BLIND OMNIDIRECTIONAL VIDEO SUPER-RESOLUTION VIA GEOMETRY-CONDITIONED PROBABILISTIC DEGRADATION MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Blind Video Super-Resolution (BVSR) promotes generalization across diverse imaging conditions by explicitly accounting for unknown degradations to guide reconstruction. However, its extension to OmniDirectional Video (ODV) remains underexplored. The unique fisheye imaging and EquiRectangular Projection (ERP) process of ODVs couples unknown capture degradations with projection geometry, making local degradation characteristics jointly dependent on both factors and complicating degradation inference. Therefore, we propose **OmniVaria**, the first blind ODVSR framework with a particular focus on modeling the interaction between unknown degradations and spherical geometry. In particular, we introduce Probabilistic Spherical Degradation Conditioning (PSDC), which leverages explicit spherical geometry as an additional condition in probabilistic degradation modeling, *i.e.*, representing unknown degradations through geometrically conditioned multivariate Gaussian distributions in a latent space, with posterior statistics inferred from Low-Resolution (LR) ODV observations guiding reconstruction. For PSDC optimization, we establish an information-theoretically grounded regularization term, named Geometry-Conditioned Variational Information Bottleneck (GCVIB), and derive its tractable variational form for practical optimization. Drawing on information bottleneck theory, GCVIB encourages PSDC to learn a minimally sufficient representation of LR observations for High-Resolution (HR) reconstruction, thereby promoting generalization to unseen scenes and degradation levels. Furthermore, recognizing that ERP projection transforms coherent scene motion into spatially varying displacements and deformations, complicating the exploitation of temporal context, we bypass explicit motion estimation and develop Implicit Omnidirectional Temporal Aggregation (IOTA), which jointly processes current-frame features and propagated hidden states to exploit inter-frame correlations for reconstruction. Extensive experiments on various ODV benchmarks and degradation settings demonstrate the effectiveness of our design. The code will be publicly released.

1 INTRODUCTION

Blind Video Super-Resolution (BVSR) aims to reconstruct High-Resolution (HR) videos from Low-Resolution (LR) observations with unknown degradations (Pan et al., 2021; Jeelani et al., 2023; Lee et al., 2021; Bai & Pan, 2024; Zhu et al., 2025). By accounting for unknown degradations, BVSR extends video enhancement beyond the prescribed degradation settings assumed by conventional VSR methods (Chan et al., 2021; 2022; Xu et al., 2024; Zhou et al., 2024), offering greater flexibility across diverse imaging conditions and improving its applicability to real-world scenarios. Extensive research on conventional planar videos has made significant progress by explicitly modeling degradation characteristics and exploiting inter-frame contextual information to provide additional guidance and complementary details for reconstruction. However, BVSR remains underexplored for OmniDirectional Video (ODV), an emerging format characterized by its immersive viewing experience and rapid adoption in autonomous driving, virtual reality, and live streaming. Its unique imaging process hinders extending existing methods to omnidirectional content by complicating both degradation inference and the exploitation of temporal contextual information.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

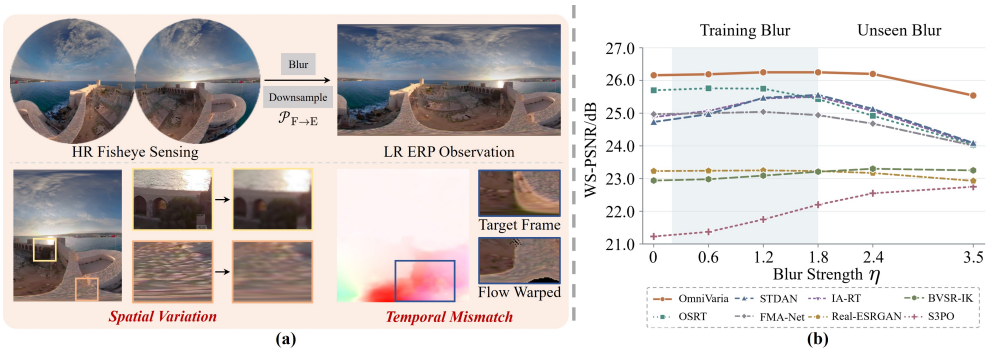

Figure 1: (a) Fisheye imaging process produces spatially varying degradation after ERP reprojection and further complicates temporal correspondence. (b) Comparison with other state-of-the-art methods on both seen and unseen blur on ODV360.

To be specific, ODVs capture a full $360° \times 180°$ field of view, commonly through fisheye cameras whose observations are stitched and mapped into the EquiRectangular Projection (ERP) format for the sake of storage and transmission. During this process, nonlinear projection and resampling transform blur and downsampling in the fisheye domain into spatially varying degradations in ERP frames, as shown in Fig. 1(a), entangling the degradation with the projection geometry. Consequently, degradation inference must account for both unknown capture degradations and their location-dependent interaction with projection geometry, making a global degradation representation insufficient. Besides, this imaging process also complicates temporal correspondence, *i.e.,* scene motion manifests as position-dependent displacements and deformations in ERP frames (Artizzu et al., 2021; Bhandari et al., 2021; Shi et al., 2023), while correspondences crossing the ERP seam introduce additional discontinuities. These geometric effects, compounded by degradation-induced loss of matching cues, hinder reliable motion estimation and temporal information aggregation. Although existing ODVSR methods offer insights into incorporating spherical geometry into reconstruction (Liu et al., 2024; Baniya et al., 2023; An et al., 2026b), they predominantly assume prescribed degradations, leaving the coupled challenges of unknown degradation inference and projection-dependent temporal correspondence insufficiently addressed.

Under these circumstances, we propose the first blind ODVSR framework, named OmniVaria, with a particular focus on the challenges posed by the unique fisheye imaging and ERP reprojection process of ODVs in modeling degradation distributions and capturing temporal correspondences.

We summarize the main contributions and corresponding modules as follows:

- To account for spatial variation and ambiguity in degradation inference, we introduce Probabilistic Spherical Degradation Conditioning (PSDC), which uses known ERP geometry as a reference to infer posterior distributions over local degradation conditions from LR observations. The inferred degradation conditions are subsequently used to adapt reconstruction features to local degradation through feature modulation gated by the posterior statistics.

- To guide PSDC learning and promote generalization, we seek to reduce the dependence of the inferred degradation conditions on scene-specific content. To this end, we formulate a Geometry-Conditioned Variational Information Bottleneck (GCVIB), which provides an information-theoretically grounded regularization objective that discourages redundant content encoding while encouraging the capture of reconstruction-relevant degradation cues.

- To exploit temporal information under projection-dependent motion and degraded matching cues, we further develop Implicit Omnidirectional Temporal Aggregation (IOTA). By jointly processing current-frame features and propagated hidden states, IOTA integrates complementary temporal information without explicit optical-flow estimation or feature warping.

- Extensive experiments on multiple ODV benchmarks demonstrate the effectiveness of OmniVaria across varying degradation levels, including improved generalization to severe blur beyond the training range.

