Parallax to Align Them All: An OmniParallax Attention Mechanism for Distributed Multi-View Image Compression

Appendix

A. Foundations of Distributed Source Coding

In this section, we present several theoretical foundations of distributed source coding (DSC), including the Slepian-Wolf theorem [11, 13], the Berger-Tung bounds [3, 10], and the Wyner-Ziv theorem [9].

A.1. Slepian-Wolf Theorem

Let X and Y be two statistically correlated discrete information sources, obtained by repeated independent drawings from a discrete bivariate distribution $p(x, y)$. The achievable rate region for separate encoding with joint decoding under lossless compression is:

$$R_X + R_Y \geq H(X, Y), \tag{1}$$

$$R_X \geq H(X \mid Y), \tag{2}$$

$$R_Y \geq H(Y \mid X), \tag{3}$$

where R_X and R_Y denote the coding rates for X and Y, respectively.

A.2. Berger-Tung Bounds

Fix (D_1, D_2). Let X and Y be two sources out of which pairs of sequences $\left(X^n, Y^n\right)$ are drawn i.i.d. $\sim p(xy)$; and let U and V be auxiliary variables defined over alphabets $\mathcal{U}$ and $\mathcal{V}$, such that there exist functions $\gamma_1 : \mathcal{U} \times \mathcal{V} \to \hat{\mathcal{X}}$ and $\gamma_2 : \mathcal{U} \times \mathcal{V} \to \hat{\mathcal{Y}}$, for which $\mathbb{E}\left[d_1\big(X, \gamma_1(UV)\big)\right] \leq D_1$ and $\mathbb{E}\left[d_2\big(Y, \gamma_2(UV)\big)\right] \leq D_2$.

Consider rates (R_1, R_2), such that $R_1 \geq I(XY \wedge U \mid V)$, $R_2 \geq I(XY \wedge V \mid U)$, and $R_1 + R_2 \geq I(XY \wedge UV)$, for some joint distribution $p(xyuv)$. Now:

- **Inner Bound**: For any $p(xyuv)$ that satisfies a Markov chain of the form $U - X - Y - V$, all rates (R_1, R_2) obtained for any such p are achievable;
- **Outer Bound**: If there exists a $p(xyuv)$ that satisfies two Markov chains of the form $U - X - Y$ and $X - Y - V$, then if we consider the union of the set of rates defined for each such $p(xyuv)$, we must have that any achievable rates are included in that union.

A.3. Wyner-Ziv Theorem

Let X and Y denote the source and side information variables, respectively, with joint distribution $p(x, y)$.

The information-theoretic fundamental limit in lossy compression is characterized by the *rate-distortion function*, given by:

$$R_{X|Y}^{WZ}(d) = \inf I(X; V \mid Y), \tag{4}$$

where $R^*(d)$ is the rate of compression (in bits per source sample) to achieve a given target average distortion d, between the input sequence $\mathbf{x} \in \mathbb{R}^n$ and its reconstruction $\hat{\mathbf{x}} \in \mathbb{R}^n$. The infimum is with respect to all auxiliary random variables V and reconstruction functions $f : \mathcal{Y} \times \mathcal{V} \to \hat{\mathcal{X}}$ that satisfy: i) V and Y are conditionally independent given X, i.e., $V - X - Y$ form a Markov chain; ii) $\mathbb{E}\left[D(X, f(V, Y))\right] \leq d$.

B. Analysis of OmniParallax Attention Mechanism

In LDMIC [14], average pooling is used to integrate information from side views. However, this strategy is suboptimal, as it assigns equal importance to all views regardless of their semantic relevance. To overcome this limitation, we propose the OmniParallax Attention Mechanism (OPAM), which efficiently explores the full two-dimensional spatial context of the side information source to provide a reliable reference and corresponding consistency for the main source. The consistency reflects the reliability of the reference and can be interpreted as the semantic relevance between the main and side sources.

OPAM consists of two complementary components: Horizontal Parallax Attention (HPA) and Vertical Parallax Attention (VPA). The key difference is that HPA performs attention along the horizontal axis, while VPA operates along the vertical axis. An overview of the OPAM is shown in Fig. 1. In this section, we first present the detailed formulation of HPA in Section B.1 and VPA in Section B.2. We then describe the two-stage parallax attention process in OPAM in Section B.3, where HPA and VPA are applied sequentially to capture the full two-dimensional spatial context.

B.1. Horizontal Parallax Attention

In this subsection, we present the detailed formulation and analysis of the Horizontal Parallax Attention (HPA).

Let the main and side information sources be denoted as $f_u, f_v \in \mathbb{R}^{B \times H \times W \times C}$. We first compute the query feature map Q and the key feature map K using a selective kernel module (SKM) [7]:

$$Q = \text{SKM}(f_u), \quad K = \text{SKM}(f_v), \tag{5}$$

where $Q, K \in \mathbb{R}^{B \times H \times W \times C}$. The row dimension of Q and K is reshaped into the batch dimension to obtain $Q_b, K_b \in$

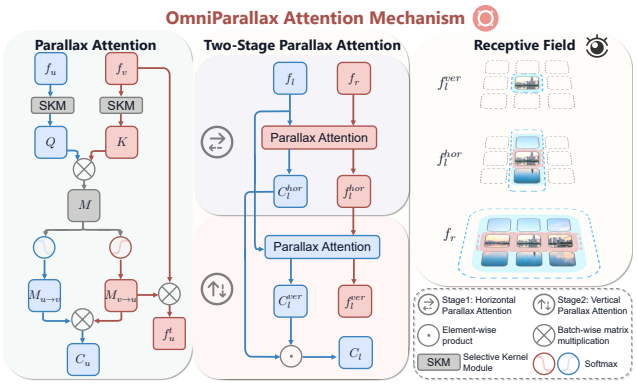

Figure 1. **Overview of OmniParallax Attention Mechanism (OPAM).** *Left:* **Parallax attention.** *Middle:* **Two-stage parallax attention in OPAM.** OPAM applies horizontal (red) and vertical (blue) parallax attention sequentially to capture the full 2D spatial context. *Right:* **Receptive fields of the aligned features.** Each position in f_l^{hor} attends to one row of f_r, and each position in f_l^{ver} attends to one column of f_l^{hor}, allowing each position in f_l^{ver} to attend to the entire 2D spatial domain of f_r.

$\mathbb{R}^{BH\times W\times C}$. The horizontal cross-correlation map $M^{hor} \in \mathbb{R}^{BH\times W\times W}$ is then computed as:

$$M^{hor} = Q_b \otimes K_b^{\top}, \tag{6}$$

where $\otimes$ denotes batch-wise matrix multiplication and $K_b^{\top}$ represents the transpose of K_b over its last two dimensions. The matrix M^{hor} represents the matching scores between positions in f_u and f_v within the same row.

Computationally, forming M^{hor} requires (BH) batch multiplications of matrices with dimensions $(W \times C)$ and $(C \times W)$, resulting in (BHW^2C) multiply-accumulate operations (MACs).

