent inductive bias.

2.2. Variational formulation

Formally, we approximate the true posterior$p(\tilde{\boldsymbol{w}}|\boldsymbol{x})$ with a variational density $q(\tilde{\boldsymbol{w}}|\boldsymbol{x})$ by minimizing the expected KL divergence over the data distribution $p(\boldsymbol{x})$ [2, 17, 27]:

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} D_{KL}[q||p_{\tilde{\boldsymbol{w}}|\boldsymbol{x}}] \tag{2}$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} \mathbb{E}_{\tilde{\boldsymbol{w}}\sim q}\left[\log \frac{q(\tilde{\boldsymbol{w}}|\boldsymbol{x})}{p_{\tilde{\boldsymbol{w}}|\boldsymbol{x}}(\tilde{\boldsymbol{w}}|\boldsymbol{x})}\right] \tag{3}$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} \mathbb{E}_{\tilde{\boldsymbol{w}}\sim q}\left[\log q(\tilde{\boldsymbol{w}}|\boldsymbol{x}) - \log p_{\tilde{\boldsymbol{w}}|\boldsymbol{x}}(\tilde{\boldsymbol{w}}|\boldsymbol{x})\right]. \tag{4}$$

Applying Bayes' rule,

$$p_{\tilde{\boldsymbol{w}}|\boldsymbol{x}}(\tilde{\boldsymbol{w}}|\boldsymbol{x}) = \frac{p_{\boldsymbol{x}|\tilde{\boldsymbol{w}}}(\boldsymbol{x}|\tilde{\boldsymbol{w}}) p_{\tilde{\boldsymbol{w}}}(\tilde{\boldsymbol{w}})}{p_{\boldsymbol{x}}(\boldsymbol{x})}. \tag{5}$$

Substituting into Eq. 4 yields:

$$D_{KL}[q||p_{\tilde{\boldsymbol{w}}|\boldsymbol{x}}] \tag{6}$$

$$= \log q(\tilde{\boldsymbol{w}}|\boldsymbol{x}) - \log p(\boldsymbol{x}|\tilde{\boldsymbol{w}}) - \log p(\tilde{\boldsymbol{w}}) + \log p(\boldsymbol{x}). \tag{7}$$

We now examine each component in turn.

1. $\mathbb{E}[\log q(\tilde{\boldsymbol{w}}|\boldsymbol{x})]$**.** In rate–distortion coding, inference aligns with quantization. When quantization is modeled

as additive uniform noise [1],

$$q(\tilde{\boldsymbol{w}}|\boldsymbol{x}) = \prod_i \mathcal{U}(\tilde{\boldsymbol{w}}_i|\boldsymbol{w}_i - \tfrac{1}{2}, \boldsymbol{w}_i + \tfrac{1}{2}), \tag{8}$$

$$\boldsymbol{w} = \mathcal{F}_\theta^{-1}(\boldsymbol{x}), \tag{9}$$

whose expectation is constant and thus irrelevant to optimization.

2. $-\mathbb{E}[\log p_{\boldsymbol{x}|\tilde{\boldsymbol{w}}}(\boldsymbol{x}|\tilde{\boldsymbol{w}})]$**.** This term enforces distributional consistency between the reconstruction (induced by $\tilde{\boldsymbol{w}}$) and the original signal. If we choose a Gaussian likelihood with fixed variance, the term reduces to the mean squared error (MSE). This explains the prevalence of pixel-wise losses—while also revealing their limitation: the Gaussian assumption poorly matches the structured residuals commonly observed in INR reconstructions.
3. $-\mathbb{E}[\log p_{\tilde{\boldsymbol{w}}}(\tilde{\boldsymbol{w}})]$**.** This term measures the complexity of representing the signal in function space and corresponds to the rate in rate–distortion theory. Structured or compressible priors on $\tilde{\boldsymbol{w}}$ directly improve efficiency, linking this term to weight regularization and quantization.
4. $\mathbb{E}[\log p_{\boldsymbol{x}}(\boldsymbol{x})]$**.** This term is constant for a given sequence and can be discarded from the optimization objective.

Discarding constants, the variational objective becomes:

$$\mathcal{L} = \mathcal{L}_R + \lambda \mathcal{L}_D = -\log p(\tilde{\boldsymbol{w}}) - \log p(\boldsymbol{x}|\tilde{\boldsymbol{w}}), \tag{10}$$

where $\mathcal{L}_R$ represents the rate (representation complexity) and $\mathcal{L}_D$ represents the distortion (likelihood agreement).