2 Related Work

2.1 Omnidirectional Super-Resolution

Existing omnidirectional image super-resolution methods handle spherical geometry mainly in two ways: *Distortion-aware Adaptation* and *Projection-aware Representation*. By leveraging sphere-aware objectives, adaptive sampling, or feature modulation to compensate for ERP distortion, distortion-aware adaptation methods process ERP content directly. Ozcinar et al. (2019) augment adversarial SR with a sphere-aware loss. LAU-Net partitions an ERP image into latitude bands and assigns upscaling paths (Deng et al., 2021). OSRT (Yu et al., 2023) subsequently introduces fisheye downsampling and modulates features using distortion-aware offsets, while OPDN (Sun et al., 2023) combines deformable feature extraction with frequency-domain enhancement. FAOR (Shen et al., 2025b) adapts implicit functions using spherical priors. Projection-aware representation methods instead mitigate ERP distortion through alternative representations. SphereSR (Yoon et al., 2022), the spherical pseudo-cylindrical representation (Cai et al., 2024), and BPOSR (Wang et al., 2024) leverage continuous spherical or multi-projection representations, while OmniSSR (Li et al., 2024) transfers diffusion priors through tangent projections. More recent methods, including RealOSR (Sheng et al., 2024) and D^2R^2OSR (An et al., 2026a), further consider complex or real-world degradation, but remain image-based.

ODVSR extends these spatial adaptations by exploiting inter-frame correspondence under distortion. SMFN (Liu et al., 2024) combines single- and multi-frame branches with deformable alignment, while S3PO (Baniya et al., 2023) employs recurrent attention with spherical reconstruction optimization. STDAN (An et al., 2026b) further integrates spatio-temporal continuous alignment with interlaced multi-frame reconstruction to improve geometric and temporal coherence. These ODVSR methods improve reconstruction under spherical distortion, but primarily consider prescribed degradations, leaving degradation-adaptive temporal reconstruction underexplored.

2.2 Blind Super-Resolution

A common strategy in degradation-aware blind SR is to infer degradation information from LR observations and use it to guide reconstruction, either through explicit kernel estimation or learned implicit degradation representation. Explicit methods retain a physical degradation model: IKC (Gu et al., 2019) iteratively corrects an estimated blur kernel, DAN (Luo et al., 2020) unfolds alternating kernel estimation and restoration, and DCLS (Luo et al., 2022) combines dynamic kernel estimation with constrained filtering. Although effective under analytic blurs, these methods are limited by kernel estimation errors and predefined degradation spaces. Implicit methods instead learn an implicit representation to distinguish different degradations. DASR (Wang et al., 2021a) uses contrastive learning to distinguish degradations, while KDSR (Xia et al., 2022) distills degradation knowledge from an HR-LR teacher into an LR estimator. These representations avoid direct kernel supervision but are often deterministic and globally pooled, which can obscure local degradation variation. BVSR further models temporal degradation. DynaVSR (Lee et al., 2021) meta-learns degradation adaptation, DBVSR (Pan et al., 2021) jointly estimates blur, motion, and HR frames, FMA-Net (Youk et al., 2024) uses flow-guided dynamic filtering for spatio-temporally varying degradation, and BVSR-IK (Zhu et al., 2025) introduces a multi-scale implicit kernel dictionary for correction and alignment. These advances improve robustness to dynamic blur, yet most formulations target planar videos and prescribed kernel spaces.

2.3 Information Bottleneck for Degradation Representation

Proposed by Tishby et al. (2000), the Information Bottleneck (IB) principle seeks a compact representation that suppresses input redundancy while preserving information relevant to a target. The variational information bottleneck further makes this objective tractable through stochastic encoders and variational bounds (Alemi et al., 2016). Subsequent studies relate information minimality to

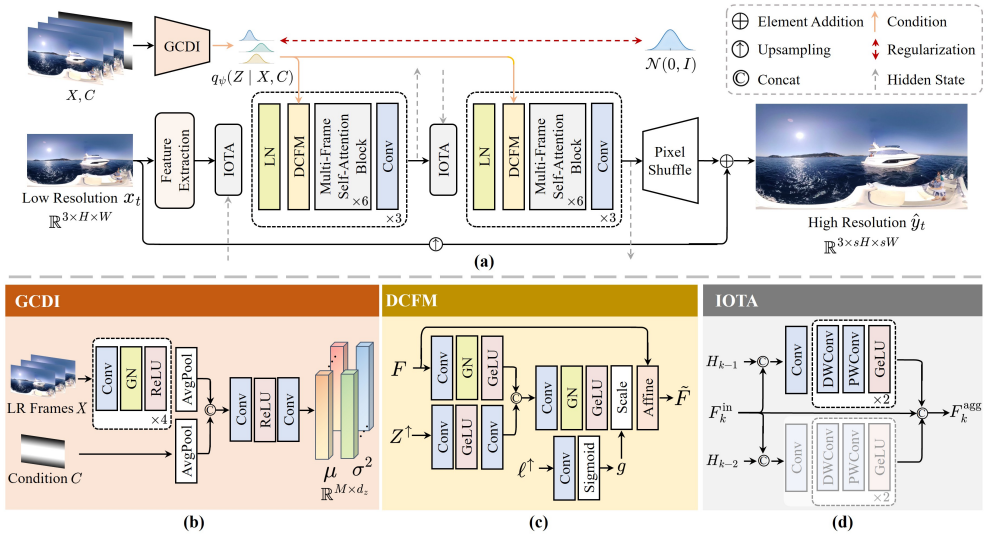

Figure 2: Overall framework of the proposed OmniVaria and detailed design of key modules.

invariance and disentanglement and clarify its behavior in deep networks (Achille & Soatto, 2018; Amjad & Geiger, 2019). Shen et al. (2025a) regularize probabilistic prompts to preserve degradation information while suppressing redundant image content, thereby aligning representations for joint compression and restoration. These studies motivate using IB to suppress redundant content from auxiliary representations while preserving information useful for restoration.

3 Methodology

3.1 Motivation and Overview

Existing ODVSR models typically assume a pre-defined degradation, *e.g.,* bicubic or fisheye downsampling. To handle unknown degradations in ODVs, we study blind ODVSR with unknown fisheye-domain blur using a fisheye-imaging-inspired degradation, as shown in Fig. 1(a). Let $X = \{x_t\}_{t=1}^{T} \in \mathbb{R}^{T\times 3\times H\times W}$ denote a group of degraded LR ERP frames, and let $Y = \{y_t\}_{t=1}^{T} \in \mathbb{R}^{T\times 3\times sH\times sW}$ be its HR counterpart, where s is the scale factor. To model blind degradation induced by panoramic imaging, we generate the LR observation through

$$x_t = \mathcal{P}_{\mathrm{F}\to\mathrm{E}}((k_\eta \otimes \mathcal{P}_{\mathrm{E}\to\mathrm{F}}(y_t)) \downarrow_s), \tag{1}$$

where $\mathcal{P}_{\mathrm{E}\to\mathrm{F}}$ and $\mathcal{P}_{\mathrm{F}\to\mathrm{E}}$ denote ERP-to-Fisheye and Fisheye-to-ERP reprojection, respectively, k_η is an unknown blur kernel parameterized by η, $\otimes$ denotes convolution, and $\downarrow_s$ denotes downsampling. The resulting LR statistics depend on scene content, fisheye-domain degradation, and ERP distortion. Degradation modeling therefore needs to interpret the observed cues in relation to the geometric properties of ERP, motivating us to use such geometric priors as a description in local condition inference and to guide recovery of features.