The horizontal parallax attention maps, denoted as $M_{v\to u}^{hor}$ and $M_{u\to v}^{hor} \in \mathbb{R}^{BH\times W\times W}$, are obtained by applying the softmax function to M^{hor} and its transpose $(M^{hor})^{\top}$ along the last dimension:

$$\begin{aligned} M_{v\to u}^{hor} &= \text{softmax}(M^{hor},\ \text{dim} = -1), \\ M_{u\to v}^{hor} &= \text{softmax}((M^{hor})^{\top},\ \text{dim} = -1), \end{aligned} \tag{7}$$

where $M_{v\to u}^{hor}[bH+g,i,j]$ denotes the correlation between $f_u[b,g,i]$ and $f_v[b,g,j]$. This value can be interpreted as the probability that $f_u[b,g,i]$ attends to information from $f_v[b,g,j]$, where b denotes the batch index. Similarly, $M_{u\to v}^{hor}[bH+g,i,j]$ denotes the correlation between $f_v[b,g,i]$ and $f_u[b,g,j]$, representing the probability that $f_v[b,g,i]$ attends to information from $f_u[b,g,j]$.

Next, the second dimension of $M_{v\to u}^{hor}$ and the third dimension of $M_{u\to v}^{hor}$ are reshaped into the batch dimension to obtain $M_{v\to u,b}^{hor} \in \mathbb{R}^{BHW\times 1\times W}$ and $M_{u\to v,b}^{hor} \in \mathbb{R}^{BHW\times W\times 1}$. These matrices are then used to compute the batch horizontal cycle consistency [12], denoted as $C_{u,b}^{hor} \in \mathbb{R}^{BHW\times 1\times 1}$:

$$C_{u,b}^{hor} = M_{v\to u,b}^{hor} \otimes M_{u\to v,b}^{hor}, \tag{8}$$

where

$$\begin{aligned} &C_{u,b}^{hor}[bHW+gW+i,0,0] = \\ &\sum_{j=0}^{W-1} M_{v\to u,b}^{hor}[bHW+gW+i,0,j] \\ &\quad\times M_{u\to v,b}^{hor}[bHW+gW+i,j,0]. \end{aligned} \tag{9}$$

Here, $M_{v\to u,b}^{hor}[bHW+gW+i,0,j]$ represents the correlation between $f_u[b,g,i]$ and $f_v[b,g,j]$, indicating the probability that $f_v[b,g,j]$ corresponds to $f_u[b,g,i]$ within row g of f_v. Similarly, $M_{u\to v,b}^{hor}[bHW+gW+i,j,0]$ represents the correlation between $f_v[b,g,j]$ and $f_u[b,g,i]$, indicating the probability that $f_u[b,g,i]$ corresponds to $f_v[b,g,j]$ within row g of f_u.

Lemma 1. *The horizontal cycle consistency $C_{u,b}^{hor}$ has an upper bound of 1.*

Proof. Since all elements of $M_{v\to u,b}^{hor}$ and $M_{u\to v,b}^{hor}$ are within the range $[0,1]$, we have:

$$\begin{aligned} &C_{u,b}^{hor}[bHW+gW+i,0,0] = \\ &\sum_{j=0}^{W-1} M_{v\to u,b}^{hor}[bHW+gW+i,0,j] \\ &\times M_{u\to v,b}^{hor}[bHW+gW+i,j,0] \\ &\le \Big(\sum_{j=0}^{W-1} M_{v\to u,b}^{hor}[bHW+gW+i,0,j]\Big) \times 1 \\ &= 1. \end{aligned} \tag{10}$$

The equality is achieved when $M_{v\to v,b}^{hor}[bHW+gW+i,j,0] = 1$ for any index j. □

A high value of the product $M_{v\to u,b}^{hor}[bHW+gW+i,0,j] \times M_{u\to v,b}^{hor}[bHW+gW+i,j,0]$ suggests strong mutual correspondence between $f_u[b,g,i]$ and $f_v[b,g,j]$, implying that $f_v[b,g,j]$ can reliably provide information for reconstructing $f_u[b,g,i]$. Conversely, when $M_{v\to u,b}^{hor}$ is high but $M_{u\to v,b}^{hor}$ is low, it indicates potential ambiguity in the matching within row g of f_v, meaning that $f_u[b,g,i]$ may have been mismatched to $f_v[b,g,j]$, while another position in f_u might correspond more accurately to $f_v[b,g,j]$. The same reasoning applies when $M_{v\to u,b}^{hor}$ is low but $M_{u\to v,b}^{hor}$ is high.

Therefore, the product of $M_{v\to u,b}^{hor}$ and $M_{u\to v,b}^{hor}$ mitigates mismatches between the two sources. Consequently, $C_{u,b}^{hor}[bHW+gW+i,0,0]$ represents the reliability of reconstructing $f_u[b,g,i]$ using all possible positions $f_v[b,g,j]$, i.e., all locations in the g-th row of f_v.

Then, $C_{u,b}^{hor}$ is reshaped to obtain the horizontal cycle consistency $C_u^{hor} \in \mathbb{R}^{B\times H\times W}$, where $C_u^{hor}[b,g,i] = C_{u,b}^{hor}[bHW + gW + i, 0, 0]$. The horizontal cycle consistency C_u^{hor} measures the reliability of reconstructing each position in f_u using all positions in the same row of f_v.

Finally, the row dimension of $f_v \in \mathbb{R}^{B\times H\times W\times C}$ is reshaped into the batch dimension to obtain $f_{v,b} \in \mathbb{R}^{BH\times W\times C}$. The batch horizontally aligned feature $f_{u,b}^{hor} \in \mathbb{R}^{BH\times W\times C}$ is then computed as:

$$f_{u,b}^{hor} = M_{v\to u}^{hor} \otimes f_{v,b}, \tag{11}$$

where

$$\begin{aligned} f_{u,b}^{hor}[bH+g,i,k] = &\sum_{j=0}^{W-1} M_{v\to u}^{hor}[bH+g,i,j] \\ &\times f_{v,b}[bH+g,j,k]. \end{aligned} \tag{12}$$

Here, $M_{v\to u}^{hor}[bH+g,i,j]$ denotes the correlation between $f_u[b,g,i]$ and $f_v[b,g,j]$, and $f_{v,b}[bH+g,j,k]$ denotes the k-th channel value of $f_v[b,g,j]$. Therefore, $f_{u,b}^{hor}[bH+g,i,k]$ represents the reconstructed k-th channel of $f_u[b,g,i]$ using all $f_v[b,g,j]$, i.e., all positions in the g-th row of f_v. Then, $f_{u,b}^{hor}$ is reshaped into the horizontally aligned feature $f_u^{hor} \in \mathbb{R}^{B\times H\times W\times C}$, where $f_u^{hor}[b,g,i,k] = f_{u,b}^{hor}[bH+g,i,k]$. The horizontally aligned feature f_u^{hor} represents the transformation of the side source into the perspective of the main source, aggregating information from all positions in the corresponding row of f_v.