2.3. Distributional Assumptions in Pixel-wise Losses.

Pixel-wise losses correspond to explicit distributional assumptions on the reconstruction error [9, 14, 22]. More precisely, minimizing an ℓ_p loss is equivalent to maximum likelihood estimation (MLE) under a generalized Gaussian distribution (GGD) [7], whose probability density function is given by:

$$p(e) = \frac{p}{2\alpha\Gamma(1/p)} \exp\left(-\left|\tfrac{e}{\alpha}\right|^p\right), \tag{11}$$

where $e = \boldsymbol{x} - \tilde{\boldsymbol{x}}$ denotes the reconstruction error, $\alpha > 0$ is the scale parameter controlling dispersion, $p = \beta$ is the shape parameter determining tail heaviness and peak sharpness, and $\Gamma(\cdot)$ denotes the Gamma function, defined as

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \quad z > 0. \tag{12}$$

Special cases of the GGD include: (a) $p = 2$: Gaussian distribution $\mathcal{N}(0, \sigma^2)$, leading to mean squared error (MSE); (b) $p = 1$: Laplace distribution $\mathrm{Laplace}(0, b)$, leading to ℓ_1 loss; (c) $p < 1$: heavy-tailed distributions with higher robustness to outliers; (d) $p > 2$: sharper-peaked distributions emphasizing small errors. For instance, Gaussian error modeling leads to:

$$\max \log p_{\boldsymbol{x}|\tilde{\boldsymbol{w}}}(\boldsymbol{x}|\tilde{\boldsymbol{w}}) \tag{13}$$

$$= -\min \log \mathcal{N}(\boldsymbol{x}|\tilde{\boldsymbol{x}}, \sigma^2) \tag{14}$$

$$= -\min \log \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}||\boldsymbol{x} - \tilde{\boldsymbol{x}}||^2\right) \tag{15}$$

$$= \min \frac{1}{2\sigma^2}||\boldsymbol{x} - \tilde{\boldsymbol{x}}||^2, \tag{16}$$

while Laplace error modeling corresponds to:

$$\max \log p_{\boldsymbol{x}|\tilde{\boldsymbol{w}}}(\boldsymbol{x}|\tilde{\boldsymbol{w}}) \tag{17}$$

$$= -\min \log \mathrm{Laplace}(\boldsymbol{x}|\tilde{\boldsymbol{x}}, b) \tag{18}$$

$$= -\min \log \frac{1}{2b} \exp\left(-\frac{1}{b}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_1\right) \tag{19}$$

$$= \min \frac{1}{b}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_1. \tag{20}$$

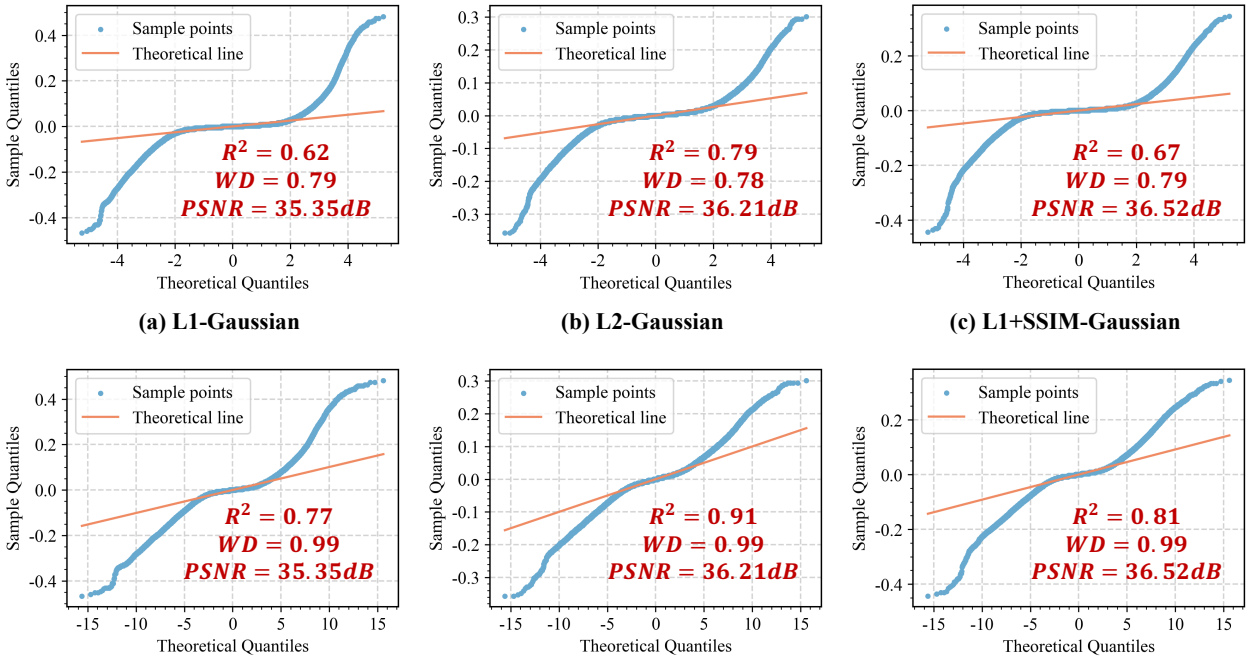
Thus, adopting a pixel-wise loss is equivalent to committing to a fixed parametric error model. However, in video-INRs such assumptions rarely hold. First, reconstruction errors deviate significantly from Gaussian or Laplacian distributions due to strong temporal dependencies and structured spatial patterns. Second, error statistics are highly video-dependent: high-motion videos often produce heavier-tailed residuals, while static scenes yield more concentrated error profiles. Consequently, a single parametric assumption (e.g., Gaussian residuals) is inherently unreliable and leads to sub-optimal optimization objectives. This motivates the need for alternative formulations that relax fixed distributional assumptions and better capture the true statistics of INR reconstruction errors.