As illustrated in Fig. 2(a), OmniVaria combines geometry-conditioned probabilistic degradation modeling with implicit temporal aggregation in a second-order recurrent Transformer-based VSR framework. PSDC comprises Geometry-Conditioned Degradation Inference (GCDI) and Degradation-Conditioned Feature Modulation (DCFM), connecting probabilistic degradation inference to feature adaptation. Given LR ODV X and ERP geometry prior C, GCDI predicts the parameters (μ, σ^2) of the geometry-conditioned posterior $q_\psi(Z \mid X, C)$ over local degradation conditions Z, yielding guidance for reconstruction. At step k along either propagation direction, IOTA aggregates the current-frame input F_k^{in} with the preceding hidden states H_{k-1} and H_{k-2} to obtain F_k^{agg}. Within the subsequent propagation backbone, DCFM expands the inferred posterior parameters to apply gated affine modulation before multi-frame self-attention. The HR output $\hat{Y} = \{\hat{y}_t\}_{t=1}^{T}$ is reconstructed after bidirectional propagation. During training, GCVIB regularizes the posterior

learning to encourage compact, reconstruction-relevant conditions rather than unrestricted encoding of LR observations.

3.2 Probabilistic Spherical Degradation Conditioning

Geometry-Conditioned Degradation Inference. GCDI uses known ERP geometry as a reference for interpreting degradation cues in LR observations. We construct a latitude-dependent geometry map $C \in \mathbb{R}^{H\times W}$ with $C(u,v) = \cos\phi_v$, where $\phi_v = (v - H/2 + 0.5)\pi/H,\ v \in \{0,\dots,H-1\}$ is the latitude corresponding to row v, providing a prior reflecting ERP geometry. To capture latitude-dependent degradation, we organize $Z = \{z_m\}_{m=1}^{M}$, $z_m \in \mathbb{R}^{d_z}$, over M latitude anchors. An encoder G_ψ predicts $(\mu, \sigma^2) = G_\psi(X, C)$, parameterizing Gaussian posteriors to represent degradation conditions and inference ambiguity:

$$q_\psi(Z \mid X, C) = \prod_{m=1}^{M} \mathcal{N}(z_m; \mu_m, \mathrm{diag}(\sigma_m^2)), \tag{2}$$

where $\mu_m \in \mathbb{R}^{d_z}$ and $\sigma_m^2 \in \mathbb{R}^{d_z}$ denote the posterior distribution parameters of z_m. During training, Z is sampled from the posterior $q_\psi(Z \mid X, C)$ and we use $Z = \mu$ at inference, while posterior variances provide the gating signal in both cases.

Degradation-Conditioned Feature Modulation. DCFM leverages statistics predicted by GCDI to adapt the input features F using a gated affine modulation. Let $\ell = \log\sigma^2$ denote the posterior log-variances. Z and ℓ are first spatially expanded through linear interpolation along latitude and broadcasting along longitude, yielding $Z^{\uparrow}$ and $\ell^{\uparrow}$. Learnable projections then produce affine modulation parameters from $(F, Z^{\uparrow})$ and a gating signal from $\ell^{\uparrow}$, giving

$$\begin{aligned} [\alpha, \beta] &= \mathrm{proj}_{\mathrm{mod}}(F, Z^{\uparrow}), \\ g &= \mathrm{sigmoid}(\mathrm{proj}_{\mathrm{gate}}(\ell^{\uparrow})), \\ \tilde{F} &= F \odot (1 + g \odot \alpha) + g \odot \beta. \end{aligned} \tag{3}$$

3.3 Geometry-Conditioned Variational Information Bottleneck

Reconstruction supervision alone may encourage the inferred conditions to encode content information in addition to degradation cues. GCVIB guides PSDC toward a minimally sufficient representation of LR observations for HR reconstruction by balancing information compression and reconstruction relevance. For the variational analysis, let $\mathcal{F} = f(X)$ denote the reconstruction context, with f independent of ψ. With geometry supplied through C, we formulate the optimization objective as:

$$\mathcal{J}_{\mathrm{GCVIB}} = \lambda_{\mathrm{IB}} I(Z; X \mid C) - I(Z; Y \mid \mathcal{F}, C), \tag{4}$$

where the first term compresses input information retained by Z given C, whereas the second preserves relevant information about Y beyond that provided by $\mathcal{F}$ and C, and λ_{IB} denotes the trade-off parameter.

To obtain a tractable objective, we introduce the variational conditional likelihood $p_\theta(Y \mid \mathcal{F}, Z, C)$, parameterized by the reconstruction network. The relevance term is lower-bounded as:

$$\begin{aligned} I(Z; Y \mid \mathcal{F}, C) &= H(Y \mid \mathcal{F}, C) - H(Y \mid Z, \mathcal{F}, C) \\ &\geq H(Y \mid \mathcal{F}, C) + \mathbb{E}_{p(X,Y,C)q_\psi(Z\mid X,C)}[\log p_\theta(Y \mid \mathcal{F}, Z, C)], \end{aligned} \tag{5}$$

where $p(X, Y, C)$ is the data distribution. For a fixed feature mapping producing $\mathcal{F}$, the entropy term is independent of ψ, motivating reconstruction supervision for Z.

For the compression term, the aggregated posterior $q_\psi(Z \mid C)$ averages $q_\psi(Z \mid X, C)$ over $p(X \mid C)$, which is intractable. We introduce a shared reference prior:

$$r(Z) = \prod_{m=1}^{M} \mathcal{N}(z_m; 0, I). \tag{6}$$

The expected posterior-to-prior divergence then provides an upper bound:

$$\begin{aligned}&\mathbb{E}_{p(X,C)} D_{\mathrm{KL}}[q_\psi(Z \mid X, C) \| r(Z)] \\ &= I(Z; X \mid C) + \mathbb{E}_{p(C)} D_{\mathrm{KL}}[q_\psi(Z \mid C) \| r(Z)] \\ &\geq I(Z; X \mid C).\end{aligned} \tag{7}$$

Since the posterior and prior factorize over the local conditions z_m, their joint KL divergence decomposes into a sum. Averaging over the M conditions gives the regularizer:

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{p(X,C)}[\frac{1}{M} \sum_{m=1}^{M} D_{\mathrm{KL}}[q_\psi(z_m \mid X, C) \| \mathcal{N}(0, I)]]. \tag{8}$$

3.4 Implicit Omnidirectional Temporal Aggregation

Projection-dependent motion deformation and degraded matching cues can impair optical-flow estimation on ERP observations. IOTA instead learns temporal associations directly in feature space. At propagation step k, IOTA combines the current-frame input with the hidden states H_{k-1} and H_{k-2} as

$$F_k^{\mathrm{agg}} = \mathcal{A}(F_k^{\mathrm{in}}, H_{k-1}, H_{k-2}), \tag{9}$$

where $\mathcal{A}$ denotes the learnable aggregation mapping. As illustrated in Fig. 2(d), the current input is combined with each hidden state and processed by a shared hybrid convolutional mapping. The outputs are concatenated with the current input F_k^{in} to form F_k^{agg}. F_k^{agg} is then processed by the corresponding propagation backbone, whose output is propagated as H_k.

3.5 Training Objective

The variational analysis concerns the sampled conditions Z under the reconstruction context. Motivated by the analysis, we jointly optimize the posterior encoder and the full reconstruction network using

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{IB}} \mathcal{L}_{\mathrm{KL}}, \tag{10}$$

where λ_{IB} controls the strength of posterior regularization.

4 Experiments

4.1 Experimental Setting

Dataset. We train on the training set of ODV360 (Cao et al., 2023), a dataset of 200 ERP videos at 2160×1080 covering indoor and outdoor scenes. The evaluation is conducted on the ODV360 validation set, MiG Panorama (Liu et al., 2024), and ODV-SR test set (An et al., 2026b). All $\times 4$ LR data are generated using the fisheye-imaging-inspired degradation in Fig. 1(a), which applies isotropic Gaussian blur and downsampling in the fisheye domain before ERP projection. During training, the blur standard deviation η is sampled from $[0.2,\ 1.8]$, whereas testing covers $\{0.0,\ 1.2,\ 1.8,\ 2.4,\ 3.5\}$ to provide both in- and out-of-distribution data.