The aggregation step in Eq. 11, i.e., ($f_{u,b}^{\text{hor}} = M_{v\to u}^{\text{hor}} \otimes f_{v,b}$), adds another (BHW^2C) MACs. Meanwhile, the cycle consistency in Eq. 8 contributes (BHW^2) MACs, which is negligible when $C \gg 1$. Therefore, the overall computational complexity of the parallax attention mechanism is $\mathcal{O}(N^3)$, where $N = \max(H,W)$.

B.2. Vertical Parallax Attention

In this subsection, we present the detailed formulation and analysis of the Vertical Parallax Attention (VPA). The process of VPA is similar to that of HPA, except that the reshape operations applied to the row dimension in HPA are instead applied to the column dimension in VPA.

Let the main and side information sources be denoted as $f_u, f_v \in \mathbb{R}^{B\times H\times W\times C}$. We first compute the query feature map Q and key feature map K using a selective kernel module (SKM) [7]:

$$Q = \text{SKM}(f_u), \quad K = \text{SKM}(f_v), \tag{13}$$

where $Q, K \in \mathbb{R}^{B\times H\times W\times C}$. The column dimension of Q and K is reshaped into the batch dimension to obtain $Q_b, K_b \in \mathbb{R}^{BW\times H\times C}$. The vertical cross-correlation map $M^{ver} \in \mathbb{R}^{BW\times H\times H}$ is then computed as:

$$M^{ver} = Q_b \otimes K_b^\top, \tag{14}$$

where $\otimes$ denotes batch-wise matrix multiplication and $K_b^\top$ represents the transpose of K_b over its last two dimensions. The matrix M^{ver} represents the matching scores between positions in f_u and f_v within the same column. Analogously, building M^{ver} costs (BWH^2C) MACs.

The vertical parallax attention maps, denoted as $M_{v\to u}^{ver}$ and $M_{u\to v}^{ver} \in \mathbb{R}^{BW\times H\times H}$, are obtained by applying the softmax function to M^{ver} and its transpose $(M^{ver})^\top$ along the last dimension:

$$\begin{aligned} M_{v\to u}^{ver} &= \text{softmax}(M^{ver},\ \text{dim} = -1), \\ M_{u\to v}^{ver} &= \text{softmax}((M^{ver})^\top,\ \text{dim} = -1), \end{aligned} \tag{15}$$

where $M_{v\to u}^{ver}[bW+g,i,j]$ denotes the correlation between $f_u[b,i,g]$ and $f_v[b,j,g]$. This value can be interpreted as the probability that $f_u[b,i,g]$ attends to information from $f_v[b,j,g]$. Similarly, $M_{u\to v}^{ver}[bW+g,i,j]$ denotes the correlation between $f_v[b,i,g]$ and $f_u[b,j,g]$.

Next, the second dimension of $M_{v\to u}^{ver}$ and the third dimension of $M_{u\to v}^{ver}$ are reshaped into the batch dimension to obtain $M_{v\to u,b}^{ver} \in \mathbb{R}^{BWH\times 1\times H}$ and $M_{u\to v,b}^{ver} \in \mathbb{R}^{BWH\times H\times 1}$. These matrices are then used to compute the batch vertical cycle consistency [12], denoted as $C_{u,b}^{ver} \in \mathbb{R}^{BWH\times 1\times 1}$:

$$C_{u,b}^{ver} = M_{v\to u,b}^{ver} \otimes M_{u\to v,b}^{ver}, \tag{16}$$

where

$$\begin{aligned} &C_{u,b}^{ver}[bWH+gH+i,0,0] = \\ &\sum_{j=0}^{H-1} M_{v\to u,b}^{ver}[bWH+gH+i,0,j] \\ &\quad\times M_{u\to v,b}^{ver}[bWH+gH+i,j,0]. \end{aligned} \tag{17}$$

Here, $M_{v\to u,b}^{ver}[bWH+gH+i,0,j]$ represents the correlation between $f_u[b,i,g]$ and $f_v[b,j,g]$, indicating the probability that $f_v[b,j,g]$ corresponds to $f_u[b,i,g]$ within column g of f_v. Similarly, $M_{u\to v,b}^{ver}[bWH+gH+i,j,0]$ represents the correlation between $f_v[b,j,g]$ and $f_u[b,i,g]$, indicating the probability that $f_u[b,i,g]$ corresponds to $f_v[b,j,g]$ within column g of f_u.

Lemma 2. *The vertical cycle consistency $C_{u,b}^{ver}$ has an upper bound of 1.*

Proof. Since all elements of $M_{v\to u,b}^{ver}$ and $M_{u\to v,b}^{ver}$ are

within the range $[0, 1]$, we have:

$$
\begin{aligned}
&C^{ver}_{u,b}[bWH + gH + i, 0, 0] = \\
&\sum_{j=0}^{H-1} M^{ver}_{v\to u,b}[bWH + gH + i, 0, j] \\
&\times M^{ver}_{u\to v,b}[bWH + gH + i, j, 0] \\
&\leq \Big(\sum_{j=0}^{H-1} M^{ver}_{v\to u,b}[bWH + gH + i, 0, j] \Big) \times 1 \\
&= 1.
\end{aligned} \tag{18}
$$

The equality is achieved when $M^{ver}_{u\to v,b}[bWH + gH + i, j, 0] = 1$ for any index j. □

A high value of the product $M^{ver}_{v\to u,b}[bWH + gH + i, 0, j] \times M^{ver}_{u\to v,b}[bWH + gH + i, j, 0]$ suggests strong mutual correspondence between $f_u[b, i, g]$ and $f_v[b, j, g]$, implying that $f_v[b, j, g]$ can reliably provide information for reconstructing $f_u[b, i, g]$. Conversely, when $M^{ver}_{v\to u,b}$ is high but $M^{ver}_{u\to v,b}$ is low, it indicates potential ambiguity in the matching within column g of f_v, meaning that $f_u[b, i, g]$ may have been mismatched to $f_v[b, j, g]$, while another position in f_u might correspond more accurately to $f_v[b, j, g]$. The same reasoning applies when $M^{ver}_{v\to u,b}$ is low but $M^{ver}_{u\to v,b}$ is high.

Therefore, the product of $M^{ver}_{v\to u,b}$ and $M^{ver}_{u\to v,b}$ mitigates mismatches between the two sources. Consequently, $C^{ver}_{u,b}[bWH + gH + i, 0, 0]$ represents the reliability of reconstructing $f_u[b, i, g]$ using all possible positions $f_v[b, j, g]$, i.e., all locations in the g-th column of f_v.

Then, $C^{ver}_{u,b}$ is reshaped to obtain the vertical cycle consistency map $C^{ver}_u \in \mathbb{R}^{B\times H\times W}$, where $C^{ver}_u[b, i, g] = C^{ver}_{u,b}[bWH + gH + i, 0, 0]$. The vertical cycle consistency C^{ver}_u measures the reliability of reconstructing each position in f_u using all positions in the same column of f_v.