3. Error Distribution Under Different Losses

To further examine the statistical behavior of reconstruction errors, we conduct a cross-over study using three representative pixel-level losses: ℓ_1, ℓ_2 (MSE), and a hybrid ℓ_1+SSIM. Fig. 1 reports the corresponding error distributions via Q–Q plots, where ℓ_1 and ℓ_2 implicitly assume Laplace and Gaussian residuals, respectively.

Within a narrow error range $[-0.05, 0.05]$, the empirical residuals moderately align with these parametric models. However, in the distribution tails, all cases exhibit pronounced heavy-tailed behavior, with Wasserstein Distance (WD) consistently exceeding 0.7, revealing clear model–data mismatches.

Another noteworthy observation is the divergence between fidelity and distributional fit. Although ℓ_1+SSIM delivers the highest PSNR (36.52 dB), its distributional alignment under both Gaussian and Laplace evaluations remains inferior. This indicates that better distortion metrics (e.g.,



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

(d) L1-Laplace **(e) L2-Laplace** **(f) 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 [32]. 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 [24], where larger values indicate stronger distributional mismatch. Perfect distributional alignment would result in points lying along the diagonal reference line.

PSNR) do not necessarily imply better statistical consistency. Instead, it suggests that ℓ_1+SSIM implicitly induces a residual distribution that differs from both p_{Gaussian} and p_{Laplace}, offering a closer approximation to the true error distribution p_{true} and yielding improved convergence relative to the purely parametric ℓ_1 and ℓ_2 losses.

In summary, pixel-level losses inherently enforce rigid Gaussian or Laplace priors that are systematically violated by single-video INR reconstructions. The resulting error distributions are heavy-tailed, video-dependent, and structurally correlated, making simplistic parametric assumptions unreliable. Unlike amortized models such as VAEs—where dataset-level averaging smooths out sample-specific fluctuations—single-video INRs exhibit highly idiosyncratic residual statistics. These findings motivate the exploration of alternative optimization domains or perceptually aligned objectives that more faithfully capture the intrinsic structure of INR reconstruction errors.

4. Architecture

Our model follows an encoder–decoder paradigm with temporal conditioning. The encoder extracts compact content embeddings from input frames, while the decoder reconstructs frames conditioned on both content and temporal indices. The overall architecture is illustrated in Fig. 2.

4.1. Encoder

The encoder consists of cascaded down-sampling blocks, each composed of a ConvNeXt block [20] followed by a strided convolution for resolution reduction. Given a ground-truth frame $\boldsymbol{x}_t \in \mathbb{R}^{H\times W\times 3}$ at time t, the encoder produces a content embedding:

$$\boldsymbol{e}_t = f_{\text{enc}}(\boldsymbol{x}_t) \in \mathbb{R}^{C'\times H'\times W'}, \tag{21}$$

where H' and W' denote the reduced spatial resolution, and C' is the embedding dimension. This formulation is similar to HNeRV [6], but with modified building blocks.

4.2. Temporal-Conditional Decoder

The content embedding $\boldsymbol{e}_t$ is fed into a lightweight convolutional decoder to reconstruct the frame:

$$\hat{\boldsymbol{x}}_t = f_{\text{dec}}(\boldsymbol{e}_t) \in \mathbb{R}^{H\times W\times 3}. \tag{22}$$

Following observations from Zhang et al. [31], we introduce a modulation mechanism to adapt the reconstruction