Benchmark. As for the benchmark, we compare with four planar baselines: IA-RT (Xu et al., 2024), FMA-Net (Youk et al., 2024), BVSR-IK (Zhu et al., 2025), and Real-ESRGAN (Wang et al., 2021b), representing generic VSR, joint video deblurring and SR, BVSR, and real-world image SR, respectively. For omnidirectional SR, OSRT (Yu et al., 2023) serves as a geometry-aware, frame-wise omnidirectional image super-resolution baseline, whereas S3PO (Baniya et al., 2023) and STDAN (An et al., 2026b) represent temporal ODVSR methods under prescribed degradations. All methods are fine-tuned on the same dataset as ours for fair comparisons, and we report WS-PSNR (Sun et al., 2017) and WS-SSIM (Zhou et al., 2018).

Implementation Details. We train OmniVaria for 300k iterations on 6-frame clips for $\times 4$ SR, with a batch size of 4 and 256×256 HR crops. The geometry map C is cropped with the frames to provide ERP latitudes. We use the Charbonnier loss (Lai et al., 2017) with the KL regularizer averaged over $M = 8$ local conditions of dimension $d_z = 128$. The KL weight λ_{IB} increases

Table 1: Quantitative comparison on ODV360, MiG Panorama and ODV-SR for $\times 4$ blind ODVSR under different fisheye-domain blur levels. Each entry reports WS-PSNR / WS-SSIM. At $\eta = 0$, only fisheye downsampling is applied. **Bold** and underlined values indicate the best and second-best results for each metric, respectively.

Method	$\eta = 0.0$	$\eta = 1.2$	$\eta = 1.8$	$\eta = 2.4$	$\eta = 3.5$
			ODV360		
IA-RT	24.88/0.7958	25.45/0.7974	25.50/0.7878	25.07/0.7681	24.06/0.7295
FMA-Net	24.97/0.7813	25.04/0.7791	24.94/0.7731	24.68/0.7623	24.01/0.7343
BVSR-IK	22.94/0.7236	23.09/0.7248	23.21/0.7245	23.30/0.7220	23.25/0.7101
Real-ESRGAN	23.23/0.7247	23.25/0.7233	23.23/0.7211	23.17/0.7174	22.93/0.7061
S3PO	21.23/0.6851	21.75/0.6946	22.20/0.7011	22.55/0.7039	22.75/0.6980
OSRT	25.70/0.7916	25.75/0.7853	25.43/0.7725	24.92/0.7554	24.04/0.7245
STDAN	24.73/0.7989	25.47/0.8015	25.56/0.7912	25.13/0.7700	24.09/0.7298
OmniVaria (Ours)	**26.16/0.8058**	**26.25/0.8058**	**26.25/0.8043**	**26.20/0.8000**	**25.54/0.7748**
			MiG Panorama		
IA-RT	26.47/0.7659	27.15/0.7756	27.43/0.7751	27.27/0.7640	26.46/0.7329
FMA-Net	27.85/0.8010	28.02/0.7988	27.92/0.7919	27.61/0.7797	26.75/0.7467
BVSR-IK	25.14/0.7317	25.34/0.7339	25.51/0.7345	25.65/0.7325	25.68/0.7211
Real-ESRGAN	25.00/0.7198	25.03/0.7190	25.06/0.7178	25.04/0.7150	24.83/0.7039
S3PO	23.22/0.6868	23.79/0.6980	24.30/0.7061	24.72/0.7104	24.99/0.7050
OSRT	**28.35/0.8144**	**28.40**/0.8063	28.07/0.7916	27.55/0.7721	26.59/0.7369
STDAN	27.57/0.8041	28.01/0.8030	28.01/0.7934	27.61/0.7756	26.58/0.7378
OmniVaria (Ours)	28.29/0.8092	28.36/**0.8086**	**28.35/0.8067**	**28.29/0.8023**	**27.78/0.7781**
			ODV-SR		
IA-RT	26.01/0.8077	26.69/0.8121	26.73/0.8044	26.24/0.7858	25.11/0.7456
FMA-Net	26.54/0.8085	26.65/0.8062	26.49/0.7990	26.11/0.7859	25.19/0.7526
BVSR-IK	23.91/0.7402	24.09/0.7414	24.24/0.7410	24.34/0.7380	24.27/0.7249
Real-ESRGAN	24.01/0.7388	24.03/0.7375	24.03/0.7352	23.98/0.7314	23.74/0.7194
S3PO	21.92/0.6944	22.49/0.7053	23.00/0.7131	23.41/0.7169	23.64/0.7107
OSRT	27.09/0.8168	27.07/0.8096	26.68/0.7956	26.10/0.7767	25.09/0.7418
STDAN	26.21/0.8215	27.00/0.8243	27.01/0.8141	26.43/0.7924	25.18/0.7487
OmniVaria (Ours)	**27.42/0.8263**	**27.52/0.8259**	**27.52/0.8241**	**27.44/0.8195**	**26.72/0.7935**

linearly to 1×10^{-4} over the first 20k iterations. Conditions are sampled during training using Gaussian reparameterization. We use Adam (Kingma & Ba, 2014) with an initial learning rate of 1×10^{-4}, halved at 80k, 130k, and 220k iterations. The model is implemented in PyTorch and trained on 1 NVIDIA A100 GPU.

4.2 Experimental Results

Quantitative Comparison. As displayed in Table 1, OmniVaria achieves the best quantitative metrics on almost all benchmarks, demonstrating consistent performance gains across datasets and degradation levels. The advantage is particularly pronounced beyond the training blur range: OmniVaria leads both metrics on all three datasets at $\eta = 2.4$ and $\eta = 3.5$. At $\eta = 3.5$, it exceeds the best competing method, FMA-Net, by 1.53 dB and 0.0409 on ODV-SR. At $\eta = 0$, OmniVaria ranks first on ODV360 and ODV-SR and second on MiG Panorama, showing that its blind degradation modeling remains effective in the blur-free setting.

Qualitative Comparison. As shown in Fig. 3, OmniVaria recovers clearer textures and more regular structures across different ERP latitudes. Near the equator, it better preserves the thin diagonal

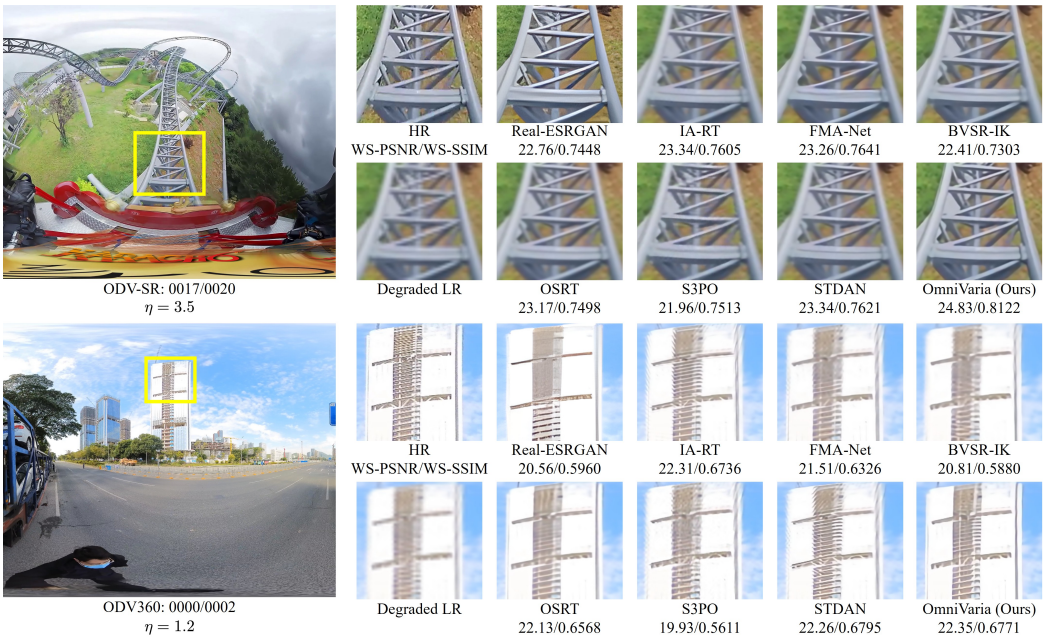

Figure 3: Visual comparisons of $\times 4$ SR results on ODV-SR and ODV360.