Finally, the column dimension of $f_v \in \mathbb{R}^{B\times H\times W\times C}$ is reshaped into the batch dimension to obtain $f_{v,b} \in \mathbb{R}^{BW\times H\times C}$. The batch vertically aligned feature $f^{ver}_{u,b} \in \mathbb{R}^{BW\times H\times C}$ is then computed as:

$$
f^{ver}_{u,b} = M^{ver}_{v\to u} \otimes f_{v,b}, \tag{19}
$$

where

$$
\begin{aligned}
f^{ver}_{u,b}[bW + g, i, k] = &\sum_{j=0}^{H-1} M^{ver}_{v\to u}[bW + g, i, j] \\
&\times f_{v,b}[bW + g, j, k].
\end{aligned} \tag{20}
$$

Here, $M^{ver}_{v\to u}[bW + g, i, j]$ denotes the correlation between $f_u[b, i, g]$ and $f_v[b, j, g]$, and $f_{v,b}[bW + g, j, k]$ denotes the k-th channel value of $f_v[b, j, g]$. Therefore, $f^{ver}_{u,b}[bW + g, i, k]$ represents the reconstructed k-th channel of $f_u[b, i, g]$ using all $f_v[b, j, g]$, i.e., all positions in the g-th column of f_v. Then, $f^{ver}_{u,b}$ is reshaped into the vertically aligned feature $f^{ver}_u \in \mathbb{R}^{B\times H\times W\times C}$, where $f^{ver}_u[b, i, g, k] = f^{ver}_{u,b}[bW + g, i, k]$. The vertically aligned feature f^{ver}_u represents the transformation of the side source into the perspective of the main source, aggregating information from all positions in the corresponding column of f_v.

The aggregation step in Eq. 19, i.e., $(f^{ver}_{u,b} = M^{ver}_{v\to u} \otimes f_{v,b})$, adds (BWH^2C) MACs. Meanwhile, the cycle consistency in Eq. 16 contributes (BWH^2) MACs, which is negligible when $C \gg 1$. Therefore, the overall computational complexity of the vertical parallax attention mechanism is $\mathcal{O}(N^3)$, where $N = \max(H, W)$.

B.3. Two-Stage Parallax Attention in OPAM

After introducing the processes of HPA and VPA, we now present the overall process of OPAM, which sequentially applies HPA and VPA to capture the full two-dimensional spatial context.

Given the input features f_l and f_r, the HPA is first applied to capture long-range dependencies along the row dimension, yielding the horizontally aligned feature f^{hor}_l and the horizontal cycle consistency C^{hor}_l. The formulation of the horizontally aligned feature f^{hor}_l is given by:

$$
\begin{aligned}
f^{hor}_l[b, g, i, k] = &\sum_{j=0}^{W-1} M^{hor}_{r\to l}[bH + g, i, j] \\
&\times f_r[b, g, j, k],
\end{aligned} \tag{21}
$$

where $M^{hor}_{r\to l}$ denotes the horizontal parallax attention map from f_r to f_l.

Next, f_l and f^{hor}_l are fed into VPA, which models vertical dependencies and produces the vertically aligned feature f^{ver}_l along with the vertical cycle consistency C^{ver}_l. The formulation of the vertically aligned feature f^{ver}_l is given by:

$$
\begin{aligned}
f^{ver}_l[b, i, g, k] = &\sum_{j=0}^{H-1} M^{ver}_{r\to l}[bW + g, i, j] \\
&\times f^{hor}_l[b, j, g, k],
\end{aligned} \tag{22}
$$

where $M^{ver}_{r\to l}$ denotes the vertical parallax attention map from f^{hor}_l to f_l.

The overall cycle consistency C_l is computed by combining both horizontal and vertical consistency:

$$
C_l = C^{hor}_l \odot C^{ver}_l, \tag{23}
$$

where $\odot$ denotes the element-wise product.

By combining Equations 21 and 22, the final transformation can be expressed as:

Table 1. **BDBR comparison relative to LDMIC across different datasets.** The best results are highlighted in **bold**. The number of input views is indicated in parentheses. A dash (–) denotes that the corresponding model cannot be evaluated on that dataset.

Methods	InStereo2K(2)		Cityscapes(2)		WildTrack(3)		WildTrack(6)		Mip-NeRF 360(3)	
	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM
LDMIC	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
VVC	48.68%	80.69%	54.33%	101.92%	49.47%	115.64%	25.16%	120.67%	7.14%	96.47%
MV-HEVC	84.84%	182.11%	106.27%	202.95%	31.84%	176.86%	10.01%	168.04%	41.15%	171.15%
ECSIC	25.65%	11.99%	50.51%	41.12%	–	–	–	–	–	–
BiSIC	-2.95%	-4.65%	11.40%	-7.11%	–	–	–	–	–	–
BiSIC-fast	7.37%	-3.09%	18.65%	-7.00%	–	–	–	–	–	–
CAMSIC	1.05%	4.98%	8.64%	-5.02%	–	–	–	–	–	–
LMVIC	–	–	–	–	–	–	–	–	-14.30%	93.73%
LDMIC-fast	13.15%	-1.85%	3.90%	24.58%	10.83%	9.24%	10.51%	9.67%	25.52%	16.05%
ParaHydra	**-6.92%**	**-5.95%**	**-1.42%**	**-7.15%**	**-19.72%**	**-7.07%**	**-24.18%**	**-6.23%**	**-18.20%**	**-4.53%**

$$f_l^{ver}[b,i,j,k] = \sum_{g=0}^{H-1} M_{r\to l}^{ver}[bW+j,i,g] f_l^{hor}[b,g,j,k] = \sum_{g=0}^{H-1}\sum_{s=0}^{W-1} M_{r\to l}^{ver}[bW+j,i,g] M_{r\to l}^{hor}[bH+g,j,s] f_r[b,g,s,k], \tag{24}$$

which demonstrates that by sequentially applying horizontal and vertical parallax attention, the parallax attention is no longer constrained to aggregation along a single epipolar line. Instead, it can exploit the full two-dimensional spatial context (Fig. 1 *Right*), providing a reliable reference f_l^{ver} and corresponding consistency C_l. The consistency C_l captures the joint reliability of context information provided by f_r for reconstructing f_l, guided by both horizontal and vertical parallax attention. It can be interpreted as the semantic relevance between f_r and f_l.

Since the complexity of one stage of parallax attention is $\mathcal{O}(N^3)$, the overall complexity of OPAM remains $\mathcal{O}(N^3)$, which is far more efficient than full two-dimensional self-attention with complexity $\mathcal{O}(N^4)$.

C. Experimental Details

C.1. Dataset

We evaluate the proposed framework on two public stereo image datasets, **InStereo2K** [1] and **Cityscapes** [5], and two multi-view datasets, **WildTrack** [4] and **Mip-NeRF 360** [2].

- **InStereo2K.** The InStereo2K dataset contains 2060 image pairs of close-view indoor scenes, with 2010 pairs used for training and 50 pairs for testing. Each image has a resolution of 1080×860.
- **Cityscapes.** The Cityscapes dataset comprises 5000 stereo image pairs depicting outdoor street scenes. The dataset is split into 2975 training pairs, 500 validation pairs, and 1525 testing pairs. Each image has a resolution of 2048×1024.
- **WildTrack.** For the WildTrack dataset, we use `FFmpeg` to extract frames from seven HD 1080p videos at one frame per second [14]. The first 2000 frames from each view are used for training, and the remaining 51 frames are used for testing. Each image has a resolution of 1920×1080.
- **Mip-NeRF 360.** The Mip-NeRF 360 dataset consists of real-world 360° scenes captured under varying exposure conditions. Many scenes reach 4K resolution. This dataset is particularly challenging due to its unbounded environments, complex lighting, and wide range of viewpoints.