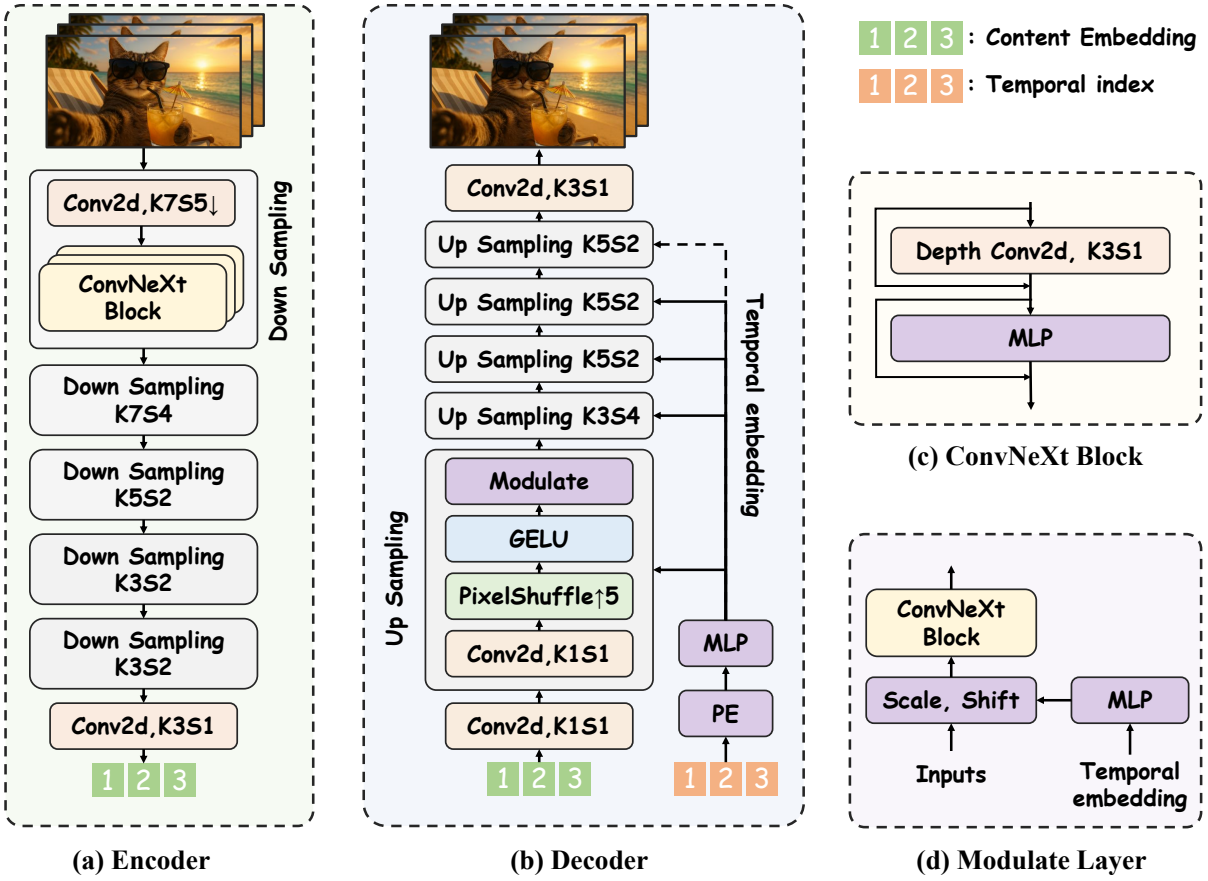


Figure 2. Overview of the proposed architecture. The encoder cascades ConvNeXt [20]-based down-sampling blocks to extract compact content embeddings $\boldsymbol{e}_t$ from ground-truth frames. The decoder reconstructs frames from $\boldsymbol{e}_t$ with temporal conditioning: intermediate features are modulated by sinusoidal positional encodings of the time index, enabling adaptive reconstruction across different frames. For efficiency, the last up-sampling layer avoids ConvNeXt operations to handle high-resolution outputs. Strides for different resolutions can be adjusted adaptively.

to temporal variations. Given intermediate feature maps f_t, temporal modulation is defined as:

$$\text{modulate}(f_t \mid \alpha_t, \beta_t) = \alpha_t(\gamma(t)) \cdot f_t + \beta_t(\gamma(t)), \quad (23)$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ are learned nolinear transformations from a temporal embedding $\gamma(t)$. We adopt sinusoidal positional encodings [5, 21]:

$$\begin{aligned} &\gamma(t) \in \mathbb{R}^{2l} \quad (24)\\ &= \left[\sin(b^0\pi t), \cos(b^0\pi t), \ldots, \sin(b^{l-1}\pi t), \cos(b^{l-1}\pi t)\right], \end{aligned}$$

where l controls the number of frequency components and b is a frequency scaling factor. This encoding mitigates the spectral bias [29] of neural networks, allowing the decoder to better capture high-frequency temporal variations.

To reduce computational cost at high resolutions, the final up-sampling layer of the decoder deletes ConvNeXt block, which we found significantly reduces inference latency without degrading quality.

4.3. Rate in INR

The rate in INR-based compression is fundamentally determined by two factors: (1) *quality*, referring to the bit-width used to quantize each parameter, and (2) *quantity*, referring to the total number of parameters to be transmitted. Increasing bit-width improves numerical precision and reconstruction fidelity, but directly increases the coding rate. Similarly, enlarging the network size (i.e., parameter count) enhances reconstruction capacity, but also increases the total number of bits to encode. In addition, higher rate inevitably introduces higher computational complexity, since larger models and higher-precision arithmetic both slow down inference and training.

As discussed in Appendix 2.2, this naturally leads to a rate–distortion trade-off. For a fixed architecture, reducing rate (via fewer parameters or lower precision) often sacrifices reconstruction accuracy, while increasing rate im-

proves fidelity at the cost of storage and complexity.

In our implementation, we vary the rate primarily by adjusting the channel dimension of the decoder, while keeping the bit-width fixed, consistent with most prior works [6, 16]. Although recent studies [27] demonstrate that dynamically adjusting bit-width is an effective alternative for rate control, we leave this direction as promising future work.