Table 2: Ablation of (a) degradation conditioning with IOTA fixed and (b) temporal aggregation. Cond. denotes PSDC trained with GCVIB. Disabling IOTA uses flow-warp-based alignment.

(a) Degradation conditioning

Variant	ODV360	MiG
w/o PSDC + GCVIB	25.48/0.7757	27.41/0.7647
w/o GCVIB	26.02/0.7966	28.02/0.7937
w/o Geometry C	25.73/0.7882	28.13/0.7978
w/o Gating	25.70/0.7863	28.06/0.7962
w/o Prob.	25.73/0.7879	27.78/0.7793
Global Condition	25.98/0.7943	27.74/0.7760
OmniVaria	**26.08/0.7981**	**28.21/0.8010**

(b) Temporal aggregation

Cond.	IOTA	ODV360	MiG
✗	✗	24.97/0.7576	27.02/0.7486
✗	✓	25.48/0.7757	27.41/0.7647
✓	✗	25.87/0.7925	27.31/0.7650
✓	✓	**26.08/0.7981**	**28.21/0.8010**

beams and their intersections, whereas competing methods tend to blur or distort these structures. The advantage becomes more evident in the higher-latitude region, where several baselines exhibit severe aliasing patterns, jagged boundaries, or broken facade patterns. In contrast, OmniVaria maintains clearer details and straighter boundaries, demonstrating more consistent reconstruction under ERP degradation.

4.3 Ablation Studies

Table 2 evaluates degradation conditioning and temporal aggregation on ODV360 and MiG Panorama, averaging WS-PSNR and WS-SSIM over the five test blur levels. Variant configurations are detailed in Appendix C.

Probabilistic Spherical Degradation Conditioning. Table 2(a) examines the conditioning branch with IOTA fixed. Ablating conditioning lowers WS-PSNR by 0.60 dB on ODV360 and 0.80 dB on MiG. A global condition ($M = 1$) retains C and recovers part of this gain by providing global degradation modeling, while using eight local conditions further improves both datasets by 0.10 dB

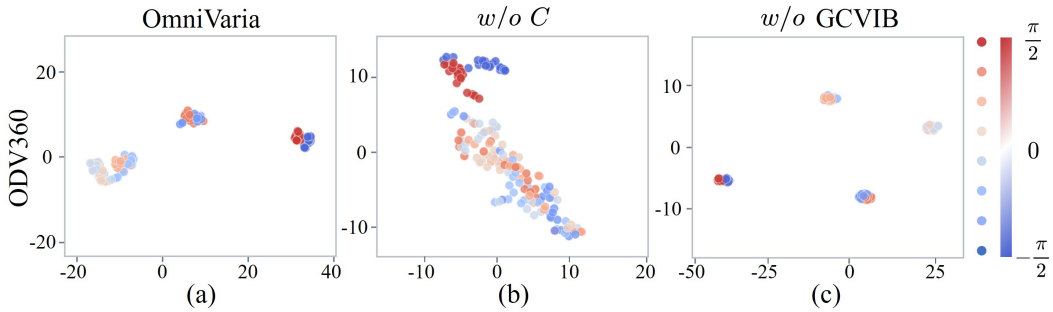

Figure 4: Visualization of posterior means μ at $\eta = 1.8$ by t-SNE. Each point denotes the μ of one latitude-anchor condition inferred from a complete video, colored by latitude.

and 0.47 dB, respectively. Removing C reduces performance despite retaining the latitude anchors, supporting an explicit geometric reference for condition inference. Deterministic conditions and ungated modulation also underperform the full model.

Geometry-Conditioned Variational Information Bottleneck. Removing GCVIB retains the probabilistic representation but reduces WS-PSNR by 0.06 dB on ODV360 and 0.19 dB on MiG. The larger gap on MiG supports its contribution to cross-dataset generalization, consistent with limiting input-specific redundancy beyond probabilistic parameterization alone.

Implicit Omnidirectional Temporal Aggregation. Table 2(b) shows that IOTA improves reconstruction over flow-warp-based alignment both with and without degradation conditioning. With the complete conditioning branch, IOTA increases WS-PSNR by 0.21 dB on ODV360 and 0.90 dB on MiG. Without conditioning, the corresponding gains are 0.51 dB and 0.39 dB. Although the marginal gains vary across datasets, the combined configuration performs best on both, supporting the complementary contributions of spatial degradation adaptation and flow-free temporal aggregation.

Organization of Local Conditions. Fig. 4 visualizes the posterior means using t-SNE (Van der Maaten & Hinton, 2008). OmniVaria separates polar conditions from lower-latitude conditions, while some neighboring low-latitude conditions occupy overlapping regions. Removing C produces more intermingled condition groups. Without GCVIB, the conditions form four compact groups corresponding to opposite-latitude pairs, dominated by latitude or content differences and overlooking the proximity of degradations in low-latitude regions.

5 Conclusion

We present OmniVaria, a blind ODVSR framework that combines geometry-conditioned probabilistic degradation modeling with implicit temporal aggregation. PSDC uses known ERP geometry to infer posterior distributions over latitude-wise degradation conditions and translates the inferred conditions and posterior statistics into gated feature modulation. GCVIB regularizes posterior learning to encourage compact, reconstruction-relevant guidance, while IOTA exploits current-frame features and propagated hidden states without optical-flow estimation or feature warping. Evaluations on three benchmarks under five blur levels demonstrate the effectiveness of this design, including improved reconstruction at the unseen blur levels $\eta = 2.4$ and 3.5. These results support jointly accounting for unknown capture degradation and projection geometry when learning conditions for spatiotemporal reconstruction.

AI use statement

In this work, we used LLMs to assist with manuscript drafting, translation, literature review, and mathematical derivations. The authors reviewed all material and take full responsibility for the final manuscript.

Reproducibility Statement

The degradation protocol, model components, and training objective are described in Sec. 3, with the additional variational derivation of GCVIB provided in Appendix A. Sec. 4.1 specifies the datasets, degradation ranges, evaluation metrics, and training configuration. Additional temporal consistency and efficiency evaluations are presented in Appendix B. Detailed ablation variant configurations and additional alignment comparisons are provided in Appendix C. The source code will be publicly released.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Rana Ali Amjad and Bernhard C Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2225–2239, 2019.

Hongyu An, Xinfeng Zhang, Xu Fan, Shijie Zhao, Li Zhang, and Ruiqin Xiong. D^2r^2osr: Degradation-disentangled representation for real-world omnidirectional image super-resolution. *arXiv preprint arXiv:2606.29314*, 2026a.

Hongyu An, Xinfeng Zhang, Shijie Zhao, Li Zhang, and Ruiqin Xiong. Spatio-temporal distortion aware omnidirectional video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2309–2317, 2026b.