Training and testing settings on the datasets follow previous works [8, 14] to ensure a fair comparison. During evaluation, each image in the InStereo2K dataset is minimally cropped so that both its height and width are multiples of 64. For the Cityscapes dataset, we follow the cropping procedure described in [14] to remove rectification artifacts and the ego-vehicle. Specifically, 64, 256, and 128 pixels are cropped from the top, bottom, and sides of each image, respectively.

C.2. Implementation Details

Implementation of the Baseline Codecs. We employ the evaluation scripts provided by CompressAI to obtain the results of traditional codecs. Images are first converted into YUV 4:4:4 video format, followed by compression using the corresponding codec implementations. The coding efficiency of MV-HEVC is evaluated with the HTM-16.3 reference software, whereas the performance of VVC is assessed using the VTM-23.6 reference implementation.

For DNN-based image codecs, we reimplement all methods to ensure a fair comparison.

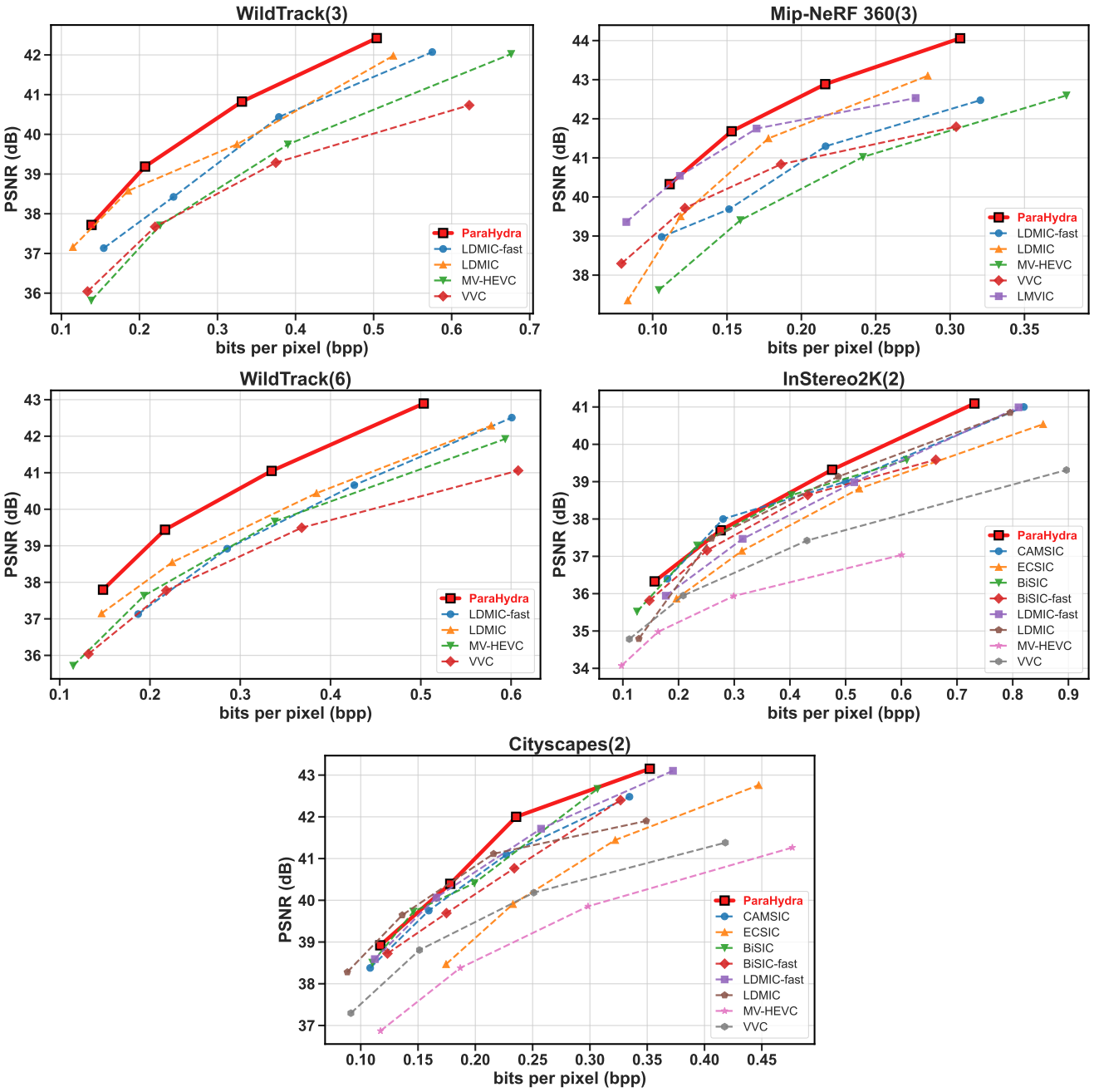

Figure 2. Rate-distortion curves of the proposed method compared with competitive baselines, where distortion is measured in PSNR.

Implementation of the Proposed Framework. The implementation follows that of [6, 7, 14]. Specifically, the number of channels for the latent representations y_k is set to $M = 192$, and the number of channels for the latent representations z_k and f_k is set to $N = 192$. The number of channel slices l is set to 8, resulting in $s_c = 24$ channels per slice. The window size w is set to 5.

All learning-based models are trained with the trade-off parameter $\lambda = 1024, 2048, 4096, 8192\,(32, 64, 128, 256)$ under MSE (MS-SSIM). We train our models for 1400 epochs on multi-view datasets with a batch size of 8 and for 3000 epochs on stereo image datasets with a batch size of 16. During training, images are randomly cropped to a resolution of 256×256 [14, 15]. The AdamW optimizer is employed with a learning rate of 10^{-4}. Experiments are conducted on a single NVIDIA A30 GPU.