5. Multi-Vision Models Representation

We visualize the feature sensitivity of representative pretrained models on different categories of sequences, including animal, nature, food, building, and face. As shown in Fig. 3, relying on a single pretrained model introduces strong inductive biases toward specific visual patterns. For example, AlexNet [15] consistently emphasizes the main subject region, while VGG [28] places greater weight on high-frequency edges and contours. This, to some extent, explains why VGG-based perceptual losses often correlate better with human judgments of visual fidelity, as also observed in Zhang et al. [30]. However, such attention can fail on smooth regions, e.g., VGG tends to under-emphasize facial regions (Fig. 3 (e)) where structural cues are subtle but perceptually critical. In contrast, DINOv2 [23] captures intra-object variability more robustly and yields feature maps that are less biased toward low-level edges or single-object saliency.

These observations highlight that no single pretrained model provides a universally reliable perceptual space. We aggregate feature-based losses from multiple pretrained models. This ensemble strategy reduces the risk of suboptimal alignment that may arise when optimization is guided solely by one model's representational prior.

6. Experiments

6.1. Sequence-level Evaluation.

While conventional frame-level metrics such as PSNR, MS-SSIM, LPIPS, and DISTS provide a direct assessment of spatial fidelity and perceptual quality, they operate independently on individual frames and thus fail to capture temporal dynamics and holistic video realism. To address this limitation, we adopt the sequence-level evaluation protocol from VBench [11], which integrates multiple complementary indicators beyond frame fidelity. For clarity, we briefly restate the metrics considered in this work:

1. *Subject Consistency.* Evaluates whether the appearance of a primary subject (e.g., a person, vehicle, or animal) remains stable across frames. This is measured by computing feature similarity with DINO [4], which is particularly sensitive to identity variations.
2. *Background Consistency.* Measures temporal stability of backgrounds by comparing CLIP [26] embeddings across frames.
3. *Motion Smoothness.* Complements the above appearance-based metrics by quantifying whether object motion follows physically plausible dynamics. This is estimated with motion priors from a video frame interpolation model [19].
4. *Imaging Quality.* Evaluates perceptual frame-level fidelity by detecting distortions such as over-exposure, noise, or blur. It adopts MUSIQ [13], trained on SPAQ [8]. Although frame-based, it is included here since it forms part of the VBench protocol.

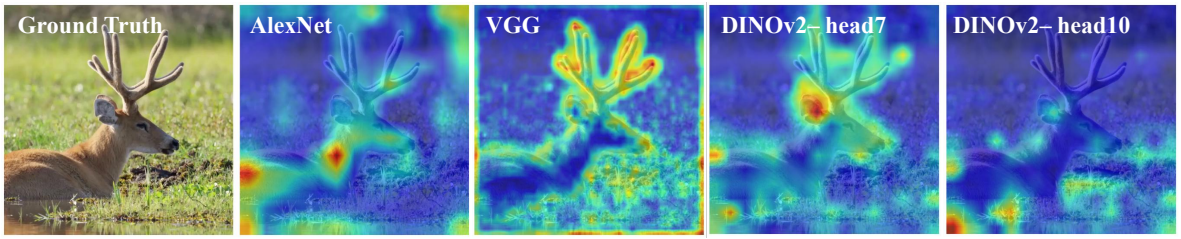
Note: We note that sequence-level metrics are sensitive to the number of frames evaluated, since several indicators rely on comparisons relative to the first frame or across temporal neighborhoods. Therefore, consistent evaluation requires using either the full sequence or a standardized subset of frames. In our experiments, we adopt the full set of frames from UVG to ensure reliable and reproducible results.

Here, we provide a detailed sequence-level evaluation in Table 1. The results are consistent with intuitive expectations: sequences with large motion impose significant challenges on maintaining content consistency and motion smoothness. For example, *Jockey* and *ReadySetGo* exhibit substantially lower subject consistency (76.8% and 68.0%, respectively) compared to relatively static sequences such as *HoneyBee* (99.7%) and *Beauty* (96.0%). This corresponds to a reduction of more than 20% in subject stability. A similar trend is observed in background consistency, where motion-heavy sequences (*ReadySetGo* at 83.3%) fall behind stable ones (*HoneyBee* at 98.9%). In contrast, motion smoothness remains consistently high across all sequences (above 97.7%), indicating that INR-based reconstruction preserves local dynamics well, even under large temporal variations.

On the other hand, it is important to recognize that these metrics are all no-reference, and their scores can be influenced by the quality of the original source videos. Nevertheless, they provide valuable insights into temporal behaviors that frame-level metrics overlook. For example, VAE-based methods often exhibit noticeable fluctuations in frame quality, which remain hidden under conventional PSNR or SSIM evaluations. This limitation becomes even more critical under perceptual optimization, where human observers are especially sensitive to temporal inconsistencies. Incorporating sequence-level evaluation thus provides a more holistic and reliable understanding of reconstruction quality in video representation and compression.