Charles-Olivier Artizzu, Haozhou Zhang, Guillaume Allibert, and Cédric Demonceaux. Omniflownet: a perspective neural network adaptation for optical flow estimation in omnidirectional images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2657–2662. IEEE, 2021.

Haoran Bai and Jinshan Pan. Self-supervised deep blind video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4641–4653, 2024.

Arbind Agrahari Baniya, Tsz-Kwan Lee, Peter W Eklund, and Sunil Aryal. Omnidirectional video super-resolution using deep learning. *IEEE Transactions on Multimedia*, 26:540–554, 2023.

Keshav Bhandari, Ziliang Zong, and Yan Yan. Revisiting optical flow estimation in 360 videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8196–8203. IEEE, 2021.

Qing Cai, Mu Li, Dongwei Ren, Jun Lyu, Haiyong Zheng, Junyu Dong, and Yee-Hong Yang. Spherical pseudo-cylindrical representation for omnidirectional image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 873–881, 2024.

Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Gen Li, Ying Shan, Radu Timofte, et al. Ntire 2023 challenge on 360 omnidirectional image and video super-resolution: Datasets, methods and results. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1731–1745. IEEE, 2023.

Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4945–4954. IEEE, 2021.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5962–5971. IEEE, 2022.

Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9185–9194. IEEE, 2021.

Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1604–1613. IEEE, 2019.

Mehran Jeelani, Noshaba Cheema, Klaus Illgner-Fehns, Philipp Slusallek, Sunil Jaiswal, et al. Expanding synthetic real-world degradations for blind video super resolution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1199–1208. IEEE, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5835–5843. IEEE, 2017.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *European conference on computer vision*, pp. 179–195. Springer, 2018.

Suyoung Lee, Myungsub Choi, and Kyoung Mu Lee. Dynavsr: Dynamic adaptive blind video super-resolution. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2092–2101. IEEE, 2021.

Runyi Li, Xuhan Sheng, Weiqi Li, and Jian Zhang. Omnissr: Zero-shot omnidirectional image super-resolution using stable diffusion model. In *European Conference on Computer Vision*, pp. 198–216. Springer, 2024.

Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. Netflix Technology Blog, 2016. URL `https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652`. Accessed: September 20, 2026.

Hongying Liu, Wanhao Ma, Zhubo Ruan, Chaowei Fang, Fanhua Shang, Yuanyuan Liu, Lijun Wang, Chaoli Wang, and Dongmei Jiang. A single frame and multi-frame joint network for 360-degree panorama video super-resolution. *Engineering Applications of Artificial Intelligence*, 134:108601, 2024.

Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, 33: 5632–5643, 2020.

Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17621–17631. IEEE, 2022.

Cagri Ozcinar, Aakanksha Rana, and Aljosa Smolic. Super-resolution of omnidirectional images using adversarial learning. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6. IEEE, 2019.

Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4791–4800. IEEE, 2021.

Xuelin Shen, Quan Liu, Jiayin Xu, and Wenhan Yang. Prompt-guided alignment with information bottleneck makes image compression also a restorer. *Advances in Neural Information Processing Systems*, 38:3424–3444, 2025a.

Xuelin Shen, Yitong Wang, Silin Zheng, Kang Xiao, Wenhan Yang, and Xu Wang. Fast omnidirectional image super-resolution: Adapting the implicit image function with pixel and semantic-wise spherical geometric priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6833–6841, 2025b.

Xuhan Sheng, Runyi Li, Bin Chen, Weiqi Li, Xu Jiang, and Jian Zhang. Realosr: Latent unfolding boosting diffusion-based real-world omnidirectional image super-resolution. *arXiv e-prints*, pp. arXiv–2412, 2024.

Hao Shi, Yifan Zhou, Kailun Yang, Xiaoting Yin, Ze Wang, Yaozu Ye, Zhe Yin, Shi Meng, Peng Li, and Kaiwei Wang. Panoflow: Learning 360 optical flow for surrounding temporal understanding. *IEEE Transactions on Intelligent Transportation Systems*, 24(5):5570–5585, 2023.

Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Qiufang Ma, Xuhan Sheng, Ming Cheng, Haoyu Ma, Shijie Zhao, Jian Zhang, Junlin Li, et al. Opdn: Omnidirectional position-aware deformable network for omnidirectional image super-resolution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1293–1301. IEEE, 2023.

Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters*, 24(9):1408–1412, 2017.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Jiangang Wang, Yuning Cui, Yawen Li, Wenqi Ren, and Xiaochun Cao. Omnidirectional image super-resolution via bi-projection fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5454–5462, 2024.

Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10576–10585. IEEE, 2021a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914. IEEE, 2021b.

Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. Knowledge distillation based degradation estimation for blind super-resolution. *arXiv preprint arXiv:2211.16928*, 2022.

Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. Enhancing video super-resolution via implicit resampling-based alignment. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2546–2555. IEEE, 2024.

Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. Spheresr: 360° image super-resolution with arbitrary projection via continuous spherical image representation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5667–5676. IEEE, 2022.

Geunhyuk Youk, Jihyong Oh, and Munchurl Kim. Fma-net: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 44–55. IEEE, 2024.

Fanghua Yu, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, and Chao Dong. Osrt: Omnidirectional image super-resolution with distortion-aware transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13283–13292. IEEE, 2023.

Xingyu Zhou, Leheng Zhang, Xiaorui Zhao, Keze Wang, Leida Li, and Shuhang Gu. Video super-resolution transformer with masked inter&intra-frame attention. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25399–25408. IEEE, 2024.

Yufeng Zhou, Mei Yu, Hualin Ma, Hua Shao, and Gangyi Jiang. Weighted-to-spherically-uniform SSIM objective quality evaluation for panoramic video. In *Proceedings of the IEEE International Conference on Signal Processing*, pp. 54–57, 2018.

Qiang Zhu, Yuxuan Jiang, Shuyuan Zhu, Fan Zhang, David Bull, and Bing Zeng. Blind video super-resolution based on implicit kernels. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10971–10981. IEEE, 2025.

A VARIATIONAL DERIVATION OF GCVIB

This section provides the variational derivation of the Geometry-Conditioned Variational Information Bottleneck. The probabilistic local degradation conditions Z and the ERP geometry condition C follow the definitions in Sec. 3.2. In particular,

$$q_\psi(Z \mid X, C) = \prod_{m=1}^{M} \mathcal{N}(z_m; \mu_m, \mathrm{diag}(\sigma_m^2)), \tag{11}$$

where $Z = \{z_m\}_{m=1}^{M}$ contains one condition per reference latitude. All expectations below are taken under the joint distribution induced by the data and the posterior encoder:

$$\pi_\psi(X, Y, C, Z) = p(X, Y, C)\, q_\psi(Z \mid X, C). \tag{12}$$

For this analysis, $\mathcal{F} = f(X)$ is a fixed deterministic context with f independent of the posterior parameters ψ. The information terms concern the sampled conditions $Z \sim q_\psi(Z \mid X, C)$.