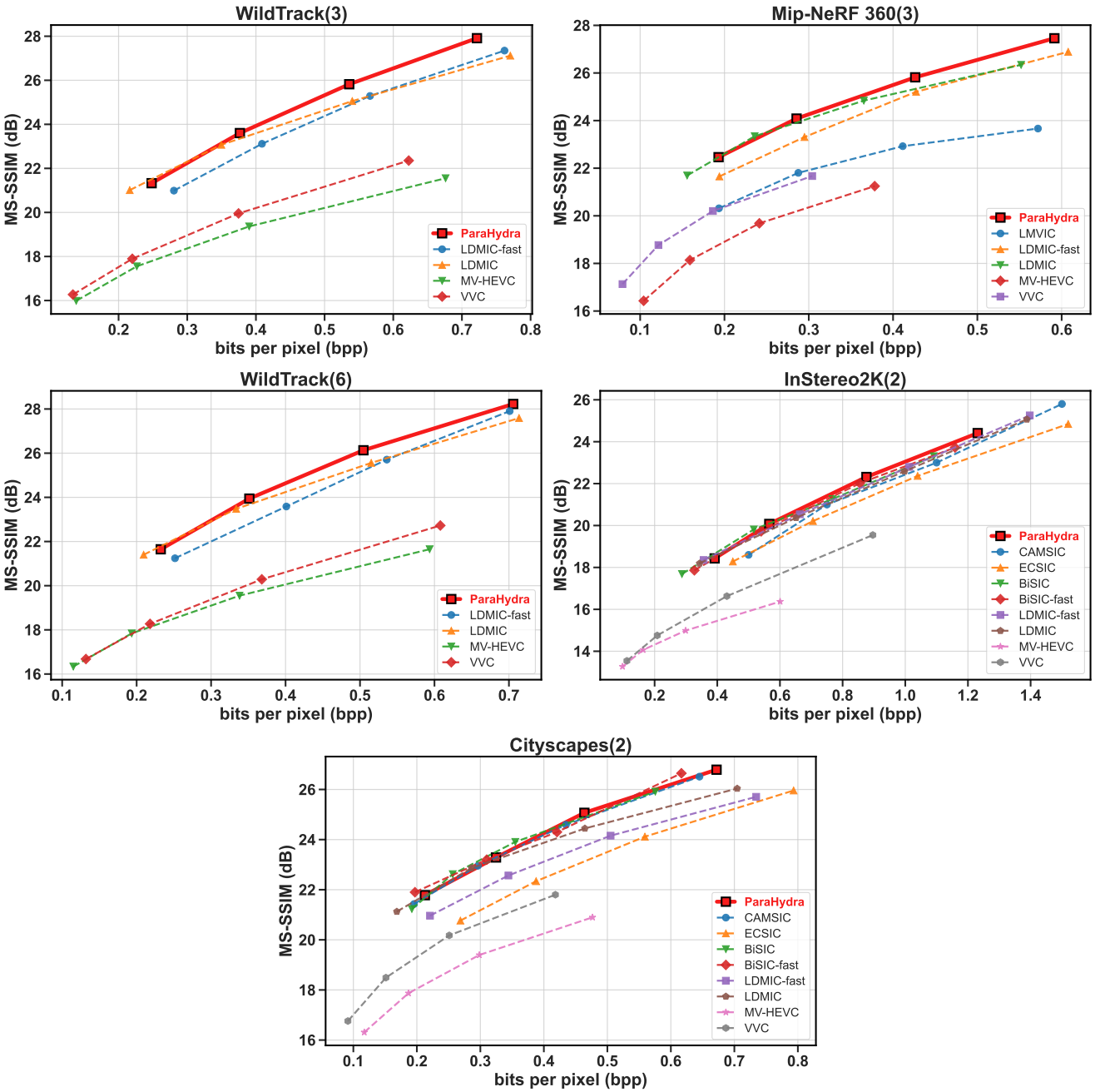

Figure 3. Rate-distortion curves of the proposed method compared with competitive baselines, where distortion is measured in MS-SSIM.

Ablation Study Details. We conduct ablation study on the WildTrack dataset using three input views to evaluate the contribution of each proposed module. All models are optimized with the MSE loss. Based on the proposed framework, we insert two PMIFM into the encoder to construct the *Joint Enc-Dec* variant, allowing both the encoder and decoder to access inter-view contextual information. For the *Sep Enc-Dec* variant, the two PMIFM in the decoder are removed, making it equivalent to single image compression.

To assess the effectiveness of OPAM, we replace it within the PMIFM with the JCT operation [14] to obtain *JCT*, and with 2D self-attention to obtain *2D Attn*. We further remove the HPA stage in OPAM to obtain *w/o HPA*, and remove the VPA stage to obtain *w/o VPA*.

To evaluate the effectiveness of the proposed Para-EM in exploiting multiple contexts within entropy models, we

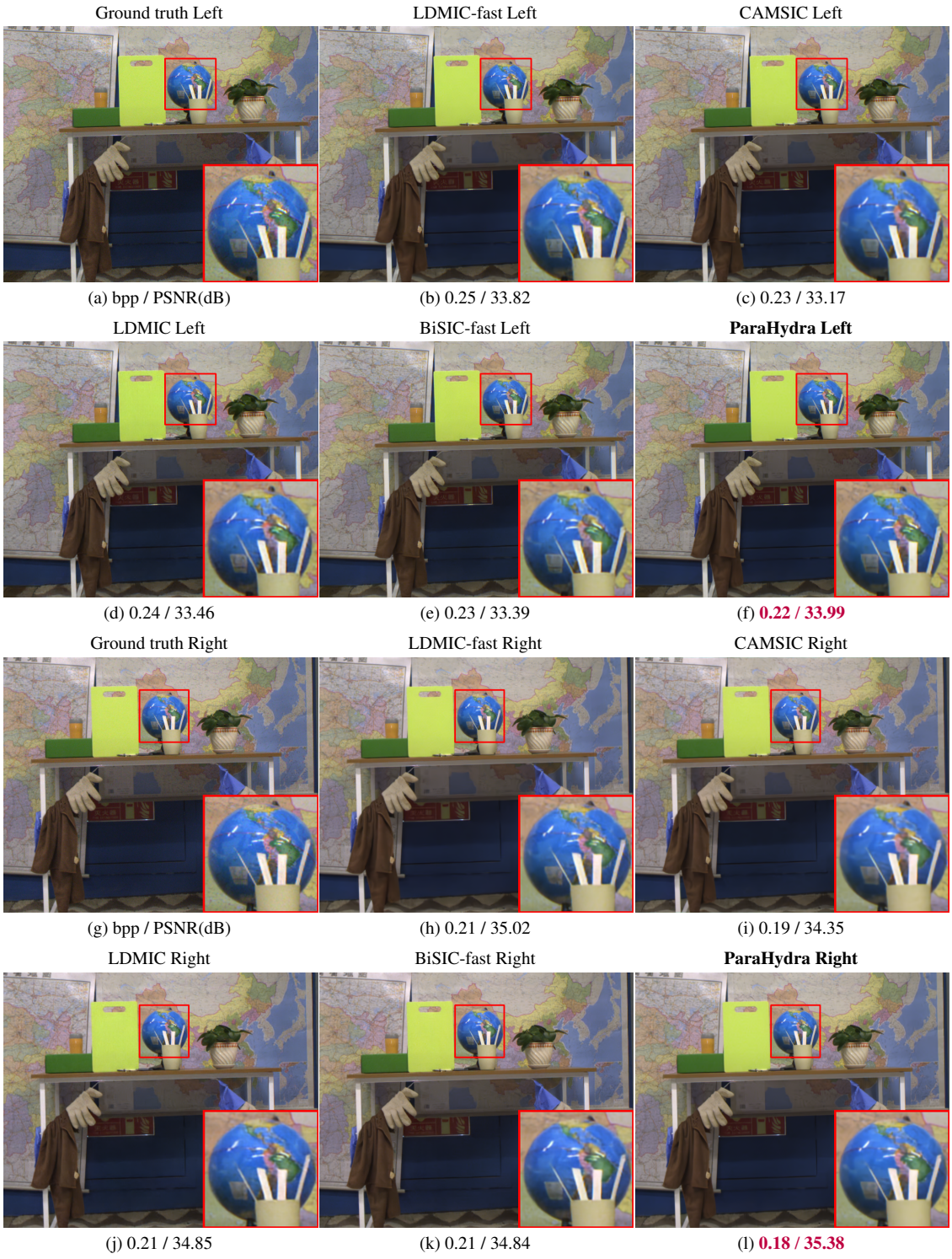

Figure 4. **Qualitative comparison on the InStereo2K dataset.** Both left and right view reconstructions are shown, with their corresponding bpp and PSNR values. Red boxes denote zoomed-in regions for detailed visual inspection. The best result is highlighted in **bold**.