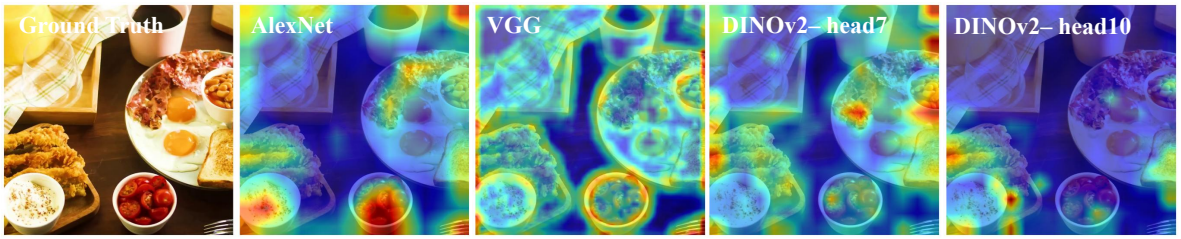
6.2. Decoding Complexity

We evaluate sequential decoding complexity in terms of frames per second (FPS) at 1080p resolution, as shown in Fig. 4. The comparison includes representative baselines: DCVC-RT [12], the fastest VAE-based codec out-



(a) Animal

(b) Nature



(c) Food

(d) Building

(e) Face

Figure 3. Illustrative heatmaps of feature sensitivity from different pretrained vision models on samples from YouHQ [32]. From left to right: AlexNet [15], VGG [28], and DINOv2 [23]. Each model exhibits distinct inductive biases, emphasizing different spatial or semantic structures, which in turn influences the optimization objective.

Table 1. Sequence-level evaluation on UVG using VBench [11] with 0.01bpp.

Sequence	Subject Consistency	Background Consistency	Motion Smoothness	Image Quality	Average Score
Beauty	96.03%	94.60%	99.45%	56.90%	86.75%
Bosph.	94.69%	93.55%	99.71%	69.21%	89.29%
Honey.	99.71%	98.94%	99.21%	63.82%	90.42%
Jockey	76.76%	87.65%	97.74%	54.86%	79.25%
Ready.	67.95%	83.32%	97.98%	61.49%	77.69%
Shake.	74.86%	88.76%	99.67%	65.93%	82.31%
Yacht.	81.44%	78.97%	99.55%	70.01%	82.49%
Avg.	**84.49**$_{\pm 1.5}$**%**	**89.40**$_{\pm 1.0}$**%**	**99.04**$_{\pm 0.5}$**%**	**63.17**$_{\pm 2.4}$**%**	**84.03**$_{\pm 2.8}$**%**
Empirical Min	14.62%	26.15%	70.60%	0.00%	/
Empirical Max	100%	100%	99.75%	100%	/

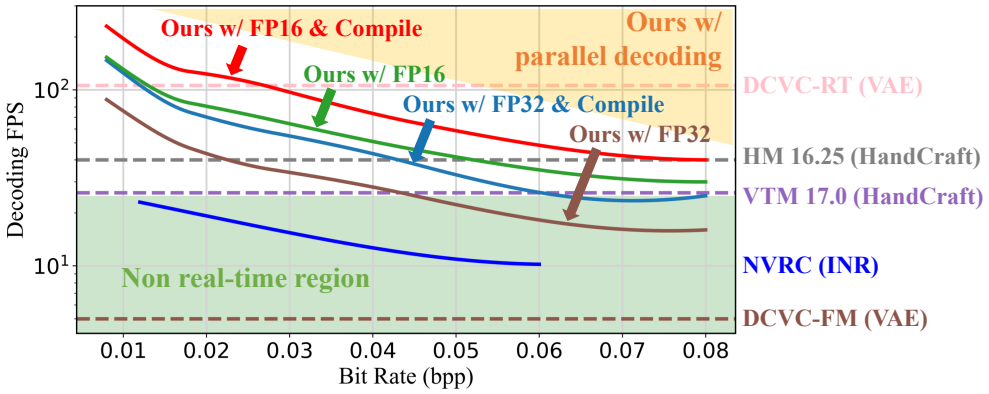


Figure 4. Sequential decoding FPS (frames per second) for 1080p resolution. The reported FPS with *compile* is measured with PyTorch's `torch.compile`, which applies ahead-of-time compilation to optimize model execution by fusing operators and reducing Python overhead. This generally improves runtime performance without altering model accuracy.

performing VVC; DCVC-FM [18], a state-of-the-art VAE-based model; VVC and HEVC as classical references; and NVRC [17], the strongest INR-based codec to date.

Unlike VAE-based codecs—with decoding FPS remaining nearly constant across bitrates due to fixed latent transforms—INR-based methods exhibit complexity that scales with network evaluation. NVRC attains strong PSNR through rate–distortion optimization and hierarchical grids, but its computational burden results in substantially lower FPS. In contrast, our method adopts a lightweight feed-forward INR trained with perceptual supervision, achieving significantly higher decoding speed while maintaining competitive perceptual fidelity. For example, our FP16 implementation with `torch.compile` exceeds 100 FPS, comfortably surpassing real-time requirements, whereas NVRC remains below 25 FPS.

A key advantage of our formulation is the absence of inter-frame dependencies, which enables full-frame parallel decoding. In the ideal scenario, all frames can be reconstructed within a single forward pass, reducing total latency to that of one model evaluation. As illustrated in Fig. 4, parallel decoding further amplifies the achievable speedup, with the exact gain determined by the implementation (e.g., multi-threading or multi-GPU deployment).

6.3. Visualization

To qualitatively assess the impact of different optimization strategies, we provide comparisons between pixel-domain and feature-domain supervision.