A.1 VARIATIONAL UPPER BOUND ON LATENT COMPRESSION

The conditional mutual information is:

$$I(Z; X \mid C) = \mathbb{E}_{p(X,C)} D_{\mathrm{KL}}[q_\psi(Z \mid X, C) \,\|\, q_\psi(Z \mid C)], \tag{13}$$

where the geometry-conditioned aggregated posterior is

$$q_\psi(Z \mid C) = \int q_\psi(Z \mid X, C)\, p(X \mid C)\, \mathrm{d}X. \tag{14}$$

Because this marginal is generally intractable, we introduce the factorized reference prior

$$r(Z) = \prod_{m=1}^{M} r(z_m), \qquad r(z_m) = \mathcal{N}(z_m; 0, I). \tag{15}$$

Define the posterior-to-prior divergence

$$\mathcal{U}_{\mathrm{comp}} = \mathbb{E}_{p(X,C)} D_{\mathrm{KL}}[q_\psi(Z \mid X, C) \,\|\, r(Z)]. \tag{16}$$

By inserting $q_\psi(Z \mid C)$ into the density ratio, we obtain

$$\begin{aligned}
\mathcal{U}_{\mathrm{comp}} &= \mathbb{E}_{\pi_\psi(X,C,Z)}[\log \frac{q_\psi(Z \mid X, C)}{r(Z)}] \\
&= \mathbb{E}_{\pi_\psi(X,C,Z)}[\log \frac{q_\psi(Z \mid X, C)}{q_\psi(Z \mid C)}] \\
&\quad + \mathbb{E}_{\pi_\psi(X,C,Z)}[\log \frac{q_\psi(Z \mid C)}{r(Z)}] \\
&= I(Z; X \mid C) + \mathbb{E}_{p(C)} D_{\mathrm{KL}}[q_\psi(Z \mid C) \,\|\, r(Z)].
\end{aligned} \tag{17}$$

The second term in Eq. 17 is non-negative. Therefore,

$$I(Z; X \mid C) \le \mathcal{U}_{\mathrm{comp}}. \tag{18}$$

This is a variational upper bound induced by the reference prior.

A.2 VARIATIONAL LOWER BOUND ON RECONSTRUCTION RELEVANCE

The reconstruction-relevance term can be expressed as

$$\begin{aligned}
I(Z; Y \mid \mathcal{F}, C) &= H(Y \mid \mathcal{F}, C) - H(Y \mid Z, \mathcal{F}, C) \\
&= H(Y \mid \mathcal{F}, C) + \mathbb{E}_{\pi_\psi}[\log \pi_\psi(Y \mid Z, \mathcal{F}, C)].
\end{aligned} \tag{19}$$

For the variational decoder $p_\theta(Y \mid \mathcal{F}, Z, C)$ introduced in Sec. 3.3, the conditional KL divergence satisfies

$$\mathbb{E}_{\pi_\psi(Z,\mathcal{F},C)} D_{\mathrm{KL}}[\pi_\psi(Y \mid Z, \mathcal{F}, C) \,\|\, p_\theta(Y \mid \mathcal{F}, Z, C)] \ge 0, \tag{20}$$

which implies

$$I(Z;Y \mid \mathcal{F}, C) \geq H(Y \mid \mathcal{F}, C) + \mathbb{E}_{p(X,Y,C)q_\psi(Z|X,C)}[\log p_\theta(Y \mid \mathcal{F}, Z, C)]. \tag{21}$$

The marginal distribution of $(Y, \mathcal{F}, C)$ does not depend on $q_\psi(Z \mid X, C)$. Hence, $H(Y \mid \mathcal{F}, C)$ does not affect the optimization of the latent posterior. Substituting Eqs. 18 and 21 into Eq. 4 yields

$$\mathcal{J}_{\mathrm{GCVIB}} \leq \lambda_{\mathrm{IB}}\mathcal{U}_{\mathrm{comp}} - \mathbb{E}_{p(X,Y,C)q_\psi(Z|X,C)}[\log p_\theta(Y \mid \mathcal{F}, Z, C)] - H(Y \mid \mathcal{F}, C). \tag{22}$$

For the fixed context mapping, the entropy term is independent of ψ. This bound motivates reconstruction supervision of the sampled conditions. In the implemented architecture, DCFM also receives $\ell = \log \sigma^2$ as deterministic gating statistics. We jointly optimize the posterior encoder and reconstruction backbone using the reconstruction surrogate

$$\mathcal{L}_{\mathrm{rec}} = \mathbb{E}_{p(X,Y,C)q_\psi(Z|X,C)}[\rho(\hat{Y} - Y)], \tag{23}$$

where $\hat{Y}$ is the HR output of the full architecture, including variance-gated modulation, and $\rho(\cdot)$ denotes the Charbonnier loss.

A.3 Latitude-Wise KL Decomposition

Because the posterior and reference prior both factorize over z_m, their joint KL divergence decomposes exactly:

$$\begin{aligned} &D_{\mathrm{KL}}[q_\psi(Z \mid X, C) \,\|\, r(Z)] \\ &= \mathbb{E}_{q_\psi(Z|X,C)}[\log \frac{\prod_{m=1}^{M} q_\psi(z_m \mid X, C)}{\prod_{m=1}^{M} r(z_m)}] \\ &= \sum_{m=1}^{M} D_{\mathrm{KL}}[q_\psi(z_m \mid X, C) \,\|\, r(z_m)]. \end{aligned} \tag{24}$$

For the diagonal Gaussian posterior in Eq. 11 and the unit-Gaussian prior, the divergence for each z_m has the closed form

$$\begin{aligned} &D_{\mathrm{KL}}[q_\psi(z_m \mid X, C) \,\|\, \mathcal{N}(0, I)] \\ &\quad = \frac{1}{2}\sum_{j=1}^{d_z}(\mu_{m,j}^2 + \sigma_{m,j}^2 - \log \sigma_{m,j}^2 - 1). \end{aligned} \tag{25}$$

Accordingly, the compression surrogate can be written as

$$\mathcal{U}_{\mathrm{comp}} = \mathbb{E}_{p(X,C)}[\sum_{m=1}^{M} D_{\mathrm{KL}}[q_\psi(z_m \mid X, C) \,\|\, \mathcal{N}(0, I)]]. \tag{26}$$

We normalize the regularizer by M:

$$\begin{aligned} \mathcal{L}_{\mathrm{KL}} &= \frac{1}{M}\mathcal{U}_{\mathrm{comp}} \\ &= \mathbb{E}_{p(X,C)}[\frac{1}{M}\sum_{m=1}^{M} D_{\mathrm{KL}}[q_\psi(z_m \mid X, C) \,\|\, \mathcal{N}(0, I)]]. \end{aligned} \tag{27}$$

Since M is fixed, this normalization only rescales the compression coefficient. Reusing λ_{IB} for the rescaled weight, the joint training surrogate is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{IB}}\mathcal{L}_{\mathrm{KL}}. \tag{28}$$

The factorization in Eq. 24 applies to the variational KL regularizer. All local degradation conditions still jointly condition the reconstruction network, and no independence assumption is made for HR reconstruction across latitude regions.

Table 3: Temporal consistency and perceptual quality comparison on ODV360 and MiG Panorama.