individually replace the proposed channel-wise and global spatial context modules with their counterparts from [6]. Specifically, replacing the PCCM with the channel-wise context module [6] yields *w/o PCCM*, while replacing the

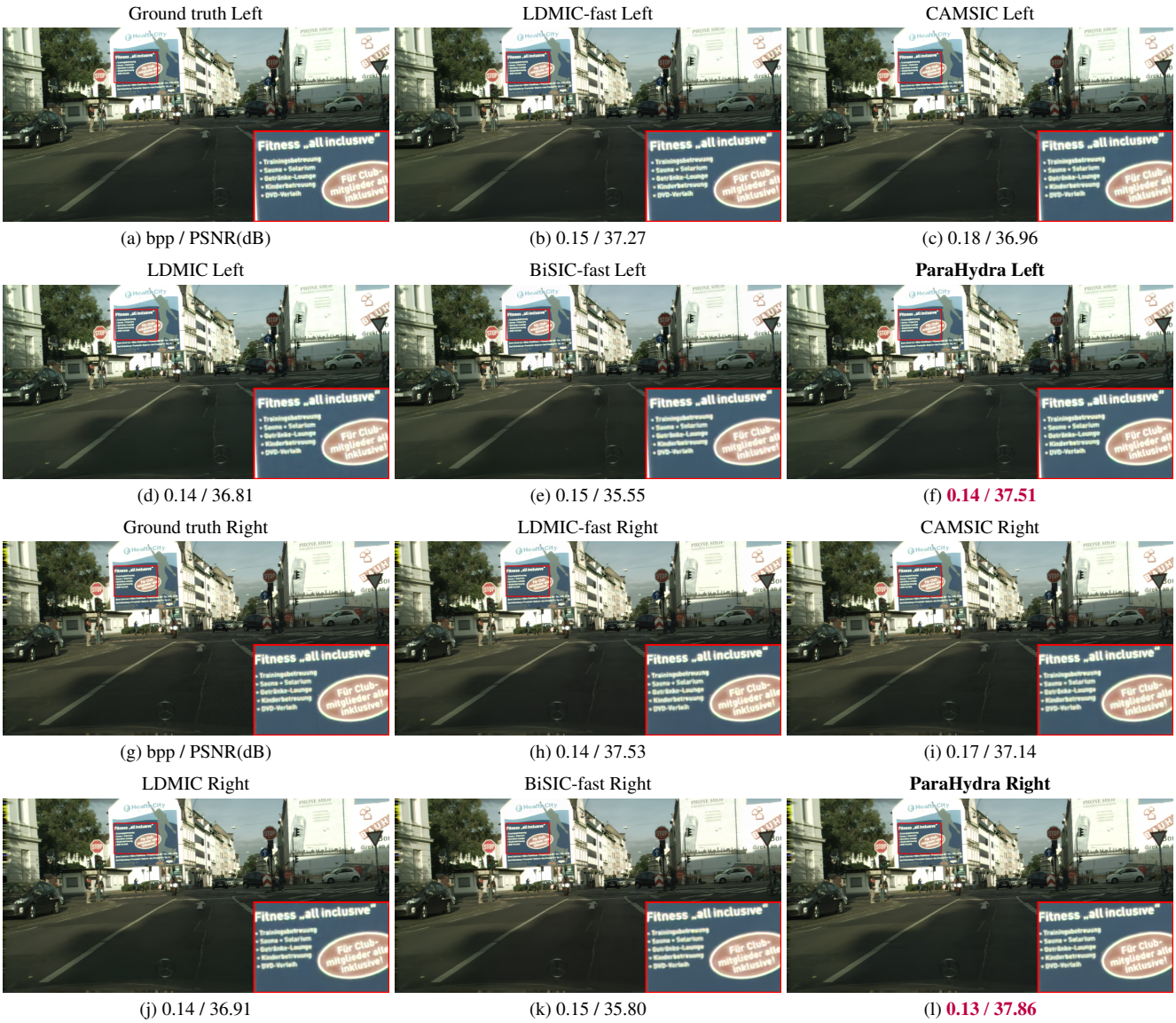

Figure 5. **Qualitative comparison on the Cityscapes dataset.** Both left and right view reconstructions are shown, with their corresponding bpp and PSNR values. Red boxes denote zoomed-in regions for detailed visual inspection. The best result is highlighted in **bold**.

PGCM with the intra-slice global context module [6] yields *w/o PGCM*. Apart from these changes, all other components remain identical to the original model. Furthermore, the *MLIC* variant is obtained by replacing the entire Para-EM with the MLIC module from [6].

D. Quantitative Results

Fig. 2 and Fig. 3 show the rate-distortion (RD) performance of various codecs under PSNR and MS-SSIM metrics, respectively, while Tab. 1 presents the corresponding BDBR values relative to LDMIC. Across all datasets and configurations, the proposed ParaHydra consistently demonstrates clear advantages over both traditional and learning-based codecs, highlighting its robustness and generalization across different scene types and view settings.

Table 2. **BDBR comparison relative to LDMIC on Mip-NeRF 360.**

Views	VVC	MV-HEVC	LDMIC-fast	LMVIC	**ParaHydra**
3	7.14%	41.15%	25.52%	-14.30%	**-18.20%**
4	53.06%	91.30%	14.31%	17.27%	**-16.84%**

Figure 6. **Qualitative comparison on the WildTrack dataset.** Reconstructions from three camera views are shown with their corresponding bpp and PSNR values. Red boxes denote zoomed-in regions for detailed visual inspection. The best result is highlighted in **bold**.

Overall Performance. On both the WildTrack and Mip-NeRF 360 datasets, ParaHydra achieves substantial bitrate savings, reaching up to **24.18%** saving on WildTrack(6) and **18.20%** saving on Mip-NeRF 360(3) relative to LDMIC. These savings indicate that the OmniParallax Attention Mechanism (OPAM) and Parallax Multi Information Fusion Module (PMIFM) can effectively exploit semantic correlations among multiple views. These improvements are achieved *without any inter-view side priors at the encoder side*, emphasizing the strength of the distributed coding paradigm.