Pixel Optimization vs. Feature Optimization. As shown in Fig. 5, under the same architecture, perceptual (feature-domain) optimization produces reconstructions with sharper edges, richer textures, and more faithful visual realism. By contrast, pixel-domain supervision tends to over-smooth fine structures, suppressing high-frequency

details to minimize pixel-wise error. These qualitative observations align with our quantitative findings and demonstrate the effectiveness of perceptual optimization in retaining semantically meaningful structures.

Comparison with Other Methods. We visualize reconstructions on the *Beauty* and *YachtRide* sequences in Fig. 6 and Fig. 7. Due to limited availability of certain prior works (e.g., incomplete open-sourcing or non-reproducible training pipelines), we compare against representative methods with accessible implementations. Results are presented in order of increasing PSNR for clarity.

As shown, our approach delivers the most visually faithful reconstructions despite achieving lower PSNR, reflecting the well-known mismatch between pixel-wise metrics and perceptual quality. By contrast, DCVC-FM yields the highest PSNR but produces noticeably over-smoothed textures, illustrating that numerical fidelity does not necessarily translate into perceptual realism.

We also observe that perceptual optimization may attenuate certain fine-grained patterns—such as subtle textures or small facial features—reflecting a trade-off between perceptual sharpness and pixel-level accuracy.

References

[1] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*, 2017. 2

[2] 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] 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

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[5] 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. 4

[6] 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. 3, 5

[7] 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. 2

[8] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686, 2020. 5

[9] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016. 2

[10] Boris Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10):992, 2019. 1

[11] 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. 5, 7

[12] 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. 5

[13] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 5

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5, 6

[16] 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. 5

[17] 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. 1, 7

[18] 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. 7

[19] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 5

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3, 4

[21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

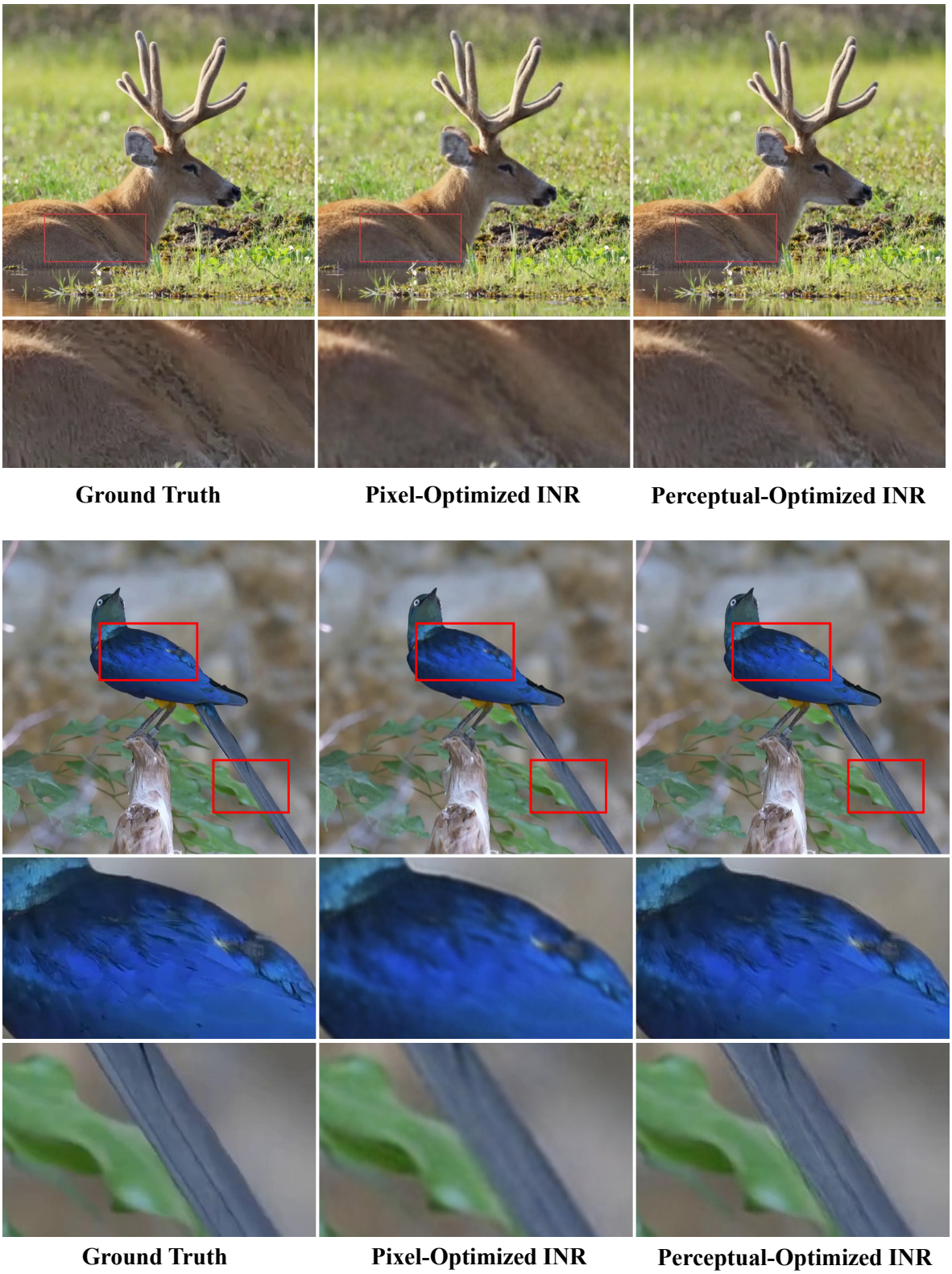


Figure 5. The visualization of our 0.6MB model optimized with different supervision on sample from YouHQ.