Method	$\eta = 0$		$\eta = 1.2$		$\eta = 1.8$		$\eta = 2.4$		$\eta = 3.5$	
	$E^*_{\text{warp}} \downarrow$	VMAF ↑	$E^*_{\text{warp}} \downarrow$	VMAF ↑	$E^*_{\text{warp}} \downarrow$	VMAF ↑	$E^*_{\text{warp}} \downarrow$	VMAF ↑	$E^*_{\text{warp}} \downarrow$	VMAF ↑
Dataset	ODV360									
IA-RT	1.9767	65.41	1.4756	58.69	1.1311	47.74	0.8838	34.32	0.6472	13.92
FMA-Net	1.9119	55.31	1.6078	51.61	1.3755	45.77	1.1681	37.65	0.8761	20.83
BVSR-IK	1.7371	31.66	1.5141	30.03	1.3126	27.39	1.1073	23.12	0.8055	12.28
Real-ESRGAN	4.1816	54.28	4.0142	54.70	3.8809	54.47	3.7757	53.53	3.7200	50.75
S3PO	2.8282	28.12	2.1775	25.57	1.7091	21.92	1.3480	16.94	0.9645	7.00
OSRT	2.3636	65.59	1.6703	59.41	1.2304	49.05	0.9494	36.56	0.6861	16.56
STDAN	2.3312	67.43	1.6594	61.20	1.2101	49.58	0.9121	35.37	0.6502	14.18
OmniVaria (Ours)	1.9728	71.28	1.8492	72.09	1.7768	72.34	1.6526	70.42	1.1416	54.21
Dataset	MiG Panorama									
IA-RT	0.0208	42.08	0.0143	36.97	0.0097	27.47	0.0062	14.71	0.0033	0.86
FMA-Net	0.0364	40.27	0.0277	35.44	0.0203	27.64	0.0139	17.38	0.0069	2.72
BVSR-IK	0.0161	12.71	0.0135	11.33	0.0105	9.01	0.0082	5.70	0.0051	0.56
Real-ESRGAN	0.0715	33.29	0.0691	33.51	0.0678	33.20	0.0685	32.36	0.0836	28.93
S3PO	0.0367	11.04	0.0256	8.42	0.0172	5.40	0.0113	2.25	0.0692	0.00
OSRT	0.0228	43.93	0.0133	35.32	0.0081	23.69	0.0053	11.40	0.0033	0.90
STDAN	0.0300	44.34	0.0197	37.75	0.0125	26.72	0.0073	13.17	0.0035	0.83
OmniVaria (Ours)	0.0173	45.08	0.0152	45.34	0.0141	45.34	0.0124	43.25	0.0075	24.74

Table 4: Efficiency comparison for $\times 4$ ODVSR on 20 ERP frames.

Method	Param. (M)	Latency (ms/frame)	Peak Mem. (GiB)
IA-RT	13.41	2950.01	13.07
FMA-Net	9.77	648.21	5.33
BVSR-IK	9.30	738.46	6.63
Real-ESRGAN	16.70	121.42	1.23
S3PO	9.32	51.86	1.59
OSRT	11.93	1506.46	1.57
STDAN	28.56	265.99	10.50
OmniVaria (Ours)	11.77	1552.51	10.93

B More Experimental Results

Temporal Consistency and Perceptual Quality Comparison. Following STDAN (An et al., 2026b), we evaluate temporal consistency and perceptual video quality on ODV360 validation and MiG Panorama under the five blur levels used in the main comparison. We report temporal warping error E_{warp} (Lai et al., 2018), which measures discrepancies between adjacent reconstructed frames after motion compensation, and VMAF Li et al. (2016), which assesses perceptual video quality relative to the HR references. For readability, the reported warping error is scaled as $E^*_{\text{warp}} = 10^3 \times E_{\text{warp}}$.

As shown in Table 3, OmniVaria achieves the highest VMAF on ODV360 at all tested blur levels and on MiG Panorama for $\eta \leq 2.4$. Its warping error is not uniformly the lowest. For several baselines, this error decreases as blur becomes stronger even while VMAF deteriorates. We therefore interpret warping error jointly with VMAF and the reconstruction metrics in Table 1, rather than treating it as a standalone measure of reconstruction quality.

Efficiency Comparison. We compare model size, inference time, and GPU memory usage on a fixed 20-frame ERP video at 2160×1080 for $\times 4$ SR on an NVIDIA A100 GPU. Table 4 reports the

Table 5: Ablation of degradation conditioning on ODV360 validation and MiG Panorama. All variants use IOTA. M denotes the number of latitude-wise conditions. Prob. and Gating denote probabilistic representation and posterior-variance gating, respectively. Each entry reports WS-PSNR / WS-SSIM. A dash denotes a component that is not applicable.

Variant	PSDC				GCVIB	WS-PSNR / WS-SSIM	
	M	C	Prob.	Gating		ODV360	MiG
w/o PSDC + GCVIB	–	–	–	–	–	25.48/0.7757	27.41/0.7647
w/o GCVIB	8	✓	✓	✓	✗	26.02/0.7966	28.02/0.7937
w/o Geometry C	8	✗	✓	✓	✓	25.73/0.7882	28.13/0.7978
w/o Gating	8	✓	✓	✗	✓	25.70/0.7863	28.06/0.7962
w/o Prob.	8	✓	✗	–	–	25.73/0.7879	27.78/0.7793
Global Condition	1	✓	✓	✓	✓	25.98/0.7943	27.74/0.7760
OmniVaria	8	✓	✓	✓	✓	**26.08/0.7981**	**28.21/0.8010**

comparison results, complementing the reconstruction comparisons by making the computational cost of degradation-adaptive spatiotemporal reconstruction explicit.

C More Detailed Ablation Studies

We specify the configurations of the degradation conditioning and temporal aggregation variants to complement the ablation studies in the main text. All results report WS-PSNR / WS-SSIM on ODV360 and MiG Panorama, averaged over $\eta \in \{0.0, 1.2, 1.8, 2.4, 3.5\}$.

Degradation conditioning configurations. All variants in Table 5 use IOTA for temporal aggregation. The full model infers $M = 8$ local conditions of dimension $d_z = 128$ from (X, C), with Gaussian posteriors, variance gating, and GCVIB regularization. The w/o PSDC + GCVIB variant bypasses condition inference and DCFM, removing the complete conditioning branch and its KL loss. Global Condition reduces M to 1 while retaining C, probabilistic representation, variance gating, and GCVIB. The w/o Geometry C variant omits the explicit cosine-latitude input to the encoder while retaining the eight latitude anchors and their spatial expansion in DCFM.

The remaining variants distinguish posterior regularization, probabilistic representation, and modulation strength. In w/o GCVIB, we set $\lambda_{\mathrm{IB}} = 0$ while retaining Gaussian posterior prediction, training-time sampling, and variance gating. In w/o Gating, the gate in Eq. 3 is fixed to one, while posterior learning and GCVIB remain active. The w/o Prob. variant directly predicts eight deterministic conditions from (X, C) and uses them for affine modulation during both training and inference. It removes variance prediction, reparameterization, variance gating, and the KL loss together, while retaining the local condition layout and geometric input.

Temporal aggregation configurations. Table 6 compares temporal aggregation with and without the complete degradation conditioning branch. The Flow Warp-based variant replaces IOTA with optical-flow estimation and bilinear warping of propagated features, following Chan et al. (2021). We additionally evaluate variants using flow-guided deformable convolution (DCN) alignment (Chan et al., 2022) for a comprehensive comparison. Each variant retains the first- and second-order temporal references of the recurrent backbone.

Table 6: Combination of degradation conditioning and temporal aggregation on ODV360 validation and MiG Panorama. Conditioning denotes PSDC trained with GCVIB regularization. Flow Warp-based and DCN-based alignment replace IOTA in the corresponding variants. Each entry reports WS-PSNR / WS-SSIM.

Conditioning	Temporal aggregation	ODV360	MiG
✗	Flow Warp-based Alignment	24.97/0.7576	27.02/0.7486
✗	DCN-based Alignment	25.28/0.7681	27.42/0.7708
✗	IOTA	25.48/0.7757	27.41/0.7647
✓	Flow Warp-based Alignment	25.87/0.7925	27.31/0.7650
✓	DCN-based Alignment	25.90/0.7940	27.23/0.7698
✓	IOTA	**26.08/0.7981**	**28.21/0.8010**