Notably, on the 4K-resolution Mip-NeRF 360 dataset, our method surpasses LMVIC, which utilizes 3D Gaussian geometric priors during encoding. As shown in Tab. 2, with three input views, ParaHydra achieves an average bitrate saving of 3.9% compared with LMVIC. When the number of views increases to four, the average bitrate saving further improves to **34.11%**. This result is particularly significant, as it demonstrates not only the robustness of ParaHydra in high-resolution settings but also the ability of the proposed DMIC framework to outperform joint encoding-decoding methods by effectively leveraging inter-view information.

Analysis on Different View Numbers. The performance gains become more pronounced as the number of input views increases. On WildTrack, the bitrate savings rise

from **19.72%** (3 views) to **24.18%** (6 views), confirming that PMIFM scales effectively with the complexity of multi-view dependencies. The adaptive nature of PMIFM enables the system to integrate useful information from multiple correlated features while suppressing redundant or inconsistent features. This trend shows the scalability of the model: as more information sources are available, the fusion mechanism aggregates them more efficiently to enhance coding performance.

When the number of input views is limited to two (as in InStereo2K and Cityscapes), the improvement margin, although smaller, remains clear. This is expected because, in two-view scenarios, the main view can only reference a single side view. As a result, the PMIFM in the joint decoding branch (Para-JD) cannot fully utilize its multi-view fusion capacity. Even in this constrained setting, ParaHydra still surpasses all baselines, achieving up to **6.92%** bitrate saving on InStereo2K and **1.42%** saving on Cityscapes compared with LDMIC. Although PMIFM in the Para-JD branch faces limited cross-view diversity under two-view settings, the Para-EM remains capable of effectively utilizing context information within latent representations to achieve high-quality image reconstruction. This compensates for the reduced inter-view reference information and demonstrates the robustness of ParaHydra even under limited view diversity.

Comparison with Traditional Codecs. Compared with traditional codecs such as VVC and MV-HEVC, ParaHydra shows clear superiority. Traditional codecs depend on hand-designed prediction and transform schemes that struggle to represent complex semantic relationships across diverse viewpoints. In contrast, ParaHydra learns to capture long-range correlations, achieving higher reconstruction quality at significantly lower bitrates. This advantage is particularly evident on high-resolution datasets like WildTrack, where geometric and photometric variations are pronounced.

Discussion. The consistent improvements across all benchmarks demonstrate that the proposed framework effectively leverages inter-view information during decoding. The OPAM and PMIFM enable ParaHydra to align and fuse informative features across sources, resulting in compact yet expressive latent representations. The small variations in BDBR under two-view setups are likely due to limited reference diversity rather than architectural limitations. Since most practical multi-camera systems operate with three or more views, this phenomenon is both expected and acceptable.

In summary, the results validate the **robustness, scalability, and generality** of ParaHydra. It not only establishes new state-of-the-art performance among distributed methods but also outperforms joint encoding-decoding approaches, setting a new benchmark for future research in multi-view image compression.

E. Qualitative Results

We present additional qualitative results in Figures 4, 5, and 6 to further validate the effectiveness of our proposed framework against representative baselines, including LDMIC, LDMIC-fast, BiSIC-fast, and CAMSIC. These comparisons are conducted on three challenging datasets: InStereo2K, Cityscapes, and WildTrack, covering diverse scenarios from indoor stereo images to complex outdoor urban environments.

As shown in the figures, our method consistently preserves *sharper edges and finer texture details* compared to competing approaches at lower bitrates.

Fig. 4 demonstrates that on the InStereo2K dataset, our proposed ParaHydra achieves higher PSNR and lower bitrates (bpp) for both left and right views compared to other methods. Furthermore, ParaHydra preserves high-fidelity structural and textural details. In geometrically discontinuous regions, such as the upper boundary of the globe (Fig. 4), our method maintains texture curves that closely match the ground truth. Unlike other methods that exhibit noticeable blurring, ParaHydra demonstrates superior capability in preserving stereo-geometric consistency.

Fig. 5 shows reconstructed images from different methods on the Cityscapes dataset. ParaHydra achieves higher-quality reconstructions for both left and right views while operating at the lowest bitrates. This performance is particularly notable compared to benchmark codecs such as CAMSIC and BiSIC-fast, which yield lower PSNR values of 36.96 dB and 35.55 dB, respectively, while ParaHydra achieves an impressive PSNR of 37.51 dB. Notably, our method produces clearer structural details than other approaches. For instance, text on distant billboards appears sharply rendered in our results, whereas other reconstructions often exhibit blurring, local distortion, and texture smoothing artifacts.

On the WildTrack dataset with multiple camera viewpoints (Fig. 6), the PMIFM fully exploits its multi-view fusion capability. The adaptive nature of our framework enables effective integration of relevant information from multiple correlated views while suppressing redundant or inconsistent features. Experimental results show that our approach generates visually consistent reconstructions across different viewpoints at approximately **0.17 bpp**, achieving up to 0.7 dB PSNR improvement over the baseline method LDMIC-fast. This validates the effectiveness of our method in capturing inter-view correlations and maintaining geometric alignment, thereby ensuring consistent visual quality across all viewpoints.

Overall, the qualitative results demonstrate that the proposed framework successfully leverages multi-view seman-

tic correlations to reconstruct fine-grained scene details more accurately, achieving higher perceptual quality and a superior rate-distortion trade-off compared to existing stereo and multi-view image compression methods.

References

[1] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63(11):212101, 2020. 5

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. 5

[3] Toby Berger. Multiterminal source coding. *The information theory approach to communications*, 1978. 1

[4] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5039, 2018. 5

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 5

[6] Wei Jiang, Jiayu Yang, Yongqi Zhai, Feng Gao, and Ronggang Wang. Mlic++: Linear complexity multi-reference entropy modeling for learned image compression. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(5), 2025. 6, 8, 9

[7] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–519, 2019. 1, 3, 6

[8] Zhening Liu, Xinjie Zhang, Jiawei Shao, Zehong Lin, and Jun Zhang. Bidirectional stereo image compression with cross-dimensional entropy model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496. Springer, 2024. 5

[9] Nitish Mital, Ezgi Özyılkan, Ali Garjani, and Deniz Gündüz. Neural distributed image compression using common information. In *Data Compression Conference (DCC)*, pages 182–191. IEEE, 2022. 1

[10] Sergio D Servetto. Multiterminal source coding with two encoders–i: A computable outer bound. *arXiv preprint cs/0604005*, 2006. 1

[11] David Slepian and Jack Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 2003. 1

[12] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2108–2125, 2022. 2, 3

[13] Jack Keil Wolf. Data reduction for multiple correlated sources. In *Proc. 5th Colloquium on Microwave Commun., Budapest, Hungary, June 1973*, pages 287–295, 1973. 1

[14] Xinjie Zhang, Jiawei Shao, and Jun Zhang. Ldmic: Learning-based distributed multi-view image coding. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 1, 5, 6, 7

[15] Xinjie Zhang, Shenyuan Gao, Zhening Liu, Jiawei Shao, Xingtong Ge, Dailan He, Tongda Xu, Yan Wang, and Jun Zhang. Camsic: Content-aware masked image modeling transformer for stereo image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 10239–10247, 2025. 6