Figure 6. The visualization of *Beauty* sequence.

thesis. *Communications of the ACM*, 65(1):99–106, 2021. 4

[22] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 2

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,

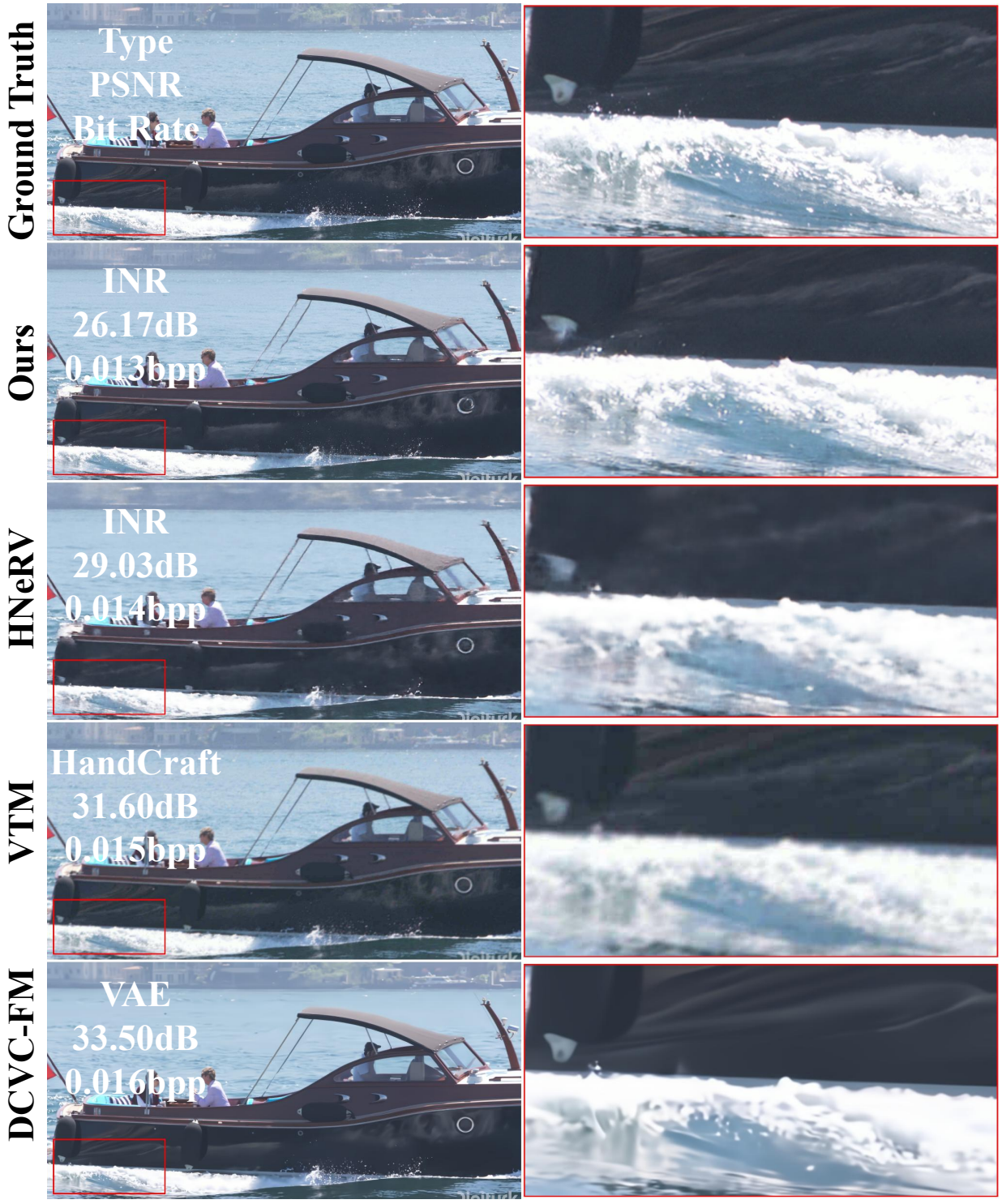


Figure 7. The visualization of *YachtRide* sequence.

Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 6

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

[25] Yash Patel, Srikar Appalaraju, and R Manmatha. Deep perceptual compression. *arXiv preprint arXiv:1907.08310*, 2019. 1

[26] 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. 5

[27] 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. 1, 5

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[29] 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. 4

[30] 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. 5

[31] 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. 3

[32] 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, 6